STATISTICAL MODELLING OF SPATIAL AND TEMPORAL PATTERNS IN HUMAN-ELEPHANT CONFLICT

A THESIS SUBMITTED TO

THE UNIVERSITY OF KENT AT CANTERBURY

IN THE SUBJECT OF STATISTICS

FOR THE DEGREE

OF MASTERS OF STATISTICS BY RESEARCH

By

Chloe Francis

September 2020

Statistical modelling of spatial and temporal patterns in human-elephant conflict

Chloe Francis

September 22, 2020

Abstract

The Assam Haathi Project is run by Chester Zoo and Eco-Systems-India. The aim of the project is to manage and reduce human-elephant conflict by making communities aware of harmless methods to stop elephants from destroying their livelihoods and prevent villagers harming elephants in retaliation. The task force educate and provide villagers with new and safer deterents to use on the elephants rather than ones they may currently use, such as guns. The project (based in Assam, India) began in 2004 as deforestation has forced elephants to venture into villages in order to search for other sources of food and shelter. Since the start of the project, trained community members have recorded crop-raiding incidents. In this thesis, we derive meaningful biological conclusions related to questions concerning the project by applying various statistical methods to the elepahant data; these methods include capture-recapture, distance sampling and generalised linear models. Capture-recapture methods are applied to estimate elephant population size and distance sampling methods are used to investigate how the probability of detecting elephants herds varies spatially. Generalised linear models are used to investigate which of the mitigations used play a positive effect at reducing human-elephant conflict across the two Assam study sites – Goalpara and Sonitpur. We go on to discover that based on this data, there is some significant evidence to suggest that spotlights and electric fences are mitigations which are likely to reduce crop loss caused by elephants.

Declaration

The work in this thesis is based on research completed at the School of Mathematics, Statistics and Actuarial Science at the University of Kent. This thesis, nor any part of it, has been submitted elsewhere for any other degree or qualification. All Chapters contain necessary background material, for which no originality is claimed. All other material is believed to be original unless stated otherwise.

Acknowledgments

I would like to thank the following people who have helped me undertake this research project, starting with my academic supervisors: Dr Diana Cole and Professor Rachel McCrea. The knowledge, support and guidance that they have both provided me with has been second to none – I really could not have asked for better supervisors. I thank them both for giving me the opportunity to study this unique research at the University of Kent and for all of the laughter along the way.

Thank you to my supervisor Simon Tollington at Chester Zoo, and Scott Wilson for going above and beyond to retrieve extra requested data and information in order to aid my analyses for this research. I'd also like to recognise and thank the Darwin Initiative, Chester Zoo, N. Hazarika and J. Chakrabarty for their contributions to the Assam Haathi Project and data collection.

I am very grateful for the smiley Claire Carter for her encouragement and advice throughout the past year as the postgraduate research support officer – you will forever be my 'office Mummy'. And to my office friends and colleagues, as well as others who have been a part of my journey at the School of Mathematics, Statistics and Actuarial Science at the University of Kent.

I would also like to thank my friends and my partner. I simply could not have done this without you all and your constant reassurance and support.

Finally, I would like to thank my family – in particular my Mum and Dad – for their unconditional love and encouragement of pursuing a fulfilling future.

Contents

Chapter 1

Introduction

The Assam Haathi Project (AHP) began in 2004 and is a collaboration between Chester Zoo, Eco-Systems-India and the Darwin Initiative; implementing conflict mitigations to attempt to reduce levels of human-elephant conflict within two sites in Assam, India. The data and project goals are explained in more detail in Section 1.1. In Section 1.2 we discuss multiple articles that have been written on the topic since the start of this project, including: Davies et al. [2011]; Wilson et al. [2009]; Wilson et al. [2015]; and Zimmermann et al. [2009]. This Chapter ends with a brief overview of the thesis and how it relates to the AHP in Section 1.3.

1.1 Elephant Data

Chester Zoo states that the forests of Assam in North-East India provide one of the last strongholds for the endangered Asian elephant (*Elephas maximus*), but these forests have some of the highest levels of human-elephant conflict in the World. Human-elephant conflict occurs because people destroy elephant forest habitat (such as deforestation), meaning that elephants must travel further to find shelter and food – often coming to villages [Darwin Initiative]. Elephants

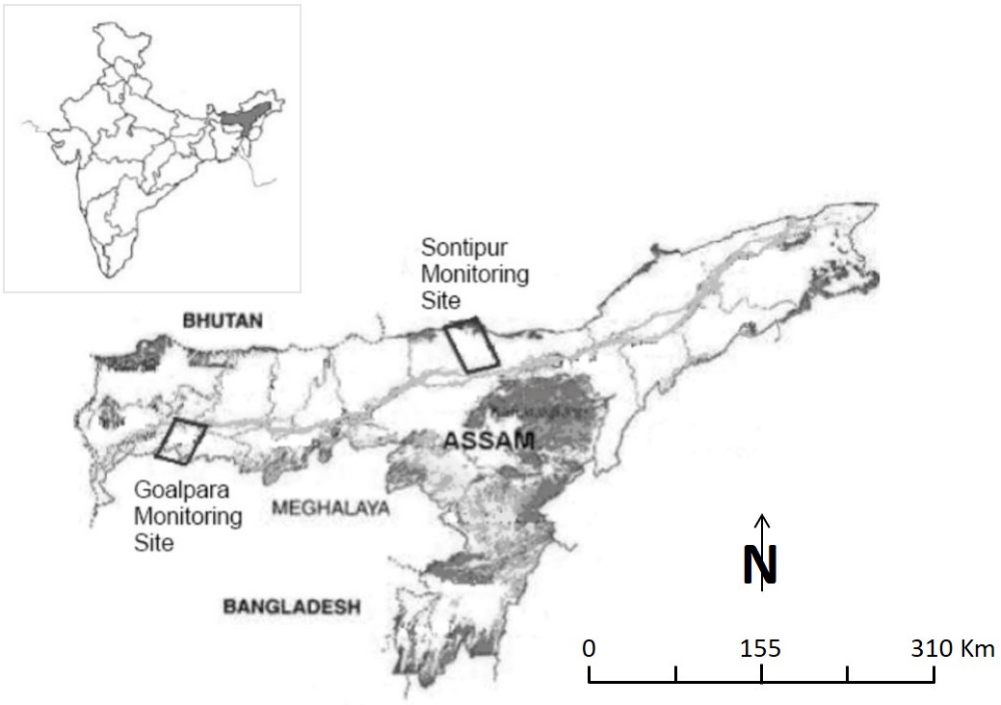


Figure 1.1.1: (Top left) Map of India, with Assam highlighted in grey. (Center) Assam with two AHP site locations emphasised with boxes – Goalpara and Sonitpur [Darwin Initiative].

become a direct and regular threat to a community which can include the safety of people and houses or their livelihoods such as crops, and as a result people retaliate by harming or even killing the elephants. These large mammals are at risk of becoming extinct and this conflict threatens elephant populations outside protected areas even further [Wilson et al., 2015].

The aim of the study is to manage and reduce human-elephant conflict by educating local communities to take responsibility for deterring elephants and encouraging them to implement other humane methods to reduce the effect of the conflict on their livelihoods. The project is based at two study sites – Goalpara and Sonitpur which can be seen in Figure 1.1.1. Wilson et al. [2009] was written with the goal to help educate individuals about the project, see Section 1.2. For data protection related reasons, monitors were assigned unique identity numbers. Figures do not contain exact location values such as longitude and latitude.

The AHP's objectives are:

1) To model observer effort in recording elephant sightings in two spatially independent regions of Assam.

2) To validate and apply the model to produce a 'surface' of sampling effort across both study sites ('surface' meaning map).

3) To use the validated model to determine relative abundance indices of elephants over the study period in both study sites.

4) To use the validated model to predict population-level responses of elephants to conflict mitigation strategies, based on levels of conflict and elephant sightings before and after the implementation of deterrent interventions.

5) To use the outcomes of this research to inform future design of surveys and monitoring.

Due to data restrictions, we will learn throughout this thesis that some of these questions cannot be answered precisely and so we attempt to answer related questions such as 'what is the probability that Monitors sight an elephant herd given that it is present at that exact time?'.

Tables 1.1.1, 1.1.2, and 1.1.3 are extracts from the AHP elephant data; randomly selected from a total sample size of 5895 unique identification entries (UID). To establish a reliable and independent conflict reporting system, a team of community members were trained as monitors to enumerate crop-raiding incidents [Davies et al., 2011]. Monitors were stationed to ensure complete coverage of both study areas, recording sightings from their home area which included observations from themselves and reportings to them from other villagers. The trained monitors would each visit all incidents within their assigned area to verify, quantify and record the location using a GPS unit. Sightings were mainly recorded of herds causing conflict; however, it was also encouraged to record sightings of elephants in non-conflicting situations, for example a herd passing through the

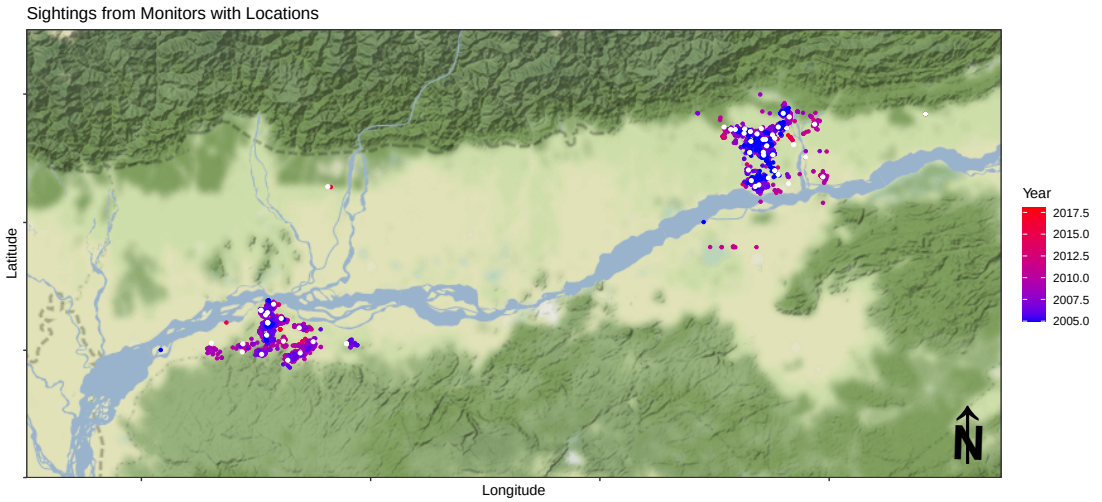


Figure 1.1.2: Visual representation of monitoring observations on a map. White dots are monitor locations, observations by year are on a scale of blue (2005) to red (2014).

village peacefully. See Figure 1.1.2 for a visual representation of all observations from monitors over the length of this project.

For data protection, we have given each of the monitors a unique idenitification under the data column 'MonitorID'. Next in the table is the 'District': this contains information about which study site the observation was made, either Goalpara or Sonitpur. Year, Season, Month, Day and Time are recordings respectively of what year, season, month, date and time of day observations were made. 'Latitude' and 'Longitude' columns represent the recorded locations of the sighting using the monitors GPS unit – due to data protection reasons values in this table have been replaced with random numbers for illustrative purposes only. 'HerdID' consists of the herd identification that the monitor belives the sighting of elephants to be a part of, for example using defined features or scars of a particular elephant. The final column of Table 1.1.1 is 'Count' which is the number

of elephants sighted at that given time. The continued table on page 8 (Table 1.1.2) first contains 'CropCost' and 'CropLoss'. CropCost is the estimated cost of crop loss from that particular elephant observation. CropLoss is directly related to CropCost, stating a 1 if there was any crop cost and a 0 if the crop loss was zero. The following column is 'Mitigation' and declares TRUE or FALSE as to whether a mitigation was used during this particular elephant sighting. The next eighteen columns relate to whether that specific mitigation was put into play or not, with answers being TRUE or FALSE. We will go onto to discuss each of these mitigations in the next paragraph. 'OtherMit Info' tells us which other mitigation was used if column 'OtherMit' reads TRUE, otherwise the entry reads N/A. Finally, the last column of Table 1.1.3 is 'Distance'. Distance is the calculated distance in meters from the monitors location to the elephant sighting (we will discuss how this is calculated in Chapter 3). Note, due to data protection we have hidden columns containing the monitors home location.

There are a total of 17 mitigations used plus the addition of an 'other mitigation' column. The implemented mitigations from both the project and villagers with generalised descriptions and whether they are reactive (employed in the presence of elephants) or more permanent (always in place even when elephants haven't been sighted) mitigiations include:

- Noise – a general noise such as banging or shouting (reactive).
- Fire stick – sticks with a rag, or similar, wrapped around the end and then soaked in a flammable liquid (reactive).
- Torch light – a normal household torch (reactive).
- Cracker – a small firework similar to a bird scarer (reactive).
- Drum tin – villager endoursed banging of pots and pans (reactive).
- Siren – battery operated siren (reactive).

- Watchtower – some watchtowers were built as an early warning system to allow people to spot elephants coming earlier, they would then be scared off nore strategically using one of the other methods (more permanent).
- Kunkie – a mahout which is a person riding a semi-domesticated elephant in an effort to scare others way. This was owned by the Forest department and not local villagers or a part of the AHP's mitigation methods (reactive).
- Chillismoke – dried chilli mixed with tobacco or straw and burnt to create smoke that deterred elephants (reactive).
- Catapult – not a project-endorsed method (reactive).
- Chillifence – dried chilli mixed with used grease and spread on a rope fence (more permanent).
- AHP spotlight – a large spotlight bought and issued by the project (reactive).
- Other spotlight – a homemade version of the AHP spotlight made by villagers (reactive).
- Arrows – homemade bow and arrows, not project-endorsed (reactive).
- eFence – an electric fence often errected around people's homes (more permanent).
- Tripwire – a trip wire connected to something that makes a noise, for example tin cans (more permanent).
- Other mitigation – includes stones, guns and spears/harpoons; the type used is stated in the 'OtherMit Info' column of the data table (reactive).

UID	MonitorID	District	Year	Season	Month	Day	Time	Latitude	Longitude	Herd_ID	Count
1	Monitor17	Sonitpur	2005	POST-MONSOON	12	16	Forenoon	26.612	92.823	SP02	13
2	Monitor17	Sonitpur	2005	POST-MONSOON	12	17	Forenoon	26.822	92.845	SP02	13
3	Monitor17	Sonitpur	2005	POST-MONSOON	12	18	Forenoon	26.562	92.735	SP02	13
4	Monitor17	Sonitpur	2005	POST-MONSOON	12	19	Forenoon	26.792	92.892	SP02	13
5	Monitor03	Goalpara	2006	PRE-MONSOON	5	29	Early Night	26.097	90.624	GP01	23
6	Monitor03	Goalpara	2006	PRE-MONSOON	5	30	Early Night	26.026	90.717	GP01	23
7	Monitor03	Goalpara	2006	MONSOON	7	13	Late Night	26.376	90.613	GP01	24
8	Monitor03	Goalpara	2006	MONSOON	7	13		26.234	90.827	GP01	24
9	Monitor03	Goalpara	2006	MONSOON	7	28	Early Night	26.712	90.615	GP01	24
10	Monitor03	Goalpara	2006	MONSOON	7	28	Late Night	26.073	90.635	GP01	24
11	Monitor03	Goalpara	2006	MONSOON	7	29	Early Night	26.093	90.582	GP01	24
12	Monitor03	Goalpara	2006	MONSOON	7	30	Early Night	26.062	90.635	GP01	24
13	Monitor22	Goalpara	2006	MONSOON	8	3	Early Night	25.828	90.423	GP01	22
14	Monitor22	Goalpara	2006	MONSOON	8	5	Late Night	25.961	90.555	GP01	22
15	Monitor22	Goalpara	2006	MONSOON	8	8	Early Night	25.932	90.433	GP01	22
16	Monitor22	Goalpara	2006	MONSOON	8	14	Late Night	25.961	90.322	GP01	22
17	Monitor22	Goalpara	2006	MONSOON	8	17	Early Night	25.988	90.532	GP01	22
18	Monitor46	Goalpara	2006	MONSOON	8	19	Early Night	25.914	90.689	GP01A	19
19	Monitor54	Goalpara	2006	MONSOON	8	19	Evening	26.011	90.544	GP01	22
20	Monitor03	Goalpara	2006	MONSOON	8	20	Early Night	26.133	90.583	GP01	22
21	Monitor46	Goalpara	2006	MONSOON	8	20	Early Night	25.364	90.682	GP01A	19
22	Monitor46	Goalpara	2006	MONSOON	8	20	Late Night	25.990	90.601	GP01A	19
23	Monitor54	Goalpara	2006	MONSOON	8	20	Evening	26.168	90.319	GP01	22

Table 1.1.1: An extract from the Assam Haathi Project (AHP) elephant data, containing observations of every elephant sighting as a unique identification entry (UID).

UID	CropCost	CropLoss	Mitigation	Fire stick	Chillismoke	Torch light	AHP spotlight	Other spotlight	Noise	Drum tin
1	3000	1	TRUE	TRUE	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE
2	3000	1	TRUE	TRUE	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE
3	3000	1	TRUE	TRUE	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE
4	3000	1	TRUE	TRUE	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE
5	10000	1	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
6	500	1	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
7	1500	1	TRUE	TRUE	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE
8	4000	1	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
9	4000	1	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
10	4000	1	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
11	2500	1	TRUE	TRUE	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE
12	1000	1	TRUE	TRUE	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE
13	2500	1	TRUE	TRUE	FALSE	TRUE	FALSE	FALSE	TRUE	TRUE
14	7500	1	TRUE	TRUE	FALSE	TRUE	FALSE	FALSE	TRUE	TRUE
15	2500	1	TRUE	TRUE	FALSE	TRUE	FALSE	FALSE	TRUE	TRUE
16	2400	1	TRUE	TRUE	FALSE	TRUE	FALSE	FALSE	TRUE	TRUE
17	1200	1	TRUE	TRUE	FALSE	TRUE	FALSE	FALSE	TRUE	TRUE
18	5000	1	TRUE	TRUE	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE
19	2000	1	TRUE	TRUE	FALSE	TRUE	FALSE	FALSE	TRUE	TRUE
20	1200	1	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
21	4000	1	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
22	4000	1	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
23	500	1	TRUE	TRUE	FALSE	TRUE	FALSE	FALSE	TRUE	TRUE

Table 1.1.2: An extract from the Assam Haathi Project (AHP) elephant data, containing observations of every elephant sighting as a unique identification entry (UID).

UID	Cracker	Siren	Tripwire	Catapult	Chillifence	eFence	Arrows	Watchtower	Kunkie	OtherMit	OtherMit Info	Distance
1	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	N/A	9790.64
2	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	N/A	9790.64
3	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	N/A	9790.64
4	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	N/A	9790.64
5	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	N/A	3541.52
6	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	N/A	1682.21
7	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	N/A	1119.19
8	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	N/A	1324.10
9	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	N/A	1324.10
10	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	N/A	1013.00
11	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	N/A	1119.19
12	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	N/A	1119.19
13	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	N/A	2734.55
14	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	N/A	545.34
15	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	N/A	545.34
16	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	N/A	3327.24
17	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	N/A	2734.55
18	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	Stones	6025.57
19	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	N/A	2120.79
20	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	N/A	1574.25
21	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	N/A	7818.17
22	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	N/A	7704.17
23	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	N/A	1510.84

Table 1.1.3: An extract from the Assam Haathi Project (AHP) elephant data, containing observations of every elephant sighting as a unique identification entry (UID).

Note, although AHP classes watchtower as a mitigation it infact is a method which uses multiple deterents. It is stated in relevant analyses, later in this thesis, whether this mitigation was included or not.

1.2 Related Articles

The papers, Zimmermann et al. [2009]; Wilson et al. [2009, 2015]; Davies et al. [2011] were written on the project and the data collected from the project. Zimmermann et al. [2009] and Wilson et al. [2009] were both written nearer to the start of the Assam Haathi Project. Zimmermann et al. [2009] disscuses the study area, project details and challanges of the project formally and in depth, whereas Wilson et al. [2009] was written with a goal to help educate individuals about the project, safe vs unsafe deterents and the positive impact elephants could have on their lives.

Wilson et al. [2015] and Davies et al. [2011] have both investigated the elephant data for the Assam Haathi Project at an earlier date. Wilson et al. [2015] concludes that there are seasonal and diurnal (time of day) patterns for each study site as well as factors which potentially influence the occurrence and scale of human-elephant conflict at each site. Davies et al. [2011] investigated the efficacy of the interventions used but did not present results on whether the significant factors had a negative or positive effect on reducing human-elephant conflict.

1.3 Brief Overview of Thesis

This thesis aims is to model spatial and temporal patterns in human-elephant conflict of the Assam Haathi Project using statistical methods as discussed in Chapter 2. These methods consist of capture-recapture (Section 2.1), distance

sampling (2.2) and generalised linear models (Section 2.3); which all include illustrative examples using the elephant data. We then go on to discuss the probability of detection of the elephant herds in Section 3 using distance sampling methods. In particular we consider a selection of individual monitors (Section 3.1) and all data combined (Section 3.2), as well as assumptions and limitations of our analyses (Section 3.3). Next, we investigate the effect of mitigations of crop loss in Chapter 4 where we apply generalised linear models (Section 4.1) to the elephant data, look at a random monitor effect (Section 4.2) and visit limitations and possible extensions to this particular chapter (Section 4.3). The thesis ends with a conclusion in Chapter 5.

Chapter 2

Statistical Methods in Ecology

In this Chapter we talk about various statistical methods which are often implemented in ecological settings. Methods include capture-recapture in Section 2.1; distance sampling in Section 2.2; and generalised linear models in Section 2.3. Each talk about the fundamentals for each statistical method, as well as a handful of relevant branches, ending with an example that has been applied to the elephant data.

2.1 Capture-Recapture

Obtaining the exact population size of a given species in a region by counting individuals is often impractical and unachievable due to imperfect detection. This is a particular problem if there exists a large area in which these individuals belong. Taking into account various factors, including continuous changes in population size, can also make knowing the exact number of inhabitants difficult, e.g. births, deaths, migration, etc. Instead we can estimate population size using data from a capture, mark, release and recapture study, by fitting either *closed* or *open* models (see for example, Amstrup et al., 2005). Open models take population dynamics into account, meanwhile, closed models do not take into account change in

population abundance. We instead assume that there is an approximate equilibrium between numbers of inhabitants increasing and decreasing. Typically, closed populations have no births/immigration and deaths/emigration as the results are measured and recorded over a short time period (e.g. one week). In contrast, open populations are more often recorded over a long time period (for example a year or more) and therefore births, deaths, immigration and emigration are more likely to occur.

Capture-recapture is the process of attempting to capture and record information on successive capture occasions – it can also be called mark-recapture. A unique identification for each member of the sample is then recorded. In some cases, animals may already have a unique identification of their own, for example Great Crested Newts have unique belly patterns (see Figure 2.1.1) – much like a human has a unique fingerprint ID. For animals who do not have a recognisable individual trait, various methods are implemented which can be used to identify one from another. This includes ear notching for animals such as pigs (see Figure 2.1.2 for example); quick-drying cellulose paint for snails and arthropods; and toe clipping in frogs or toads. Another common example is a metal ring with a unique number for a bird.

Figure 2.1.1: Great crested newt with unique belly pattern (©Steven Allain).

Figure 2.1.2: Pig with unique ear notching (©Courtney Verk).

2.1.1 Methods for Closed Populations

Initially a sample of n_1 individuals are caught and marked. The next step is to release the n_1 individuals and after a given time frame a new sample of size n_2 should be captured and recorded. We are particularly interested in the number of individuals that have been captured in both samples, m_1. This 'overlapping' can be thought of visually by a Venn-Diagram in a universal set, N, which is equivalent to the total population size that is being estimated. See example in Figure 2.1.3, where the total population size being estimated is $N = 17$. On the first capture occasion $n_1 = 7$. For the second occasion $n_2 = 8$ elephants were captured. Three of the n_2 elephants were also captured in the original n_1 capture, so we have $m_1 = 3$.

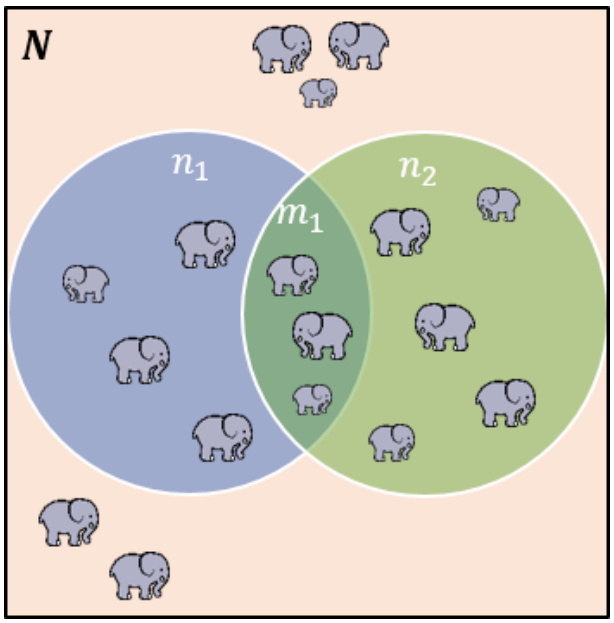


Figure 2.1.3: Venn diagram showing visual representation of how notation is related to a possible two sample capture-recapture study. (Elephant image: ©NiceClipart.com)

Individuals can be captured and recaptured for a maximum number of $k \geq 2$ occasions. The total number of elephants captured from the start of the study is

	Occasion, j				
Individual, i	1	2	3	$\cdots$	k
1	1	0	1	$\cdots$	0
2	1	0	0	$\cdots$	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
R_k	0	0	0	$\cdots$	1

Table 2.1.1: Table showing whether inidividual $i = 1, .., R_k$ was captured or not captured in occasions $j = 1, .., k$.

denoted by R_k where

$$R_k = \sum_{j=2}^{k} r_j = \sum_{j=2}^{k} n_{j-1} + n_j - m_{j-1},$$

where n_{j-1} is the capture occassion at time $j-1$; m_{j-1} is the number of individuals recaptured from in ocassion n_j from occassion n_{j-1}; and r_j is the total number of individuals captured between occasions n_{j-1} and n_j. We can clearly see that R_k is the sum of r_j, where r_j is the total number of distinct inidividuals captured within the time interval $(j-1, j)$ for $j = 2, ..., k$. Note, $R_1 = n_1$. Applying this to Figure 2.1.3:

$$R_2 = r_2 = n_1 + n_2 - m_1 = 7 + 8 - 3 = 12.$$

We can record individual capture histories in binary form where a one represents that the animal was captured at a given time and a zero representing the individual was not captured. For example, animal i with capture history $x = 010010$ has been caught on both the second and fifth occassion but not in the first, third, fourth or sixth occassion. These capture histories can then be listed in a table with each row representing an individual $i = 1, ..., R_k$ and each column representing a time occasion, $j = 1, ..., k$. An example of this layout can be seen below in Table 2.1.1.

Similarly, we can list all individuals of a sample population in matrix form:

$$\mathbf{A} = \begin{bmatrix} X_{1,1} & X_{1,2} & X_{1,3} & \dots & X_{1,k} \\ X_{2,1} & X_{2,2} & X_{2,3} & \dots & X_{2,k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ X_{R,1} & X_{R,2} & X_{R,3} & \dots & X_{R,k} \end{bmatrix},$$

where the subscript i for $X_{i,j}$ represents the individual animal and j denotes the capture event. Applying the data from Table 2.1.1's example produces

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 1 & \dots & 0 \\ 1 & 0 & 0 & \dots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix}.$$

Generally, we can use Akaike's information criterion (AIC) for model selection which assess relative fit. We can also use AICc which is an adapted version of AIC, where the formula considers sample size which is especially useful if the sample size is small. Delta AIC looks at the difference in value compared to the best ranked model, where the best ranked model will always have a Delta AIC value of zero. Similarly, this theory also applies for Delta AICc. If we wanted to assess absolute fit of the model goodness-of-fit tests should be used. For example, Pearson chi-squared tests of observed versus expected values [McCrea and Morgan, 2014, Section 9.3].

2.1.1.1 Two-Sample Lincoln-Peterson Estimator

The *Lincoln-Peterson estimator* uses the structure as already discussed with one recapture, i.e. $k = 2$. The time interval between these two capture times must also be relatively small to ensure that we have a closed population and we must assume that no marks are lost between sampling occasions. Another assumption

is that all individuals have the same detection probability.

We estimate abundance based on the theory that the proportion of marked individuals is approximately equal to the proportions of the total population caught, i.e.

$$\frac{n_1}{N} \approx \frac{m_1}{n_2},$$

and so, we can estimate the total population by $\hat{N}$, which is expressed by rearranging the above formula:

$$\hat{N} = \frac{n_1 \cdot n_2}{m_1},$$

(see, for example, Williams et al., 2002). We can apply this estimation to our elephant example in Figure 2.1.3:

$$\hat{N} = \frac{7 \cdot 8}{3} = 18.67.$$

We can see that $\hat{N}$ is approximately equal to the true population value of $N = 17$.

However, this Lincoln-Pearson estimate has been proven to be biased (see, for example, Amstrup et al., 2005). This bias is inversely related to sample size and particularly the value of m_1. To overcome this issue, we can use the unbiased estimator proposed by Chapman [1951] which holds for $n_1 + n_2 \geq \hat{N}$:

$$\hat{N} = \frac{(n_1 + 1)(n_2 + 1)}{(m_1 + 1)} - 1.$$

The *Chapman Estimator* produces a finite estimate of population size even when the number of recaptures is zero since the denominator is now $m_1 + 1$ which is greater than zero. Evaluating the Chapman estimator for our data we have

$$\hat{N} = \frac{(7 + 1)(8 + 1)}{(3 + 1)} - 1 = 17$$

which is the true value of our population size ($N = 17$).

Both estimates of N have approximately the same variance formula given by:

$$\widehat{Var}(\hat{N}) = \frac{(n_1+1)(n_2+1)(n_1-m_1)(n_2-m_1)}{(m_1+1)^2(m_1+2)},$$

(see, for example, Amstrup et al., 2005). Hence, the variance can be used to calculate an approximate confidence interval for the data. Alternatively a likelihood-ratio based confidence region could be obtained for N which would ensure the region only included permissible values [Morgan, 2008].

The *relative abundance* of animals at two sampling locations can be calculated using estimates of population size for each of these locations. Say we have locations a and b, still with sampling occasions k, then

$$B = \frac{N_b}{N_a},$$

which is our relative abundance of the two locations. Assuming equal capture probabilities for all individuals and substituting N for $\hat{N}$, the equation simplifies to

$$\hat{B} = \frac{r_b}{r_a}.$$

For further information, see Williams et al. [2002].

2.1.1.2 K-Sample Models

The closed population capture-recapture model for $k > 2$ has the same assumptions: the population is closed; marks are neither lost nor overlooked; and capture probabilities are equal for all individuals (equiprobable captures). The simplest model assumes constant probability of capture, p, this is consistant with the two-sample Lincoln-Peterson estimator which has no variation from the original model besides there being a larger number of recaptures. This is called the M_0 model – the subscript 0 referring to no variation. Table 2.1.2 shows the probabilities of

Capture History	Probability, M_0
1 1 1	p^3
1 1 0	$p^2(1-p)$
1 0 1	$p(1-p)p$
1 0 0	$p(1-p)^2$
0 1 1	$(1-p)p^2$
0 0 1	$(1-p)^2p$
0 1 0	$(1-p)p(1-p)$
0 0 0	$(1-p)^3$

Table 2.1.2: Table showing M_0 probabilities of a capture history, given that $k = 3$.

each capture history for $k = 3$.

Amstrup et al. [2005] shows that model M_0 can fail by providing poor interval estimates to examples provided due to its simplistic nature. Relaxing particular constraints within the model leads to different model types which can improve these estimates.

Model M_b, allows for the behavioural effect of how likely an individual is to be recaptured after being caught once already. All individuals have probability p_c of first being captured and then probability p_r of being caught once it has been caught already. Model M_b, caters for animals which may become either trap happy or trap shy. To explain a little further, an individual may learn that a tasty treat awaits them – for example – and so become trap happy as they are more willing to return to the food source; increasing the probability p_r, relative to p_c. Conversely, individuals become trap shy possibly due to an unpleasant first capture say, and they learn not to return, lowering the probability p_r. Table 2.1.3 shows each of the capture history probabilities for example capture histories with $j = 3$ occasions.

The M_t model accounts for different probabilities of being captured on different occasions, regardless of whether the individual has been previously caught or not. Individuals have probability p_j of being captured on occasion j. This model is of a great benefit for animals whose behaviour is affected by weather for

Capture History	Probability, M_b	Probability, M_t
1 1 1	$p_c {p_r}^2$	$p_1 p_2 p_3$
1 1 0	$p_c p_r (1-p_r)$	$p_1 p_2 (1-p_3)$
1 0 1	$p_c (1-p_r) p_r$	$p_1 (1-p_2) p_3$
1 0 0	$p_c (1-p_r)^2$	$p_1 (1-p_2)(1-p_3)$
0 1 1	$(1-p_c) p_c p_r$	$(1-p_1) p_2 p_3$
0 0 1	$(1-p_c)^2 p_c$	$(1-p_1)(1-p_2) p_3$
0 1 0	$(1-p_c) p_c (1-p_r)$	$(1-p_1) p_2 (1-p_3)$
0 0 0	$(1-p_c)^3$	$(1-p_1)(1-p_2)(1-p_3)$

Table 2.1.3: Table showing M_b and M_t probabilities of a capture history, given that $k = 3$.

example. An animal may choose to hibernate during a particular season, reducing the probability of being captured dramatically. M_t may also be relevant if different surveyors are used on each occasion or if different survey efforts are applied across the sampling period. Table 2.1.3 also gives the probabilities corresponding to example capture histories for $j = 3$ occasions under the M_t model.

Another model, M_h, caters for individual capture heterogeneity, meaning each individual has its own probability of being caught. One way of modelling unknown heterogenity is through the use of a finite mixture model. The two group mixture model is defined as: p_1 the probability that an individual in group 1 is captured; and p_2 the probability an individual in group 2 is captured. We have γ defined as the probability that the individual belongs to group 1, and conversely $1 - \gamma$ being the probability an individual belongs to group 2. See Equation 2.1.1 for an example of capture data and probabilities of an individual when $k = 4$.

$$\begin{aligned} Pr(x = 1101) &= \gamma\{p_1 p_1 (1-p_1) p_1\} + (1-\gamma)\{p_2 p_2 (1-p_2) p_2\} \\ &= \gamma\{p_1^3 (1-p_1)\} + (1-\gamma)\{p_2^3 (1-p_2)\}. \end{aligned} \tag{2.1.1}$$

All four of these models have a likelihood of the form:

$$L(n, p; x) \propto \frac{N!}{(N-R)!} \cdot \prod_{i=1}^{R} Pr(x_i) \cdot Pr(x_0)^{N-R} \tag{2.1.2}$$

where x_i is the observed history for all individuals $i = 1, 2, \ldots, R$; $Pr(x_i)$ denotes the probability of the capture history of individual i; and $Pr(x_0)$ the associated probability of not being captured, i.e. $Pr(x_0) = Pr(x = 000)$. See for example King and McCrea [2019].

The last single variation model is based on group variation, M_g. For example groups by gender: those who are female would have capture probability p_f and males would have capture probability p_m, with probabilities occurring much like the original model M_0. However, the difference between this model and the four already listed is how the likelihood is calculated. For this case, likelihoods of both groups need to be calculated separately using the original likelihood formula (2.1.2) with N_m and N_f as the population size of males and females respectively. This would create likelihood values for each group: L_m and L_f. These values are then used in the following formula to calculate the overall likelihood for M_g:

$$L \propto L_m \cdot L_f.$$

To summarise M_0, M_b, M_t, M_h and M_g are the basic k-sample models for closed Capture-Recapture data. However, all of these models, excluding M_0, can be combined to create more refined models suitable for estimating population size. For example, M_{tb} allows for change in temporal and behavioural differences and M_{tbh} allows for change in temporal, behavioural and individual capture heterogeneity.

But how do we estimate the population size from this information? Generally, it is not possible to obtain closed form estimates for our parameters N and p and therefore we use numerical optimisation methods to obtain maximum likelihood

estimates of N and p that maximise L. Computer software including MARK [Cooch and White, 2019]; built in function 'optim' on R or packages such as 'maxLik' on R [Henningsen and Toomet, 2011] can also be used to obtain these maximum likelihood estimates.

2.1.2 Methods for Open Populations

If we are interested in estimating the abundance of individuals in a given area where population dynamics occur (i.e. births/immigration and death/emigration) we need to assume we have an open population. An *open population* is usually assumed for longer periods of time (e.g. yearly records) where births and deaths are more likely to occur. Models in which we will consider include mark-recovery as discussed in Section 2.1.2.1; Cormack-Jolly-Seber in Section 2.1.2.2; capture-recapture-recovery in Section 2.1.2.3; and the original Jolly-Seber model in Section 2.1.2.4. We will also go on to discuss an applied example in Section 2.1.3.

Open population models include similar assumptions as closed models. The models considered in this section have the following assumptions [McCrea and Morgan, 2014]:

- Every marked, alive animal has the same probability of being recaptured at a given time.
- Every marked, dead animal has the same probability of being recovered dead at a given time.
- Every marked, alive animal has the same probability of surviving until the next sampling occasion.
- All emigration from the sample area is permanent (once an animal leaves the region, it cannot re-enter)

- Sampling periods are directly after each other and recaptured animals are released instantly once recorded.
- Marks are neither lost nor over-looked.
- The fate of each animal is independent to the fate of any other animals in terms of capture and survival.

