Statistical Methods for Assessing Habitat Preferences

by

Dieter Ayers

B.Sc, University of Calgary, 1995

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

Master of Science

in

THE FACULTY OF GRADUATE STUDIES

(Department of Statistics)

we accept this thesis as conforming
to the required standard

The University of British Columbia

January 2000

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of STATISTICS

The University of British Columbia
Vancouver, Canada

Date February 10, 2000

DE-6 (2/88)

Abstract

It is often the case that samples are taken in a non-random fashion. This thesis attempts to define a methodology by which some analysis can be performed on a specific kind of non-random sample. In studies of wildlife behaviour, a common method of sampling involves tagging an animal and relocating it in subsequent time periods. Considering the number or type of animal present at sampling locations as a random sample is erroneous, as the locations were chosen by the animal and not in a random fashion. Further, with data such as this is not possible to draw conclusions about areas where no animals were observed, as it is unknown whether these areas were truly free of animals.

We treat the observed data as conditional on the presence of an animal, and then use a Bayesian approach to estimate the probability of finding animals in any given location. This results in a method that allows for mapping the propensity of a certain area to be chosen by an animal in the future.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Many wildlife biologists are interested in the way that animals use their habitat (*c.f.* Aebischer et al., 1993; Welch et al., 1990), and the way in which the animals distribute themselves over the geographical areas that they inhabit (*c.f.* Osborne & Tigar, 1992; Anderson and Rongstad, 1989). Knowledge of this sort can be used in a number of other projects, such as those for conservation or resource management. There are many different reasons that an animal would display a certain behavioural pattern, and to discover the true pattern of habitat usage a detailed study is necessary. Discovering the reasons for a specific pattern of behaviour is exceedingly difficult, as the animals don't provide unequivocal reasons for why they adopt certain strategies, but with a comprehensive study it may be possible to infer reasons for their behaviour. Information on where that animal spends its time is needed, as this can be related to other variables in the spaces that the animals occupy, and the animals' habitat preferences can be deduced.

Many reasons seem obvious. The presence of a quality food source will likely increase the probability of an animal's presence at a particular point. The same is likely true of an area that offers a degree of protection from predators. Other factors are more specific to different types of animals, such as the amount of plant coverage present. For example, it is unlikely that large animals would be found in areas of a

forest that are choked with thickets.

In addition to factors affecting individual animals, are those that affect the entire population of animals. The presence of conspecific animals may be reason for an animal to inhabit an area, or it may be reason to avoid that area. If the animal under study is a prey species an individual animal would stand a greater chance of being preyed upon if it were alone, than it would in a group setting. Conversely, if the animal were a predator it may be in its best interest to act alone, rather than compete, or share, with another animal. This is also true in territorial animals, where the presence of one animal will decrease the chance of finding a second in the same area. Between the territorial and social extremes lie those animals that are relatively indifferent to the whereabouts of their ilk. These types of animals will display a random pattern of location over the area of the study.

The problem that faces the investigator in studying habitat usage and animal location is that the animal is (presumably) aware of what it is doing, but the researcher is not. A properly designed study should be able to identify preferences and patterns displayed by the animal, while discounting random behavioural fluctuations.

The current literature seems to divide into two main types of analysis and study. Some studies pertain to resource selection and habitat usage (Aebischer et al., 1993; Welch et al., 1990). Others focus on determination of the home range of an animal (Osborne & Tigar, 1992; Anderson and Rongstad, 1989). Studies about resource selection attempt to determine the environmental characteristics that animals select and use preferentially over other characteristics. These studies generally identify certain resource types that are associated with an animal, and then compare those resources to the types of resources in the surrounding area (Alldredge & Ratti 1986, 1992). Principally, tests of proportions are undertaken in an attempt to determine if the animal uses specific resources proportionally to their availability.

It is also becoming increasingly common to see more sophisticated analysis in the literature. Bayesian analysis has been used in modelling bird distributions (Tucker et al. 1997), and resource selection functions have been identified as a viable method of analysis (Manly et al. 1993). These methods are substantially more computer intensive that other methods, but this is only a minor problem with modern computing capabilities. These authors also make heavy use of GIS software.

Home range determination simply attempts to define an area in which an individual or group of individuals spends most of their time. These studies require only locations of animals, which are then used to identify the most likely area to find an individual animal.

A number of methods of analysis have been proposed for both types of study and shall be outlined in a later chapter. These two types of studies are mostly independent of one another, but there is occasional overlap. For example, some studies of resource selection compare animal usage to home range availability.