2.1.2.1 Mark-Recovery

Mark-Recovery is the process of marking individuals and then recovering those found deceased which are then recorded. Animals are tagged or marked during time $j_m = 1, ..., k$ and the number recovered dead are counted and recorded throughout times $j_r = 1, ..., t$, where k is the final sampling period, t is the final recovery period and $m \leq k$, usually measured in years. In reality, it is not necessarily possible to recover all dead individuals due to various reasons such as corpse decomposition, predator consumption, etc. or just being missed due to incomplete sampling. The purpose of this model is to be able to estimate animal survival probabilities, including discovering any factors which may influence survival for a particular species.

To explain the example from Table 2.1.4, let us say we are observing tufted ducks. In the first year $1,000$ of the ducks were marked and only two of these ducks were recovered dead in the first year. $3,500$ tufted ducks were marked in

		Year Recovered, j_r		
Year Marked, j_m	Number Marked	1	2	3
1	1000	2	1	0
2	3500		4	2
3	2150			1

Table 2.1.4: Example of Mark-Recovery data.

	Time Recovered, j_r				
Time Marked, j_m	1	2	3	...	t
1	$(1-\phi)\lambda$	$\phi(1-\phi)\lambda$	$\phi^2(1-\phi)\lambda$		$\phi^{t-1}(1-\phi)\lambda$
2		$(1-\phi)\lambda$	$\phi(1-\phi)\lambda$		$\phi^{t-2}(1-\phi)\lambda$
3			$(1-\phi)\lambda$		$\phi^{t-3}(1-\phi)\lambda$
$\vdots$				$\ddots$	
k					$(1-\phi)\lambda$

Table 2.1.5: Mark-Recovery probabilities of the number of individuals marked at time $j_m = 1, 2, ..., k$ and recovered at time $j_r = 1, 2, ..., t$, where $k \leq t$.

the second year with four of these recovered dead and one duck from the first year also recovered.

To calculate the approximate probabilities of these recoveries, we let ϕ denote the probability that an individual survives a year, where $(1-\phi)$ is the probability the animal does not survive. The parameter λ denotes the probability that a given animal dies and the mark is recovered. Again, we have $j_m = 1, 2, ..., k$ marking times and $j_r = j_m, ..., t$ recovering periods where t is the final recovery period and $k \leq t$. The data given in Table 2.1.4 has corresponding probabilities given in Table 2.1.5.

Adjustments can be made to the model to allow for survival and/or recovery probabilities to be dependant on age or time – see Brownie et al. [1985]; Freeman and Morgan [1992]; Catchpole et al. [1995]; McCrea and Morgan [2014]. An example of when this alteration can be useful is with young offspring. Young animals are usually more vulnerable to predators and are more susceptible to an early death when compared to adults of the same species. Therefore a seperate survival probability for young and adults may be applicable.

The likelihood for the Mark-Recovery method is

$$L(n, \phi, \lambda; x) \propto \prod_{j_m=1}^{k} \left(\prod_{j_r=j_m}^{t} (P_{j_m,j_r})^{F_{j_m,j_r}} \right) \cdot \prod_{j_m=1}^{k} \left(1 - \sum_{j_r=j_m}^{t} P_{j_m,j_r} \right)^{M_{j_m} - \sum_{j_r=j_m}^{t} F_{j_m,j_r}}$$

where,

F_{j_m,j_r} number of animals marked in time j_m and recovered in j_r.

P_{j_m,j_r} probability an animal marked in time j_m is recovered in j_r.

M_{j_m} number of animals marked in time j_m.

[Cole et al., 2012]. The multinomial coefficient has not been presented because it does not depend on the parameters of the model.

2.1.2.2 Cormack-Jolly-Seber Model

The *Cormack-Jolly-Seber* (CJS) model [Cormack, 1964; Jolly, 1965; Seber, 1965] is a capture-recapture model for open populations. It uses the same binary notation as with the k-sample open models, indicating whether an inidividual has been captured/recaptured at time $j = 1, ..., k$. We assign the probability that an individual survives one year as ϕ, or not surviving as $1 - \phi$. As not all individuals are caught the model also includes the probability of capture, p. The CJS model allows us to estimate animal survival probability and to discover what factors influence survival for the particular species recorded. Below we give examples of how the probability for specific capture history is formed.

$$Pr(x = 1111) = \underbrace{\phi p}_{\mathbf{2}} \cdot \underbrace{\phi p}_{\mathbf{3}} \cdot \underbrace{\phi p}_{\mathbf{4}} = \phi^3 p^3 \tag{2.1.3}$$

Equation (2.1.3) shows the probability of a history whhere the individual is captures at all occasions. The first capture is at time $j = 1$. So the probability of the individual surviving to the next year, ϕ, and being captured again, p, at time **2** is ϕp. This applies to times **3** and **4** as the individual is captured at every occasion.

$$Pr(x = 1011) = \underbrace{\phi(1 - p)}_{\mathbf{2}} \cdot \underbrace{\phi p}_{\mathbf{3}} \cdot \underbrace{\phi p}_{\mathbf{4}} = \phi^2(1 - \phi)p^3 \tag{2.1.4}$$

However, Equation (2.1.4) shows an example where capture the inidividual but

then do not recapture it at each further occasion. The animal is first captured at time $j = 1$, but we do not capture it again until time $j = 3$. So, at time **2** we know the animal must have survived the year, ϕ, but was not caught, $(1 - p)$, producing probability $\phi(1 - p)$. For times **3** and **4** we captured the individual on both occasions, so have probabilitiy ϕp as discussed with (2.1.3).

$$Pr(x = 0110) = \underbrace{\phi p}_{\mathbf{3}} \cdot \underbrace{[(1 - \phi) + \phi(1 - p)]}_{\mathbf{4}} \tag{2.1.5}$$

In Equation (2.1.5), we have an example of an individual who is not captured at the first sampling occasion but who is first captured at occassion $j = 2$ so our first probability is at time **3**. The animal was captured so its combined probability of surviving the year and being caught is ϕp. Our last capture time at $j = 4$ is 0, meaning the animal was not caught. So, it is unknown to us whether the animal survived the year but was not caught, $\phi(1 - p)$, or if it in fact died, $1 - \phi$. So we allow for all these possible outcomes to produce probability $(1 - \phi) + \phi(1 - p)$ for time **4**. See Lebreton et al. [1992] for further details and applications.

The likelihood for the model is

$$L(n, \phi, p; x) \propto \prod_{i=1}^{R} Pr(x_i) \tag{2.1.6}$$

where once again, optimisation methods can be applied to estimate the values of unknown parameters to maximise L and used to calculate an estimate of total population size. This model can be fitted using software MARK [Cooch and White, 2019] or the Marked R package [Laake et al., 2013].

2.1.2.3 Capture-Recapture-Recovery

It is also possible to record live recaptures as well as recovery of deceased individuals, this is *Capture-Recapture-Recovery* (CRR). For the CRR model, we record the

type of encounter with the individual, i.e. dead, alive or not encountered. Each individual has encounter history x consisting of a string of zeros, ones and twos, where: 2 means individual encountered dead; 1 is individual encountered alive; and 0 meaning the individual was not encountered at all. Of course, in the case of 0 at time k (last encounter time), it is unknown whether the animal is dead or alive assuming it proceeds an alive sighting. Note, a string of zeros will always follow a two.

Much like the Cormack-Jolly-Seber model: the probability that an individual survives one year is ϕ and p is the probability that assuming the inidiviudal survives the year it is then caught. However, the Capture-Recapture-Recovery model in Section 2.1.2.2 contains an additional parameter of λ being the probability that an individual who dies is recovered (as in the Mark-Recovery method of Section 2.1.2.1).

Again, we do not include the initial capture probability. See the CRR model applied to some examples below.

$$Pr(x = 11111) = \phi p \cdot \phi p \cdot \phi p \cdot \phi p = \phi^4 p^4 \tag{2.1.7}$$

Example 2.1.7 is very similar to equation 2.1.3 from the Cormack-Jolly-Seber model as the individual is captured at each occassion. Once the animal is captured at time $j = 1$, the animal has probability of surviving to the next year, ϕ, multiplied by the probability, p, that it is also caught.

$$\begin{aligned} Pr(x = 10110) &= \underbrace{\phi(1-p)}_{\mathbf{2}} \cdot \underbrace{\phi p}_{\mathbf{3}} \cdot \underbrace{\phi p}_{\mathbf{4}} \cdot \underbrace{[(1-\phi)(1-\lambda) + \phi(1-p)]}_{\mathbf{5}} \\ &= \phi^3(1-p)p^2 \cdot [(1-\phi)(1-\lambda) + \phi(1-p)] \end{aligned} \tag{2.1.8}$$

Equation 2.1.8 is an example of an individual who is captured at the initial time $j = 1$ but is not encountered at time **2**. However, as they are captured again at time $j = 3$, we know that the individual must be alive at this time. So the

probability at time **2** is $\phi(1-p)$. Times **3** and **4** are alive captures with probability ϕp as seen above in 2.1.7. At time $j = 5$ the individual is not encountered and as this is our last potential capture occasion, it is unknown if the animal is alive or not. So, time **5** is the summation of both possible outcomes: the individual is dead but not recovered, $(1-\phi)(1-\lambda)$, and the individual survives but is not captured, $\phi(1-p)$.

$$Pr(x = 01200) = \underbrace{(1-\phi)\lambda}_{\mathbf{3}} \tag{2.1.9}$$

Finally, looking at Equation 2.1.9 we have an example of an animal who is first caught at occassion $j = 2$ but then is recovered dead at time $j = 3$. The individual, in this case, is first captured at time $j = 2$ so we only need to consider the probability across one year. As the individual does not survive the year and is recovered, we have probability $\mathbf{3} = (1-\phi)\lambda$.

The CRR model has likelihood

$$L(n, \phi, p, \lambda; x) \propto \prod_{i=1}^{R} Pr(x_i),$$

(see, for example, Catchpole et al. [1998]; Hubbard et al. [2014]). Note that when we combine CRR, we have to assume there is no permanent emigration – or if required, the model can be adpated to account for it (see Reynolds et al. [2009]). This is due to the fact that mark-recovery data estimates true survival where as capture-recapture estimates "apparent" survival which is cofounded with permanent emigration.

2.1.2.4 Jolly-Seber Model

The original *Jolly-Seber* model was proposed by Jolly [1965] and Seber [1965]. The model considers parameters ϕ and p which represent 'survival' rate and capture

probability respectively. The survival rate is approximated by

$$\widehat{\phi_j} = \frac{\widehat{M_{j+1}}}{\widehat{M_j} + R_j - m_j}$$

for $j = 1, ..., k-2$ [Amstrup et al., 2005] where M_j is the marked population size just before period j; R_j is the total number of animals captured at sampling occassion j that are released (e.g. aren't recovered as dead) and m_j is the number of animals captured at sampling occassion j that are marked. The approximation for p is

$$\widehat{p_j} = \frac{m_j}{\widehat{M_j}}$$

for $j = 1, ..., k-1$ [Amstrup et al., 2005]. Further information about the approximations of each variable can be found in Chapter 3 of Amstrup et al. [2005].

However, it is now more common practice to use an alternative approach to the Jolly-Seber model proposed by Schwarz and Arnason [1996]. It proposes the idea of a "*super-population*", denoted by N, which represents the pool of individuals which are available for capture at least once in the study area. Each animal may enter the site at time $i = 1, \ldots, k$ and is available for capture until time k or until their time of exit. However, once an individual leaves the study site it is assumed that they cannot re-enter.

Upon relaxing the assumptions of a closed population, likelihood 2.1.2 is now generalised to:

$$L(n, p, \beta, \phi; x) \propto \frac{N!}{(N-R)!} \cdot \prod_{i=1}^{R} Pr(x_i) \cdot Pr(x_0)^{N-R}$$

[King and McCrea, 2019] where,

β_j : The probability that an individual arrives in the study area between occasions j and $j+1$ (with the first availability for capture at time $j+1$).

ϕ_j : The probability that an individual that is in the study area at time j

remains in the study area until time $j+1$.

To account for unknown arrival and departure times, we alter $Pr(x_i)$ from the likelihood used with closed capture cases. We let f_i denote the first-time individual i is observed and l_i the last time individual i is observed. Note, due to temporary migration not being possible, individual i must be present in the study area for times $f_i, \ldots, l_i$.

Now, supposing that the capture probability is constant, similar to the M_t model for closed populations, we define:

$$Pr(x_i) = \underbrace{\sum_{b=1}^{f_i} \beta_{b-1}}_{\mathbf{1}} \underbrace{\sum_{d=l_i}^{k} (1-\phi_d)}_{\mathbf{2}} \underbrace{\prod_{j=b}^{d-1} \phi_j}_{\mathbf{3}} \underbrace{\prod_{j=b}^{d} p_j^{x_{ij}} (1-p_j)^{1-x_{ij}}}_{\mathbf{4}}.$$

The first term (**1**) corresponds to summing over the possible (unknown) arrival times for individual i; the second (**2**) relates to summing over the possible (unknown) departure times of individual i. (**3**) relates to the individuals remaining in the study area between arrival and departure times ($f_i, \ldots, l_i$ only). Finally, the last terms (**4**) relates to the probability that the individual is captured or not captured when it is in the study area.

From this we can create a formula which allows us to calculate the probability that an individual is not observed. If we let $Pr(x_0) = 1 - p^*$, where p^* denotes the probability than an individual is observed at least once within the study, and substitute into the above formula. This produces:

$$1 - p^* = \sum_{b=1}^{k} \beta_{b-1} \sum_{d=b}^{k} (1-\phi_d) \prod_{j=b}^{d-1} \phi_j \prod_{j=b}^{d} (1-p_j).$$

Pledger et al. [2009] shows that this model has been successfully extended to account for the probability that an animals leaving time from the study area may be affected by its arrival time. Due to this, the adapted superpopulation Jolly-Seber model is often referred to as the *'Stopover model'* – due to its usefulness

Elephant Apparent Survival by Heard, Monthly				
Model	AICc	Delta AICc	AICc Weight	Number of Parameters
ϕ_L, p_L	446.993	0.00	0.77033	4
ϕ, p_L	450.359	3.37	0.14320	3
ϕ_L, p	451.588	4.59	0.07745	3
ϕ, p	455.888	8.89	0.00902	2

Table 2.1.6: AICc values of monthly elephant apparent survival by heard using the Cormack-Jolly-Seber model produced by software MARK. Note, subscript L represents location such that $L = G, S$ which relate to elephant herds in locations Goalpara and Sonitpur respectively. The data is given in Table A.0.0.1 of Appendix A.

of modelling animals stopping over at breeding sites. For some applications see Matechou et al. [2013, 2014].

2.1.3 Elephant Example

The elephant data was recorded by various monitors based around two locations: Goalpara and Sonitpur. Monitors identified and recorded sightings of herds by individually recognisable characteristics of lead elephants, creating unique herd identifications (Herd ID). Capture histories, and therefore probabilities, for this data were recorded monthly. These models discussed in Section 2.1 are not appropriate for the data, this example purely demonstrates the models with the data used as an artificial case study. In this example we look at ϕ being the apparent survival rate which caters to the open population being observed (see Lebreton et al. [1992] for more information regarding apparent survival). Data used for this example can be found in Table A.0.0.1 of Appendix A.

We apply the Cormack-Jolly-Seber model from Section 2.1.2.2 to the elephant data. We note that the data set is small so only consider the constant model and whether the parameter is dependent on location, L, where $L = G, S$ for Goalpara and Sonitpur locations respectively. The AICc values for the found models considered are given in Table 2.1.6. We can see that all model's have small relative values of 'Delta AICc' but the best ranked model with AICc = 446.993 fits

all four parameters to the data (ϕ_G, ϕ_S, p_G, p_S). This best fit model has parameter values as listed in Table 2.1.7.

Parameter	Estimate	Standard Error	95% Confidence Interval
ϕ_G	0.9812	0.0093	(0.951, 0.993)
ϕ_S	0.9166	0.0335	(0.823, 0.963)
p_G	0.4779	0.0344	(0.411, 0.545)
p_S	0.2864	0.0606	(0.183, 0.418)

Table 2.1.7: Parameter estimates for the best fit model: ϕ_L, p_L. All values calculated using MARK [Cooch and White, 2019] (MARK calculates confidence intervals using asymptotic normality properties of the maximum likelihood estimate).

We can see that the apparent survival from one month to the next is relatively high in both locations with $\phi \geq 0.9$ in both cases, with higher probability in Goalpara. However, the probabilitiy of observing a herd each month is low particulary in Sonitpur with $p_S = 0.2864$, and probability $p_G = 0.4779$ in Goalpara. The confidence intervals of ϕ_L and p_L estimators marginally overlap in both cases and so we do not have a statistically significant difference between the parameter estimates of Goalpara compared to Sonitpur, however looking at Table 2.1.6 we can see the model has the lowest AIC value by 3.37 which supports the selection of this model.

Diagnostic goodness of fit tests are available to assess the appropriateness of the models – see for example McCrea and Morgan, 2014, Section 9.2. However due to the small sample size of this data set it would not be possible to have the power to detect such violations and therefore testing on a larger sample could lead to a more precise estimates.

2.2 Distance Sampling

Distance sampling is a method used to estimate population size and/or density; it is a widely used method where presence of individuals – or objects of interest – are most commonly obtained by surveying lines or points. In order to estimate

density, ideally the probability of detection is needed. However, if we do not know the exact probability of detection then we are able to calculate the detection probability using the data, provided that we know the distance from the observer to the individual. This produces an estimate of the probability for which the individual is detected and recorded, given that the individual is present in the area at that time (denoted by P_a). Therefore, fitting a distance sampling model allows us to estimate this probability from the detection function (which we will go onto discuss in Section 2.2.2) as well as estimating population size and/or density.

Some applications of distance sampling include: studying populations of many species of bird such as gamebirds [Dorgeloh, 2005]; terrestrial mammals including species of deer [Ward et al., 2004] and primates [Peres, 1999]. There have also been distance sampling studies on reptiles [Rodda and Campbell, 2002] and beetles [Didham et al., 1998]. Sampling can be made based on animals which are alive but can also be based on the discovery of dead animals. Examples of this include plant observations, inanimate objects (such as ant nests [Baccaro and Ferraz, 2012]) and even military applications [Buckland et al., 2001].

This section will go onto discuss various distance sampling methods (Section 2.2.1) as well as explaining the data anlysis process (Section 2.2.2) and the assumptions related to these distance sampling methods (Section 2.2.3).

2.2.1 Distance Sampling Methods

In all cases of distance sampling we use the same standard notation, this includes the total plot area, A, the true population abundance, N, and the true population density, D. The area covered by sampling is denoted by a, and, the number of animals observed is denoted by n.

For example, in Figure 2.2.1, we have an area of 10 meters by 10 meters. Therefore we have a total plot area of $A = 10 \times 10 = 100m^2$. Our true population abundance can be counted in this example – however, in most real life cases this

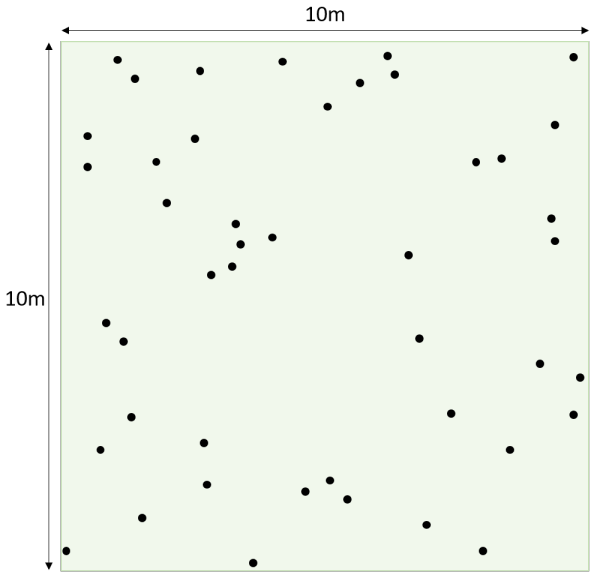


Figure 2.2.1: Example of a $10m$ by $10m$ field. Black dots represent individuals that we wish to estimate.

is not possbile, so usually it would be the value N that we are trying to estimate. Here, $N = 45$. Population density is the population size per unit area and so $D = 45/100 = 0.45$ per m^2.

2.2.1.1 Quadrat Sampling

Quadrat sampling is not quite a distance sampling method, however, various types of distance methods are extensions of quadrat sampling so we begin by describing this method [Buckland et al., 2001]. A common example of quadrat sampling is to approximate the number of buttercups or daisies in a small field which is illustrated in Figure 2.2.2 which many people first come across in GCSE Science class textbooks (for example Locke and Hulme [2016]). A study site is chosen with measured area A. Next, a number of square quadrats are placed at random across the field, covering an observed area denoted by a. The number of buttercups are then counted in observed area a of the quadrats and from this, we can calculate the estimated value of abundance, $\hat{N}$.

Assuming that all individuals, in this case buttercups, that are present in our

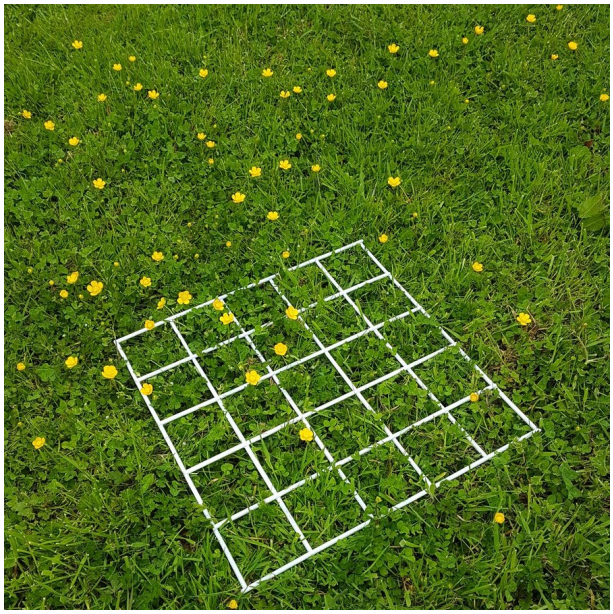


Figure 2.2.2: Quadrat sampling of buttercups in a small field (©Hermitage Academy, Science Department).

transect areas are observed (a full detection probability) we estimate abundance based on the theory that the proportion of counted individuals in an observed area is approximately equal to the proportion of population abundance in the total plot area, i.e.

$$\frac{N}{A} \approx \frac{n}{a}.$$

Now, we let $\hat{N}$ be the estimated value of the true population abundance, as a result we can produce the following estimator equation:

$$\begin{aligned} \frac{\hat{N}}{A} &= \frac{n}{a} \\ \Rightarrow \hat{N} &= \frac{n}{a} \cdot A \\ &= \hat{D} \cdot A \end{aligned} \tag{2.2.1}$$

where $\hat{D} = n/a$, representing the estimated population density.

Let us apply this to the example pictured in Figure 2.2.3. We know that $N = 45$ (the true abundance that we are trying to estimate) and that $A = 100m^2$.

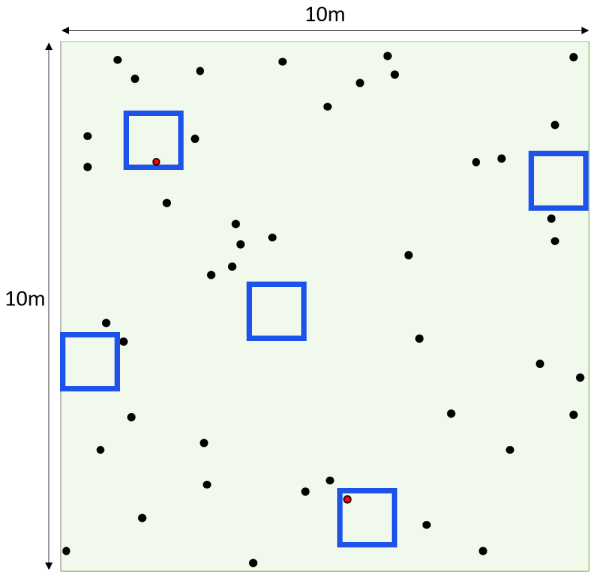


Figure 2.2.3: Example of a $10m$ by $10m$ field. Black dots represent unobserved individuals, red dots represent observed individuals and blue squares represent 1m quadrats.

Looking at the diagram we have randomly placed five $1m \times 1m$ quadrats and so we can calculate the observed area:

$$a = \underbrace{5}_{No.\ of\ quadrats} \cdot \underbrace{(1 \times 1)}_{Area\ of\ quadrat} = 5m^2$$

In the quadrats (blue squares) we observe a total of two individuals (red dots) so $n = 2$. The estimated population density is

$$\hat{D} = \frac{n}{a} = \frac{2}{5} = 0.4$$

and therefore using Formula 2.2.1:

$$\hat{N} = \hat{D} \cdot A = 0.4 \cdot 100 = 40.$$

And so, our estimate of $\hat{N} = 40$ is approximately equal to N. There is only a small discrepency between the true and estimated population abundance. Estimation consistency is guaranteed as the quadrants are chosen at random and all

quadrants in the sample size are assumed to be representative of the whole area - as mentioned earlier in this Section.

2.2.1.2 Strip Transect Sampling

In *Strip transect sampling* transects (straight lines) are used instead of using quadrats like those discussed in Section 2.2.1.1. A total of k strips can either be randomly placed, similar to the example in Figure 2.2.4 where there are $k = 2$ strips, or a systematic random design invoked (a random starting point with fixed, periodic intervals). Strip transect sampling can also be referred to as plot sampling [Buckland et al., 2015].

Once the position of survey transects are chosen, a distance where the surveyor can see everything is chosen as the width, w, observed either side from the transect. For example, surveyors may walk the strip transects i of length l_i, where

$$L = \sum_{i=1}^{k} l_i$$

is the total length of the transect lines. Surveyors can only see everything within $w = 0.5m$ of the line to ensure a perfect detection probability in this example. This is the case in Figure 2.2.4, however, in most real-life cases we do not have a full detection probability if our observed distance w is large. This is because typically, the further away an animal is from the observer, the less likely the individual will be observed. This method can be very inefficient as many individuals beyond the strip will not be included [Burnham and Anderson, 1984].

To estimate abundance for strip transect sampling we use the same formula as in quadrat sampling (Formula 2.2.1), with a new formula for a. Previously, a was the area of each quadrat multiplied by the number of quadrats used. For strip transect sampling, we must take into account the width observed from the

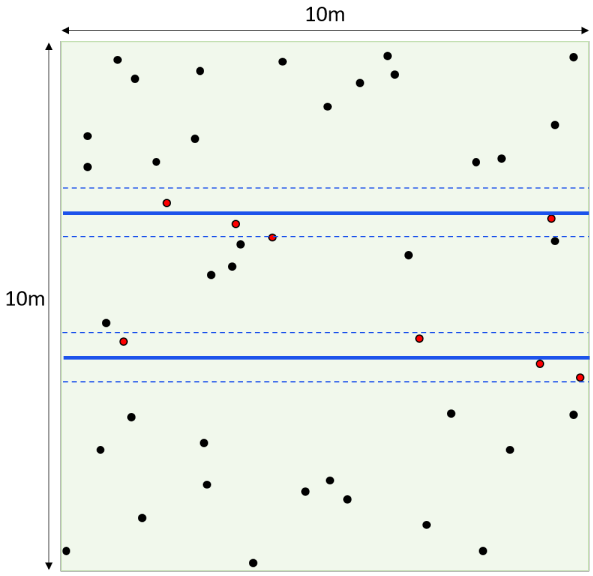


Figure 2.2.4: Example of a $10m$ by $10m$ field. Black dots represent unobserved individuals, red dots represent observed individuals and blue solid lines represent survey transects. Blue dotted lines show the distance observed from the survey transects $(0.5m)$.

transect on both sides, so we have $a = 2wL$. This gives us the formula:

$$\begin{aligned} \hat{N} &= \frac{n}{a} \cdot A \\ &= \frac{n}{2wL} \cdot A. \end{aligned} \tag{2.2.2}$$

Applying this to the method in the example of Figure 2.2.4 we have $l_1 = l_2 = 10cm$ and so $L = 20cm$; $A = 100m^2$ and observed individuals $n = 8$. We calculate the observed area:

$$a = 2 \times 0.5 \times 20 = 20m^2.$$

Using Formula 2.2.2, produces the estimated population abundance of:

$$\hat{N} = \frac{8}{20} \cdot 100 = 40.$$

In this case, $\hat{N}$ is the same estimate as in the quadrat sampling example from Section 2.2.1.1 but this may not always be the case. As a result, we can conclude

that this is another reasonable estimate of the true total population density of $N = 45$. Multiplying both sides of Formula 2.2.2 by the total plot area, A, again gives us $\hat{D} = 0.4$ which was also produced in the previous example of Section 2.2.1.1 using quadrat sampling.

2.2.1.3 Line Transect Sampling

Line transect sampling [Williams et al., 2002] is similar to strip transect sampling (Section 2.2.1.2) in the sense that the observer travels along a line detecting individuals using survey transects which are placed either at random or using a systematic random design. However, the difference is that we do not set an observed distance w from the transect where all individuals must be seen – this allows for a proportion of individuals present within a distance of w to be missed. Instead, we record any observed individuals along with their perpendicular distance from the transect (this could be any reasonable distance within the observed area). Line transect sampling is typically more efficient than strip transect sampling for sparsely distributed objects [Buckland et al., 2001]. A visual representation of this method can be seen in Figure 2.2.5.

We denote the n individual observed distances by $x_1, x_2, ..., x_n$. If recording the perpendicular distance at the time of sighting is not possible, then the radial distance can be recorded along with the angle from the transect to the sighting and then later the perpendicular distance can be calculated using trigonometry.

The probability of detecting an individual decreases the further from the transect the observer is. We are able to estimate population abundance and density by using an effective half-width strip, μ, which is the distance from the observation line such that the same number of individuals are missed before μ as that of those detected beyond μ. An example of this can be seen visually in Figure 2.2.6. It is assumed that all individuals on the transect line are observed, i.e. full detectability at distance zero, and so μ can be thought of as detecting only a

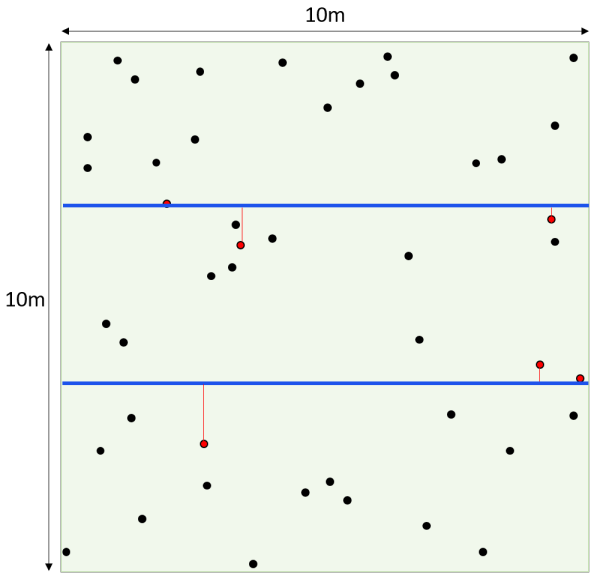


Figure 2.2.5: Example of a $10m$ by $10m$ field. Black dots represent unobserved individuals, red dots represent observed individuals and blue solid lines represent survey transects. Red lines represent the perpendicular distance from the survery transect to the individual observed.

proportion of individuals within a strip length L of width w either side. We call this the detection probability, P_a, where

$$\mu = w \cdot P_a.$$

Comparing this theory to that of strip transect sampling in Section 2.2.1.2, we can see that w from Formula 2.2.2 can now be substituded with wP_a – allowing for a proportion of individuals to be missed. Producing the formulae:

$$\begin{aligned} \hat{N} &= \frac{n}{2\mu L} \cdot A \\ &= \frac{n}{2wL\hat{P}_a} \cdot A \end{aligned} \tag{2.2.3}$$

and

$$\hat{D} = \frac{n}{2wL\hat{P}_a}$$

which are the estimated population abundance, $\hat{N}$, and density, $\hat{D}$.

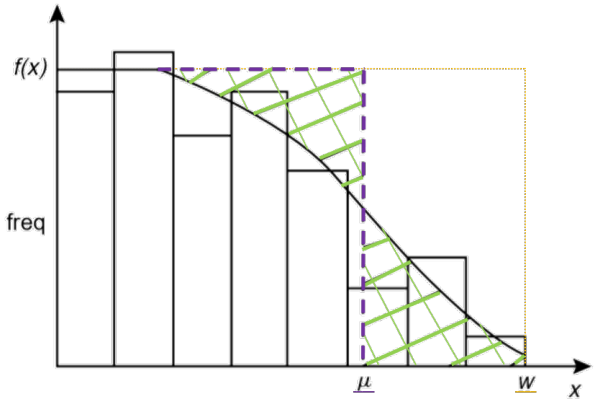


Figure 2.2.6: Diagram illustrating the effective half-width strip, μ (purple dashed), compared to the full area covered width, w (orange dotted) of a probability density function $f(x)$. Green hashed areas represent the equal number of individuals missed before and beyond the effective half-width strip. Reproduced and adapted from Thomas et al. [2002].

2.2.1.4 Point Sampling

Point sampling is similar to line transect sampling (Section 2.2.1.3) except the observer stands static at a single point for a set length of time – rather than moving along a transect. Usually, there are k points either randomly located or using a systematic random design, instantaneous observations are then made around each of these points and radial distances, r_i, are recorded for $i = 1, 2, ..., n$ observations. Similar to line transect sampling, there is only a full detection probability at distance zero from the observer and so as a result there is potential for some individuals to be overlooked. An example of this can be seen in Figure 2.2.7 – we have $k = 2$ points and a total of $n = 6$ observations. Point sampling is also sometimes referred to as 'point transect sampling' as it may be considered as a line transect of length zero, i.e. a point, however we will go onto discuss how these two sampling theories do differ from one another [Buckland et al., 2001].

We can use the radial distances recorded to estimate an effective radius, ρ, similar to the use of μ in Section 2.2.1.3. Length ρ represents a border where the number of individuals missed closer than the border equals the number observed past the border. Figure 2.2.6 can also be a visual representation for this method

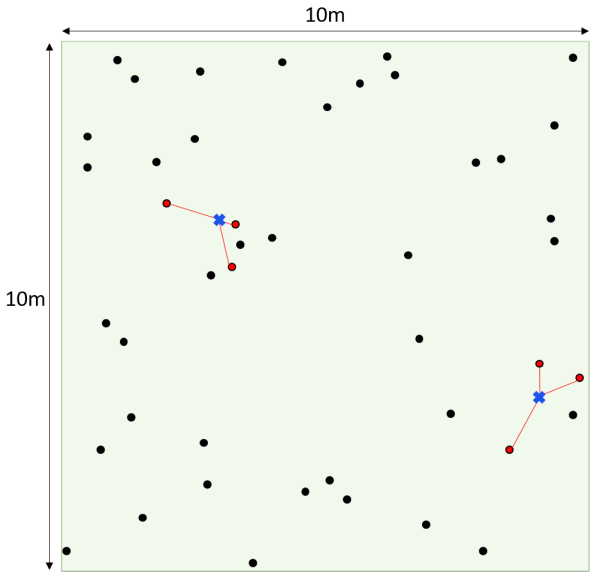


Figure 2.2.7: Example of a $10m$ by $10m$ field. Black dots represent unobserved individuals, red dots represent observed individuals and blue crosses represent observer locations. Red lines represent the distance from the observer to the individual observed.

if μ was replaced with ρ. The main difference between these two methods is that in this case we are looking at radial distances and so the total observed area is $a = k\pi\rho^2$. Substituting this into Formula 2.2.1 with the estimated value of ρ, $\hat{\rho}$, produces:

$$\hat{N} = \frac{n}{k\pi\hat{\rho}^2} \cdot A.$$

We can also think of ρ as the expected proportion of individuals detected (P_a) within radius w which we estimate using $\hat{P}_a$. Therefore, we can substitute $\rho = \sqrt{w^2\hat{P}_a}$ in the above equation to create an equivalent formula for the estimated population abundance:

$$\hat{N} = \frac{n}{k\pi w^2\hat{P}_a} \cdot A. \tag{2.2.4}$$

Both of these equations would produce the same value. The estimated probability density function for point sampling is

$$\hat{D} = \frac{n}{k\pi w^2\hat{P}_a}.$$

In point sampling, we are looking for the radial distance from the observer. On average, the number of animals available for detection will linearly increase with larger distances. This is due to the increase in area covered. However, the further the distance from the monitor, the less likely the monitor is to detect an individual. This 'give and take' idea produces a curve which is typical to point sampling in shape, starting at zero followed by a sudden incline and then slowly dropping off as the furthest distance approaches.

2.2.2 Models for Probability of Detection

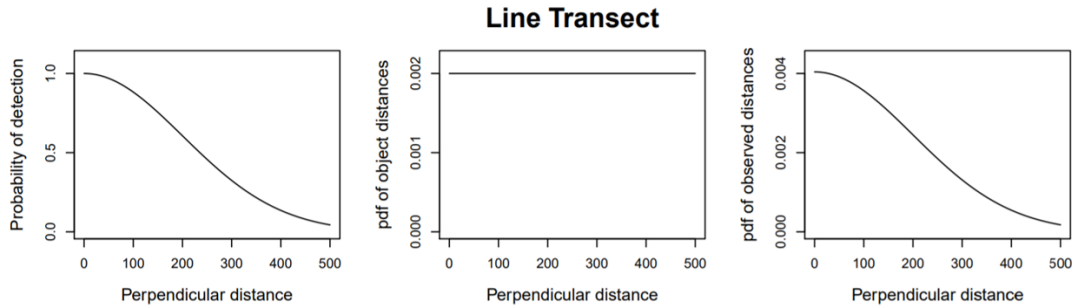
This Section discusses suitable models for the relationship between the detection probability, P_a, and the observed distances. The detection function, $g(y)$, represents the probability of detecting an object, given that it is a distance y from the random point or line – distance y being either radial (r) or perpendicular (x) [Buckland et al., 2001]. This can also be written as the following formula:

$$g(y) = Pr(\text{detection}|\text{distance}\, y).$$

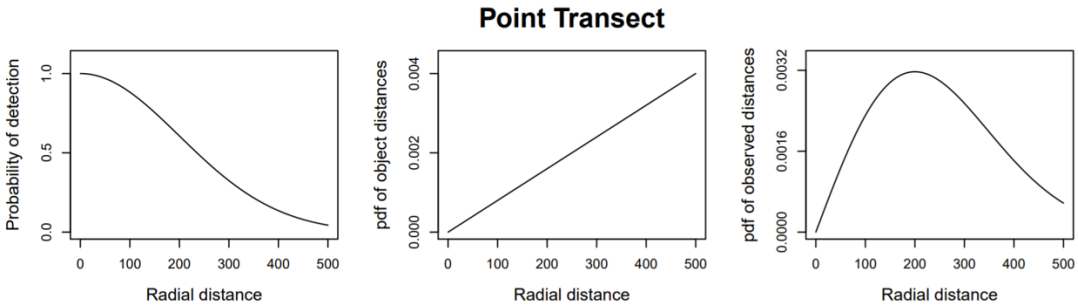
In general, histogram bars are scaled and so the function is generally increasing, $0 \geq g(y) \geq 1$ and is assumed that $g'(0) = 0$ and $g(0) = 1$ where $g'(0)$ is the gradient at point $y = 0$, i.e. perfect detectibility at distance 0. It can also be used to calculate an estimate of the detection probability, P_a, using the following formula for line transect sampling:

$$\hat{P}_a = \frac{\text{Area under curve}}{\text{Area of rectangle}} = \frac{\int_0^w \hat{g}(y)dy}{w}. \tag{2.2.5}$$

Visually, we can see how Formula 2.2.5 is formed in Figure 2.2.6 of Section 2.2.1.3 if we ignore μ annotations. Later in this thesis, we will go on to apply this to Point Sampling methods.



(a) Line Transect



(b) Point Transect

Figure 2.2.8: [Miller et al., 2019] Comparison of 'desirable' probability density functions, $f(y)$, for line and point sampling.

In reality, many factors may affect detectibility and so it is possible to include covariates in the detection function which may include factors such as species or gender – potentially increasing precision of the detection function estimate. This is called *multiple covariate distance sampling*. We assume that covariates affect the scale of the key function and not its shape, as a result we choose key functions with a scale parameter σ - which we will go on to discuss later in this Section.

Line transects (Section 2.2.1.3) have a true uniform distribution of individuals whereas the number of individuals available for detection, on average, will increase linearly for detection at larger distances for point sampling (Section 2.2.1.4). However, the percentage of how many individuals we would actually detect would decrease with distance – this is the detection function, $g(y)$. Multiplying the true distribution of animals by the detection function produces two very different looking 'desirable' observed distributions – this is called the *probability density*

function (pdf), for line and point sampling. Examples of pdf's can be seen visually in Figure 2.2.8. The pdf is the probability of observing an individual between distances y and $y + dy$, given that it was observed somewhere in $(0, w)$. The histogram bars are then scaled so that the area under the pdf, $f(y)$, is 1, i.e.

$$\int_0^w f(y)dy = 1.$$

Another useful advantage of the probability density function is that it provides another way in which we can estimate the detection probability, P_a. This relationship is discussed in Buckland et al. [2001], however, the software Distance can calculate this for you.

Distance data can be truncated (discarding of larger distances) prior to data analysis. Calculating the level of truncation, w, in practice can be difficult when some information is unknown, however, the importance lies in the smaller observed distances to the monitor which determine the detection functions shoulder at $g(0) = 1$ and therefore the area of the 'rectangle' – see Formula 2.2.5. This results in a near perfect detection close to the line/point and not just at distance 0 as well as estimators generally tending to perform better [Williams et al., 2002]. Williams et al. [2002] also states that robust modelling of the dection function leads to robust modelling of the probability density function – which we will go on to discuss later in this Section. As a 'rule of thumb', Buckland et al. [2001] suggest that typically, line transects should be truncated at around 5% and point transects at around 10%. So as a result, it is also important that we consider different levels of truncation, i.e. is truncating 5 or 10 percent reasonable for this particular data set?

When planning the analysis of data, before model selection can begin, it can be useful to explore the data that we have and any potential covariates that we may like to add to the model [Buckland et al., 2001]. This can include visually

on a map, assessing the shape of data collected in the form of a histogram and by exploring any potential patterns in covariates using box plots. This could include looking for outliers or, for example, the differences in particular years which we will see applied later in Section 3.1.1.1.

[Buckland et al., 2001] states that the modelling process can be thought of in two steps: choosing a 'key function' and then a 'series expansion'. As discussed in Burnham et al. [1979, 1980], the final model should fulfill a criteria for robust estimation which consists of these summarised four main points to ensure a good fitting model:

- *Model robustness* – use a model that will fit a wide variety of suitable shapes for the detection function, $g(y)$.
- *Pooling robustness* – use a model for the average detection function as not every individual has the same detection probability, P_a, due to various factors affecting detectibility (such as weather).
- *Shape criterion* – use a model with a 'shoulder' in the detection function, i.e. $g'(0) = 0$ and $g(0) = 1$.
- *Estimator efficiency* – use a model that will lead to a precise estimator of density.

There are four commonly used key functions, which we will breifly discuss, these include: uniform, half-normal, hazard-rate and negative exponential. These key functions are all available in the software package Distance [Thomas et al., 2010]. The uniform distribution,

$$h(y) = \frac{1}{w} \qquad \text{for } 0 \ \leq \ y \leq w \, ,$$

has 0 parameters and satisfies the shape criterion. However, it is not flexible to fit different types and shapes of data so is not model robust. The half-normal

function,

$$h(y) = \exp\left(\frac{-y^2}{2\sigma^2}\right) \qquad \text{for } y \leq w\,,$$

has one parameter, σ, which effects the scale of the curve. It satisfies the shape criterion but is not a robust model as it always assumes the same shape regardless of the scale of the curve. The hazard-rate function,

$$h(y) = 1 - \exp\left\{-\left(\frac{y}{\sigma}\right)^{-\beta}\right\} \qquad \text{for } y \leq w\,,$$

has two parameters, σ and β, which effect the scale and shape respectively. This is the only function from these four that satisifies both the shape criterion and model robustness critera. The final model is the negative exponential,

$$h(y) = 1 - \exp\left(\frac{-y}{\sigma}\right) \qquad \text{for } y \leq w\,,$$

which has only the one scale parameter, σ. This model does not fulfill the shape criterion as it is not flat at $g(0) = 1$ and is not a robust model as the curve can only take one shape (the negative exponential does not contain a shape parameter, β). Note that although the software Distance has the option of the negative exponential, as this model does not satify the shape criterion, we do not go on to discuss this model further and is not used in later data applications of Section 3.

The second step is selecting a series expansion – also referred to as adjustments

Key Function	Form	Adjustment Series	Form
Uniform	$1/w$	Cosine Simple polynomial	$\sum_{j=1}^{m} a_j cos(j\pi y/w)$ $\sum_{j=1}^{m} a_j (y/w)^{2j}$
Half-normal	$\exp(\frac{-y^2}{2\sigma^2})$	Cosine Hermite polynomial	$\sum_{j=2}^{m} a_j cos(j\pi y/w)$ $\sum_{j=2}^{m} a_j H_{2j}(y/\sigma)$
Hazard-rate	$1 - \exp(-(\frac{y}{\sigma})^{-\beta})$	Cosine Simple polynomial	$\sum_{j=2}^{m} a_j cos(j\pi y/w)$ $\sum_{j=2}^{m} a_j (y/w)^{2j}$

Table 2.2.1: [Miller et al., 2019] Modelling options for key plus adjustment series models for the detection function with adjustments of order m.

terms or series adjustments. Adding a series to a model can make it much more robust. The number of adjustment terms need to be chosen and then the final model is scaled so that $g(0) = 1$. Distance allows cosine, Hermite polynomial and simple polynomial adjustment terms. Possible modelling options for key and adjustments of order m are given in Table 2.2.1 - parameters are estimated using maximum likelihood which can be calculated by using the computer package Thomas et al. [2010]. The key to this second step is selecting just the right number of parameters to ensure that the model has an improved fit but also to ensure that it is not too flexible that it also describes the random noise in the data. This can also be described as a trade-off between bias and variance [Buckland et al., 2001]. The correct series length is often determined by comparison of AIC for each degree, followed by another comparison of AIC for each model to determine the best fitting model from those tested. However, the best selected model may still not be a good fit to the data and so a goodness-of-fit test should also be carried out such as the Cramer-von Mises test. See Burnham and Anderson [1998] for more details on critea that model selection should satify and methods that allow selection between fitted models. We also note that it is not typical to fit both adjustment and covariate terms to a model but, when proceeded with caution, is possible Miller et al. [2019]. Therefore, we will not go onto discuss how the two can be combined but further information can be found in Buckland et al. [2004].

2.2.3 Assumptions

Below we detail the assumptions for distance sampling.

- All models assume that all individuals that are on the line or point are detected, i.e. $g(0) = 1$, however this is may not always be the case. Similar to how we assume full detectibility in quadrat sampling and strip transect sampling.

- We also assume that individuals are detected at or close to their initial location, prior to any movement in response to the observer.

- Rounding errors in measurements tend to lead to data being grouped to some degree, but these must be analysed as if the data was recorded accuratley or grouped further to try to reduce the effects of rounding bias. So we must assume that distances are measured accuratley for un-grouped distance data or that individuals are correctly allocated to distance intervals for grouped data.

- Point sampling must be recorded instantaneously and as a result is more subject to bias than line transect sampling. The count of individuals and their distance from the point cannot always be instantaneous particulary when there are multiple individuals around at one given time.

- When conducting observations, the individuals that are being observed are spatially distributed according to some stochastic process with rate parameter, D (number per unit area). Meanwhile, randomly placed lines or points are surveyed and a sample of objects are detected, measured and recorded. It is therefore not necessary that individuals be randomly distributed (i.e. Poisson), but instead must be placed randomly with respect to the distribution of individuals.

[Buckland et al., 2001]

2.2.4 Elephant Example

Various monitors based around Goalpara and Sonitpur recorded radial sightings of elephants in relation to their home location and so for this example we will be using point sampling methods from Section 2.2.1.4. In this example we will look at just one monitor chosen at random, Monitor32, and so we have $k = 1$ points. Each

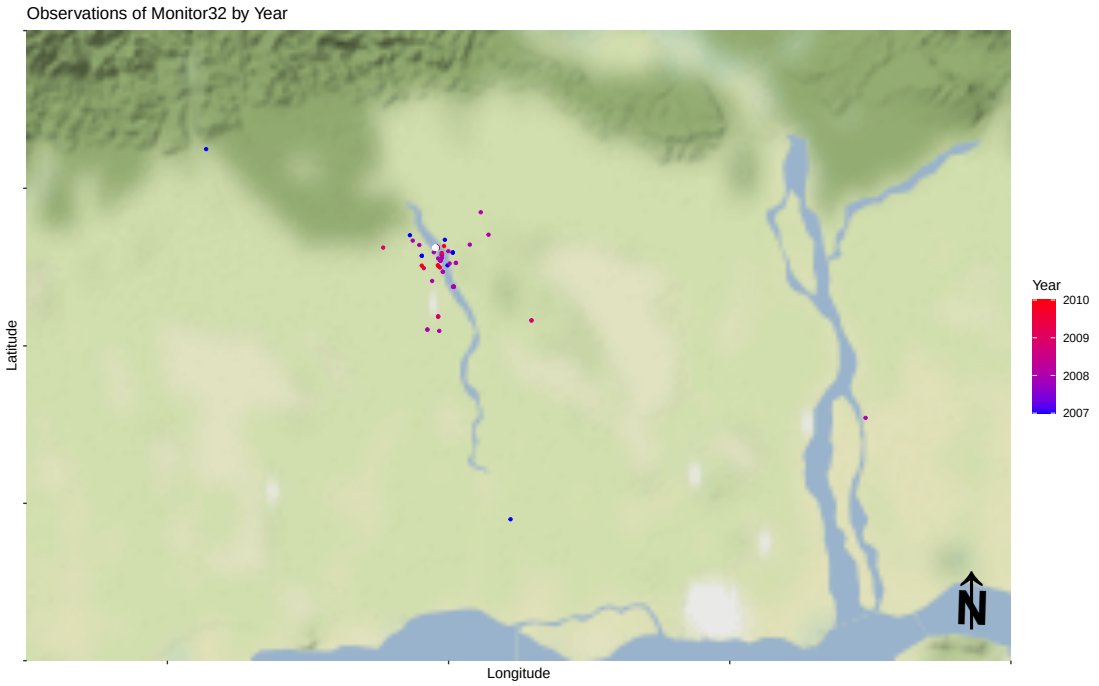


Figure 2.2.9: Visual representation of Monitor32's observations on a map. The white dot is Monitor32's location, observations by year are on a scale of blue (2007) to red (2010). Tick intervals: latitude=0.1 and longitude=0.2.

elephant group sighting by this monitor has a location which can be seen in Figure 2.2.9. Figure 2.2.10 shows the distance from the monitor to the sighted group of elephants (in meters) of all $n = 71$ observations made by Monitor32. Here, we will class an elephant 'group' as a sighting of elephants greater or equal to one. In the histogram there are several outliers. Using the Buckland et al. [2015] 'rule of thumb' for truncation, as stated in Section 2.2.2, we choose to truncate the upper 10% of the data as this is the recommendation for point transect sampling. The truncated data can be seen in Figure 2.2.11. We make the assumption to set this distance as the effective radius ρ from the monitor, where the number of individuals missed closer than the border equals the number observed past the border (discussed in Section 2.2.1.4). Let $\rho = 5779\text{m} = 5.779\text{km}$.

Some assumptions of distance sampling are violated in this data, including the

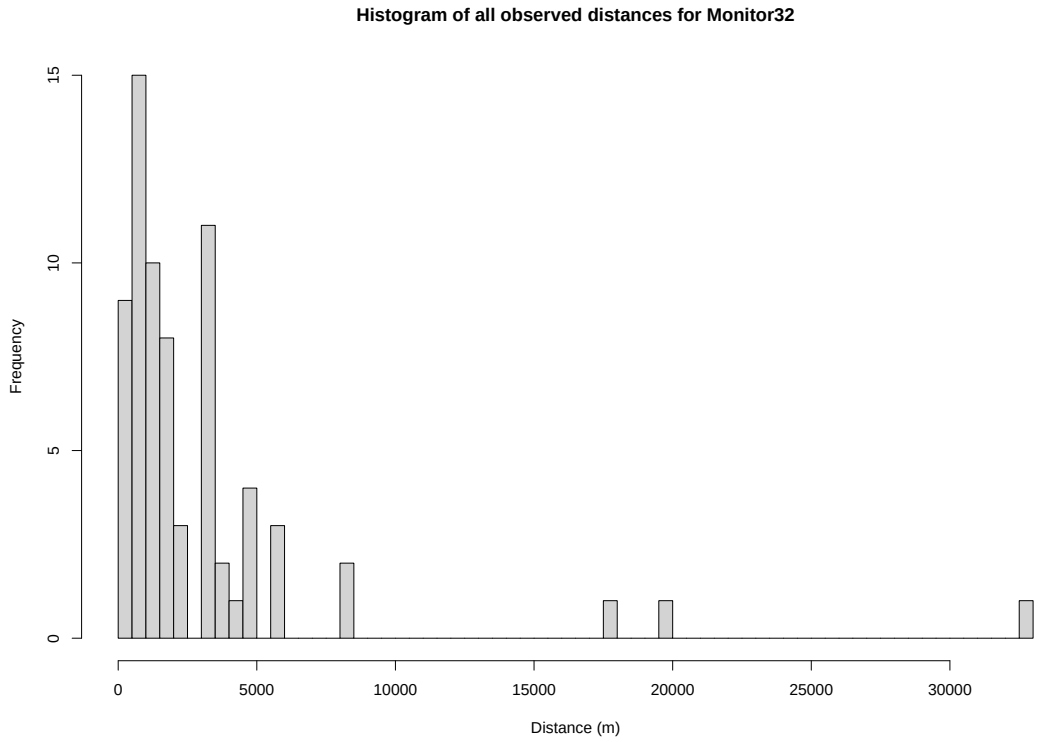


Figure 2.2.10: Histogram of Monitor32's observations.

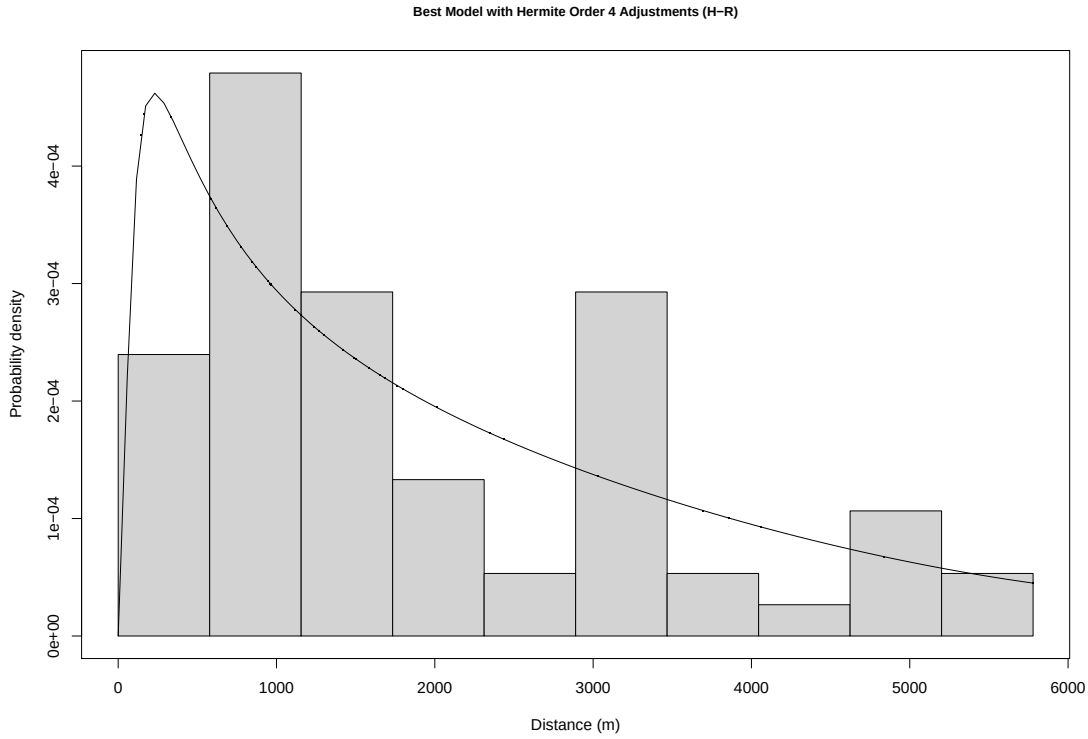


Figure 2.2.11: Histogram of Monitor32's truncated observations at 10% with the estimated probability density function, $\hat{f}(x)$, (line) using the hazard-rate distribution with Hermite adjustments of order 4.

assumption that recordings are made instantaneously over a set time period and potentially the assumption that all elephants at a distance of zero from the monitors house have been detected (see Section 2.2.3). For this illustrative example, we suppose that they have not been violated and understand that this may mean that the results are not accurate. We then estimate the density by substituting the above values into Equation 2.2.4 and rearranging for $\hat{N}/A = \hat{D}$.

$$\begin{aligned}\hat{D} &= \frac{n}{k\pi\rho^2} \\ &= \frac{71}{1 \cdot \pi \cdot 5.779^2} \\ &= 0.677.\end{aligned}$$

If assumptions were met, we could conclude that there is an approximate density of 0.677 elephant groups per km^2 in the surrounding region of Monitor32.

2.3 Generalised Linear Models in Ecology

Standard linear regression models assume that the dependent variable follows a normal distribution, however in many ecological settings this is not appropriate [Zuur et al., 2009]. In this section we discuss how GLMs can be applied to ecological data, considering standard generalised linear models in Section 2.3.1; zero truncated GLMs in Section 2.3.2; generalised linear mixed models in Section 2.3.3 and look at an applied elephant example in Section 2.3.4.

2.3.1 Standard Generalised Linear Models

Generalised Linear Models (GLM) are an extension of classic linear models – also known as regression models. In classic linear models we have

$$y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + ... + \beta_p x_{pj} + \epsilon_j$$

where y_j is the dependent variable (also called the response variable), x_{kj} are covariates, β_j are unknown parameters, $k = 1, ..., p$, $j = 1, ..., n$ and ϵ_j is the error term which follows $N(0, \sigma^2)$. This means that y_j is normally distributed with mean $E(y_j) = \mu_j = \beta_0 + \beta_1 x_{1j} + ... + \beta_p x_{pj}$ and variance σ^2.

In GLMs, the dependent variable is assumed to follow an exponential family distribution (such as a Poisson or gamma distribution) which has the form

$$f(y_j; \theta_j, \phi) = \exp\left(\frac{y_j\theta_j - a(\theta_j)}{b(\phi)} + c(y_j, \phi)\right),$$

for $j = 1, ..., n$. Where θ_j is the canonical parameter which represents the location, ϕ is a scale parameter and a, b and c are known functions. A link function, $g(\cdot)$ is used to transform the mean response $E(y) = \mu$ and a linear combination of the covariates: $\eta = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$. This can be written as

$$g(\mu) = \eta = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p \tag{2.3.1}$$

[Faraway, 2006; McCullagh and Nelder FRS, 1999]. The canonical link function is $g(\mu) = \theta$.

Common exponential family models, the type of data they are used for and the canonical link functions are:

- Normal – continous and symmetric data which can take on any value.

  Identity: $g(\mu) = \mu$

- Poisson – discrete data (counts) with no fixed upper limit.

  Log: $g(\mu) = \log(\mu)$

- Binomial – discrete data with a fixed upper limit.

$$\text{Logit:} \quad g(\mu) = \log\left(\frac{\mu}{(1-\mu)}\right)$$

- Gamma – continuous data that can take on any value greater than zero.

$$\text{Inverse:} \quad g(\mu) = \frac{1}{\mu}$$

There are two types of covariates: continuous and factors. For example, count of elephants observed would be a continuous covariate and a true/false mitigation would be a factor. Note that covariates used cannot be correlated with one another. For a specific exponential family distribution, we use maximum likelihood to estimate the parameters – in this thesis we use the R function 'glm' to fit generalised linear models.

The best covariates for a data set can be determined by a range of model selection techniques such as AIC or Bayesian information criterion (BIC). In practice, the only difference between these two techniques is the extent in which parameters and therefore model complexity is penalised - where BIC penalises more heavily for larger, more complex models. More can be read about the assumptions and performance of AIC and BIC in Kuha [2004]. Alternatively two nested models (a model that is a subset of another) can be compared using likelihood ratio, Wald and score tests or by comparing deviance. The R package 'MASS' [Venables and Ripley, 2002] contain functions which perform an automated stepwise model selection, we will go on to discuss this later in Section 4.1.

Examples of cases when GLMs were applied to ecological data include tuberculosis in wild boar [Vicente et al., 2006] (Figure 2.3.1) and parasites in cod [Hemmingsen et al., 2005].

Figure 2.3.1: Wild boar (©Getty images).

2.3.2 Zero Truncation

When fitting a GLM (Section 2.3.1) it is sometimes the case that the count data that we are trying to model cannot have a value of zero but is from a distribution which would usually include values of zero. For example: in medical data, the duration of patients' visits to the hospital emergency department [Karaca et al., 2012] (Figure 2.3.2); in ecological data, dolphin and porpoise group size [Gygax, 2002] (Figure 2.3.3) is an example of data which cannot have the response value of zero and therefore, a *zero truncated* Poisson or negative-binomial distribution would be an appropriate distribution for this type of data. We give an illustrative example in Section 2.3.4 for our AHP elephant data, the dependent variable is the count of elephants sighted however, it is not possible to have a sighting of zero elephants and so the data takes the form of a zero truncated Poisson distribution.

Figure 2.3.2: Hospital emergency department (©Stuart Harrison).

Figure 2.3.3: Bottlenose dolphin pod (©Louise Murray).

In this Section we will briefly discuss zero truncated Poisson and negative-binomial distributions, more information regarding how zero truncated distributions are calculated can be found in Zuur et al. [2009]. We first adjust the probability functions to exclude the probability of a zero observation for the Poisson:

$$f(y_j; \mu_j | y_j > 0) = \frac{\mu^{y_j} \cdot \exp(-\mu_j)}{(1 - \exp(-\mu_j)) \cdot y_j!}$$

and negative-binomial distribution:

$$f(y_j; \mu_j | y_j > 0) = \frac{\Gamma(y_j + k)}{\Gamma(k)\Gamma(y_j + 1)} \left(\frac{k}{\mu_j + k}\right)^k \left(1 - \frac{k}{\mu_j + k}\right)^{y_j} \Big/ \left(1 - \left(\frac{k}{\mu_j + k}^{y_j}\right)\right).$$

The log-likelihood for the zero truncated Poisson distribution is:

$$\log(L) = \sum_j \log\left(\frac{\mu^{y_j} \cdot \exp(-\mu_j)}{(1 - \exp(-\mu_j)) \cdot y_j!}\right)$$

and the zero truncated negative-binomial distribution is:

$$\log(L) = \log(L_{NB}) - \log\left(1 - \left(\frac{k}{\mu_j + k}\right)^k\right)$$

which is a function of the regression parameters.

Various softwares exist in order to fit these models in R. For example, the package 'VGAM' [Yee and Moler, 2020] contains the 'vglm' function which, for example, fits the zero-truncated Poisson model using the family input positive Poisson where '$family = pospoisson()$'. This is a strictly positive Poisson distribution so will not include any zeros in the Poisson distribution. The function fits a very flexible class of models called vector generalised linear mondels (VGLMs) to a wide range of assumed distributions (see Yee [2015] for more details).

2.3.3 Generalised Linear Mixed Models

Fixed effects are constant across individuals, where as random effects vary. This allows for correlation of the data in addition to the usual fixed effects, i.e. there is often more than one source of random variability in the data [Harrison et al., 2018]. Fixed effects are unknown constants across individuals that we try to estimate from the data, whereas random effects are random variables. Faraway [2006] describes randoms effects as not something that we try to estimate, but instead we try to estimate the parameters that describe the distibution of this random effect. A generalised linear model (Section 2.3.1) with random effects is known as a *generalised linear mixed model* (GLMM).

Let $\mathbf{y}$ be a vector of observations and $\mathbf{r}$ be a vector of random effects. The probability density conditional on the random effects, $f(\mathbf{y}|\mathbf{r})$, is an exponential family model. The random effects $\mathbf{r}$ have probability function $g(\mathbf{r})$. Similar to Equation 2.3.1, the linear combination of covariates is

$$\eta_{ij} = \beta_0 + \beta_1 x_{1ij} + ... + \beta_p x_{pij} + r_i$$

where $k = 0, 1, ..., p$ is the covariate number, i is the random effect number, j is the observation for the ith random effect and added random effect – frequently $b_i \sim N(0, \sigma^2)$. Fixed effect parameters are $\beta_0, \beta_1, ..., \beta_p$ and the random effect parameter is σ^2. Calculating an explicit solution for GLMMs are not possible (unless we assume a Normal distribution for both dependent variable and for random effects) and so to estimate the parameters we need to find:

$$f(\mathbf{y}) = \int f(\mathbf{y}|\mathbf{r}) g(\mathbf{r}) d\mathbf{r}.$$

R packages exist which calculate these estimated parameters, including 'lme4' [Bates et al., 2015] and 'glmmTMB' [Brooks et al., 2017]. It also still applies that

GLMMs cannot have correlated covariates.

Here we use the Likelihood Ratio Test (LRT) to test whether the random effect parameter is needed or not and then the best model can be determined by using AIC or BIC to select the best model between GLMs and GLMMs. A suggested strategy is to fit the best GLM model (Section 2.3.1) and then complete a single LRT to compare the best model with and without the random effect.

2.3.4 Elephant Example

The Assam Haathi Project recorded counts of elephants alongside mitigations that were put into place to try to reduce human-elephant conflict, such as electric fences. In this example, we are interested in how the number of elephants sighted is effected by which mitigation was used, here the count of elephants is treated as the dependant variable and the mitigations are the covariates. For ease of analysis, mitigations were grouped and analysed by type: Sound, Fire, Visual, Physical and Other. Each of these groups consisted of the following covariates:

- Sound – Cracker, Noise, Drum tin, Siren, Tripwire
- Fire – Fire stick, Chillismoke
- Visual – Torch light, AHP spotlight, Other spotlight
- Physical – Catapult, Chillifence, eFence, Arrows
- Other – Kunkie, Other mitigiation

For descriptions of each mitigation please see Chapter 1, note, 'Watchtower' was not included as this would have been a correlated mitigation (villages would use the watchtower to sight the elephants and then use another form of mitigation to attempt to deter them). Distance in meters from the observed elephants to the monitors location was also included as a covariate, calculated by R package

Model Name	K	AIC	ΔAIC	BIC	ΔBIC	Log-lik
FDSVOP	7	89853.29	0.00	89900.07	0.00	-44919.65
FDSVO	6	89885.36	32.07	89925.45	25.38	-44936.68
FDSV	5	89918.97	65.67	89952.37	52.30	-44954.48
FDS	4	89956.47	103.17	89983.20	83.13	-44974.23
FD	3	90518.63	665.33	90538.67	638.60	-45256.31
F	2	91282.87	1429.58	91296.24	1396.17	-45639.44
Constant	1	92918.43	3065.13	92925.11	3025.04	-46458.21

Table 2.3.1: Comparison of stepwise GLM regression from each stage, ranked by AIC and BIC. Best fitting model is 'FDNVOP' which includes all covariates: Fire, Distance, Sound, Visual, Other and Physical. Models were abbreviated to the single first letter of each type of mitigation, for example Fire is F and Visual is V and so FV would be the model consisting of Fire and Visual mitigations. Listed for each model: K, number of parameters (K=$p+1$); the AIC; the AIC difference, ΔAIC; the BIC; the BIC difference, ΔBIC; and the log-likelihood, Log-lik.

'Geosphere' [Hijmans et al., 2019] which we will go on to discuss in Chapter 3. There were no recordings of zero elephants sighted and so this particular data requires the use of a zero-truncated Poisson where we used the R package VGAM – as discussed in Section 2.3.2.

To calculate the best fitting Generalised Linear Model for this data we used AIC (Section 2.3.1), starting with the constant model and comparing this to all possible single covariate models. The best selected model at this stage was Fire which can be seen in Table 2.3.1. This process was then repeated, using the best model and then comparing this to all possible models with unselected covariates added. As there were only six covariates present, we were able to code each model and compare AIC values by hand, however later in Section 4.1 we go on to use a stepwise regression instead – R package [Venables and Ripley, 2002]. Note, this particular package is not compatible with the R package VGAM.

By AIC and BIC, the best fitting model to the data was the model with all covariates present, 'FDSVOP', with log-likelihood of -44919.65. Table 2.3.2 shows the summary for this model. We can see that when Sound and Physical mitigations were used there was a reduction in the number of elephants seen,

whereas when Fire, Visual and Other mitigations were used there was an increase in the number of elephants sighted. We can also conclude that the more elephants sighted, the smaller the distance between the observed elephants and the monitors location.

Coefficients	Estimate	Std. Error	z-value	p-value
(Intercept)	2.44	0.0070	347.29	< 0.0001
Fire (TRUE)	0.46	0.013	36.65	< 0.0001
Sound (TRUE)	-0.34	0.016	-21.36	< 0.0001
Visual (TRUE)	0.10	0.015	6.49	< 0.0001
Other (TRUE)	0.14	0.021	6.78	< 0.0001
Physical (TRUE)	-0.056	0.0096	-5.82	< 0.0001
Distance	-0.000008	0.00000039	-20.42	< 0.0001

Table 2.3.2: Results from the best fitting model 'FDSVOP' which includes all covariates: Fire, Distance, Sound, Visual, Other and Physical. Listed for each model: estimate, standard error, z-value is the test statistic for a hypothesis test of whether the coefficient value is zero; and p-value is the probability that the z-value is non-zero.

Chapter 3

Probability of Detection of Elephant Herds

In this Section, we talk about how distance sampling methods from Section 2.2 were applied to data from the Assam Haathi Project. The aim of this analysis was to discover the probability that elephants were detected – given that they were present – by the monitors and nearby villagers, and so we will particularly look at the estimated probability of detection ($\hat{P}_a$). We note for this section that although some assumptions were violated, we acknowledge that there are data limitations and so investigating potential factors (such as observations being skewed towards water sources) which may have caused a bias was not possible. This could be investigated in future work if monitors recorded further information relating to these variables.

The projects data is in the form of longitude and latitude, therefore to apply distance sampling methods we needed to calculate the distance from the monitor who made the observation to the observed location of elephants. The R package 'Geosphere' [Hijmans et al., 2019] was used to make these calculations with the default distance value in meters. The Geosphere package describes the function 'distVincetyEllipsoid' as calculating the shortest distance between two

points along the surface of a sphere i.e. the great-circle-distance, using an ellipsoid (a quadratic surface) to create very acurate results [Hijmans et al., 2019]. An example of the code used can be found in B.0.0.2 of Appendix B.

3.1 Individual Monitors

Figure 3.1.1 shows the number of monitor observations for individual monitors. Some monitors do not have enough data to be considered inidividually, therefore in this Section we consider and apply distance sampling methods to four monitors – two from each region (Goalpara and Sonitpur) – to obtain probabilities of detection. Monitor01, Monitor03, Monitor16 and Monitor46 were selected as they had a large range of observations over most years and therefore were more likely to have consistent recordings. Analysing these monitors alone allowed for comparison between the sites, the individual monitors, and then a comparison to the overall model with all monitors included. The analytical methods carried out were the same for all monitors – let us first discuss Monitor16 (Section 3.1.1).

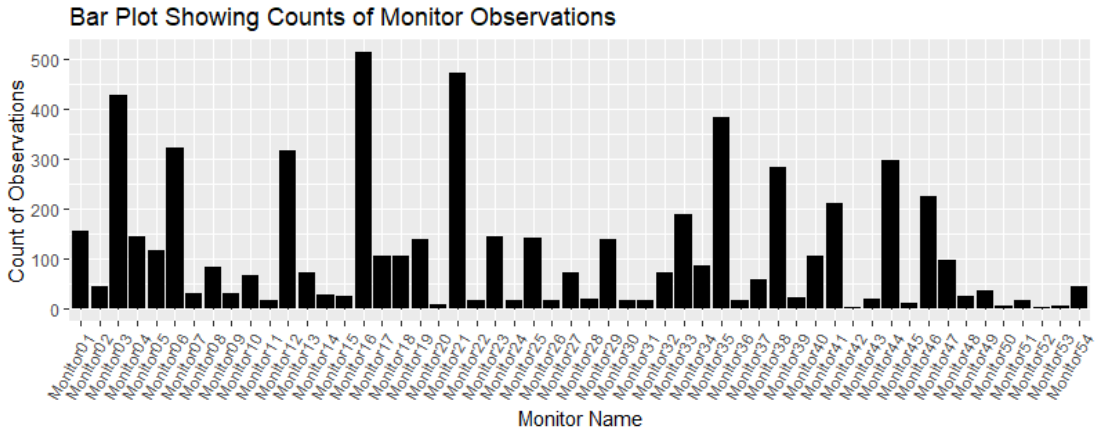


Figure 3.1.1: Histogram showing the frequency of sightings for all monitors.

3.1.1 Probability of Detection for Monitor16

Monitor16 was located in Sonitpur with a total of 512 data entries with year of sighting ranging from 2005 to 2014. See Figure 3.1.2 for a visual representation of these observations by year on a map.

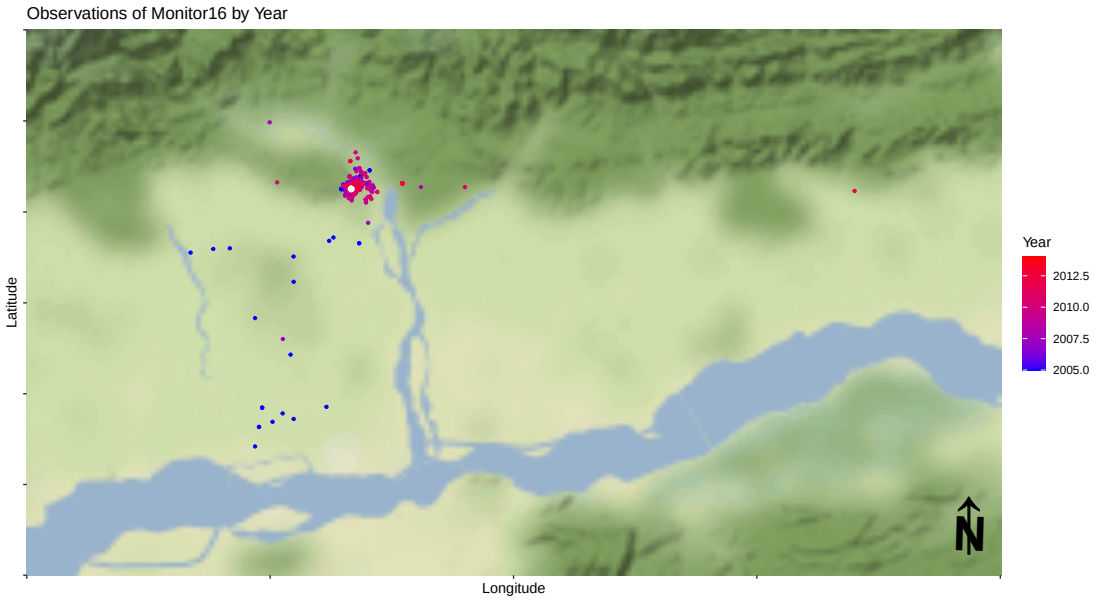


Figure 3.1.2: Visual representation of Monitor16's observations on a map. The white dot is Monitor16's location, observations by year are on a scale of blue (2005) to red (2014). Tick intervals: lat=0.1 and lon=0.3.