Complementing the number of techniques for analysis are the many different methods for collecting data. Each method has it's strengths and weaknesses, and some are more applicable to certain organisms than others. The type of analysis is obviously dependent upon the type of sampling method used to collect the data. The quality and format of the data can be immensely affected by the sampling method, and as a consequence, it is necessary to acknowledge the methods prior to analysis. For a comprehensive review of ecological sampling methods the reader is referred to Krebs (1989). It is from this reference that the following overview is taken.

The two most commonly used methods for sampling animals have in the past been quadrat sampling and transect sampling. Quadrat sampling involves dividing the study area into small units (quadrats) and sampling these. Transect sampling consists of defining distinct paths through the study area (transects). These paths are then followed and samples are taken at prespecified locations on the path. The

actual sampling is still governed by standard sampling theory. Simple random sampling or more complex stratified designs can be implemented. Under both of these sampling schemes numbers of animals or presence/absence can be recorded. The methods involving random sampling of geographical areas lend themselves to standard statistical analysis. In these situations, the number of animals in each area is a random variable. This can be modeled with an appropriate distribution, such as the binomial (for presence/absence) or the Poisson (for counts). Models incorporating covariates and spatial correlation can be constructed relatively easily.

More recently, remote sensing techniques have allowed for individual animals to be followed and their locations recorded at the times of sampling (White & Garrott, 1990). The data that is collected from remotely sensed animals is of a different nature than the data obtained from the previous methods. The animal can be thought of as having a true "trajectory", that is a path taken over the study period. The data are then a group of points on this trajectory. As the number of sampling points increases, the true trajectory becomes more completely described. These data represent a not-so-random sample of locations in the study area. The not-so-randomness is due to the fact that the animal has selected those locations based on some set of criteria, that may be based on the vegetation at those sites, the presence of other individuals, the presence of predators or prey, or other factors unknown to the investigator. From these locations it is possible to identify the states of any number of variables of interest. Thus, in this situation the presence of the animal is not a random variable (indeed, it is a necessity), but the state of other variables at the sampled locations are random variables.

In this thesis we outline techniques that can be used to map the locations of wildlife, while taking into account the habitat preferences of the wildlife. Thus, these methods are really an amalgamation of resource selection methods and mapping methods. These new methods should also be able to circumvent common problems associated with the analysis of data found in such studies.

We aim to produce techniques for mapping probabilities that accurately reflects the propensity of a bear to occupy a given area. This map should be constructed from both the information about where bears have been located and the covariates in each location. The attainment of this goal is not straightforward due to the methods used for data collection. The mapping technique would ideally allow for an assessment of the usefulness of each variable in the area (significance of terms in the model).

This thesis is organized as follows: Firstly, we discuss common methods of analysis, both for mapping and resource utilization. Reasons for their inapplicability to the present data are identified. After the old methods are outlined new methods with considerable promise are presented. The grizzly bear data is then presented and analysed via the new method. The thesis concludes by explaining drawbacks to these new methods.

Chapter 2

Description of the data

The data used to illustrate the methods of analysis was collected as part of the Ph.D. research of Rob Wielgus, formerly a student in the Department of Forest Sciences at the University of British Columbia. The data were initially used in a study of the sexual differences in habitat usage in Grizzly bears (*Ursus arctos*) (Wielgus, 1993). The data consists of information about 26 grizzly bears. In addition to the data pertaining to the bears themselves, there is information on the locations of bears, and the types of surrounding habitat and landscape. This data were made available to me by Mr. Pierre Vernier, a Research Associate in the Center for Applied Conservation Biology at UBC.

The data were collected in the Selkirk mountains of British Columbia and Idaho. The Selkirks are located in the interior eastern region of B.C., and are characterised by long cold winters and cool summers. The primary vegetation in this region is coniferous trees. The location of the study area can be seen in Figure 2.1.

The original experiment was set up to determine how different bears use the habitat that is available to them. To achieve this end, data was collected that identified the location of individual bears at given times. As bears were included in the study information regarding their age, sex, number of cubs, and other such