3.1.1.1 Exploratory Data Analysis

In Figure 3.1.2 it can be seen that the majority of observations are close to the monitors location which is what we would expect to see in point sampling (Section 2.2.1.4). We can see an expected scatter of observations with a higher density closer to the monitors location, however, we note that observations appear to be skewed towards the right near where there is a water source.

Figure 3.1.3 is a histogram of the whole data set for Monitor16. This shows

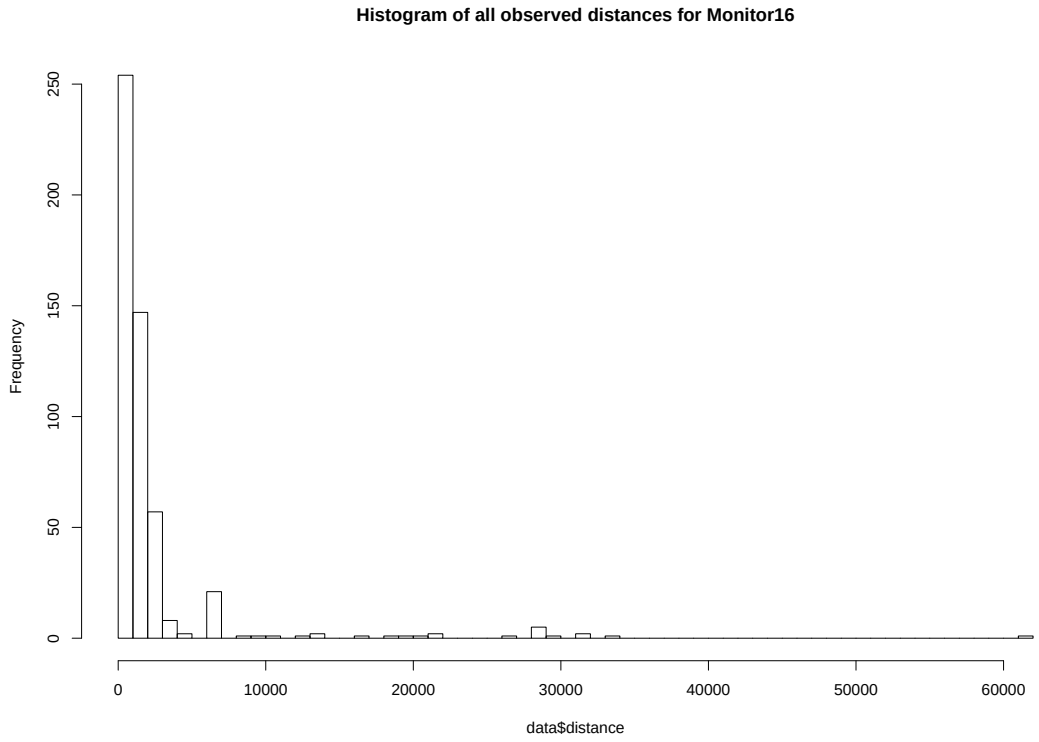


Figure 3.1.3: Histogram showing the frquency of all observed sightings by distance (m) for Monitor16.

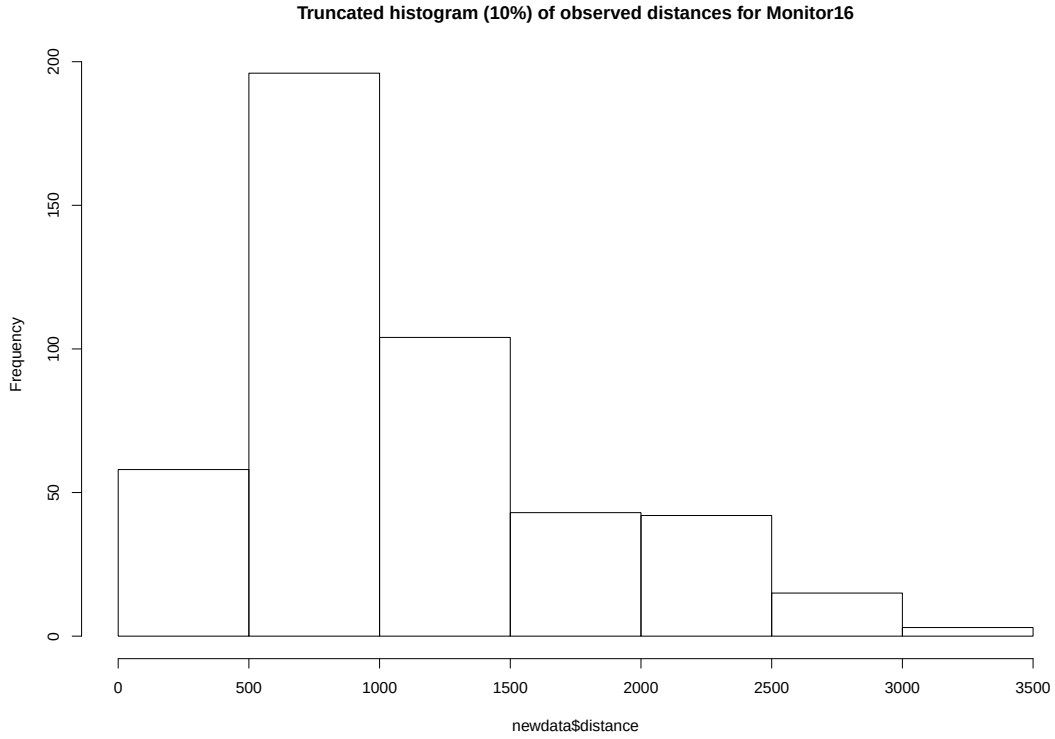


Figure 3.1.4: Histogram showing the frquency of sightings by distance (m) for Monitor16 (10% truncation).

that there are many outliers, a large distance from the monitors location. In this example, we are trying to estimate the probability of detection by solving Formula 2.2.4 from Section 2.2.1.4 where

$$\hat{P}_a = \frac{n}{k\pi w^2 \hat{N}} \cdot A.$$

An estimate of $\hat{N}$ would be biased and so as a result we cannot precisely approximate w. However, the importance lies in the smaller observed distances to the monitor which determine the detection functions shoulder as discussed in Section 2.2.2. Here we use the recommendation from Buckland et al. [2001] and truncate at 10%, this recommendation is used for all monitors as we consider the radial observed distances which are associated with point sampling methods. Truncating at this level results in a new sample size of 461 observations.

Figure 3.1.4 shows that our maximum observed distance has reduced from approximately $60,000m$ in the full data set (Figure 3.1.3) to around $3,500m$. This has removed any possible outliers and improved the shape of our histogram dramatically to resemble that of a typical point sampling example. As discussed in Section 2.2.1.4, this shape is typical of the observed distribution from point sampling data where the highest frequency of observations do not lie in the closest bin to the monitor, but instead lie usually in the next region.

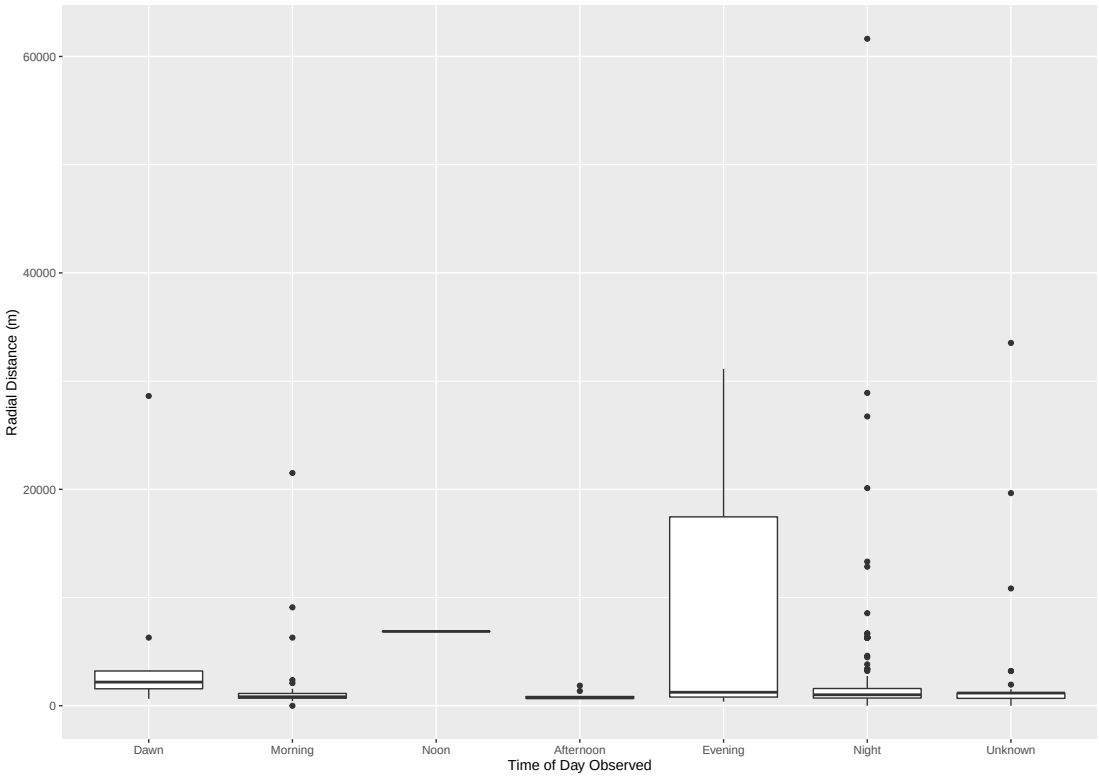
The next step is to explore potential patterns in covariates. In a similar study by Wilson et al. [2015] a diurnal pattern across a 3 year study showed incidents of crop loss or property damage caused by elephants was observed in Goalpara and Sonitpur with peak times ranging between 18:01-22:00. The same study also discovered a seasonal pattern, mentioning how the pattern differed in 2006 due to severe flooding in Assam – this impeded the movement of elephants into both study areas and also disrupted the usual agricultural calendar. For individual Monitors we looked into three covariates: time of day (TOD), year (Year) and

season (Szn), all of which were treated as discrete data. The year covariate allowed for variations year by year for cases such as 2006 when Assam had severe flooding. Each data entry relates to an observation made in a village at a specific time. As a result of data limitations we have TOD categories that include dawn, morning, noon, afternoon, evening, night, whole day and unknown. Seasons were classed by pre-monsoon (April-June), monsoon (July-Sept), post-monsoon (Oct-Dec) and winter (Jan-March).

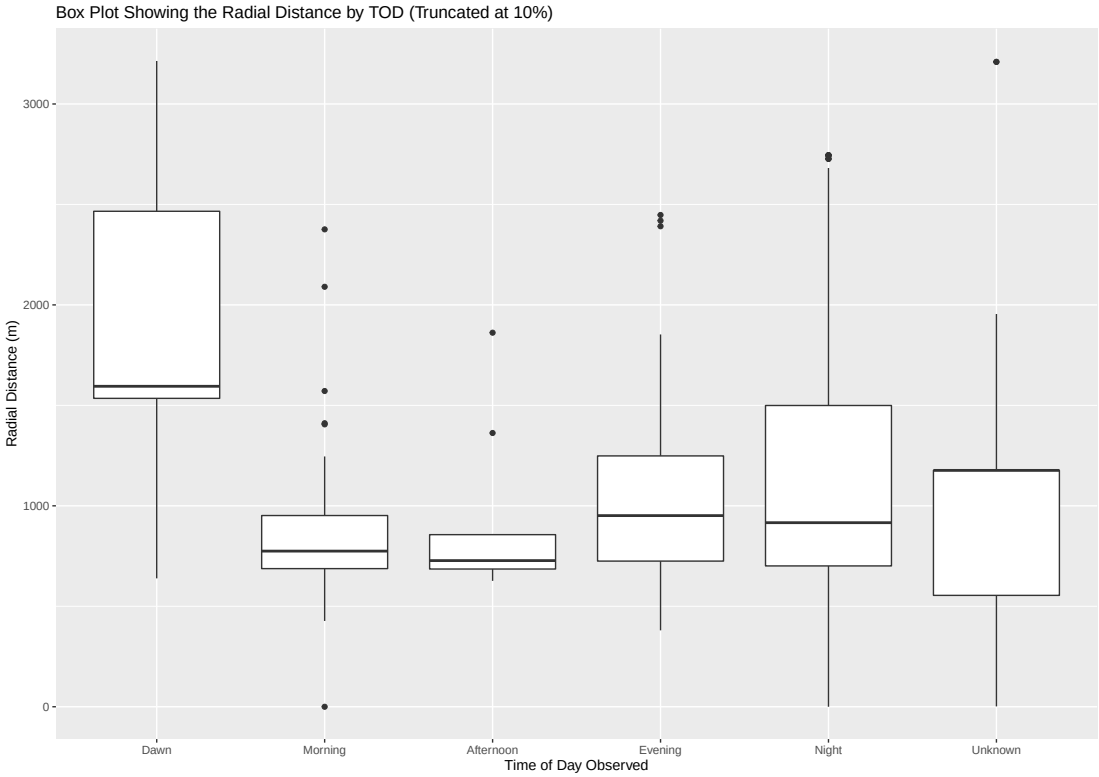
Boxplots were used to explore potential patterns of covariates TOD, Year and Szn as seen in Figure 3.1.5a, Figure 3.1.6a and Figure 3.1.7a respectively. 10% truncated versions of the box plot were also produced for ease of interpretation (Figure 3.1.5b, Figure 3.1.6b and Figure 3.1.7b).

Looking at the truncated data in Figure 3.1.5b for time of day, we can see that there is some evidence to suggest that during day-light hours (morning, noon and afternoon) elephants were observed with a shorter range of distances – in particular around noon where Monitor16 made no observations of elephants. In all observed categories except dawn, we observe a similar lower limit of the box (first quartile), however the first quartile of observations at the time of dawn are around twice the distance of the rest. Comparing the full data set to the truncated version, it is clear to see that the evening observations seem to be most effected.

Monitor16 only made one observation in the year 2014 which can be seen in Figure 3.1.6b – the reason for this is unknown, for example the monitor may have stopped working for the project. The monitor has conistently observed elephants at a similar range of distances for most years 2005-2013, however we note 2010, 2011 and 2013 have a smaller range. The project as a whole has a noticeably large number of entries for the year 2005, it was suggested that this was due to an increased effort from the Monitors for this first year in which the project was launced. As a result, we have also chosen to include a covariate for observations made in the year '2005' vs 'not 2005'.

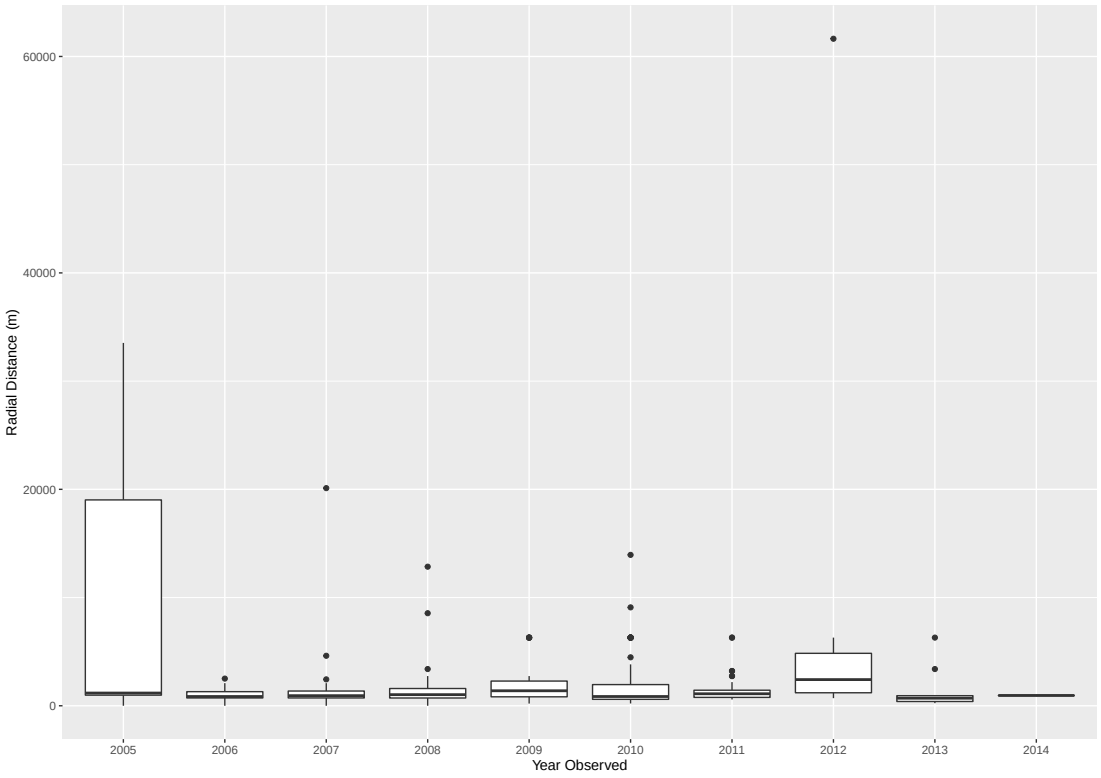


(a) Time of day (TOD)

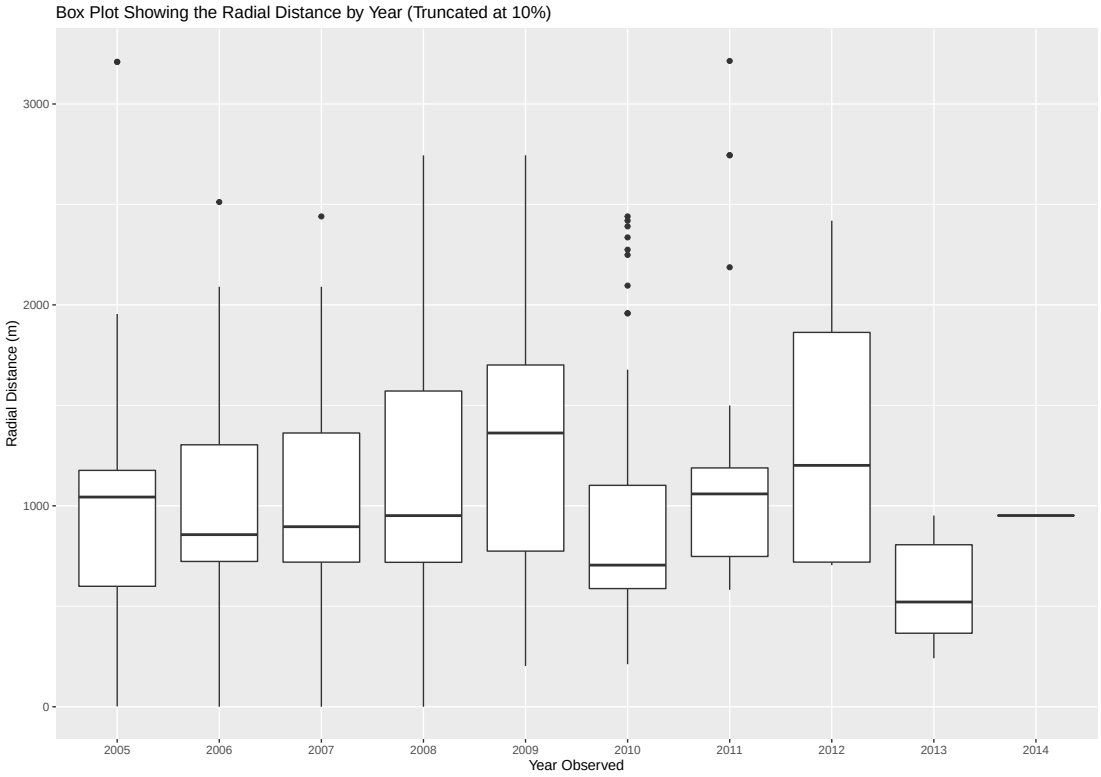


(b) TOD, 10% truncation

Figure 3.1.5: Comparison of truncated boxplot showing the radial distance of observed sightings by covariates for all monitors.

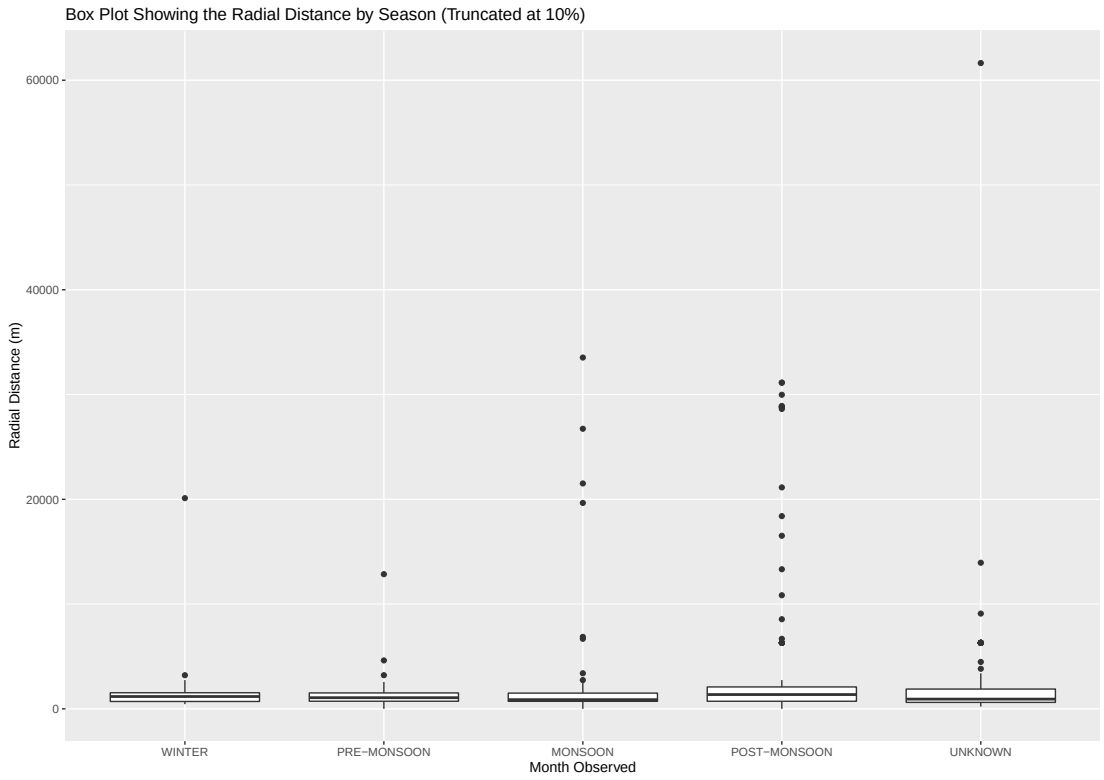


(a) Year

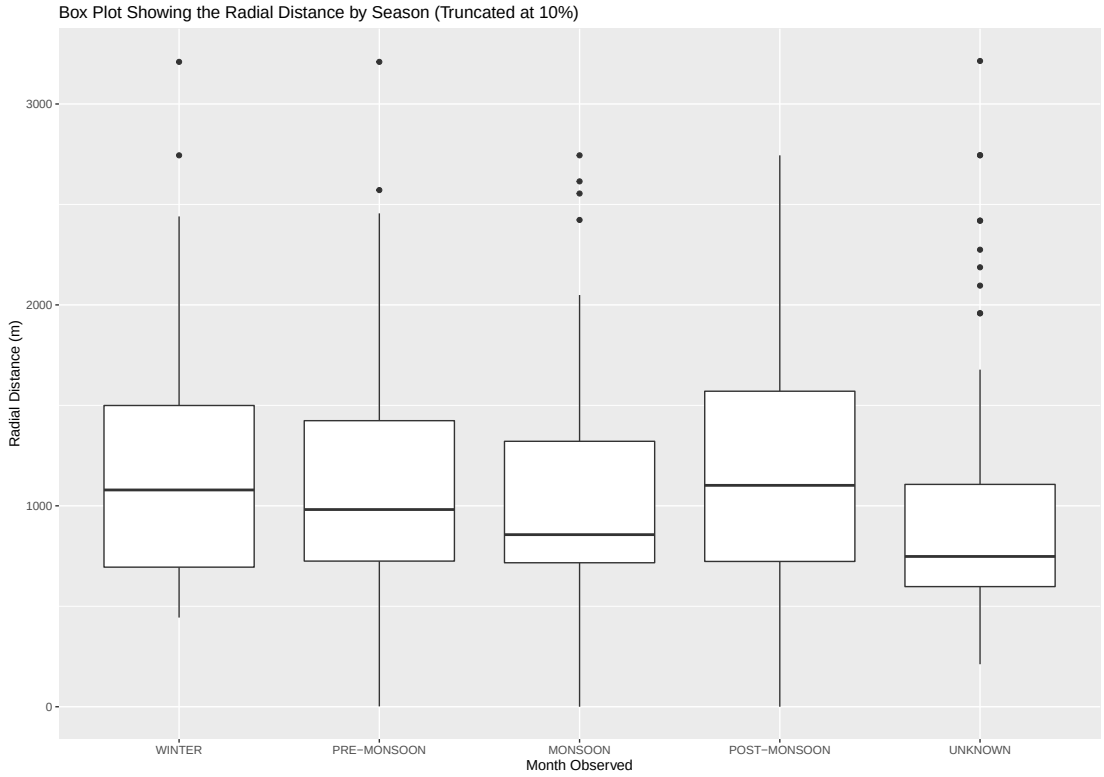


(b) Year, 10% truncation

Figure 3.1.6: Comparison of truncated boxplot showing the radial distance of observed sightings by covariates for all monitors.



(a) Season (Szn)



(b) Szn, 10% truncation

Figure 3.1.7: Comparison of truncated boxplot showing the radial distance of observed sightings by covariates for all monitors.

Seasons appears to have a more consistent spread of observation distances across each part of the year.

3.1.1.2 Best Model with Adjustments

As explained in Section 2.2.2, the detection function, $g(x)$, describes the relationship between probability of detection and distance. First, we start off by fitting a model to the data and then choose the best model by AIC. We fit a variety of models without covariates to the data, which are summarised in Table 3.1.1. We used the 'Distance' R package [Thomas et al., 2010] to fit these models. The models are then compared using AIC. The ΔAIC is given in Table 3.1.1. Note that for each key-function and adjustment type the Distance R package determines the best order of adjustments using AIC. It is also important to check that the best model is also a good fit to the data by using a goodness of fit test, in this case, Cramer-von Mises was used. Table 3.1.1 gives the p-value for this test. Taking into account the criteria for a robust model estimation from Section 2.2.2, models were fitted using hazard-rate, half-normal and uniform distributions with adjustments that included: cosine, simple polynomial and Hermite as shown in Table 3.1.1. Note, the model with key function half-normal and Hermite adjustment failed to fit.

In Table 3.1.1, the best fitting model according to AIC was '$Mon.unif.cos$'. The model is composed of the Uniform distribution with cosine adjustment terms of order 3 however although this is the best fitting model from those tested, it is not a good fit to the data by the Cramer-von Mises goodness of fit test. As the model has a p-value $= 0.0425 < 0.05$ we would reject the hypothesis that the model is a good fit to the data. Both the detection function and probability density function estimates can be seen in Figures 3.1.8 and 3.1.9 respectively.

Observe that estimates of $\hat{P}_a$ – even across the better fitting models tested – are not approximately equal, suggesting that adjustment terms alone do not

Model Name	Key Function & Adjustment	C-vM p-value	$\hat{P}_a$	se($\hat{P}_a$)	AIC	ΔAIC
Mon.unif.cos	Uniform with cosine adjustment terms (order 3)	0.0425	0.1224	0.0063	7454.21	0.00
Mon.unif.poly	Uniform with simple polynomial adjustment terms (order 8)	0.0000	0.1935	0.0128	7493.77	39.56
Mon.hr.poly	Hazard-rate*	0.1389	0.1573	0.0114	7560.44	106.23
Mon.hr.cos	Hazard-rate*	0.1389	0.1573	0.0114	7560.44	106.23
Mon.hr.herm	Hazard-rate*	0.1389	0.1573	0.0114	7560.44	106.23
Mon.hn.cos	Half-normal with cosine adjustment terms (order 4)	0.0858	0.1523	0.0263	7567.67	113.46
Mon.hn.poly	Half-normal*	0.0003	0.1606	0.0066	7596.93	142.72
Mon.unif.herm	Uniform with Hermite polynomial adjustment term (order 4)	0.0407	0.4187	0.0019	7729.04	274.83

Table 3.1.1: Monitor16 Adjustment Model Comparison. Key functions denoted with an asterix (*) denote best models with adjustments of order 0, i.e. no adjustment terms were added to the model, resulting in three identical hazard-rate models. Listed for each model: Cramer-von Mises, C-vM p-value; average detectability estimate, $\hat{P}_a$; standard error of the estimated detectibility, $se(\hat{P}_a)$; AIC; and the AIC difference, ΔAIC.

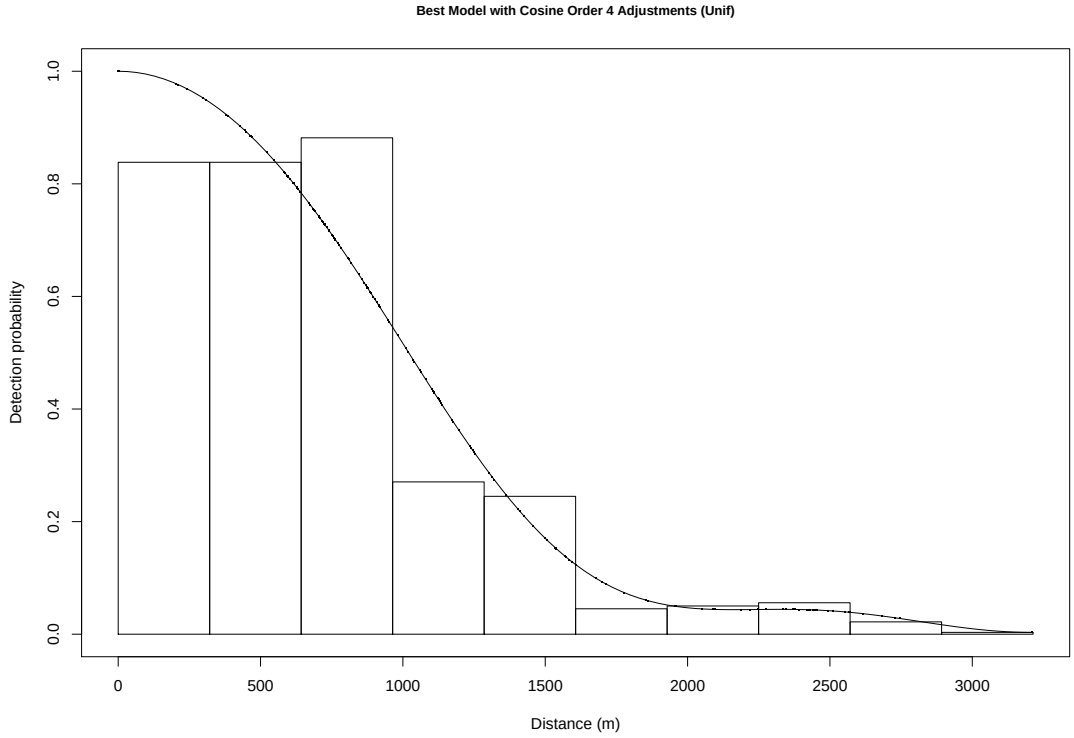


Figure 3.1.8: Histogram of true data with the detection function, $\hat{g}(x)$, (line) for Monitor16 of the best fitting model with adjustment terms – $Mon.unif.cos$.

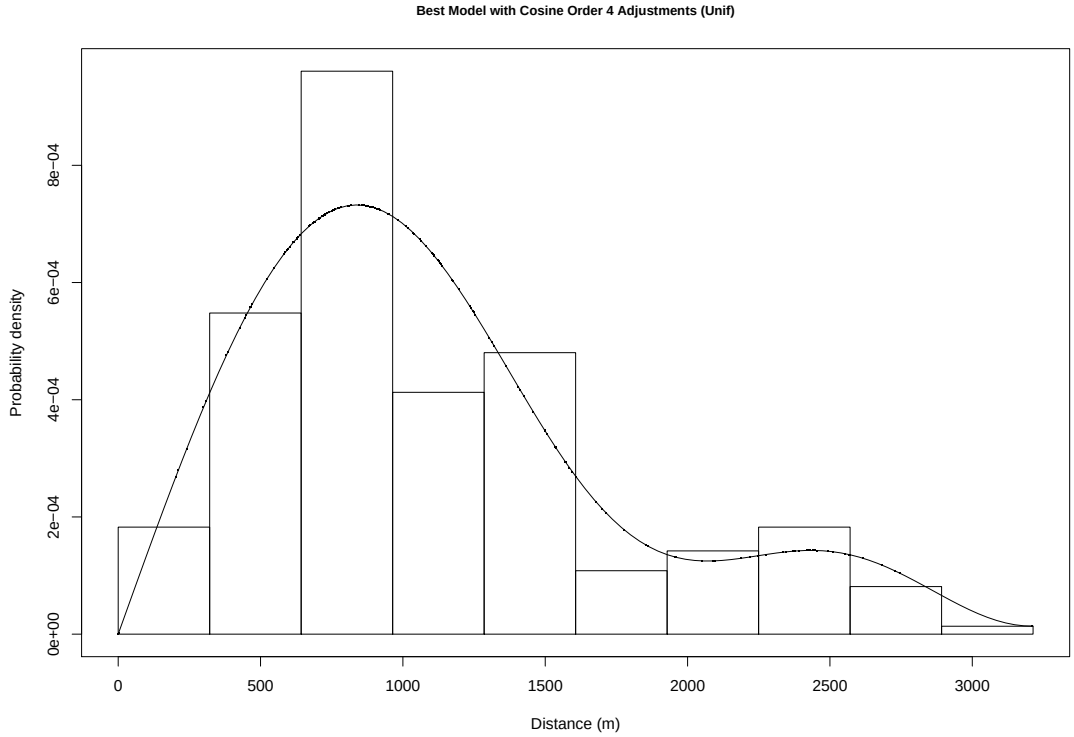


Figure 3.1.9: Histogram of true data with the probability density function, $\hat{f}(x)$, (line) for Monitor16 of the best fitting model with adjustment terms – $Mon.unif.cos$.

appear to produce a good or consist fit to the data.

3.1.1.3 Best Model with Covariates

For covariate models we do not consider adjustments, only different combinations of relevant covariates added to key functions of hazard-rate, half-normal and uniform distributions. Both adjustments and covariates can be added to a model but this is not typical (see Section 2.2.2). Further information can be found in Buckland et al. [2004]. Models are ranked by AIC and the best model is tested as a good fit to the data or not by the Cramer-von Mises goodness of fit test.

All top four models have approximately the same probability of detection ($\hat{P}_a = 16\%$), which can be seen in Table 3.1.2 – this gives us confidence that this is likely to be a well-estimated value for this monitor. The best covariate model fitted to the data is '$hr.SznYearTOD$'. This model consists of the hazard-rate distribution as the key function with Season, Year and TOD as covariates, it has a Cramer-von Mises p-value of 0.4275 which is much greater than 0.05 meaning that we can accept this model as a good fit to our data. The Q-Q Plot in Figure 3.1.10 shows the goodness of fit for this best model.

As a result, we can say that the probability of Monitor16 detecting elephants, given that they are present at that time is 16%. Both the detection function and probability density function estimates from the best fitting covariate model can be seen in Figures 3.1.11 and 3.1.12 respectively. Note that dotted lines in both of these figures represent the individual covariates included in the model. We can also now observe a much flatter, wider and therefore more desirable shoulder in Figure 3.1.11 than that of the best model with adjustments from the previous section in Figure 3.1.8.

In point sampling, we are looking for the radial distance from the observer, we expect to see that the further the distance from the monitor, the less likely the monitor is to detect an individual (Section 2.2.1.4). See Figure 3.1.13a for contour

Model Name	Key Fnct.	Covariates	C-vM p-value	$\hat{P}_a$	se($\hat{P}_a$)	AIC	ΔAIC
hr.SznYearTOD	H-R	Szn+Year +TOD	0.4274	0.1603	0.1514	7523.79	0.00
hr.SznYear	H-R	Szn+Year	0.5115	0.1599	0.1552	7533.86	10.07
hr.YearTOD	H-R	Year+ TOD	0.3702	0.1580	0.1405	7546.87	13.01
hr.Year	H-R	Year	0.3397	0.1561	0.1311	7541.95	18.16
hn.SznYearTOD	H-N	Szn+Year +TOD	0.0287	0.1338	0.4499	7546.69	22.90
hr.Szn05TOD	H-R	Szn+2005 +TOD	0.5358	0.1635	0.0115	7550.18	26.39
hr.SznTOD	H-R	Szn+TOD	0.4925	0.1631	0.0116	7550.19	26.39
hr.Szn	H-R	Szn	0.3039	0.1593	0.0115	7553.17	29.38
hn.SznYear	H-N	Szn+Year	0.0412	0.1381	0.3805	7553.81	30.02
hr.Szn05	H-R	Szn+2005	0.2807	0.1592	0.0115	7554.90	31.11
hr.model0	H-R	1	0.1388	0.1573	0.0114	7560.44	36.65
hr.TOD	H-R	TOD	0.1829	0.0117	0.0263	7560.95	37.16
hn.YearTOD	H-N	Year+ TOD	0.0182	0.1397	0.3481	7561.50	37.71
hr.05TOD	H-R	2005+ TOD	0.2061	0.1603	0.0116	7561.93	38.14
hr.05	H-R	2005	0.1376	0.1573	0.0115	7561.93	38.14
hn.Year	H-N	Year	0.0088	0.1447	0.3034	7568.85	45.06
hn.Szn05TOD	H-N	Szn+2005 +TOD	0.0195	0.1456	0.0072	7570.66	46.87
hn.SznTOD	H-N	Szn+TOD	0.0137	0.1469	0.0068	7571.83	48.04
hn.Szn	H-N	Szn	0.0034	0.1534	0.0066	7582.69	58.90
hn.Szn05	H-N	Szn+2005	0.0049	0.1529	0.0066	7583.05	59.26
hn.05TOD	H-N	2005+ TOD	0.0022	0.1550	0.0074	7592.88	69.09
hn.TOD	H-N	TOD	0.0017	0.1562	0.0068	7593.49	69.70
hn.model0	H-N	1	0.0003	0.1606	0.0066	7596.93	73.14
hn.05	H-N	2005	0.0003	0.1606	0.0067	7598.90	75.11

Table 3.1.2: Mon16 Covariate Model Comparison. Key functions include: hazard-rate (H-R) and half-normal (H-N). Covariates making up the formula include: Year, time of day (TOD), year 2005 only (2005) and no covariate included (1). Listed for each model: Cramer-von Mises, C-vM p-value; average detectability estimate, $\hat{P}_a$; standard error of the estimated detectibility, $se(\hat{P}_a)$; AIC; and the AIC difference, ΔAIC.

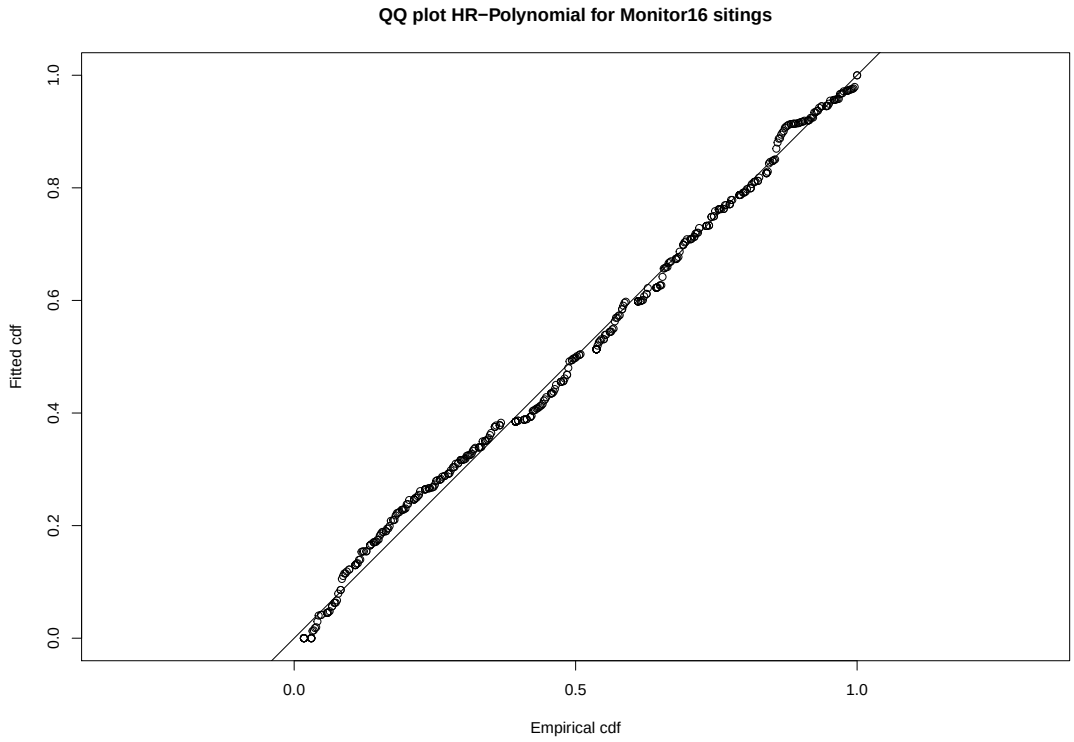


Figure 3.1.10: Q-Q plot for the best fitting model with covariate terms.

plot of Monitor16's truncated observations based on the results and detection probability found in this section against all of Monitor16's sightings and a close up visual in Figure 3.1.13b. The colour red represents the estimated full detection probability, $\hat{P}_a = 1$, found closest to the monitor's location. This is a gradient colour scale which progresses to the colour blue which represents the detection probability of zero, $\hat{P}_a = 0$ (not detectable). As we would expect, the further we move from Monitor16's location, the less likely they were able to detect elephants. This 'give and take' idea produces our curve which is typical to point sampling, starting at zero followed by a sudden incline and then slowly dropping off as the furthest distance approaches, much like our histogram in Figure 3.1.4. As a result, we can confirm that our truncated data follows this shape reassuring us that 10% truncation was a sufficient choice.

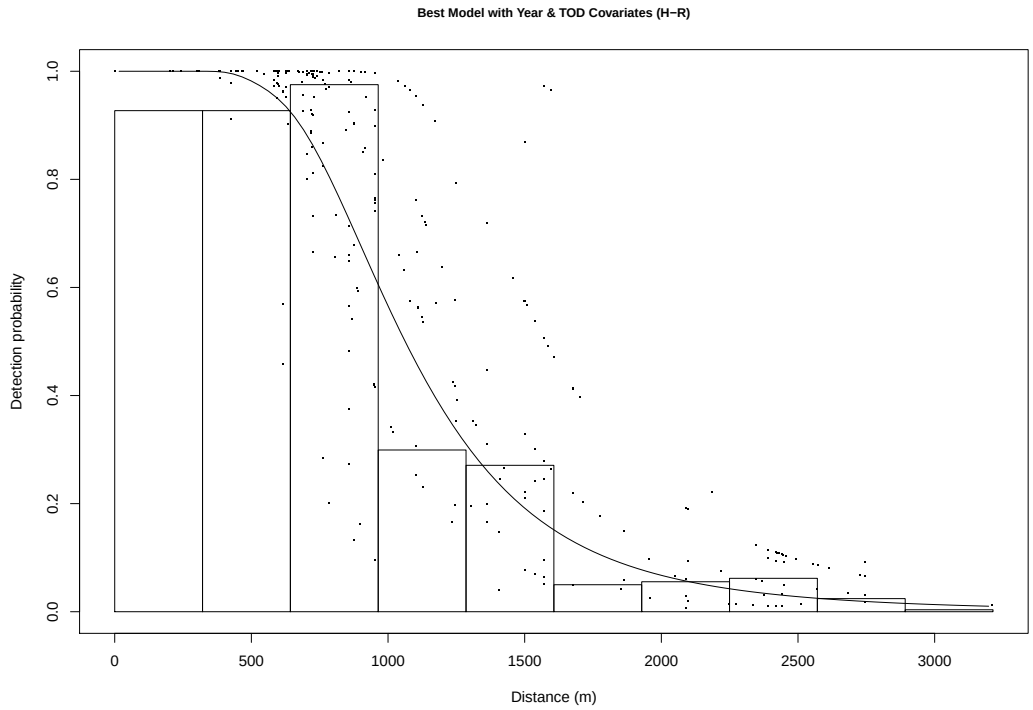


Figure 3.1.11: Histogram of true data with the detection function, $\hat{g}(x)$, (line) for Monitor16 of the best fitting model with covariate terms – $hr.SznYearTOD$. Faint dotted lines represent individual covariates.

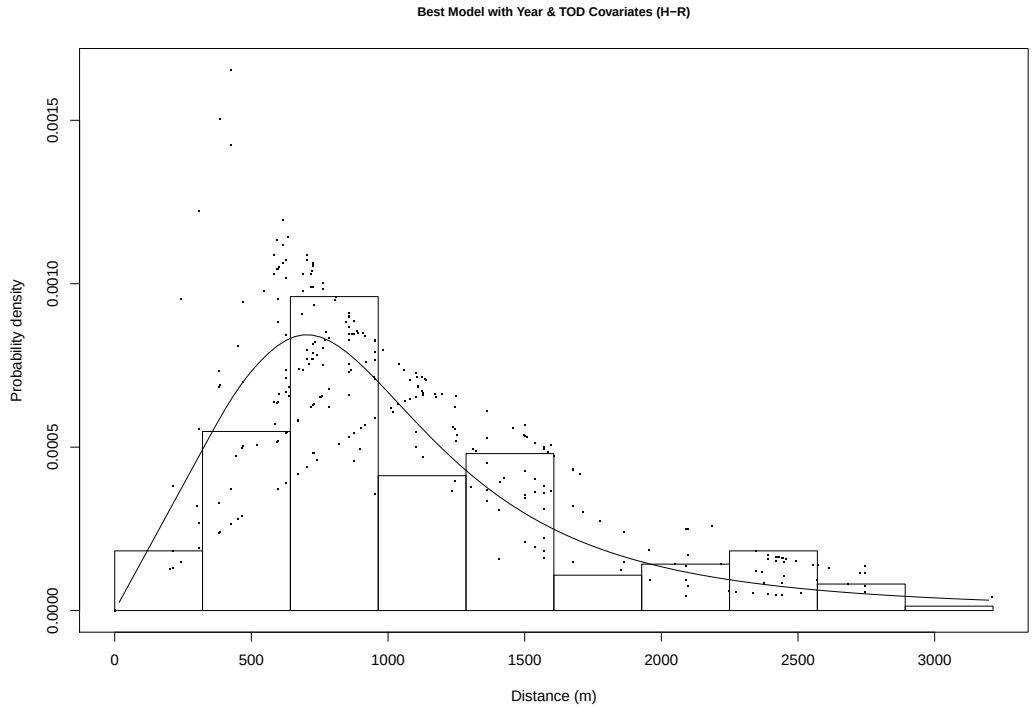
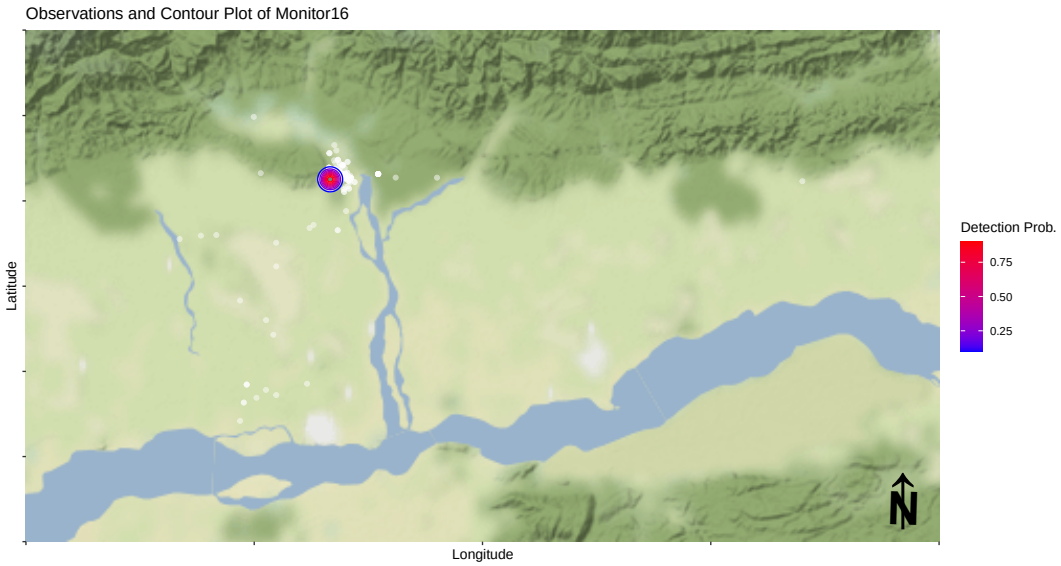
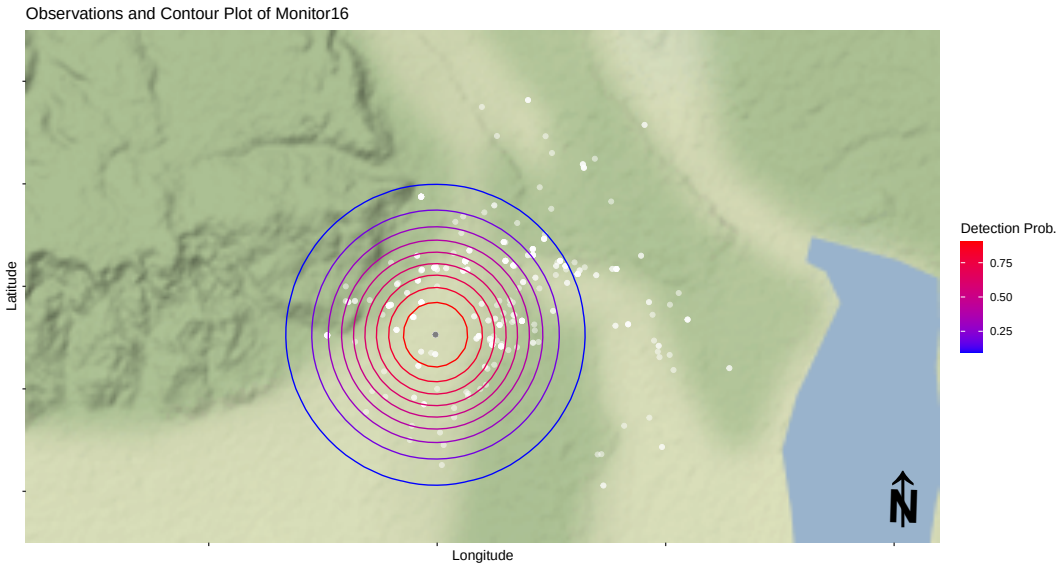


Figure 3.1.12: Histogram of true data with the probability density function, $\hat{f}(x)$, (line) for Monitor16 of the best fitting model with covariate terms – $hr.SznYearTOD$. Faint dotted lines represent individual covariates.



(a) Contour plot with all observations made by Monitor16. Tick intervals: lat=0.1 and lon=0.3.



(b) Homed in view of Monitor16's contour plot without truncated observations. Tick intervals: lat=0.01 and lon=0.025.

Figure 3.1.13: Contour plots of Monitor16 with observations (white), colour scale relates to the detection probability – probability given by the legend to the right of the diagram.

3.1.2 Probability of Detection for Monitor01

Monitor01 was located in Sonitpur with a total of 154 data entries with year of sighting ranging from 2005 to 2014. See Figure 3.1.14 for a visual representation of these observations by year on a map.

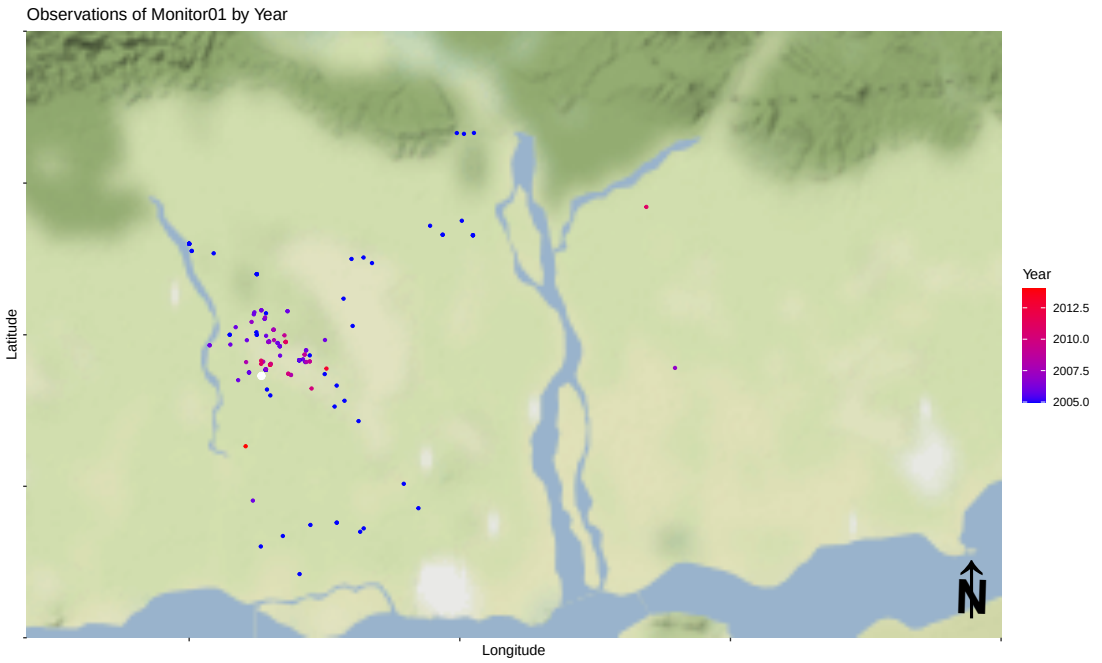


Figure 3.1.14: Visual representation of Monitor01's observations on a map. The white dot is Monitor01's location, observations by year are on a scale of blue (2005) to red (2014). Tick intervals: lat=0. and lon=0..

After exploratory analysis, data was truncated by 10%. We considered the covariates: year, time of day (TOD), the year '2005' vs 'not 2005' (Year05) and season (Szn). All covariates as described in Section 3.1.1.1.

The Distance R package was used to fit a variety of models as in 3.1.1.2 and determines the best order of adjustment models using AIC. The top three models are summarised in 3.1.3 with '$Mon.hr.cos$' as the best model where $\hat{P}_a = 9.5\%$ – composed of the hazard-rate function with cosine adjustments of order two. Table 3.1.3 gives the p-value for the Cramer-von Mises test, $p = 0.2642$. As the model

Model Name	Key Function & Adjustment	C-vM p-value	$\hat{P}_a$	se($\hat{P}_a$)	AIC	ΔAIC
Mon.hr.cos	Hazard-rate with cosine adjustment term (order 2)	0.2642	0.0948	0.0098	2551.18	0.00
Mon.hn.cos	Half-normal with cosine adjustment term (order 2)	0.1602	0.0867	0.0085	2556.36	5.18
Mon.unif.cos	Uniform with cosine adjustment terms (order 4)	0.1714	0.0906	0.0105	2556.51	5.33

Table 3.1.3: Monitor01 Adjustment Model Comparison of top three models. Listed for each model: Cramer-von Mises, C-vM p-value; average detectability estimate, $\hat{P}_a$; standard error of the estimated detectibility, $se(\hat{P}_a)$; AIC; and the AIC difference, ΔAIC.

Model Name	Key Fnct.	Covariates	C-vM p-value	$\hat{P}_a$	se($\hat{P}_a$)	ΔAIC	AIC
hr.SznTOD05	H-R	Szn+TOD +2005	0.1236	0.0824	2.5873	2501.77	0.00
hn.SznTOD05	H-N	Szn+TOD +2005	0.3457	0.0562	7.7730	2507.11	5.34
hn.Szn05	H-N	Szn+ Year05	0.1294	0.0611	8.9796	2507.35	5.58

Table 3.1.4: Mon01 Covariate Model Comparison of the top three models. Key functions include: hazard-rate (H-R) and half-normal (H-N). Covariates making up the formula include: season (Szn), time of day (TOD) and year 2005 only (2005). Listed for each model: Cramer-von Mises, C-vM p-value; average detectability estimate, $\hat{P}_a$; standard error of the estimated detectibility, $se(\hat{P}_a)$; AIC; and the AIC difference, ΔAIC.

has $p > 0.05$ we can accept the hypothesis that the model is a good fit to the data.

The same method was used with the Distance R package for covariate models as with adjustment models (explained in Section 3.1.1.3), the top three models can be seen in Table 3.1.4. The best ranked covariate model by AIC is '$hr.SznTOD05$' with $\hat{P}_a = 8.2\%$. This model consists of the hazard-rate function with season, time of day and year 2005 covariates; it also has a Cramer-von Mises p-value of 0.1236 which is much greater than 0.05 meaning that we can accept this model as a good fit to our data. The Q-Q Plot in Figure 3.1.17 shows the goodness of fit for this best model.

As a result, we can say that the probability of Monitor01 detecting elephants, given that they are present at that given time is 8%. Both the detection function and probability density function estimates from the best fitting covariate model can be seen in Figures 3.1.15 and 3.1.16 respectively. Note that the dotted lines in both of these figures represent the individual covariates included in the model.

See Figure 3.1.18 for a contour plot of Monitor01's truncated observations based on the results and detection probability found in this section against all of Monitor01's sightings. The colour red represents a full detection probability, $\hat{P}_a = 1$, with a gradient progressing to the colour blue which represents the detection probability of zero (not detectable).

3.1.3 Probability of Detection for Monitor03

Monitor03 was located in Goalpara with a total of 428 data entries with year of sighting ranging from 2005 to 2012. See Figure 3.1.19 for a visual representation of these observations by year on a map.

After exploratory analysis, data was truncated by 10% and the covariates chosen to be included in the model selection process were: year, time of day (TOD), the year '2005' vs 'not 2005' (Year05) and season (Szn). All covariates as

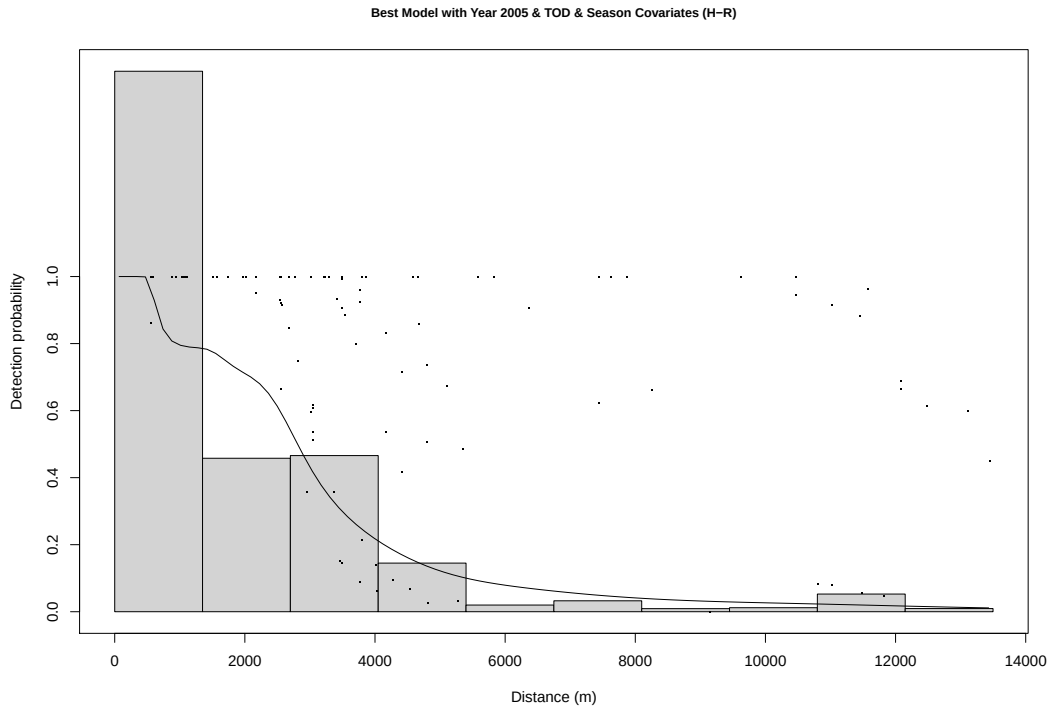


Figure 3.1.15: Histogram of true data with the detection function, $\hat{g}(x)$, (line) for Monitor01 of the best fitting model with covariate terms – $hr.SznTOD05$. Faint dotted lines represent individual covariates.

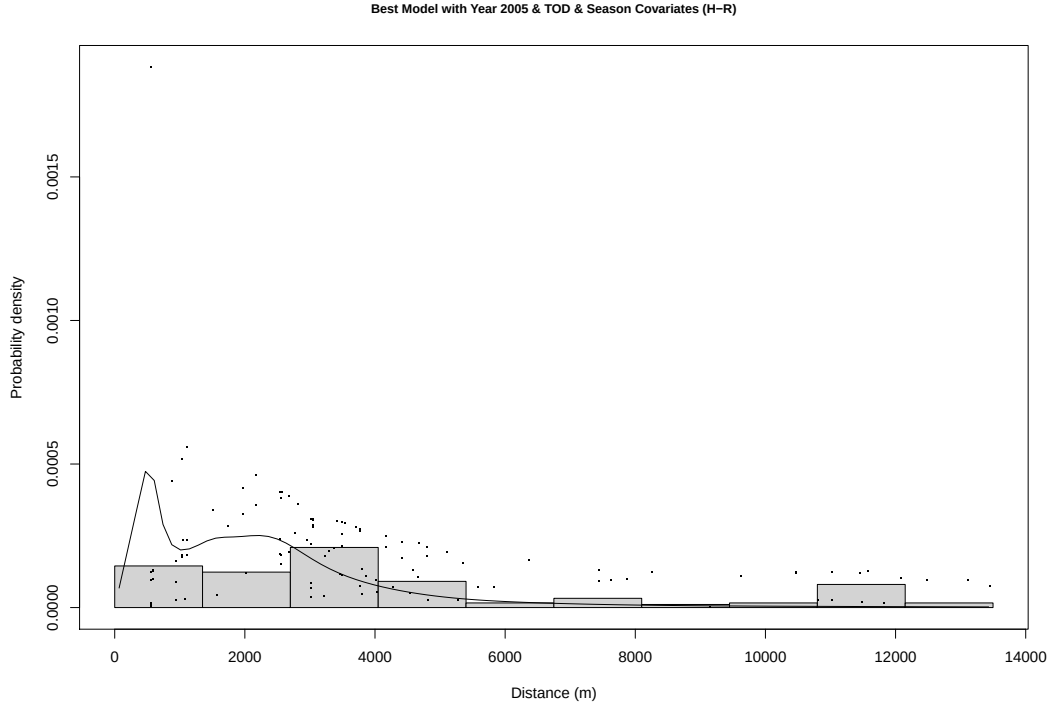


Figure 3.1.16: Histogram of true data with the probability density function, $\hat{f}(x)$, (line) for Monitor01 of the best fitting model with covariate terms – $hr.SznTOD05$. Faint dotted lines represent individual covariates.

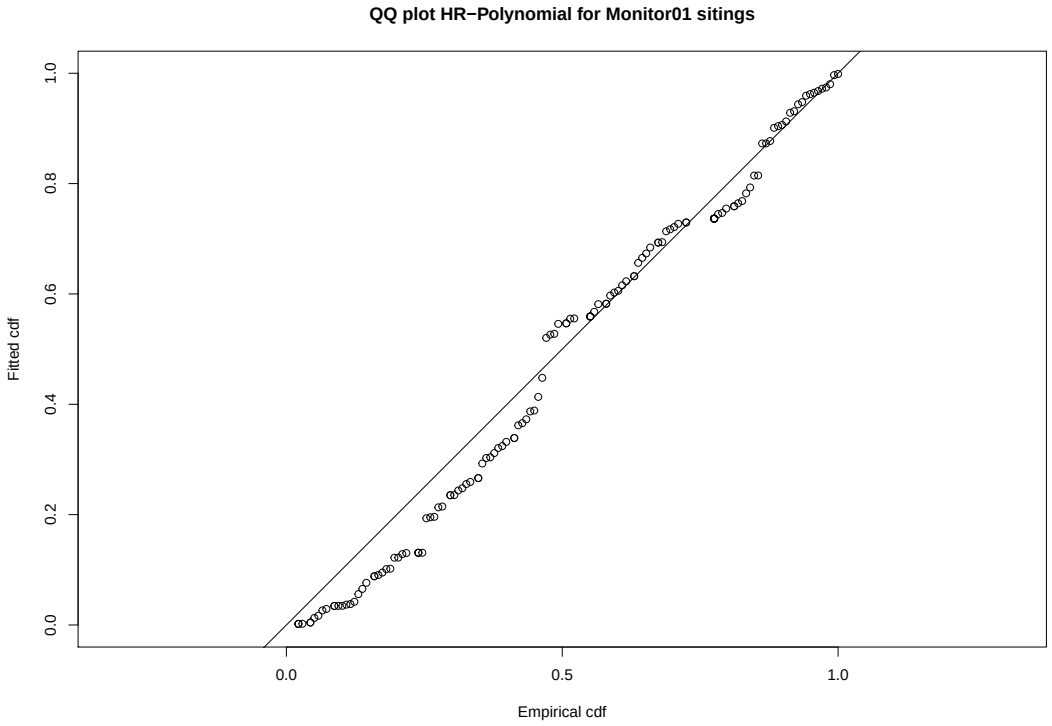


Figure 3.1.17: Q-Q plot for the best fitting model with covariate terms.

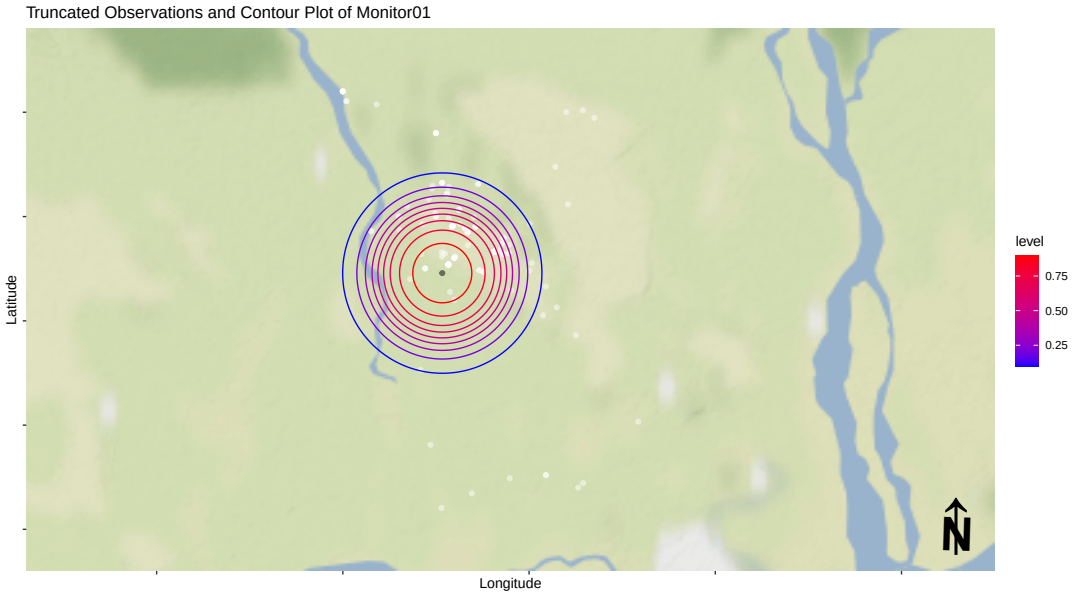


Figure 3.1.18: Homed in view of contour plot of Monitor01's observations (white), colour scale relates to the detection probability – probability given by the legend to the right of the diagram. (Tick intervals: lat=0.5 and lon=0.1)

described in Section 3.1.1.1.

The Distance R package was used to fit a variety of models as in 3.1.1.2 and determines the best order of adjustment models using AIC. The top three models are summarised in 3.1.5 with '$Mon.hr.cos$' as the best model where $\hat{P}_a = 19.8\%$ – composed of the hazard-rate function with cosine adjustments of order two. Table 3.1.5 gives the p-value for the Cramer-von Mises test, $p = 0.0327$. However, as the model has $p < 0.05$ we reject the hypothesis that the model is a good fit to the data.

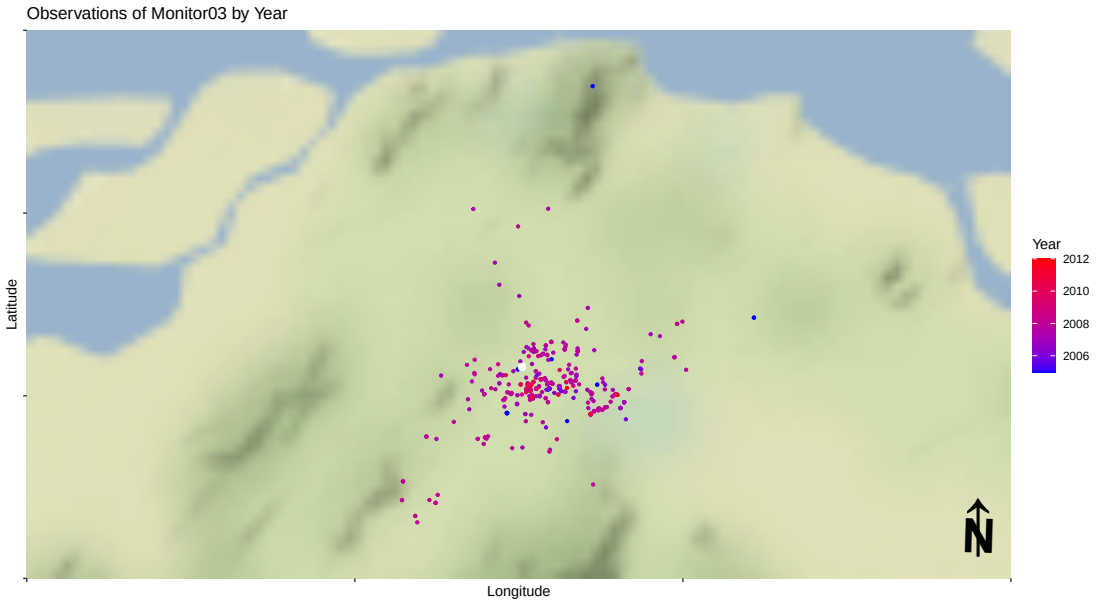


Figure 3.1.19: Visual representation of Monitor03's observations on a map. The white dot is Monitor03's location, observations by year are on a scale of blue (2005) to red (2012). Tick intervals: lat=0.05 and lon=0.1.

The same method was used with the Distance R package for covariate models as with adjustment models (explained in Section 3.1.1.3), the top three models can be seen in Table 3.1.6. The best ranked covariate model by AIC is '$hr.SznYearTOD$' with $\hat{P}_a = 20.1\%$. This model consists of the hazard-rate function with season,

Model Name	Key Function & Adjustment	C-vM p-value	$\hat{P}_a$	se($\hat{P}_a$)	AIC	ΔAIC
Mon.hr.cos	Hazard-rate with cosine adjustment terms (order 3)	0.0327	0.1975	0.0335	6088.84	0.00
Mon.hn.cos	Half-normal with cosine adjustment terms (order 3)	0.0266	0.2041	0.0303	6094.33	5.49
Mon.hr.poly	Hazard-rate*	0.1356	0.2040	0.0221	6104.31	15.47

Table 3.1.5: Monitor03 Adjustment Model Comparison of top three models. Key functions denoted with an asterix (*) denote best models had adjustments of order 0, i.e. no adjustment terms were added to the model. Listed for each model: Cramer-von Mises, C-vM p-value; average detectability estimate, $\hat{P}_a$; standard error of the estimated detectibility, $se(\hat{P}_a)$; AIC; and the AIC difference, ΔAIC.

Model Name	Key Fnct.	Covariates	C-vM p-value	$\hat{P}_a$	se($\hat{P}_a$)	ΔAIC	AIC
hr.SznYearTOD	H-R	Szn+Year +TOD	0.0236	0.2013	0.0934	6008.61	0.00
hr.YearTOD	H-R	Year+ TOD	0.0065	0.2010	0.0958	6016.37	7.76
hr.SznYear	H-R	Szn+Year	0.0040	0.1987	0.1022	6024.56	15.95

Table 3.1.6: Mon03 Covariate Model Comparison of the top three models. Key functions include: hazard-rate (H-R) and half-normal (H-N). Covariates making up the formula include: year, time of day (TOD) and season (Szn). Listed for each model: Cramer-von Mises, C-vM p-value; average detectability estimate, $\hat{P}_a$; standard error of the estimated detectibility, $se(\hat{P}_a)$; AIC; and the AIC difference, ΔAIC.

year and time of day covariates; it also has a Cramer-von Mises p-value of 0.0236 which is less than 0.05 meaning that we can reject this model as a good fit to our data. The Q-Q Plot in Figure 3.1.22 shows the goodness of fit for this best model.

Both the detection function and probability density function estimates from the best fitting covariate model can be seen in Figures 3.1.20 and 3.1.21 respectively. Note that the dotted lines in both of these figures represent the individual covariates included in the model.

See Figure 3.1.23 for a contour plot of Monitor03's truncated observations based on the results and detection probability found in this section against all of Monitor03's sightings. The colour red represents a full detection probability, $\hat{P}_a = 1$, with a gradient progressing to the colour blue which represents the detection probability of zero (not detectable).

3.1.4 Probability of Detection for Monitor46

Monitor46 was located in Goalpara with a total of 225 data entries with year of sighting ranging from 2006 to 2014. See Figure 3.1.24 for a visual representation of these observations by year on a map.

After exploratory analysis, data was truncated by 10% and the covariates chosen to be included in the model selection process were: year, time of day (TOD) and season (Szn). All covariates as described in Section 3.1.1.1.