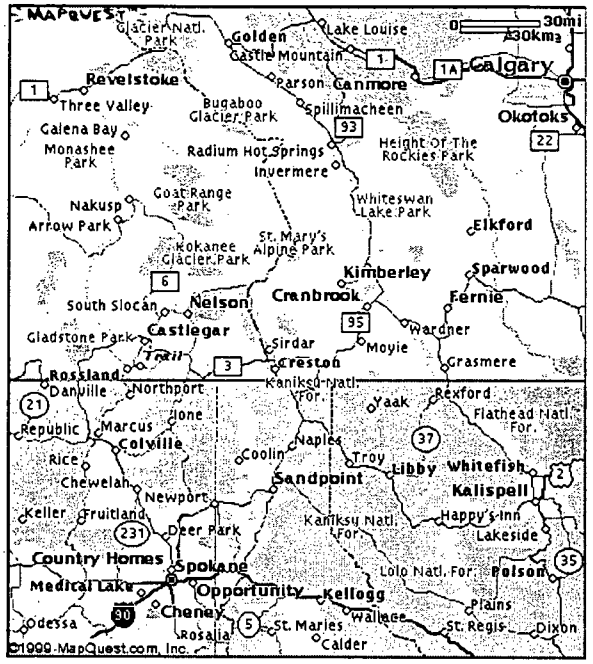

Figure 2.1: Location of study area

variables were recorded.

The entire data set contains information on the following variables: bears, broad ecological classification, presence of roads, and the forest classification. The geographical data were initially contained in ArcView database files, and the auxiliary information on the bears was contained in a Microsoft Excel spreadsheet. The data was exported as ASCII to facilitate numerical calculations. I had initially planned to present the final analysis in ArcView as well, but certain hardships precluded this. As a consequence, I have engaged Splus as a surrogate GIS. Thus, maps and other graphical methods of presentation are all done in Splus.

2.1 Roads

This area of the Columbia valley is used extensively for both logging and recreational purposes. As a consequence, many roads have been constructed throughout the area. These roads tend not to follow easy terrain such as valley bottoms, as it is necessary to enter fairly remote areas to access valuable timber and recreational areas. Such roads tend to be wildly convoluted, and subsequently cover a great deal of ground.

The roads in the study area are not explicitly mapped, but rather, zones of different proximity are presented. Three different zones are identified. The different zones are defined primarily by convenience, although there may be some biological justification for the use of these arbitrary boundaries. The zones are as follows: 0-250m from road, 250-500m from road, and >500m from road. The data from the grizzly bears is quite sparse, and very few observations were recorded in each of the first two categories. To facilitate analysis, the first two categories were collapsed. Thus, in all analysis to follow, the first two categories are grouped together, creating a binary variable with levels <500m and >500m from a road.

A map of the roads can be seen in Figure 2.2. This map was constructed in Splus, based on an ASCII file that was outputted from ArcView. Proportions of observations in each road category are presented in Table 2.1.

distance from road (m)	proportion
< 250	0.2508
$250 - 500$	0.1295
> 500	0.6197

Table 2.1: Proportion of area in road categories

2.2 BEC

Another variable contained in the GIS maps is the "Broad Ecological Classification". This is a label describing the biogeoclimatic zone. These zones are named