The Distance R package was used to fit a variety of models as in 3.1.1.2 and determines the best order of adjustment models using AIC. The top four models are summarised in 3.1.7. It appears that there are three best models however, as these top three models are hazard-rate with adjustments of order zero, they are all the same models as each other. Therefore, we will call the best model '$Mon.hr$' with $\hat{P}_a = 65.4\%$ – composed of the hazard-rate function with cosine adjustments of order two. Table 3.1.7 gives the p-value for the Cramer-von Mises test, $p = 0.0678$. As the model has $p > 0.05$ we can accept the hypothesis that

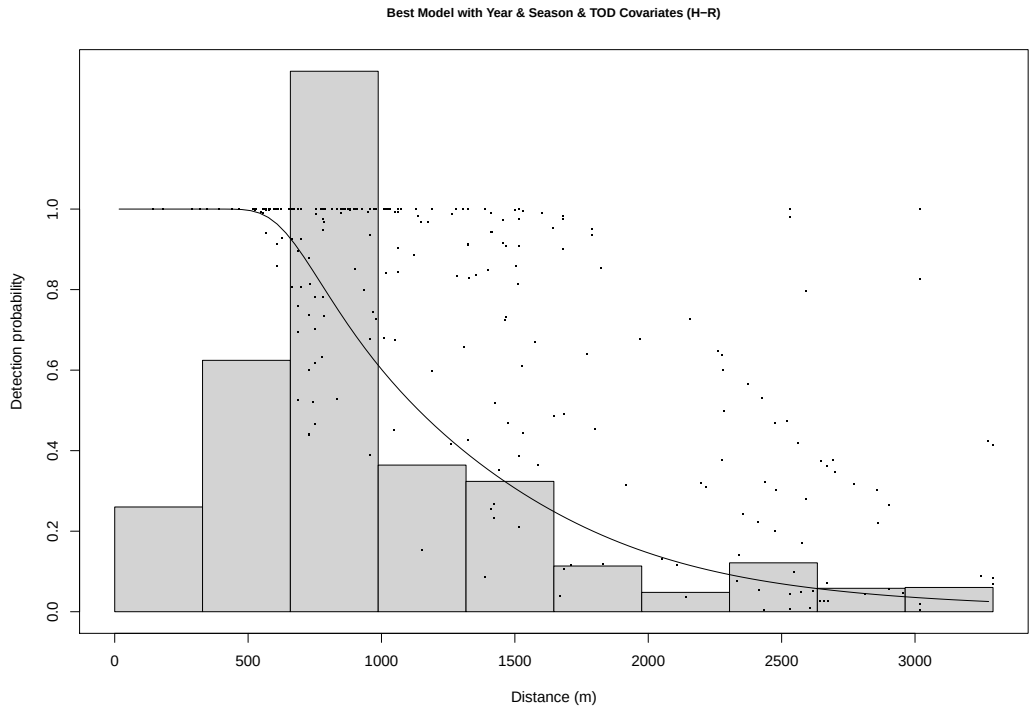


Figure 3.1.20: Histogram of true data with the detection function, $\hat{g}(x)$, (line) for Monitor03 of the best fitting model with covariate terms – $hr.SznYearTOD$. Faint dotted lines represent individual covariates.

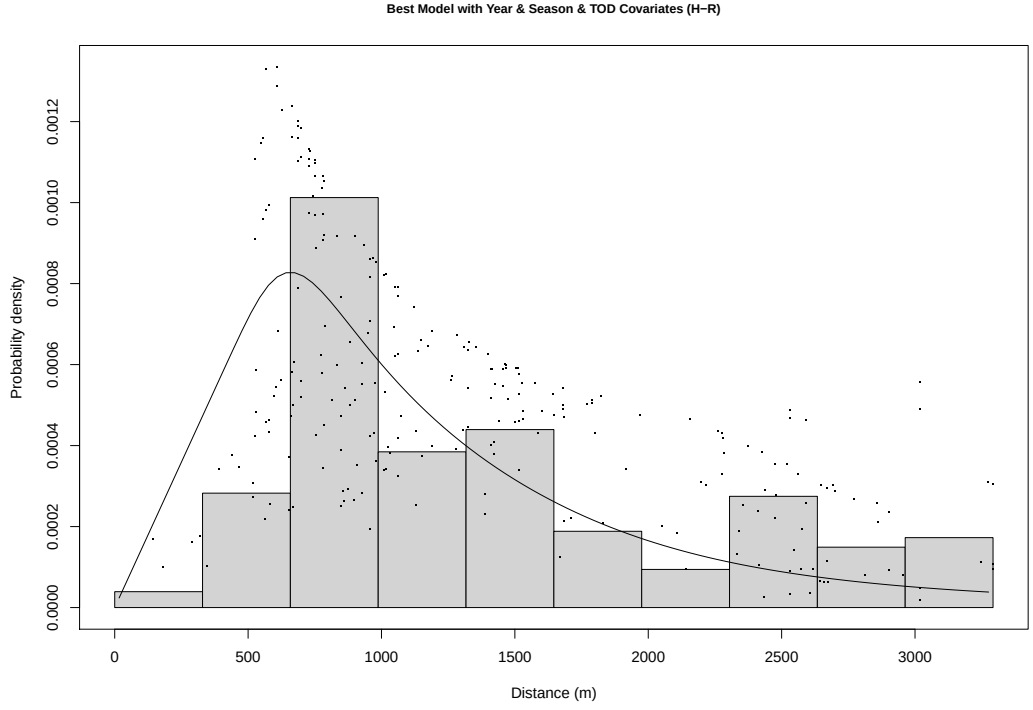


Figure 3.1.21: Histogram of true data with the probability density function, $\hat{f}(x)$, (line) for Monitor03 of the best fitting model with covariate terms – $hr.SznYearTOD$. Faint dotted lines represent individual covariates.

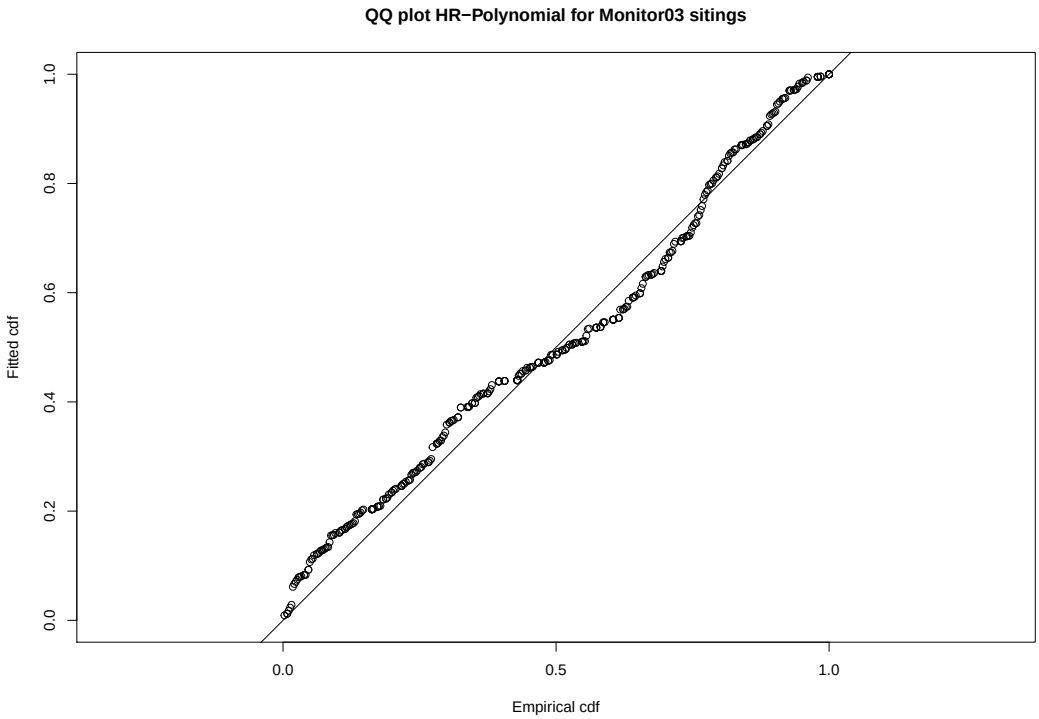


Figure 3.1.22: Q-Q plot for the best fitting model with covariate terms.

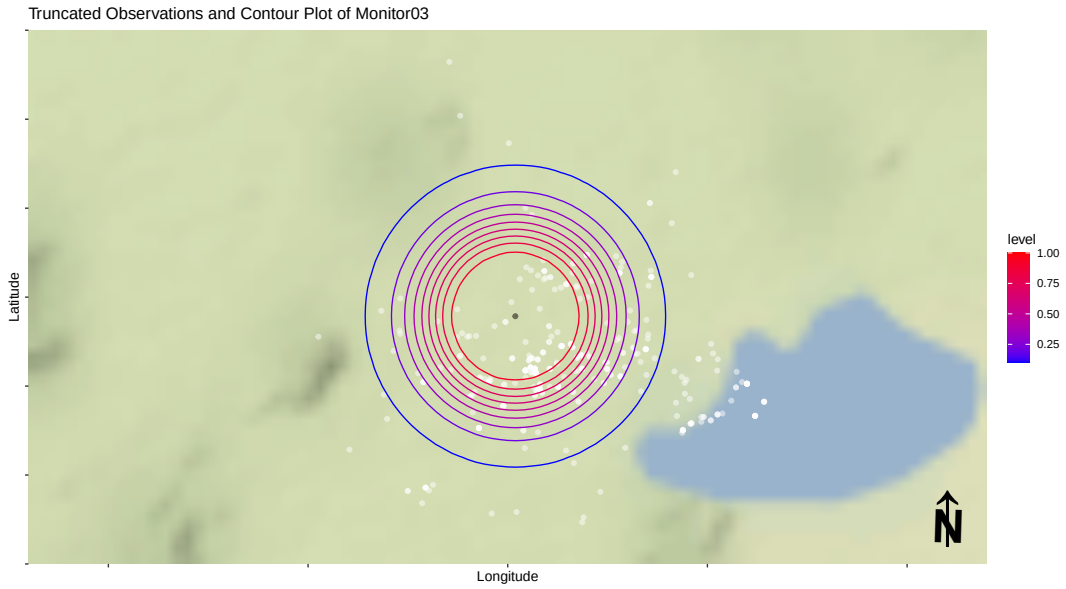


Figure 3.1.23: Homed in view of contour plot of Monitor03's observations (white), colour scale relates to the detection probability – probability given by the legend to the right of the diagram. (Tick intervals: lat=0.01 and lon=0.025)

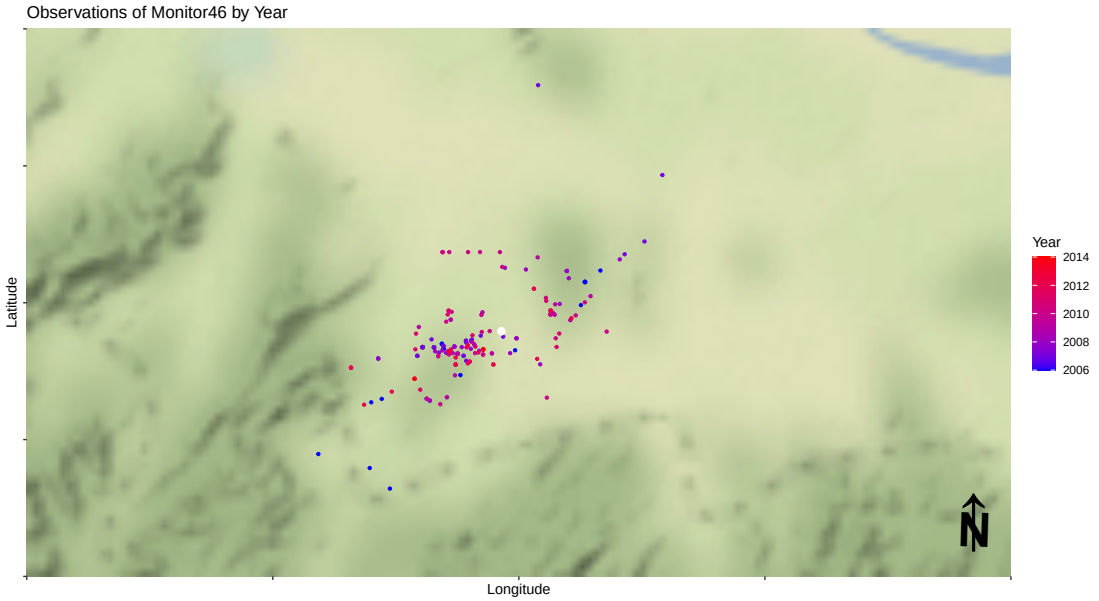


Figure 3.1.24: Visual representation of Monitor46's observations on a map. The white dot is Monitor46's location, observations by year are on a scale of blue (2006) to red (2014). Tick intervals: lat=0. and lon=0..

the model is a good fit to the data.

The same method was used with the Distance R package for covariate models as with adjustment models (explained in Section 3.1.1.3), the top three models can be seen in Table 3.1.8. The best ranked covariate model by AIC is '$hr.Year$' with $\hat{P}_a = 58.2\%$. This model consists of the hazard-rate function with year as the only covariate – note that both top models agree on the probability of detection to 3 significant figures. The model has a Cramer-von Mises p-value of 0.0275 which is less than 0.05 meaning that we reject this model as a good fit to our data. The Q-Q Plot in Figure 3.1.27 shows the goodness of fit for this best model.

As a result, we can say that the probability of Monitor46 detecting elephants, given that they are present at that given time is 58.2%. Both the detection function and probability density function estimates from the best fitting covariate

Model Name	Key Function & Adjustment	C-vM p-value	$\hat{P}_a$	se($\hat{P}_a$)	AIC	ΔAIC
Mon.hr.poly	Hazard-rate*	0.0678	0.6540	0.0661	3298.51	0.00
Mon.hr.cos	Hazard-rate*	0.0678	0.6540	0.0661	3298.51	0.00
Mon.hr.herm	Hazard-rate*	0.0678	0.6540	0.0661	3298.51	0.00
Mon.unif.cos	Uniform with cosine adjustment terms (order 2)	0.0399	0.5915	0.1394	3300.55	2.04

Table 3.1.7: Monitor46 Adjustment Model Comparison of top three models. Key functions denoted with an asterix (*) denote best models had adjustments of order 0, i.e. no adjustment terms were added to the model, resulting in three identical hazard-rate models. Listed for each model: Cramer-von Mises, C-vM p-value; average detectability estimate, $\hat{P}_a$; standard error of the estimated detectibility, $se(\hat{P}_a)$; AIC; and the AIC difference, ΔAIC.

Model Name	Key Fnct.	Covariates	C-vM p-value	$\hat{P}_a$	se($\hat{P}_a$)	ΔAIC	AIC
hr.Year	H-R	Year	0.0275	0.5817	0.9191	3285.59	0.00
hr.SznYear	H-R	Szn+Year	0.0301	0.5816	0.0710	3288.87	3.28
hr.YearTOD	H-R	Year+ TOD	0.0203	0.5583	0.0862	3288.93	3.34

Table 3.1.8: Mon46 Covariate Model Comparison of the top three models. Key functions include: hazard-rate (H-R) and half-normal (H-N). Covariates making up the formula include: year, season (Szn), time of day (TOD). Listed for each model: Cramer-von Mises, C-vM p-value; average detectability estimate, $\hat{P}_a$; standard error of the estimated detectibility, $se(\hat{P}_a)$; AIC; and the AIC difference, ΔAIC.

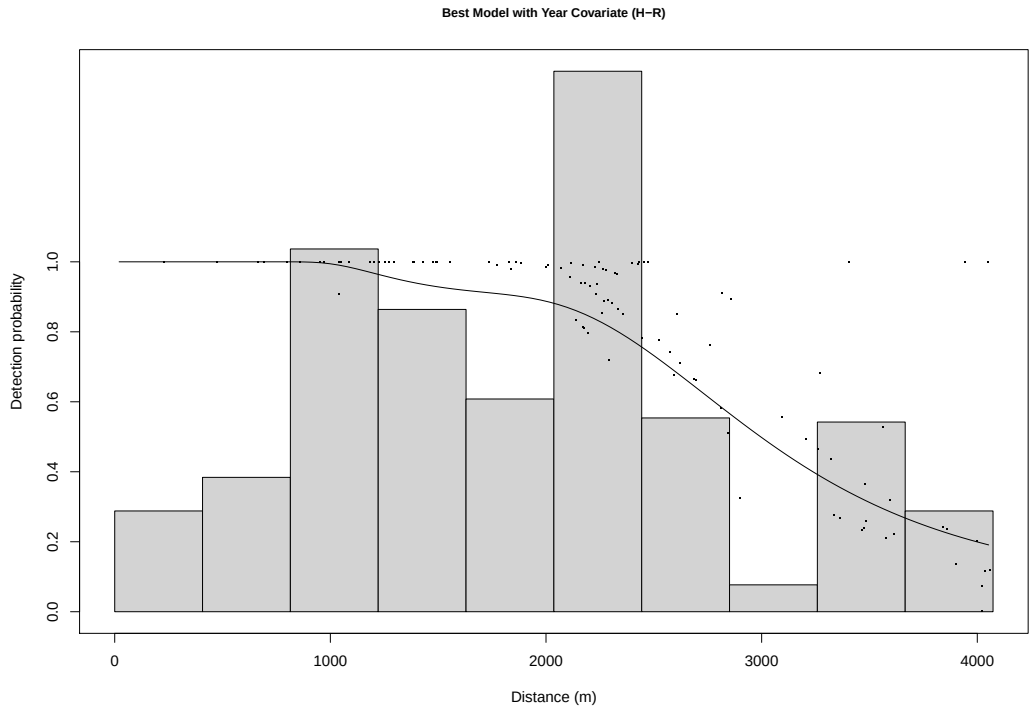


Figure 3.1.25: Histogram of true data with the detection function, $\hat{g}(x)$, (line) for Monitor46 of the best fitting model with covariate terms – $hr.Year$. Faint dotted lines represent individual covariates.

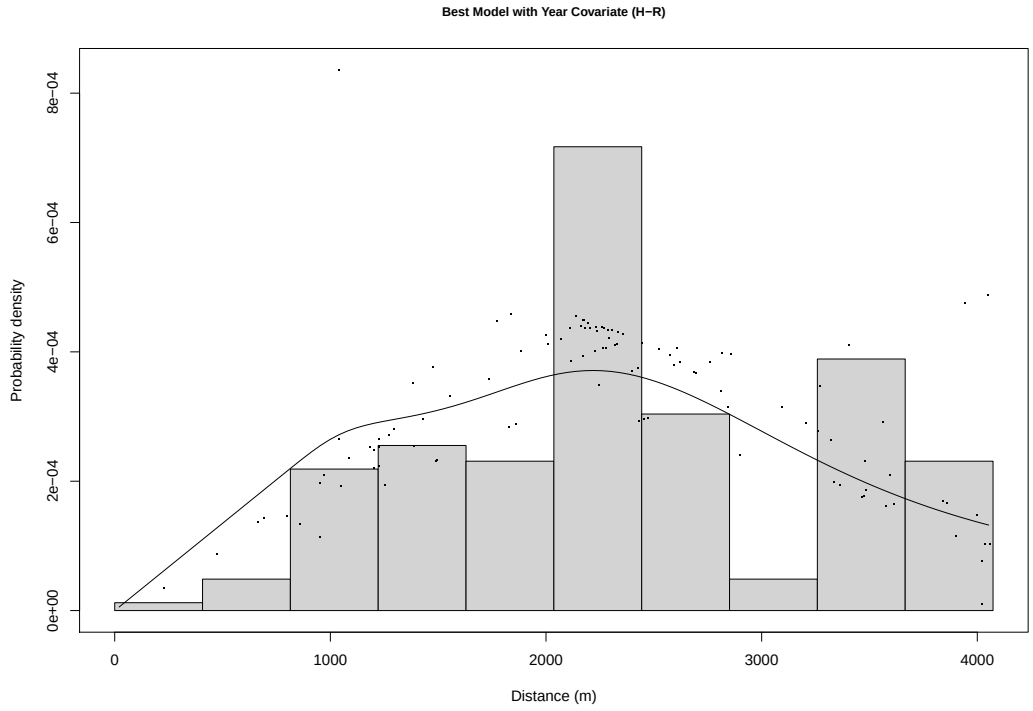


Figure 3.1.26: Histogram of true data with the probability density function, $\hat{f}(x)$, (line) for Monitor46 of the best fitting model with covariate terms – $hr.Year$. Faint dotted lines represent individual covariates.

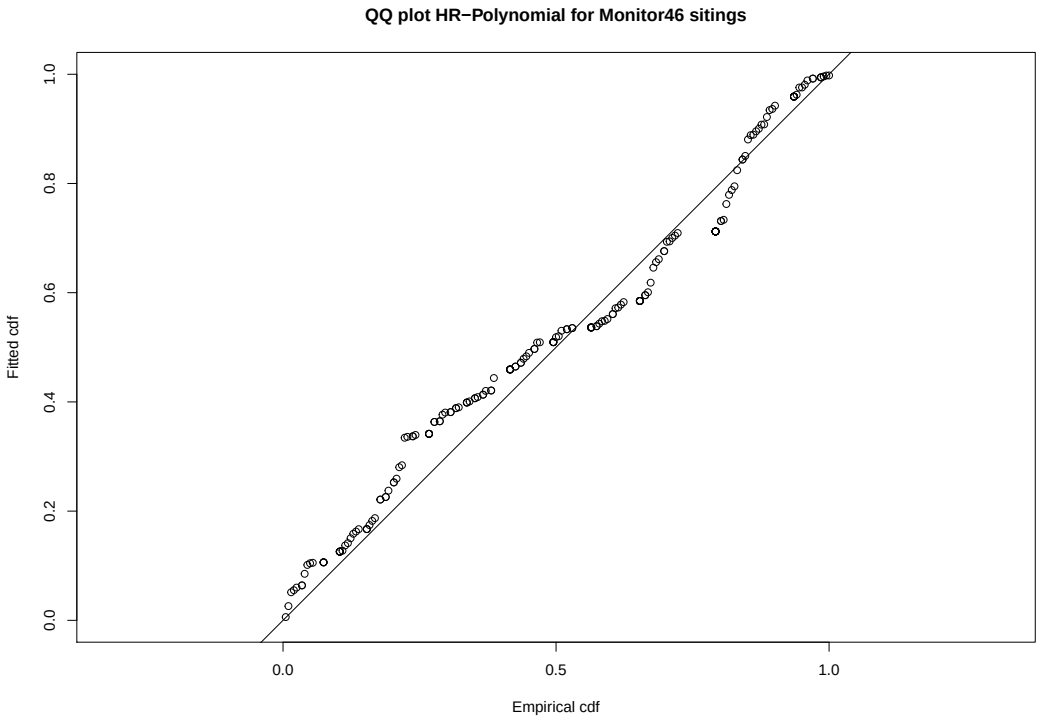


Figure 3.1.27: Q-Q plot for the best fitting model with covariate terms.

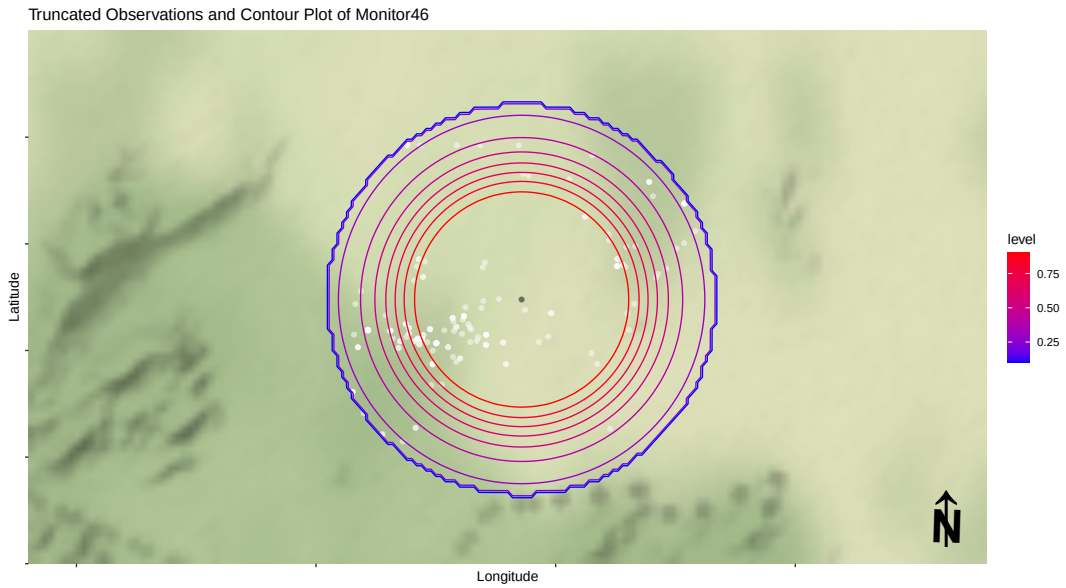


Figure 3.1.28: Homed in view of contour plot of Monitor46's observations (white), colour scale relates to the detection probability – probability given by the legend to the right of the diagram. (Tick intervals: lat=0.02 and lon=0.05)

model can be seen in Figures 3.1.25 and 3.1.26 respectively. Note that the dotted lines in both of these figures represent the individual covariates included in the model.

See Figure 3.1.28 for a contour plot of Monitor46's truncated observations based on the results and detection probability found in this section against all of Monitor46's sightings. The colour red represents a full detection probability, $\hat{P}_a = 1$, with a gradient progressing to the colour blue which represents the detection probability of zero (not detectable).

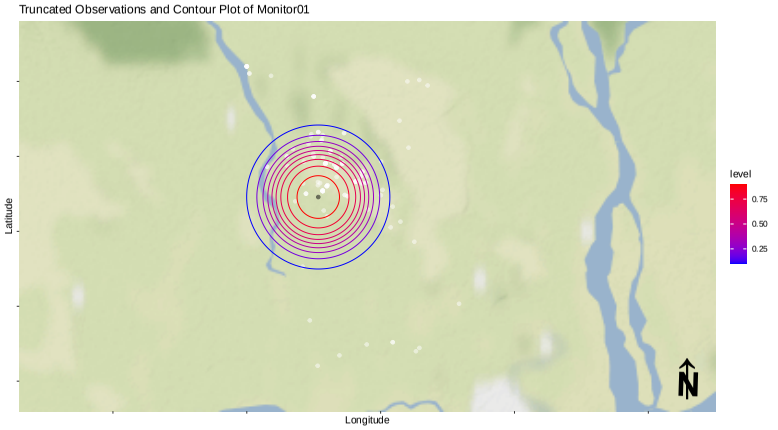
3.1.5 Comparison of Individual Studies

Only two of the four models tested for individual monitors had covariate models which were accepted as a good fit to the data collected using the Cramer-von Mises goodness of fit test (Monitor01, Section 3.1.2 and Monitor16, Section 3.1.1) – both best models were '$hr.SznYearTOD$'. There were only two accepted adjustment models from Monitor01, Section 3.1.2 and Monitor46, Section 3.1.4. Monitor03 (Section 3.1.3) had no models accepted as a good fit to the data.

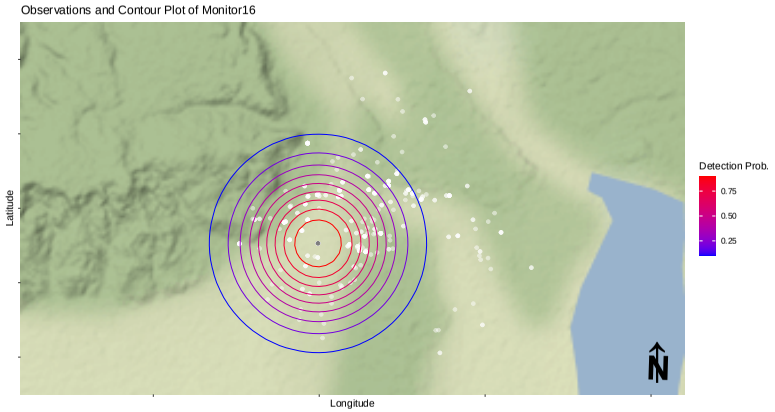
See Figure 3.1.29 for contour plots of each individual monitor tested based on the results and detection probabilities found in each analysis of the monitor observations.

3.2 All Data Combined

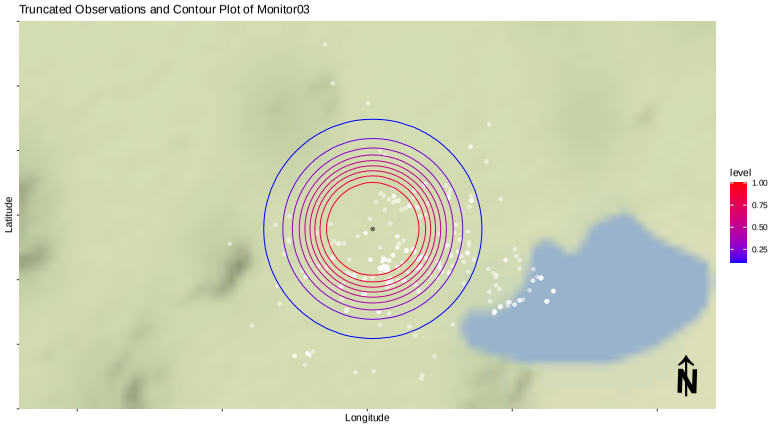
For this section we are going to discuss models fitted to the data set as a whole with all monitors included. All methods discussed in the individual studies (Section 3.1.1) are applied to this section on combined study. All models, adjuments and covariates previously fitted are the same in this case for models fitted with adjustments and for models with covariates, however, we have the addition of a 'Site' covariate which is the area that the monitor is located: Goalpara or Sonitpur.



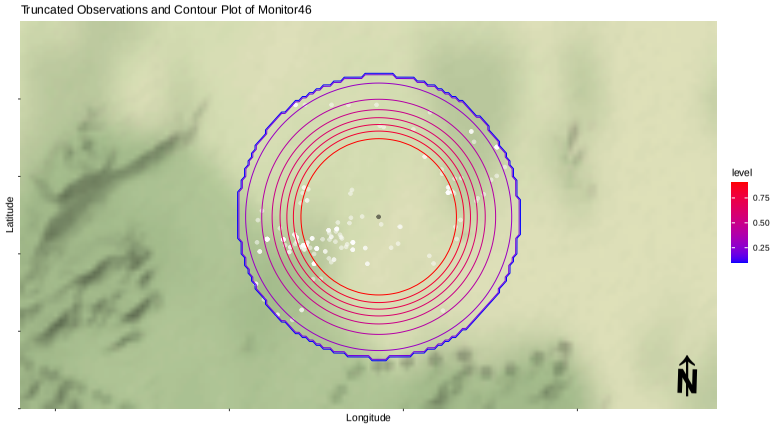
(a) Montior01 (Lat=0.05 & Lon=0.1)



(c) Montior16 (Lat=0.01 & Lon=0.025)



(b) Montior03 (Lat=0.01 & Lon=0.025)



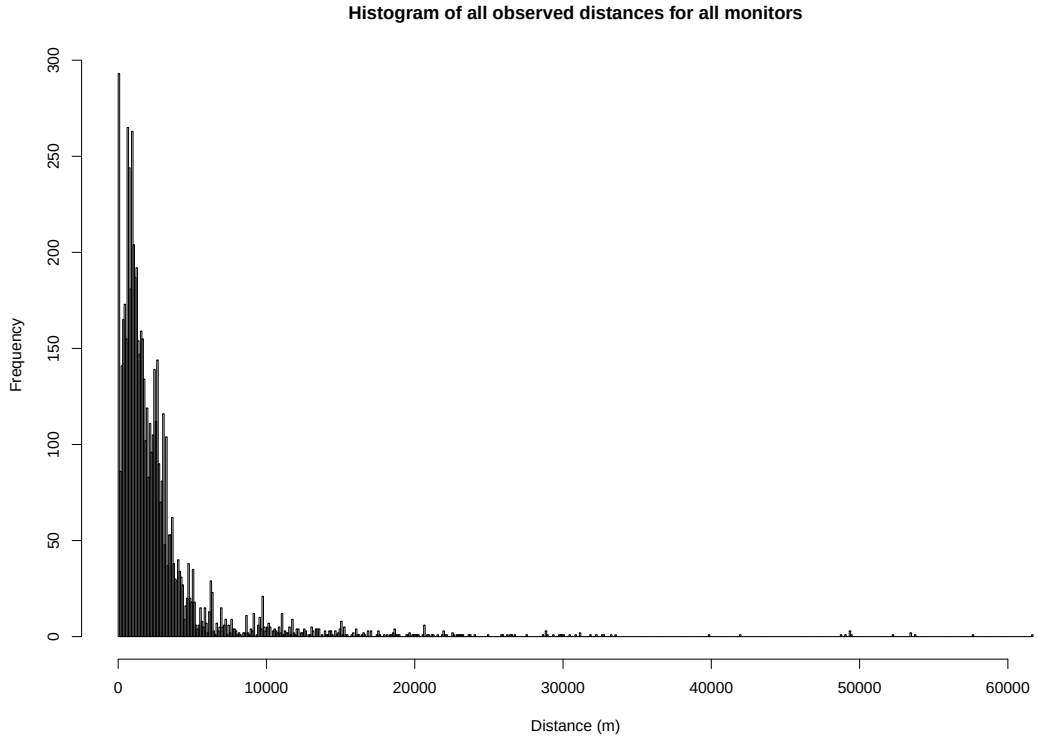
(d) Montior46 (Lat=0.02 & Lon=0.05)

Figure 3.1.29: Contour plots using detection probabilities for each montior (white) based on their 10% truncated observations. Colour scale relates to the detection probability – probability given by the legend to the right of the diagram. (Tick intervals are listed for each monitor visual)

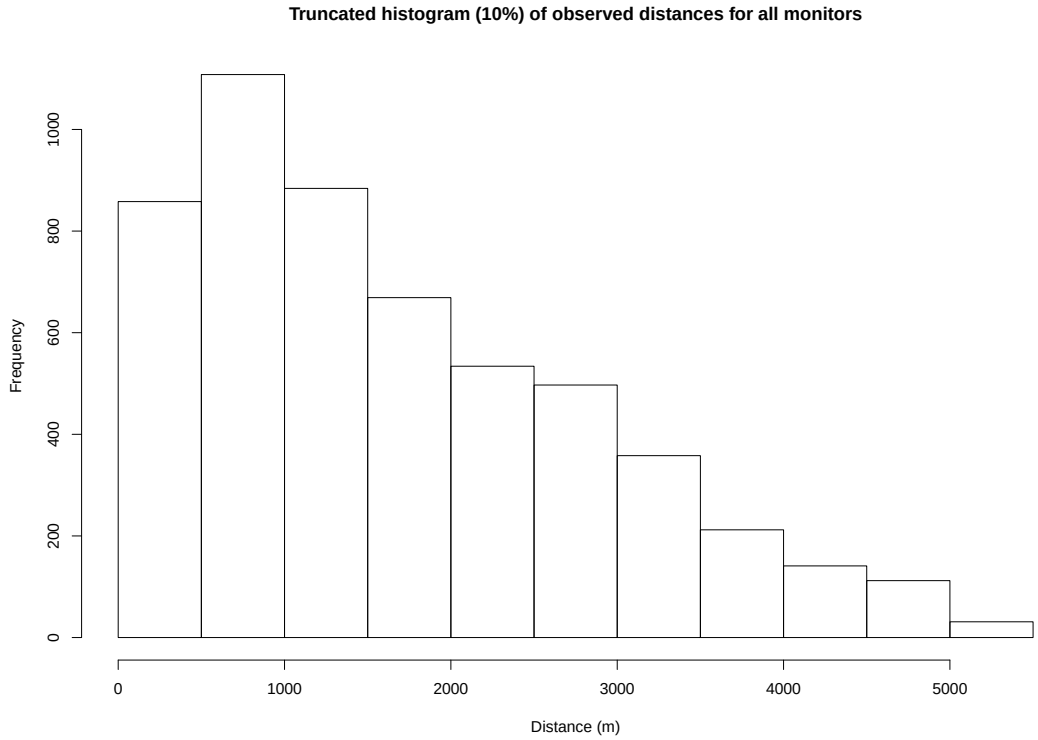
First, exploratory data analysis was undertaken. Looking at Figure 1.1.2 of monitor locations and observations on a map, we can see that observations tend to be spread reasonably close to monitor locations. We can see this spread of observations by distance in the histogram of Figure 3.2.1a, truncating this produces the histogram shown in Figure 3.2.1b which follows the point sampling shape that we desire. Finally, we look at histograms of covariates – these are displayed (pre and post truncation) in Figures 3.2.3, 3.2.4, 3.2.5 and 3.2.6. See Figure 3.2.3b for the truncated version of the 'Year' boxplot, we can see that 2018 has a much smaller range of distances – both before and after truncation – when compared to other years, however this is likely because we have less observations for this year. We also note a considerably large range of distances in the year 2005 (pre-truncation) in Figure 3.2.3a, likely because there are considerably more observations for this year than any others, possibly due to an increased effort from monitors as this was the first year that the project was fully launched. Based on this observation, it was decided that the covariate '2005' vs. '*not* 2005' should be included in the model fitting process – also used as a covariate with individual monitors in Section 3.1. As for the rest of the covariates, there appears to be no significant differences in the range of observation distances.

A covariate for 'Monitor' was attempted, but due to the complexity of having 57 monitors, this was not possible due to limitations of the data. Figures 3.2.2a and 3.2.2b show the huge range in distances of observations for each monitor before and after truncation. As a result, (taking care to not use Year and 2005 in conjunction with one another) covariates carried forward into the analysis were: Year, TOD, Season and 2005.

Models with adjustments were investigated but all lead to very poor fits to the data. The best model was the Uniform distribution with cosine adjustment terms of order 3, with an estimated detection probability of 12.8%. The Cramer-von Mises p-value was 0.02926. As a result this model and all other models with

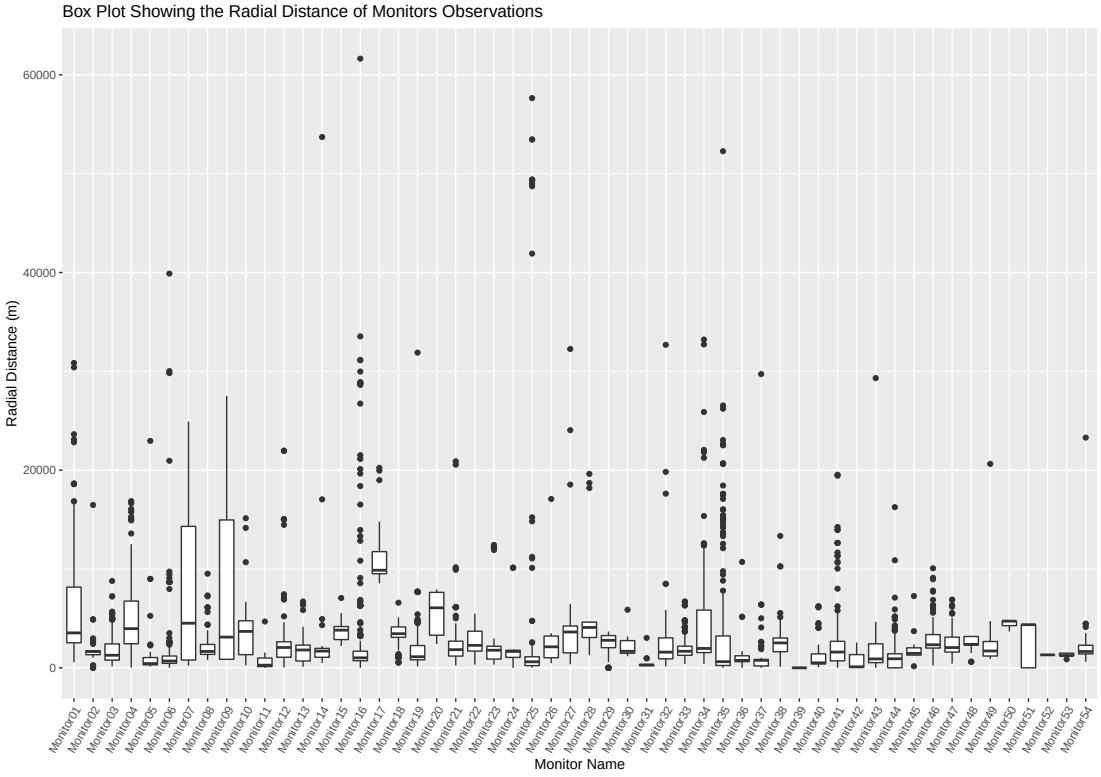


(a) No truncation

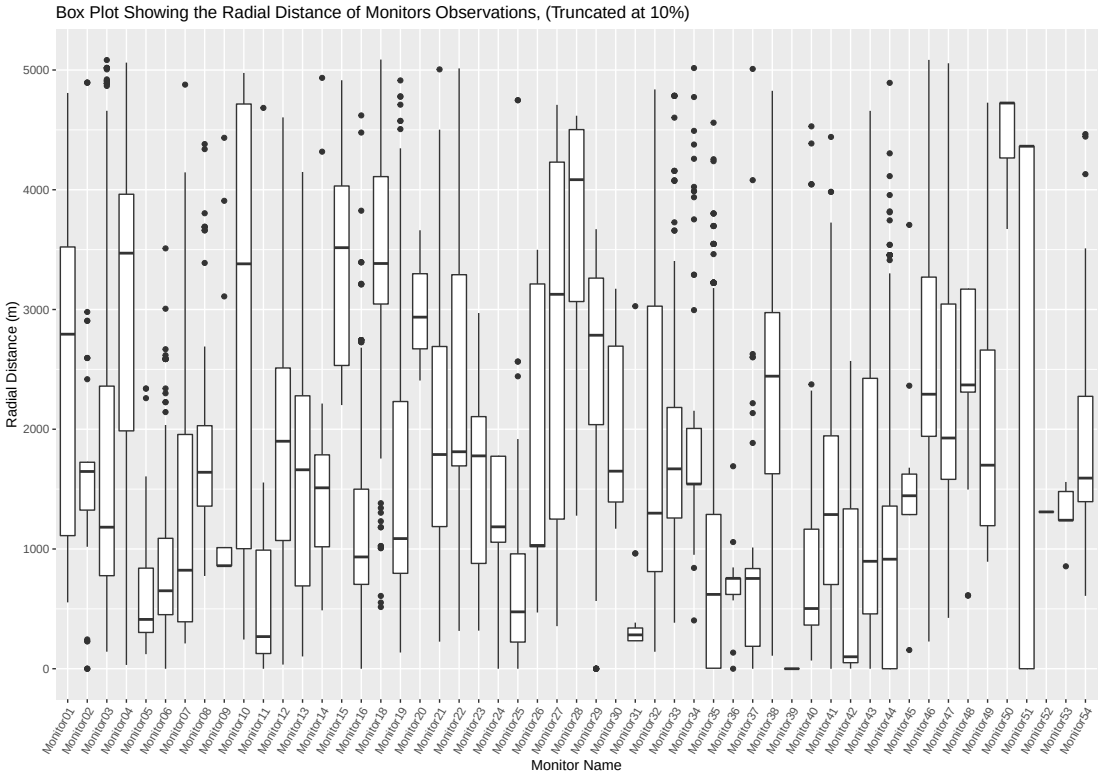


(b) 10% truncation

Figure 3.2.1: Histograms showing the frequency sightings by distance (m) for all data.

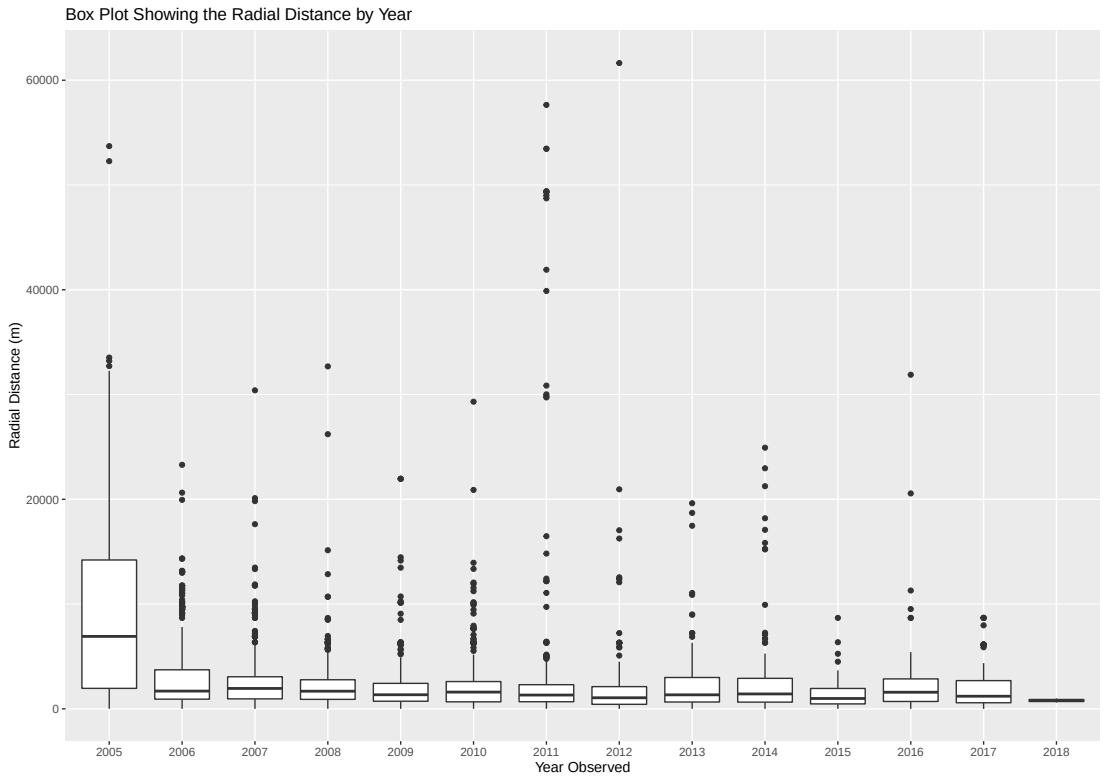


(a) No truncation

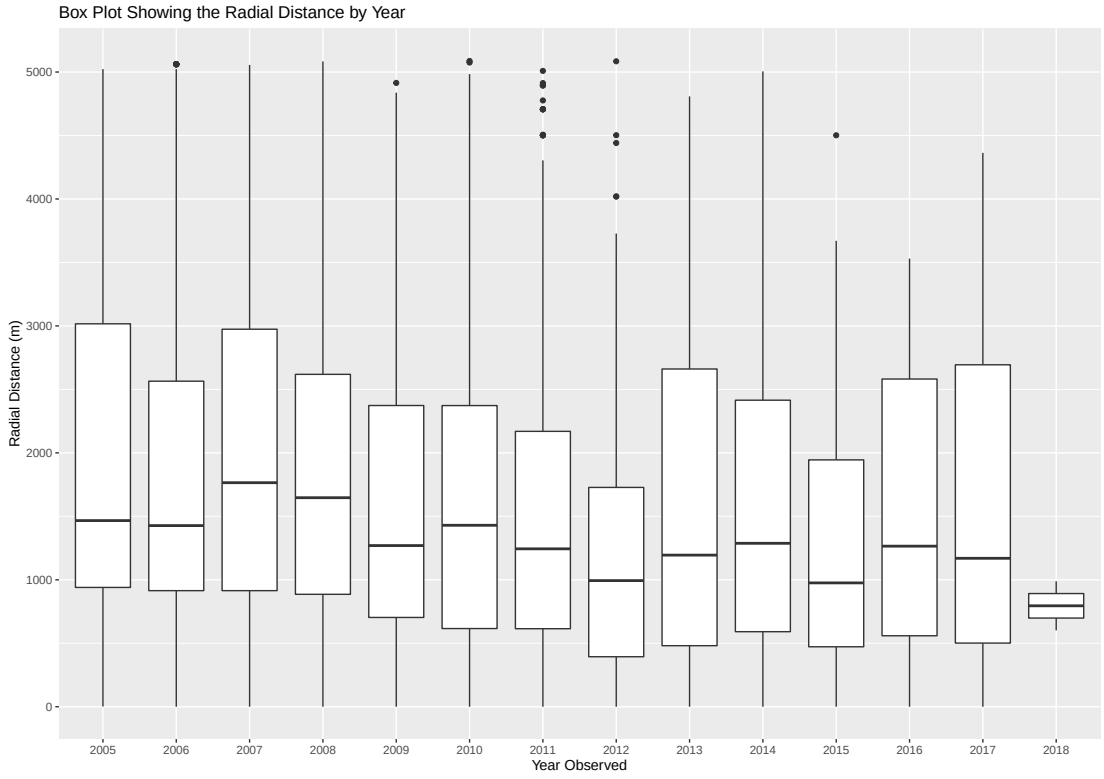


(b) 10% truncation

Figure 3.2.2: Boxplot showing the radial distance of all observed sightings by Monitor.

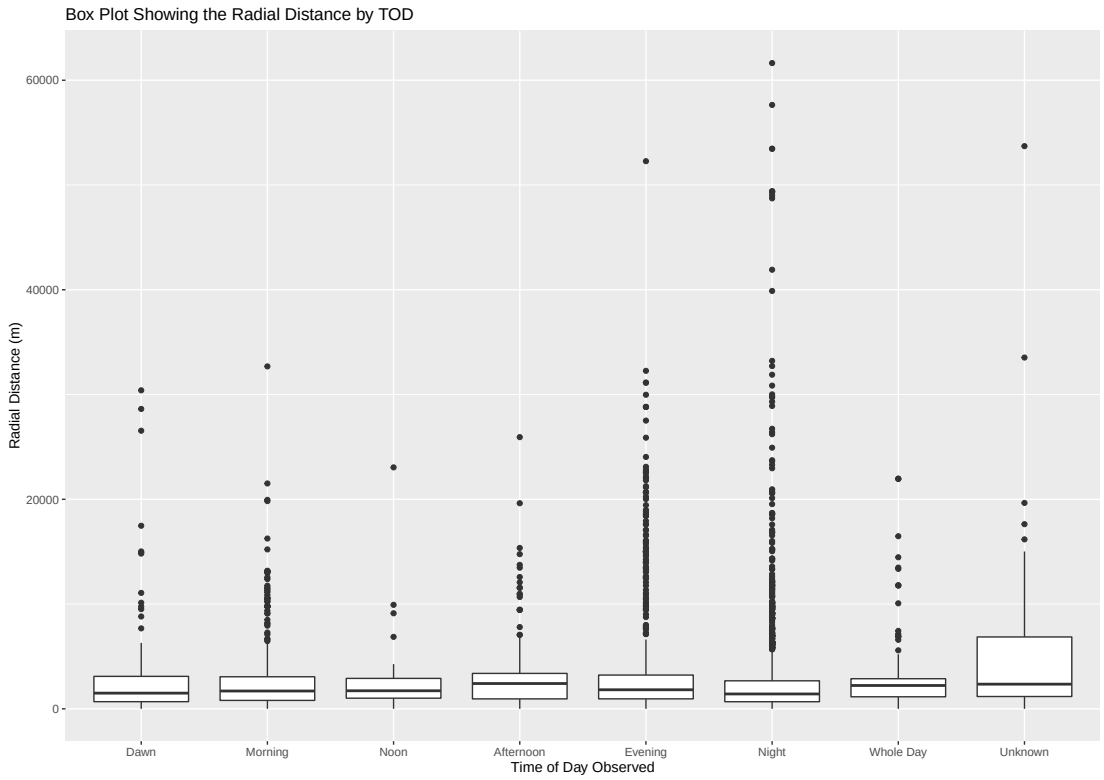


(a) Year

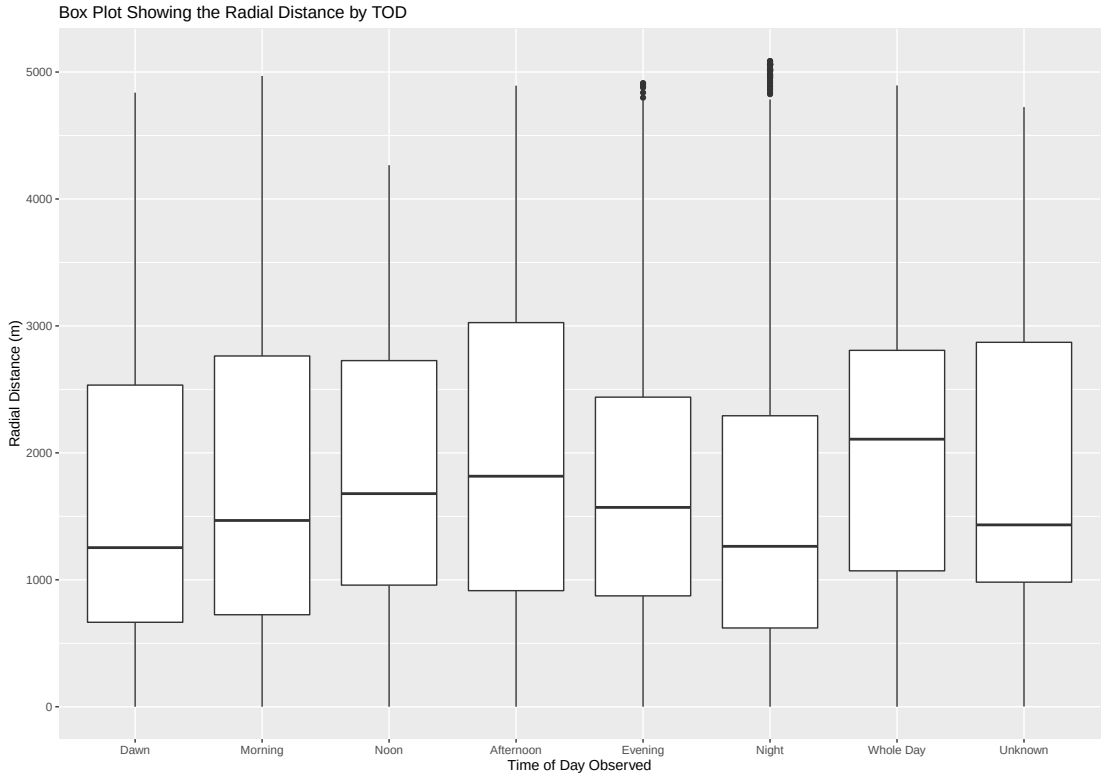


(b) Year, 10% truncation

Figure 3.2.3: Comparison of truncated boxplot showing the radial distance of observed sightings by covariates for all monitors.

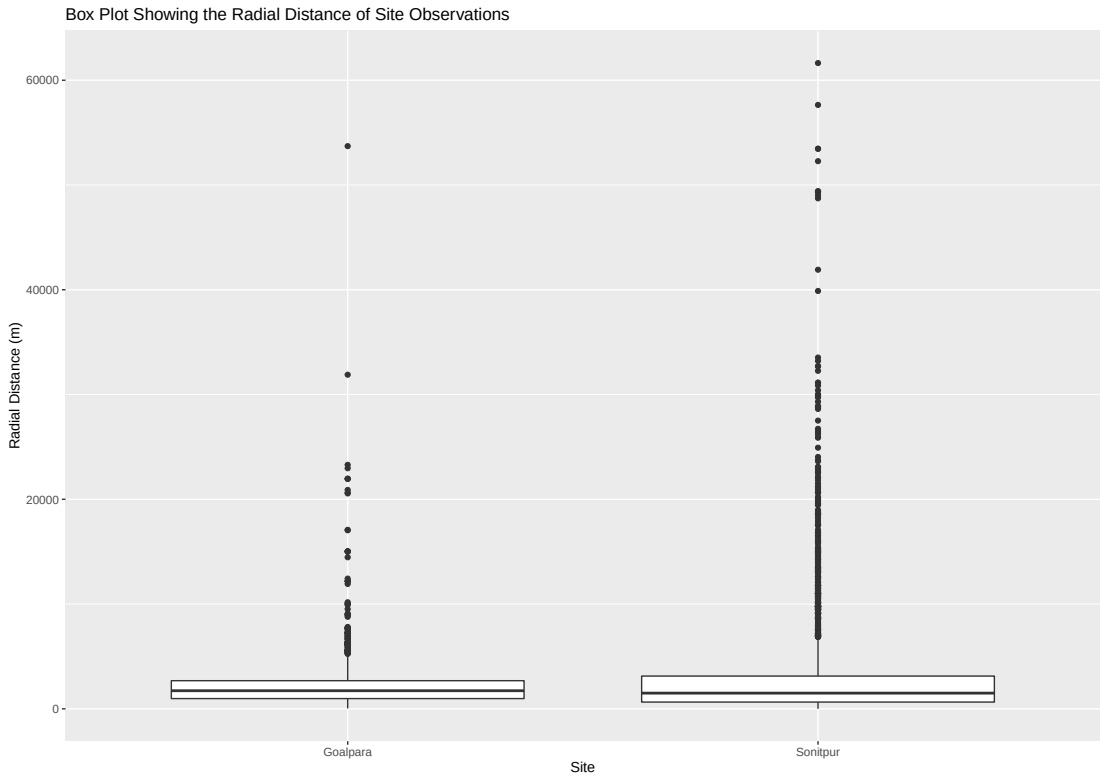


(a) Time of Day (TOD)

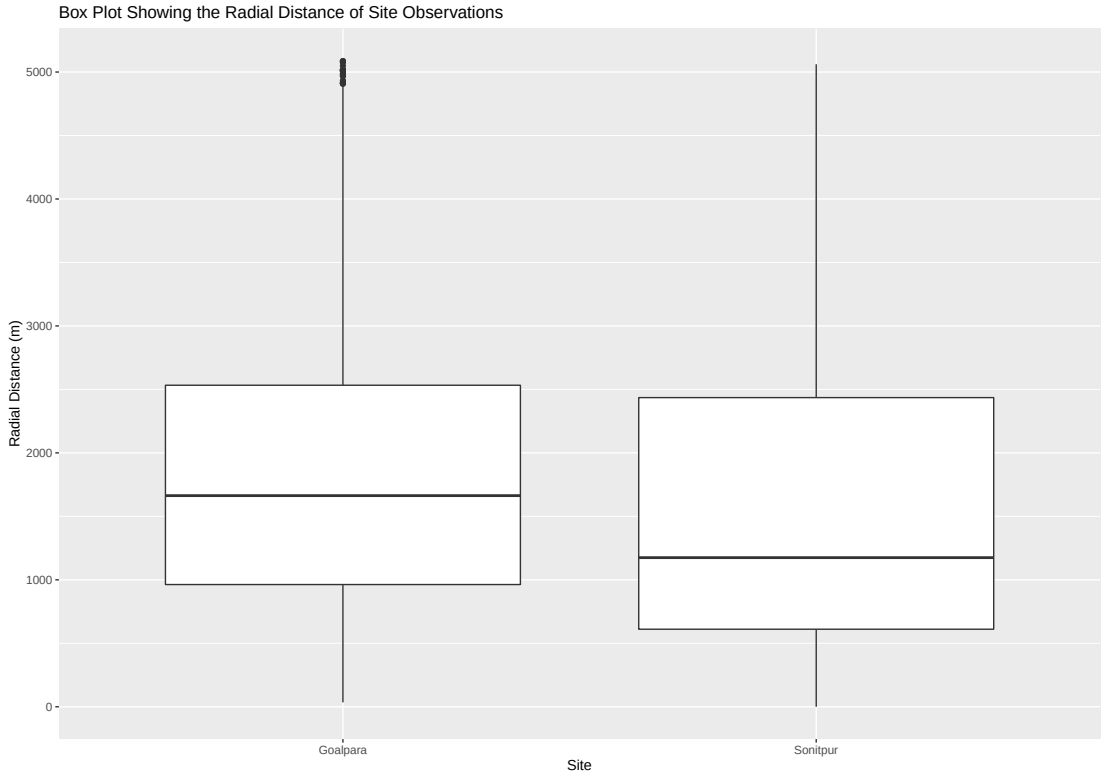


(b) TOD, 10% truncation

Figure 3.2.4: Comparison of truncated boxplot showing the radial distance of observed sightings by covariates for all monitors.

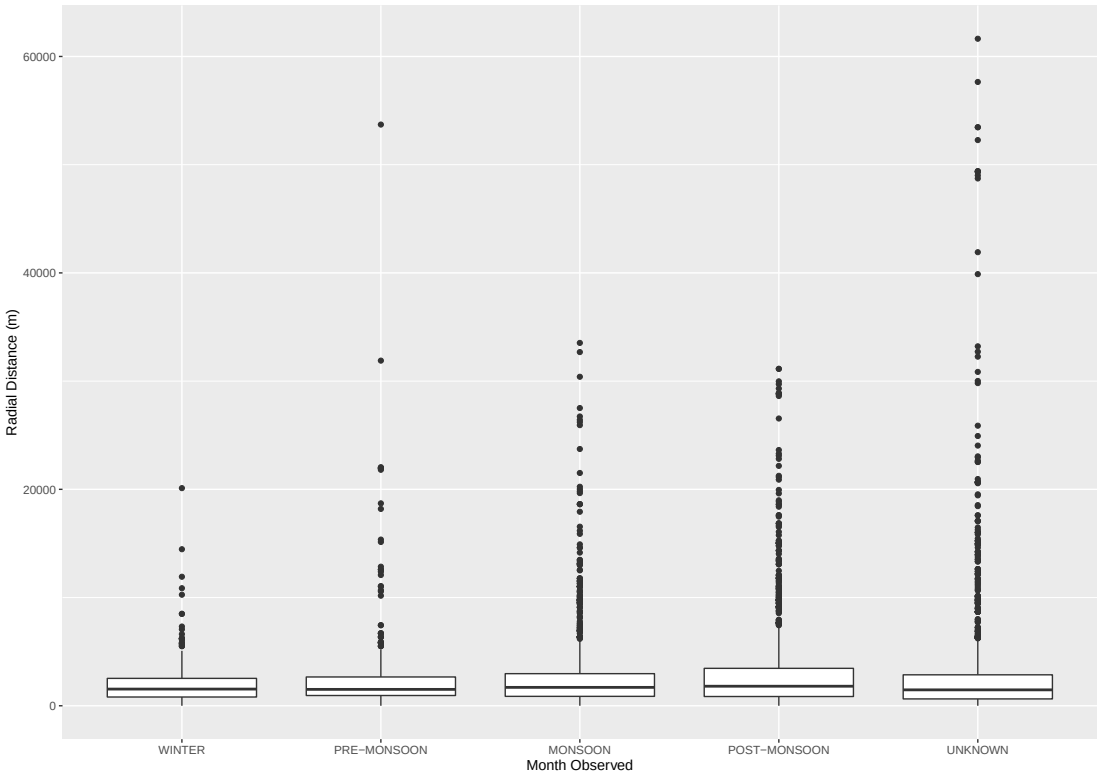


(a) Site

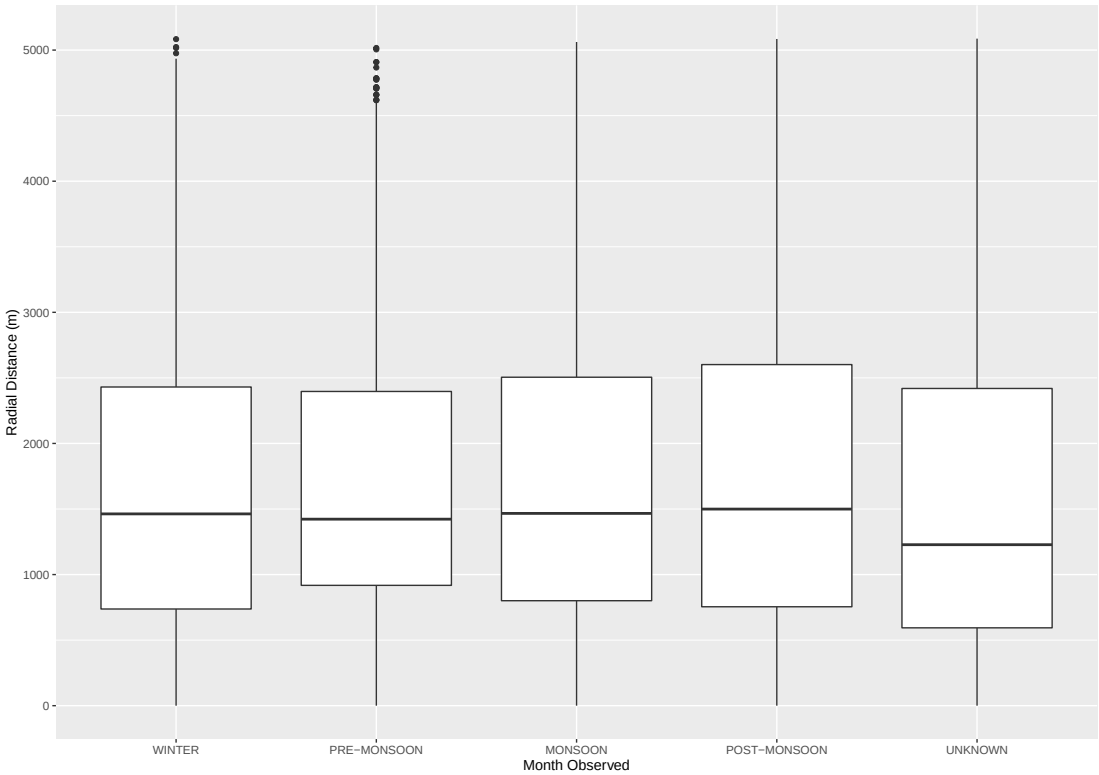


(b) Site, 10% truncation

Figure 3.2.5: Comparison of truncated boxplot showing the radial distance of observed sightings by covariates for all monitors.



(a) Season



(b) Season, 10% truncation

Figure 3.2.6: Comparison of truncated boxplot showing the radial distance of observed sightings by covariates for all monitors.

adjustments were rejected as a good fit to the data.

Table 3.2.1 shows a comparison of all covariate models which managed to fit to the data. Not all models were able to fit, there were some variance-covariance matrix elements which were 'NA' with possible numerical problems. As a result, only the detection function was estimated and we were unable to compare these with other models which accuratley fitted to the data.

The best covariate model was '$hr.SznTOD$' according to AIC. This model consists of the hazard-rate distribution as the key functions with Season and TOD as covariates. It had a Cramer-von Mises p-value of 4.71×10^{-7} which meant that this model, although the best fitting of all the models, was not a good fit to the data. A possible explanation for lack of fit could include some possible data entries being rounded to distance zero (as we can see a line of horizontal dots on the bottom left-hand side in Figure 3.2.9); an inadequate estimated level of truncation; the large sample size; or potentially the difference between the individual monitors which has not been captured in the model. See the Q-Q plot in Figure 3.2.9 for this model – the plot visually appears to be a good fit to the data but due to large number of data entries, this created a small p-value and resulted in the model being rejected. The detection function and probability density function estimates for this model can be seen in Figures 3.2.7 and 3.2.8 respectively.

Note: although all of these covariate models were rejected as a good fit the data, all top eight models ranked by AIC, agree that $\hat{P}_a \approx 0.05$. Suggesting that although there appears to be something missing from the data to fit a good model, monitors appear to have a detection probability of approximately 5% based on the information that we currently have.

Model Name	Key Fnct.	Covariates	C-vM p-value	$\hat{P}_a$	se($\hat{P}_a$)	ΔAIC
hr.SznTOD	H-R	Szn+TOD	0.0000	0.0495	0.0020	0.00
hr.SznTOD05	H-R	Szn+TOD+2005	0.0000	0.0495	0.0020	1.92
hr.TOD	H-R	TOD	0.0000	0.0518	0.0020	62.35
hr.TOD05	H-R	2005+TOD	0.0000	0.0520	0.0020	63.92
hr.Szn	H-R	Szn	0.0000	0.0466	0.0020	122.33
hr.Szn05	H-R	Szn+2005	0.0000	0.0469	0.0020	123.84
hr.05	H-R	2005	0.0000	0.0530	0.0020	239.88
hr.model0	H-R	1	0.0000	0.0523	0.0020	240.45
hn.YearTODSite	H-N	Year+TOD+Site	0.0911	0.1541	0.0019	1086.99
hn.SznYearTOD	H-N	Szn+Year+TOD	0.0888	0.1559	0.0019	1149.01
hn.YearSite	H-N	Year+Site	0.0933	0.1567	0.0019	1152.54
hn.YearTOD	H-N	Year+TOD	0.0900	0.1578	0.0019	1197.67
hn.SznYear	H-N	Szn+Year	0.0910	0.1579	0.0019	1202.52
hn.Year	H-N	Year	0.0920	0.1594	0.0019	1240.84
hn.TODSite05	H-N	2005+TOD+Site	0.0931	0.1618	0.0018	1303.11
hn.SznTOD05	H-N	Szn+TOD+2005	0.0925	0.1625	0.0018	1327.07
hn.TODSite	H-N	Site+TOD	0.0940	0.1626	0.0018	1327.71
hn.SznTOD	H-N	Szn+TOD	0.0930	0.1628	0.0018	1337.76
hn.TOD05	H-N	2005+TOD	0.0925	0.1641	0.0018	1375.19
hn.Site05	H-N	Site+2005	0.0979	0.1645	0.0018	1383.57
hn.TOD	H-N	TOD	0.0932	0.1644	0.0018	1386.08
hn.Site	H-N	Site	0.0983	0.1650	0.0018	1400.36
hn.Szn05	H-N	Szn+2005	0.0966	0.1649	0.0018	1401.53
hn.Szn	H-N	Szn	0.0970	0.1651	0.0018	1408.57
hn.05	H-N	2005	0.0964	0.1661	0.0018	1438.83
hn.model0	H-N	1	0.0968	0.1664	0.0018	1446.81

Table 3.2.1: All Monitors Covariate Model Comparison. Key functions include: hazard-rate (H-R) and half-normal (H-N). Covariates making up the formula include: Year, time of day (TOD), year 2005 only (2005) and no covariate included (1). Monitor16 Adjustment Model Comparison. Listed for each model: Cramer-von Mises, C-vM p-value; average detectability estimate, $\hat{P}_a$; standard error of the estimated detectibility, $se(\hat{P}_a)$; and the AIC difference, ΔAIC.

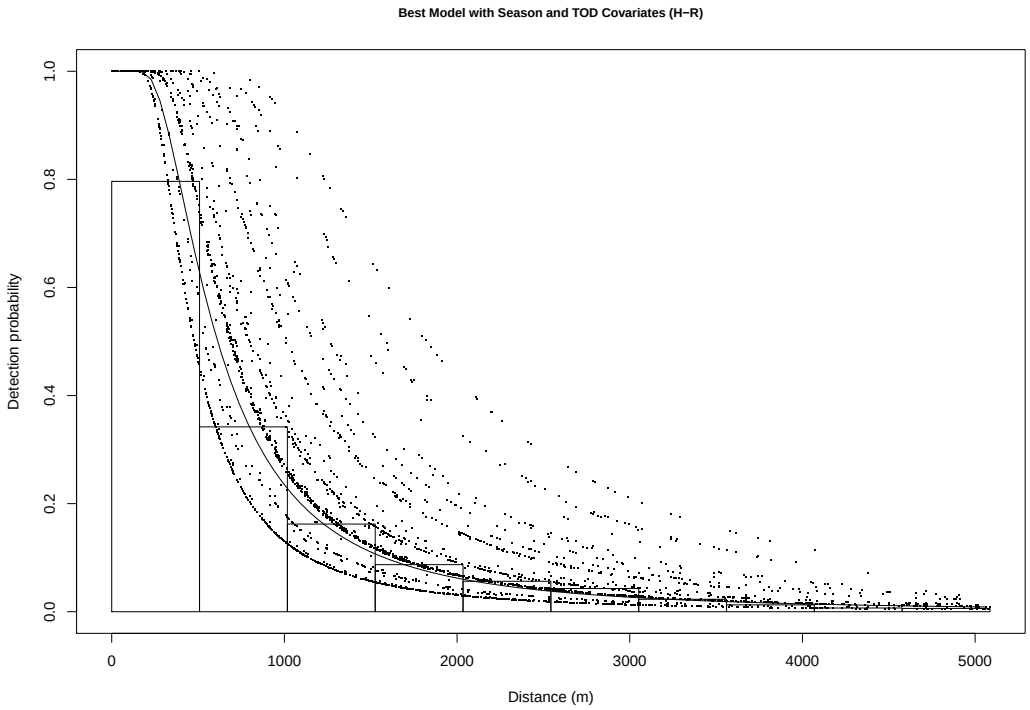


Figure 3.2.7: Histogram of true data with the detection function, $\hat{g}(x)$, (line) of the best fitting model with covariate terms only for all monitors. Dotted lines represent individual covariates.

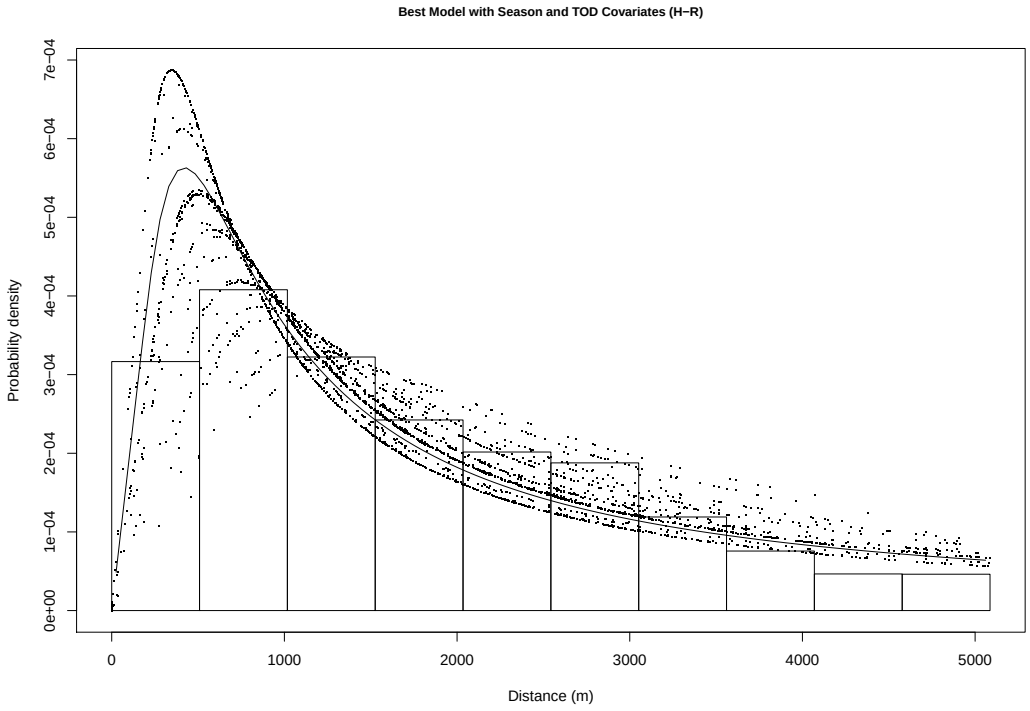


Figure 3.2.8: Histogram of true data with the probability density function, $\hat{f}(x)$, (line) of the best fitting model with covariate terms only for all monitors. Dotted lines represent individual covariates.

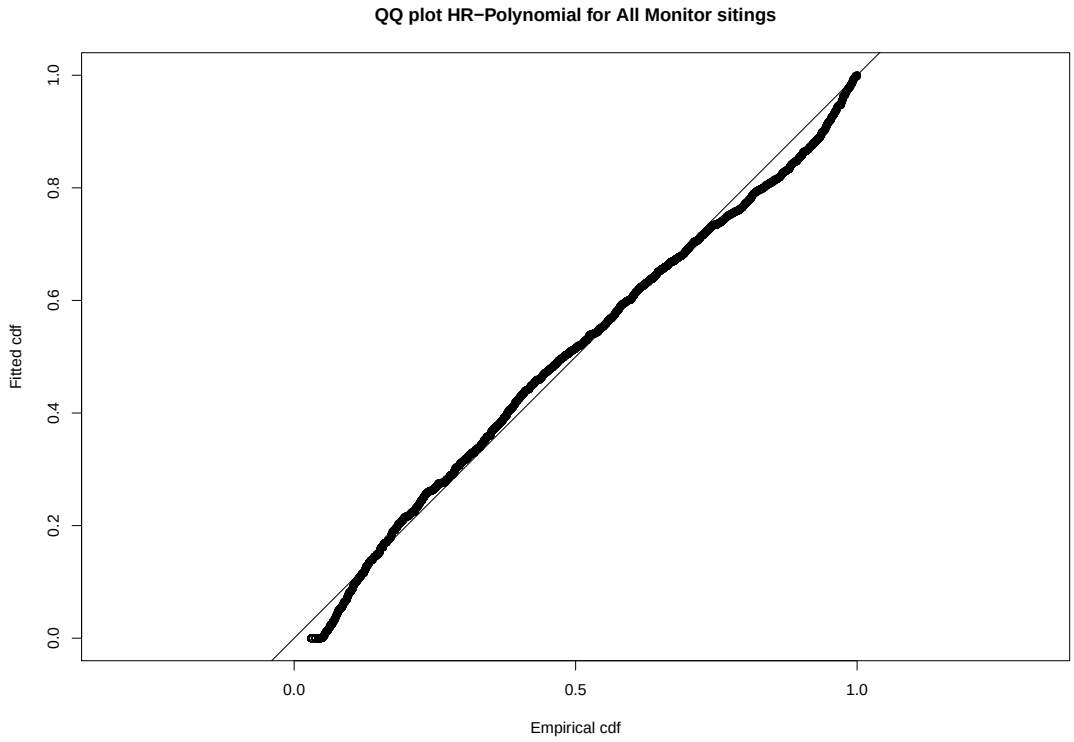


Figure 3.2.9: Q-Q plot for the best fitting model with covariate terms.