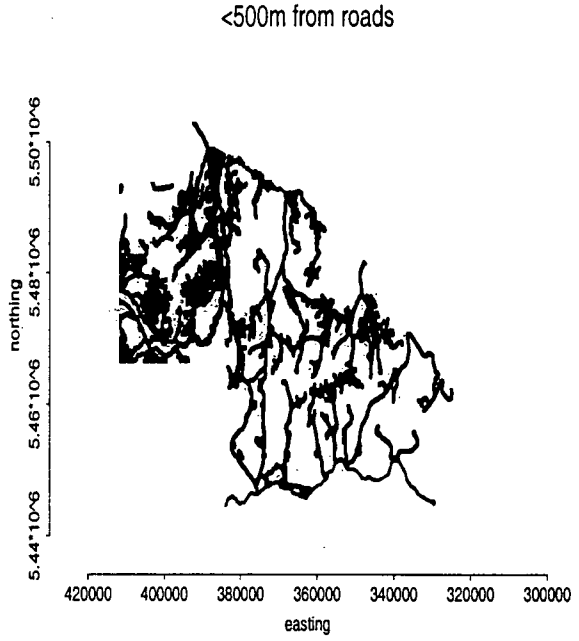

Figure 2.2: Areas less than 500m from roads

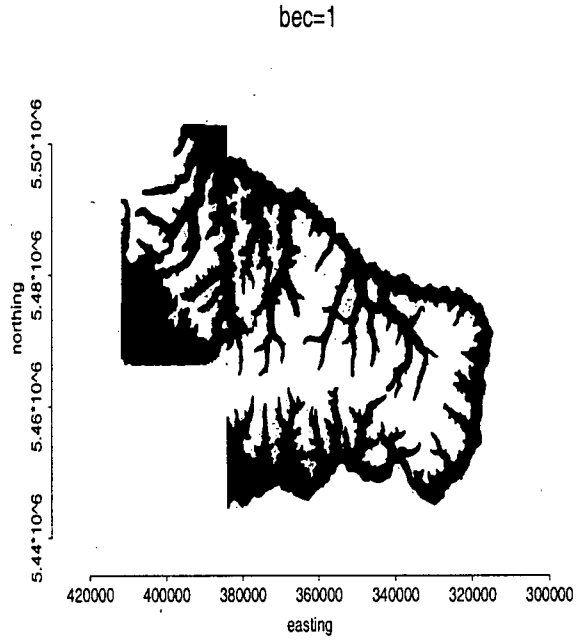

Figure 2.3: Alpine Tundra

after the dominant vegetation, and contain information on soil characteristics, vegetation types amongst other factors. There are three major zones in this study area: Engelmann Spruce/Subalpine Fir, Interior Cedar/Hemlock, and Alpine tundra. The areas occupied by these different zones can be seen in figures 2.3 (Alpine Tundra), 2.4 (Interior Cedar/Hemlock), and 2.5 (Engelmann Spruce/Subalpine fir).

This variable is related to the bear variable in that it represents areas that the bears may inhabit. The different levels of this variable may provide different quantities of food for the bears, or different food quality, or perhaps one area provides a sheltered area for sleeping. Bears may show a preference for one of these over another, and if this is the case it would be possible to predict bear presence based upon the biogeoclimatic zone at a given site.

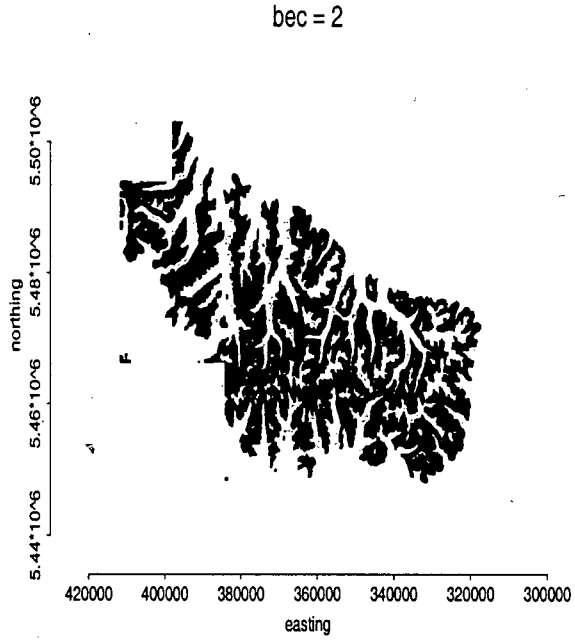

Figure 2.4: Interior Cedar/Hemlock

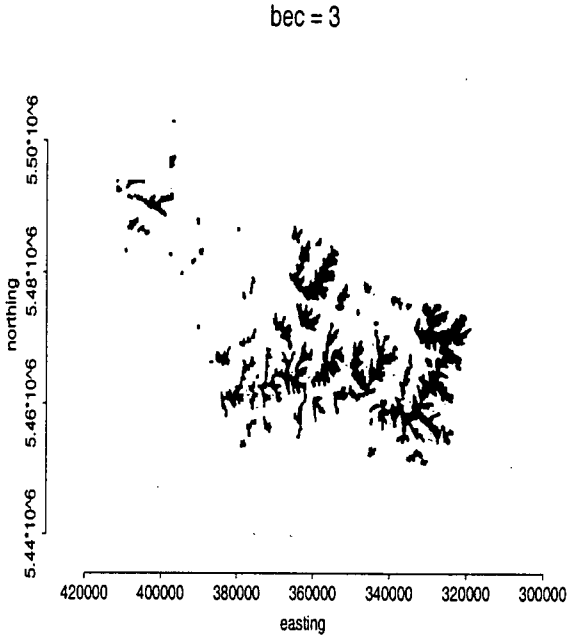

Figure 2.5: Engelmann Spruce/Subalpine fir

bec class	proportion
1	0.4867
2	0.4133
3	0.1000

Table 2.2: Proportion of area in different BEC categories

2.3 Forest

The remaining habitat variable is labeled forest, and it reflects the amount and age of the dominant vegetation of the area. The first level of this variable describes areas that are not forested. These areas could be rock outcroppings, rivers, fields, farmland, or other similar classes of land. This level is labeled "1". The second level consists of forested wildlands that are not heavily treed. This includes all areas that have undergone some recent disturbance such as logging, a fire, or an avalanche, amongst others. These areas are undergoing a regeneration of the forest, and are dominated by shrubs and quick growing vegetation. A number of young trees are establishing themselves at this point. The third class is that of "young forest". These areas are typically defined by young trees and lots of shrubs. The last possible outcome in this variable is "mature forest". These areas are covered by large old trees, and are prime candidates for logging. The understory has less shrubbery than the previous category. Figures 2.6 through 2.9 identify the parts of the study area covered by each of the different forest types.