3.3 Assumptions and Limitiations

The assumptions for this chapter overlap with those previously outlined in Section 2.2.3 for general model assumptions. As well as these, we do have other assumptions and limitations of our data set which have allowed us to fit distance sampling models, these include:

- Cases of poor data entry, e.g. wrong digits entered for longitude or latitude.
- Point sampling is in the form of circular contours only. In the Monitor16 example (discussed in Section 3.1.1) – and other individual monitor cases - the observations appear to be skewed towards natural resources such as water and forest areas. These could be added as covariates if we had more local information on these locations or could be caused by a non-uniform herd distribution which could not be resolved by adding further covariates.

- Monitors recorded sightings from their home area which included observations from themselves and reporting's to them from other villages so precision of accuracy of recordings is unkown.

- Inconsistent number of recording per monitor is noticeably different, suggesting some monitors may put in more effort than others or that elephant density varies substantially over space.

- Some violation of standard distance sampling protocols (Section 2.2.1.4) such as monitors who would not stand at one location for a set period of time and the uncertainty of perfect detectibility at distance 0.

The main goal of this study was to calculate the expected detection probability of monitors recording elephant sightings given the elephant was present at that time. To explore this further and improve the study there is potential to remove the assumption of $\hat{P}_a = 1$ at distance zero [Langrock et al., 2013] which cannot be guaranteed in this case. Monitors are not set a time frame to record individuals as typical surveyors would in distance sampling, so as a result we have uncertainty that sightings may be missed (for example when the monitor is asleep). There is also the possibility to use 'ecological distance' methods [Royle et al., 2013a; Sutherland et al., 2015] which takes into account the animals perspective of the 'distance' that they travel. For example, uneven terrain and other different locations would require much more effort and energy to travel across for an individual and so to them, they would feel they are travelling a further distance and use much more energy to do so. The method uses Spatial Capture-Recapture methods which includes some spatial point process. Connectivity could be considered along with usage of space to estimate density, using a joint modelling approach. This model could then be improved by taking connectivity into account by using a joint modelling approach. These models could then possibility be analysed over time [Royle et al., 2013b] which our elephant data would allow for.

Chapter 4

Effect of Mitigations on Crop Loss

In this Section, we talk about how the various implemented mitigations had an effect on crop loss using methods from Section 2.3. The aim of this analysis was to discover which, if any, mitigations played a positive effect on reducing human-elephant conflict in the form of crop loss impact on farmers.

Extensive cleaning for data application was necessary as some data entries had no record of whether there was crop loss or not, and so just over 25% of entries had to be removed from this analysis. We assume that these observations were unbiased in regards to whether there was crop loss or not, regardless of the amount of crop loss that had occured. Remaining entries that were non-zero were changed from cost of crop loss into binary form, entries therefore consisted of 1 if there was crop loss and 0 if there was no crop loss. Once rows of data were removed, 'Siren' and 'Tripwire' mitigations had only FALSE entries and so could not be included in the analysis. All other individual mitigations as discussed in Chapter 1 were included with the addition of Distance (from observation to monitor in meters) and Count (number of elephants sighted).

4.1 Generalised Linear Model

First GLM methods were applied from Section 2.3.1 using the function 'glm' in R. As mentioned in Sections 2.3.1 and 2.3.4, there exists the package 'MASS' [Venables and Ripley, 2002] with function 'stepAIC' which performs an automated stepwise model selection by AIC. The option to stepwise search in both directions (the default mode) was used in these analyses which performs both forwards and backward selection on the chosen full model. The default model selection method is AIC with code '$k = 2$', however, this can be altered to use BIC model selection where more parameters used results in a higher penalty. For the BIC, we use $k = \log(n)$ where n is the number of observations. In this Section we will first discuss results from the AIC stepwise selection.

Iteration	±	Covariate	Deviance	Res. Df	Res. Dev	*AIC*
1				4212	4662.69	4696.69
2	-	factor(Catapult)	0.4411	4213	4663.14	4695.14
3	-	factor(Chillismoke)	0.6965	4214	4663.83	4693.83
4	-	Distance	1.4577	4215	4665.29	4693.29
5	-	factor(Arrows)	0.8457	4216	4666.14	4692.14
6	-	factor(Chillifence)	1.4632	4217	4667.60	4691.60

Table 4.1.1: AIC table of models added/removed at each iteration of the stepwise regression. Listed for each model: iteration; whether the covariate added or removed, ±; the covariate that has been added/removed; deviance; residual degrees of freedom; residual deviance and the AIC value.

Table 4.1.1 shows a summary of the output of the AIC stepwise selection at each iteration. Iteration 1 is the starting full model and so no covariates were added or removed from the model, it had an AIC value of 4696.69. The second iteration considers removing each covariate one at a time, compared with the original model. The best model at the second iteration removes the factor Catapult. The third iteration considers either removing each covariate in turn or adding the factor catapult back into the model; the resulting best model involves also removing the factor Chillismoke. This continues for the fourth to sixth iteration

Coefficients	Estimate	Std. Error	z-value	p-value
Intercept	-2.33	0.10	-23.22	< 0.0001
Fire stick (TRUE)	0.65	0.094	6.94	< 0.0001
Torch light (TRUE)	0.38	0.084	4.56	< 0.0001
AHP spotlight (TRUE)	0.51	0.082	6.18	< 0.0001
Other spotlight (TRUE)	-0.69	0.086	-8.04	< 0.0001
Cracker (TRUE)	0.38	0.083	4.57	< 0.0001
Noise (TRUE)	0.91	0.14	6.61	< 0.0001
Drum tin (TRUE)	0.41	0.081	5.02	< 0.0001
eFence (TRUE)	-1.01	0.29	-3.53	0.0004
Kunkie (TRUE)	-0.66	0.39	-1.71	0.0877*
Other mitigation (TRUE)	1.22	0.37	3.31	0.0009
Count	0.055	0.0036	15.27	< 0.0001

Table 4.1.2: Table of covariates from the best fitting stepwise AIC model. For all covariates there is no evidence to suggest that the parameter is not zero apart from Kunkie (marked with atrix, *) which we can accept as a non-zero parameter at a 10% significance only. Listed for each model: estimate, standard error, z-value is the test statistic for a hypothesis test of whether the coefficient value is zero; and p-value is the probability that the z-value is non-zero.

Iteration	$\pm$	Covariate	Deviance	Res. Df	Res. Dev	BIC
1				4212	4662.69	4804.64
2	-	Catapult (TRUE)	0.4411	4213	4663.14	4796.73
3	-	Chillismoke (TRUE)	0.6965	4214	4663.83	4789.08
4	-	Distance	1.4577	4215	4665.29	4782.19
5	-	Arrows (TRUE)	0.8457	4216	4666.14	4774.68
6	-	Chillifence (TRUE)	1.4632	4217	4667.60	4767.80
7	-	Kunkie (TRUE)	2.9724	4218	4670.57	4762.42

Table 4.1.3: BIC table of models added/removed at each iteration of the stepwise regression. Listed for each model: iteration; was the covariate added or removed, $\pm$; the covariate that has been added/removed; deviance; residual degrees of freedom; residual deviance and the BIC.

removing Distance, Arrows and Chillifence. The final best model according the AIC stepwise selection (as seen in Table 4.1.2) is therefore one which includes: Fire stick, Torch light, AHP spotlight, Other spotlight, Cracker, Noise, Drum tin, eFence, Kunkie, Other mitigation and Count. The hypothesis test in Table 4.1.2 tests whether a parameter is significantly different from zero. All covariates except Kunkie have p-values smaller than 0.05, meaning that there is no evidence to suggest that the parameters are not zero. We can see that eFence, Kunkie and Other spotlight all have negative estimates; this suggests that only these parameters have a negative effect on crop cost i.e. these mitigations appear to reduce the probability of whether there is crop damage.

Table 4.1.3 shows a summary of the output of the BIC stepwise selection at each iteration, again we can see that the first iteration is the full model with BIC value 4804.64 with no covariates added/removed. We can see in Table 4.1.3 compared to Table 4.1.1 that there is an extra iteration where Kunkie has been removed from the model. It is not always the case that both AIC and BIC stepwise selection will choose to add/remove covariates in the same order as each other but here this is the case with the addition of the seventh iteration. The best model according to the BIC stepwise selection (as seen in Table 4.1.4) is therefore one which includes: Fire stick, Torch light, AHP spotlight, Other spotlight, Cracker, Noise, Drum tin, eFence, Other mitigation and Count. In this model, we accept that all parameters have no evidence that they are not zero.

As a result, using GLMs we can conclude that only eFence and Other spotlight mitigations appear to have a negative effect on the probability of whether there is crop loss. In more general terms, it appears to be less likely that human-elephant conflict occurs in the form of destroyed crops when eFence and Other spotlight mitigations are put into place.

Coefficients	Estimate	Std. Error	z-value	p-value
Intercept	-2.33	0.10	-23.24	< 0.0001
Fire stick (TRUE)	0.67	0.093	7.15	< 0.0001
Torch light (TRUE)	0.38	0.084	4.56	< 0.0001
AHP spotlight (TRUE)	0.51	0.082	6.26	< 0.0001
Other spotlight (TRUE)	-0.69	0.086	-7.97	< 0.0001
Cracker (TRUE)	0.37	0.083	4.51	< 0.0001
Noise (TRUE)	0.90	0.14	6.52	< 0.0001
Drum tin (TRUE)	0.41	0.081	5.03	< 0.0001
eFence (TRUE)	-1.00	0.29	-3.52	0.0004
Other mitigation (TRUE)	1.22	0.37	3.32	0.0009
Count	0.055	0.0036	15.18	< 0.0001

Table 4.1.4: Table covariates from the best fitting stepwise BIC model. For all covariates there is no evidence to suggest that the parameter is not zero. Listed for each model: estimate, standard error, z-value is the test statistic for a hypothesis test of whether the coefficient value is zero; and p-value is the probability that the z-value is non-zero.

4.2 Random Monitor Effect

In the generalised linear model we did not take account of the fact that the observations were made by several different monitors. Here, we assume that all of the data collected has been from a range of people and so include the monitor as a random effect (RME) using MonitorID from the data and methods discussed in Section 2.3.3. We note that although the random effect is at the monitor level, the monitor only makes observations across a specific location. The random effect is therefore accounting for both monitor variability as well as spatial variability - we cannot distinguish between the two in this case.

Adding this RME to the best model outcome in the previous section and using the BIC stepwise selection (Section 4.1) produces the model as seen in Table 4.2.1 with the standard deviation of random effects being 1.237. For ease of interpretation we now call the best model with random effects $RMEmodel$. Using the likelihood ratio test between the original best model and the $RMEmodel$, we can reject the null hypothesis and therefore include the random effect in the new best model (LRT=435.02 and p-value< 0.0001). Both AIC and BIC values also

Fixed effects	Estimate	Std. Error	z-value	p-value
Intercept	-2.2309	0.2128	-10.49	< 0.0001
Fire stick (TRUE)	0.5630	0.1069	5.27	< 0.0001
Torch light (TRUE)	0.3684	0.0940	3.92	< 0.0001
AHP spotlight (TRUE)	0.4021	0.0943	4.26	< 0.0001
Other spotlight (TRUE)	-0.5734	0.1017	-5.64	< 0.0001
Cracker (TRUE)	0.3954	0.0926	4.27	< 0.0001
Noise (TRUE)	0.8488	0.1524	5.57	< 0.0001
Drum.tin (TRUE)	0.3015	0.0950	3.17	0.0015
eFence (TRUE)	-0.3648	0.2914	-1.25	0.2106**
Other mitigation (TRUE)	0.4559	0.3810	1.20	0.2315**
Count	0.0530	0.0044	12.08	< 0.0001

Table 4.2.1: Table of fixed effect covariates from the best fitting stepwise BIC model with added random monitor effect. For all covariates there is no evidence to suggest that the parameter is not zero with exception of eFence and Other mitigation (marked with a double astrix, **) where there is evidence to suggest that the parameter is not zero. Listed for each model: estimate, standard error, z-value is the test statistic for a hypothesis test of whether the coefficient value is zero; and p-value is the probability that the z-value is non-zero.

Model with RME	K	AIC	ΔAIC	BIC	ΔBIC	LogLik
eFence&OtherMit removed	10	4258.70	0.00	4322.19	0.00	-2119.35
OtherMit removed	11	4259.10	0.40	4328.94	6.75	-2118.55
eFence removed	11	4259.15	0.46	4329.00	6.81	-2118.58
RMEmodel	12	4259.55	0.85	4335.74	13.55	-2117.77
Best GLM (Section 4.1)	11	4692.57	433.87	4762.42	440.23	-2335.29

Table 4.2.2: AIC and BIC comparison table containing variations of the best model from Section 4.1 with the addition of a random monitor effect. Model description from top to bottom: *RMEmodel* with both eFence and Other mitigation removed; *RMEmodel* with Other mitigation covariate removed; *RMEmodel* with eFence covariate removed; *RMEmodel* (the original best model with RME); and the original best GLM from Section 4.1.

agree that the model with random monitor effect is better. However, we can see in Table 4.2.1 that two of the fixed effect covariates are now no longer significant (eFence and Other mitigation), we also consider re-testing the $RMEmodel$ against models with either or both of these covariates removed. Results for this test can be seen in Table 4.2.2. Looking at this table, we can see that all four models have almost identical AIC values and reasonably similar BIC values suggesting that the addition of the eFence and/or the Other mitigation in the random effects model does not effect the best model an enormous amount, however the best ranked model has both eFence and Other mitigation removed. All models according to AIC and BIC are ranked better than the original best model from Section 4.1 without the random monitor effect.

Fixed effects	Estimate	Std. Error	z-value	p-value
Intercept	-2.24	0.21	-10.42	< 0.0001
Fire stick (TRUE)	0.56	0.11	5.25	< 0.0001
Torch light (TRUE)	0.38	0.094	4.03	< 0.0001
AHP spotlight (TRUE)	0.39	0.094	4.22	< 0.0001
Other spotlight (TRUE)	-0.55	0.10	-5.48	< 0.0001
Cracker (TRUE)	0.41	0.092	4.43	< 0.0001
Noise (TRUE)	0.83	0.15	5.44	< 0.0001
Drum tin (TRUE)	0.31	0.095	3.32	0.0009
Count	0.053	0.0044	12.05	< 0.0001

Table 4.2.3: Table of fixed effect covariates from the best fitting stepwise BIC model with added random monitor effect. For all covariates there is no evidence to suggest that the parameter is not zero. Listed for each model: estimate, standard error, z-value is the test statistic for a hypothesis test of whether the coefficient value is zero; and p-value is the probability that the z-value is non-zero.

To conclude, the best GLMM model for this data is the model which includes: Fire stick, Torch light, AHP spotlight, Cracker, Noise, Drum tin, Other mitigation and Count; with the random effect following a normal distribution with mean zero and standard deviation 1.253 (Table 4.2.3). In this model, we accept the hypothesis that the estimated value of all parameters are non-zero at a 0.01% significance level, i.e. p-value< 0.001. In comparison to the best GLM model

from Section 4.1, we now only have only one mitigation with a negative estimate: Other spotlight. The best GLM model had three mitigations which were more likely to cause less crop damage and therefore cost, and now with the addition of a random monitor effect we have a better fitting model which suggests that only 'Other spotlight' plays this positive role. Note, a potential reason for eFence no longer being a significant mitigation may be due to the reason that the installation of eFences were inconsistent between locations and therefore not available to every monitor – with many locations having no eFence.

4.3 Limitations and Possible Extensions

Firstly, we note the extensive data cleaning and that we converted the crop loss variable to binary which throws away a lot of information including how much the crop loss actually cost the farmer. Approximately, 25% of data entries did not contain information as to whether there was crop loss or not. We assumed that these removed observations were unbiased, however, building a model to explore this using simulation could be executed in future work. Estimated cost of crop loss ranged from 100 to $240,000$ which displays the extent to which elephants would destroy crops in relation to if a mitigation was used and which one. Potentially, an extension such as a zero-altered Gamma distribution would allow us to incorporate this information into our analyses; but other distributions such as a mixture distribution would be applicable too.

Dénes et al. [2015] compared estimating the abundance of animal populations by comparing GLMs, distance sampling and other methods. They found there was bias if imperfect detection was not taken into account. From Chapter 3 we know that there is imperfect detection and so we could use simulation to investigate whether this adds bias in this case. Potentially, an individual monitor detection probability in co-operation with the random monitor effect by use of a two-way

interaction could be used. Another possibility is to look at various emerging models like in Dénes et al. [2015].

Another question that we could ask is whether these mitigations have the same effect when used alone as they do when used in conjunction with one another, for example, would the use of Drum tin with Other spotlight at the same time result in a positive effect of reducing cost of crop loss? Davies et al. [2011] found that the use of mitigations alone had some different responses to when mitigations were used in pairs. Due to the length of time taken to run the code, we were unable to carry out this additional analysis but this may explain why so many single mitigations used appear to increase crop loss. More can be read about GLM covariate interactions in Tsai and Gill [2013].

Chapter 5

Conclusion

Let us revisit the five Assam Haathi Project objectives from Section 1.1:

1) To model observer effort in recording elephant sightings in two spatially independent regions of Assam.

2) To validate and apply the model to produce a 'surface' of sampling effort across both study sites.

3) To use the validated model to determine relative abundance indices of elephants over the study period in both study sites.

4) To use the validated model to predict population-level responses of elephants to conflict mitigation strategies, based on levels of conflict and elephant sightings before and after the implementation of deterrent interventions.

5) To use the outcomes of this research to inform future design of surveys and monitoring.

We start by looking at objective **1)**.

To model the observer effort in recording elephant sightings in two spatially independent regions of Assam (Goalpara and Sonitpur), we looked at applying distance sampling methods to estimate the probability of detecting elephant herds

in Chapter 3. Due to some monitors not having enough data to be considered individually, we first looked at the probability of detection, $\hat{P}_a$, for two individual monitors from each site. Both monitors from Sonitpur – Monitor01 ($\hat{P}_a = 16\%$) and Monitor16 ($\hat{P}_a = 8.2\%$) – had detection probabilities which were derived from covariate models that were both deemed to be a good fit to the data by the Cramer-von Mises (C-vM) goodness of fit test. However, although both monitors covariate models from Goalpara – Monitor03 ($\hat{P}_a = 20.1\%$) and Monitor46 ($\hat{P}_a = 58.2\%$) – were not deemed to be a good fit to the data by C-vM, both of the top two models did agree on the estimate for the probability of detection. We can observe from this small number of monitor analyses, monitor probability of detection for elephant herds appear to be lower in Sonitpur than in Goalpara. Note, year was a covariate in each of the monitors best models with the exception of Monitor01 which included the covariate 'year 2005' vs. 'not year 2005'. This suggests that probability of detection by monitors varies over time.

Next, we looked at distance sampling methods to estimate the probability of detection for all data combined. The comparison of boxplots showing the radial distance of observed sightings by site can be seen in Figure 3.2.5. We can see that both boxplots are relatively similar to each other suggesting there is no obvious difference in distances observed between each site. This is reflected in the covariate results table (Table 3.2.1), concluding that there is no significant difference in observer effort in recording elephant sightings between the two spatially independent regions of Assam in the best fitting model – $hr.SznTOD$.

Objective **2)** was achieved by presenting results from Objective **1)** on a map (Figure 3.1.29). We produced a 'surface' of sampling effort across both study sites in the form of the probability of detecting elephant herds for each of the individual monitors previously mentioned. Using the best model according to AIC, we have $\hat{P}_a = 5\%$ for all monitors combined.

For objective **3)** we first looked at applying capture-recapture methods to the

data, in particular the Cormack-Jolly-Seber model from Section 2.1.2.2 to the elephant example in Section 2.1.3. Due to a small data set, we were only able to consider the constant model and whether parameters were dependant on site location, and so estimating population size accuratley using this data set was not possible. The AHP data violated multiple distance sampling assumptions from Section 2.2.3 and so we were not able to estimate the population size of elephants by using this method. Numerous limitations have meant we are unable to answer objective **3)** accurately.

We have a lack of data in relation to when all permanent mitigations were first implemented and also do not have data of mitigations in place when zero elephants were sighted. As a result, we were restricted in answering objective **4)**. We looked at applying generalised linear modelling methods from Section 2.3 to model the effect of mitigations on human-elephant conflict in the form of crop loss in Chapter 4. Looking at the best fitting GLM using BIC (Table 4.1.4), 'Other spotlight' and 'eFence' were the only significant mitigations fitted to the model that had negative estimates – meaning that when implemented, they reduced crop loss. However, upon fitting a GLMM with random monitor effect we discovered that the best fitting model according to BIC (Table 4.2.3) only included 'Other spotlight' as a significant mitigation with a negative estimate. We discussed in Section 4.2 that the model containing both these mitigations with the random monitor effect was not the best model, but it was proved to be a good fit to the data if both or either of these were added to the model (Table 4.2.2). Therefore, we can conclude that there is significant evidence to suggest that there is an association between crop loss and both 'Other spotlight' and 'eFence'. The association suggests that when these mitigations are present, we see a reduction in crop loss which biologically fits with the ecological expectation.

In light of this research, to inform future design of surveys and monitoring – Objective **5)** – we suggest that the mitigiations 'Other spotlight' and 'eFence' in

particular are implemented to more villages across the two study sites in the hope that these will help to reduce elephant-conflict of crop deprivation further. For more reliable and complex types of analysis to take place in future, more information should be recorded when there are sightings of elephants. This includes, but is not limited to: a stronger effort for all monitors to record more information for each data entry column, particularly the amount of crop loss and other forms of human-elephant conflict (if applicable); train more community members to increase the number of monitors; and creating logs of when mitigations are put into and taken out of place.

Appendix A

Data Inputs

A.0.0.1 MARK

(Excel) Monthly capture-recapture elephant data by herd.

	2005			2006												2007												2008												2009												2010											
Month:	O	N	D	J	F	M	A	M	J	J	A	S	O	N	D	J	F	M	A	M	J	J	A	S	O	N	D	J	F	M	A	M	J	J	A	S	O	N	D	J	F	M	A	M	J	J	A	S	O	N	D	J	F	M	A	M	J	J	A	S	O	N	D
GP01	-	-	1	1	0	1	0	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0	1	1	0	0	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	1	0
GP01A	-	-	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	0	1	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
GP01B	-	-	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
GP01&GP02	-	-	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
GP02	-	-	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	1	0	1	0	0	0	0	1	1	0
GP02A	-	-	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GP02B	-	-	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GPMG	-	-	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	0	0	0	0	0	0	0	1	1	0	0	0	0
HP04	-	-	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
SP01	-	-	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SP02	-	-	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SP03	-	-	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SP04	-	-	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0	0	0	1	1	1	0	1	1	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	0
SP05	-	-	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SP06	-	-	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SP07	-	-	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Appendix B

R Code

B.0.0.2 Geosphere

(R) Example of how distance in meters was calculated using Hijmans et al. [2019].

```
library("Distance")
library(geosphere)
library("readxl")

# Data collected by Monitor01

#Sightings
datael<- read_excel("SitingsAll2.xlsx", sheet=1)
head(datael)

#Monitor Locations
datamon<- read_excel("GoalparaSonitpur2.xlsx", sheet=1)
head(datamon)
```

```
n <- length(datael$Latitude)
n2 <- length(datamon$Latitude)

latM <- c()
lonM <- c()
lonobs <- c()
latobs <-c()

for (i in 1:n){
  if (datael$Monitor[i]=="Monitor01") {
    for (j in 1:n2){
      if (datamon$Name[j]==datael$Monitor[i]) {
        latM <- c(latM,datamon$Latitude[j])
        lonM <- c(lonM,datamon$Longitude[j])
        latobs <- c(latobs,datael$Latitude[i])
        lonobs <-c(lonobs,datael$Longitude[i])
      }
    }
  }
}

# Geosphere package calculates distance from Monitor01's location to the obser
distance <- distVincentyEllipsoid(cbind(lonM,latM), cbind(lonobs, latobs))
```

Bibliography

S. C. Amstrup, T. L. McDonald, and B. F. J. Manly. *Handbook of Capture-Recapture Analysis*. Princeton University Press, 2005.

F. B. Baccaro and G. Ferraz. Estimating density of ant nests using distance sampling. *Insectes Sociaux*, 60:103–110, 2012.

D. Bates, M. Mächler, B. Bolker, and S. Walker. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015. doi: 10.18637/jss.v067.i01.

M. E. Brooks, K. Kristensen, K. J. van Benthem, A. Magnusson, C. W. Berg, A. Nielsen, H. J. Skaug, M. Mächler, and B. M. Bolker. glmmTMB Balances Speed and Flexibility Among Packages for Zero-inflated Generalized Linear Mixed Modeling. *The R Journal*, 9(2):378–400, 2017. doi: 10.32614/RJ-2017-066. URL `https://doi.org/10.32614/RJ-2017-066`.

C. Brownie, D. R. Anderson, K. P. Burnham, and D. S. Robson. Statistical inference from band recovery data - a handbook. *United States Department of the Interior, Fish and Wildlife Service, Washington, DC.*, 156, 1985.

S. T. Buckland, D. R. Anderson, K. P. Burnham, J. Laake, D. L. Borchers, and L. Thomas. *Introduction to Distance Sampling*. Oxford Press, 2001.

S. T. Buckland, D. R. Anderson, K. P. Burnham, J. Laake, D. L. Borchers, and L. Thomas. *Advanced Distance Sampling: Estimating Abundance of Biological Populations*. Oxford Press, 2004.

S. T. Buckland, E. A. Rexstad, T. A. Marques, and C. S. Oedekoven. *Distance Sampling: Methods and Applications*. Springer, 2015.

K. P. Burnham and D.R. Anderson. The need for distance data on transect counts. *Journal of Wildlife Management*, 48:1248–1254, 1984.

K. P. Burnham and D.R. Anderson. *Model Selection and Inference: A Practical Information-Theoretic Approach*. Springer, New York, 1998.

K. P. Burnham, D. R. Anderson, and J. L. Laake. Robust estimation from line transect data. *Journal of Wildlife Management*, 43:992–996, 1979.

K. P. Burnham, D. R. Anderson, and J. L. Laake. Estimation of density from line transect sampling for biological populations. *Wildlife Monographs*, 72:1–202, 1980.

E. A. Catchpole, S. N. Freeman, and B. J. T. Morgan. Modelling age variation in survival and reporting rates for recovery models. *Journal of Applied Statistics*, 22:597–609, 1995.

E. A. Catchpole, S. N. Freeman, B. J. T. Morgan, and M. P. Harris. Integrated recovery/recapture data analysis. *Biometrics*, 54:33–46, 1998.

D. G. Chapman. Some properties of the hypergeometric distribution with application to zoological censuses. *University of California Publications in Statistics*, 1:131–160, 1951.

Chester Zoo. Assam haathi project @ONLINE. URL `https://www.chesterzoo.org/what-we-do/our-projects/assam-haathi-project/`.

D. J. Cole, B. J. T. Morgan, E. A. Catchpole, and B. A. Hubbard. Parameter redundancy in mark-recovery models. *Biometrical Journal*, 54:507–523, 2012.

E. G. Cooch and G. C. White. *Program MARK, A Gentle Introduction*, 2019. URL `file:///C:/Users/cmf33/AppData/Local/Temp/Temp1_mark_book%20(1).zip/mark_book.pdf`.

R. M. Cormack. Estimates of survival from the sighting of marked animals. *Biometrika*, 51:429–438, 1964.

Darwin Initiative. Assam haathi project @ONLINE. URL https://www.darwininitiative.org.uk/documents/16007/18668/16-007%20AR2%20Ann3.3%20%20Project%20brochure.pdf.

T. E. Davies, S. Wilson, N. Hazarika, J. Chakrabarty, D. Das, D. J. Hodgson, and A. Zimmermann. Effectiveness of intervention methods against crop-raiding elephants. *Conservation Letters*, 4:346–354, 2011.

R. K. Didham, P. M. Hammond, J. H. Lawton, P. Eggleton, and N. E. Stork. Beetle species responses to tropical forest framgmentation. *Ecological Monographs*, 68, 1998.

F. V. Dénes, L. F. Silveira, and S. R. Beissinger. Estimating abundance of unmarked animal populations: accounting for imperfect detection and other sources of zero inflation. *British Ecological Society*, 6:543–556, 2015.

W. G. Dorgeloh. Density estimates of francolin in a Sporobolus ioclados-Acacia tortilis Savanna using distance sampling. *South African Journal of Wildlife Research*, 35:89–94, 2005.

J. J. Faraway. *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*. Chapman and Hall/CRC, 2006.

S. N. Freeman and B. J. T. Morgan. A Modelling Strategy for Recovery Data from Birds Ringed as Nestlings. *Biometrics*, 48:217–235, 1992.

L. Gygax. Evolution of group size in the dolphins and porpoises: interspecific consistency of intraspecific patterns. *Behavioral Ecology*, 13:583–590, 2002.

X. A. Harrison, L. Donaldson, M. E. Correa-Cano, J. Evans, D. N. Fisher, C. Goodwin, B. S. Robinson, D. J. Hodgson, and R. Inger. A brief introduction to mixed effects modelling and multi-model inference in ecology. *PeerJ*, 6, 2018. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5970551/.

W. Hemmingsen, P. A. Jansen, and K. MacKenzie. Crabs, leeches and trypanosomes: an unholy trinity? *Marine Pollution Bulletin*, 50:336–339, 2005.

A. Henningsen and O. Toomet. maxLik: A package for maximum likelihood estimation in R. *Computational Statistics*, 26(3):443–458, 2011. doi: 10.1007/s00180-010-0217-1. URL `http://dx.doi.org/10.1007/s00180-010-0217-1`.

R. J. Hijmans, E. Williams, and C. Vennes. geosphere: A package for spherical trigonometry in R. 2019.

B. A. Hubbard, D. J. Cole, and B. J. T. Morgan. Parameter redundancy in capture–recapture–recovery models. *Statistical Methodology*, 17:17–29, 2014.

G. M. Jolly. Explicit estimates from capture–recapture data with both death and immigration-stochastic model. *Biometrika*, 52:225–247, 1965.

Z. Karaca, H. S. Wong, and R. L. Mutter. Duration of patients' visits to the hospital emergency department. *BMC Emegency Medicine*, 12, 2012. URL `https://doi.org/10.1186/1471-227X-12-15`.

R. King and R. S. McCrea. Capture-Recapture Methods and Models: Estimating Population Size. *Statistics: Integrated Population Biology and Modelling*, 40, 2019.

J. Kuha. AIC and BIC: Comparisons of Assumptions and Performance. *Sociological Methods & Research*, 33:188–229, 2004.

J. L. Laake, D. S. Johnson, and P. B. Conn. marked: An R package for maximum-likelihood and MCMC analysis of capture-recapture data. *Methods in Ecology and Evolution*, 2013.

R. Langrock, D. L. Borchers, and H. J. Skaug. Markov-Modulated Nonhomogeneous Poisson Processes for Modeling Detections in Surveys of Marine Mammal

Abundance. *Journal of the American Statistical Association*, 108(503):840–851, 2013.

J-D. Lebreton, K. P. Burnham, J. Clobert, and D. R. Anderson. Modeling survival and testing biological hypotheses using marked animals: a unified approach with case studies. *Ecological Monographs*, 62:67–118, 1992.

J. Locke and P.G. Hulme. *OCR Gateway GCSE Biology Student Book*. Oxford University Press, 2016.

E. Matechou, B. J. T Morgan, S. Pledger, J. A. Callzo, and J. E. Lyons. Integrated analysis of capture-recapture-resighting data and counts of unmarked birds at stop-over sites. *Journal of Agricultural, Biological, and Environmental Statistics*, 18:120–135, 2013.

E. Matechou, E. Dennis, S. N. Freeman, and T. Brereton. Monitoring abundance and phenology in (multivoltine) butterfly species: a novel mixture model. *Journal of Applied Ecology*, 51:766–775, 2014.

R. S. McCrea and B. J. T. Morgan. *Analysis of Capture-Recapture Data*. Chapman & Hall / CRC Press, 2014.

P. McCullagh and J. A. Nelder FRS. *Generalized linear models*. Chapman and Hall, 1999.

D. L. Miller, E. R., L. Thomas, L. Marshall, and J. L. Laake. Distance sampling in R. *Journal of Statistical Software*, 89(1):1–28, 2019. doi: 10.18637/jss.v089.i01.

B.J.T. Morgan. *Applied Stochastic Modelling*. Chapman and Hall/CRC, 2008.

C. A. Peres. General Guidelines for Standardizing Line-Transect Surveys of Tropical Forest Primates. *Neotropical Primates*, 7:11–16, 1999.

S. Pledger, J. Pollocj M. Efford, J. Callazo, and J.Lyons. Stopover duration analysis with departure probability on unknown time since arrival. *Environmental and Ecological Statistics*, 3:1071–1082, 2009.

T. J. Reynolds, R. King, J. Harwood, M. Frederiksen, M. P. Harris, and S. Wanless. Integrated data analysis in the presence of emigration and mark loss. *Journal of Agricultural, Biological, and Environmental Statistics*, 14:411–431, 2009.

G. H. Rodda and E. W. Campbell. Distance sampling of forest snakes and lizards. *Herpetological Review*, 33:271–274, 2002.

J. A. Royle, R. B. Chandler, K. D. Gazenski, and T. A. Graves. Spatial capture–recapture models for jointly estimating population density and landscape connectivity. *Ecology - Ecological Society of America*, 94:287–294, 2013a.

J. A. Royle, R. B. Chandler, C. C. Sun, and A. K. Fuller. Integrating resource selection information with spatial capture–recapture. *Methods in Ecology and Evolution*, 4:520–530, 2013b.

C. J. Schwarz and A. N. Arnason. A General Methodology for the Analysis of Capture-Recapture Experiments in Open Populations. *Biometrics*, 52:860–873, 1996.

G. A. F. Seber. A note on the multiple-recapture census. *Biometrika*, 52:249–259, 1965.

W. J. Sutherland, D. B. Roy, and T. Amano. An agenda for the future of biological recording for ecological monitoring and citizen science. *Biological Journal of the Linnean Society*, 115:779–784, 2015.

L. Thomas, S. T. Buckland, K. P. Burnham, D. R. Anderson, J. L. Laake, D. L.

Borchers, and S. Strindberg. Distance sampling. *Encyclopedia of Environmetrics*, 1:544–552, 2002.

L. Thomas, S. T. Buckland, E. A. Rexstad, J. L. Laake, S. Strindberg, S. L. Hedley, J. R. B. Bishop, T. A. Marques, and K. P. Burnham. *Distance software: design and analysis of distance sampling surveys for estimating population size*, 2010.

T. Tsai and J. Gill. Interactions in Generalized Linear Models: Theoretical Issues and an Application to Personal Vote-Earning Attributes. *Social Sciences*, 2: 91–113, 2013.

W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. URL `http://www.stats.ox.ac.uk/pub/MASS4/`. ISBN 0-387-95457-0.

J. Vicente, U. Höfle, J. M. Garrido, I. G. Fernández-De-Mera, R. Juste, M. Barral, and C. Gortazar. Wild boar and red deer display high prevalences of tuberculosis-like lesions in Spain. *Veterinary Research*, 37:107–119, 2006.

A. I. Ward, P. C. L. White, and C. H. Critchley. Roe deer Capreolus capreolus behaviour affects density estimates from distance sampling surveys. *Journ*, 34, 2004.

B. K. Williams, J. D. Nichols, and M. J. Conroy. *Analysis and Management of Animal Populations*. Academic Press, 2002.

S. Wilson, A. Zimmermann, N. Hazarika, G. Narayan, J. Chakrabarty, D. Baruah, P. Mitra, P. J. Deka, M. Narayanan, D. J. Das, B. Hazarika, L. K. Nath, and A. Baruah. Living with Elephants in Assam: a handbook. *Assam Haati Project*, pages 1–58, 2009.

S. Wilson, T. E. Davies, N. Hazarkia, and A.Zimmermann. Understanding spatial and temporal patters of human-elephant conflict in Assam, India. *Oryx*, 2015.

T. W. Yee. *Vector Generalized Linear and Additive Models: With an Implementation in R*. Springer, New York, USA, 2015.

T. W. Yee and C. Moler. *VGAM: Vector Generalized Linear and Additive Models*, 2020. URL `https://CRAN.R-project.org/package=VGAM`. R package version 1.1-3.

A. Zimmermann, T. E. Davies, N. Hazarika, S. Wilson, J. Chakrabarty, B. Hazarika, and D. Das. Community-based human-elephant conflict management in Assam. *Gajah*, 30:34–40, 2009.

A. F. Zuur, E. N. Ieno, N. J. Walker, A. A. Saveliev, and G. M. Smith. *Mixed Effects Models and Extensions in Ecology with R*. Springer, 2009.