Again, this variable is related to the bear data because the different levels likely have different quantities or quality of resources. It should be noted that the bec variable and the forest variable do not measure the same quantities. "Bec" includes soil type, ground moisture content, and the type of vegetation, whereas "forest" is mostly a measure of quantity of vegetation.

In the analysis of this data we wish to consider the habitat type as a multinomial random variable. To do so we define the habitat type as the intersection of the previous three variables. There are then $2 \times 3 \times 4 = 24$ unique habitat types.

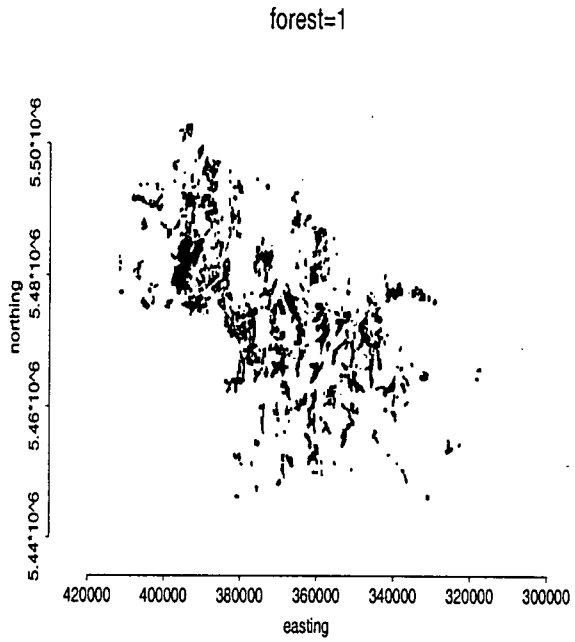

Figure 2.6: Forest area 1

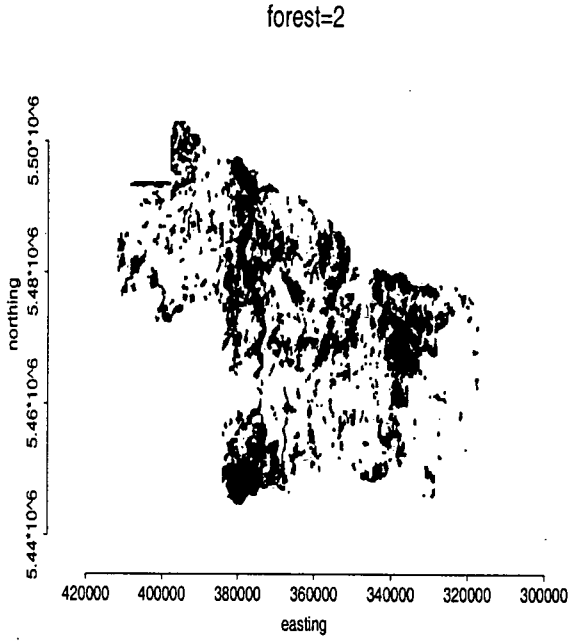

Figure 2.7: Forest area 2

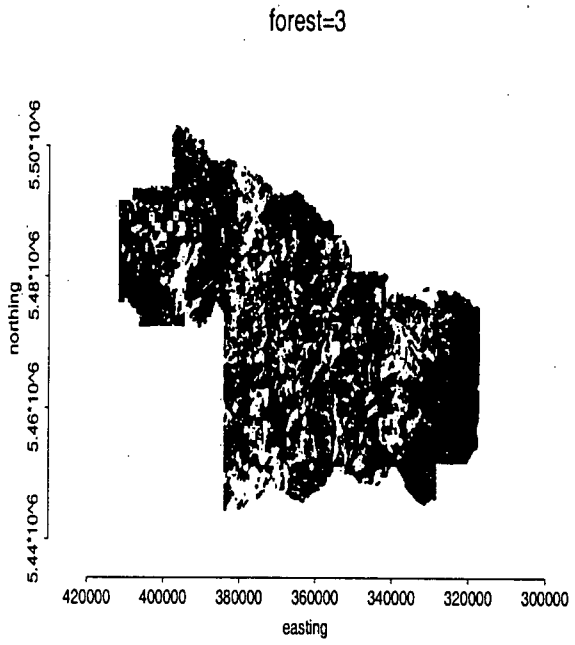

Figure 2.8: Forest area 3

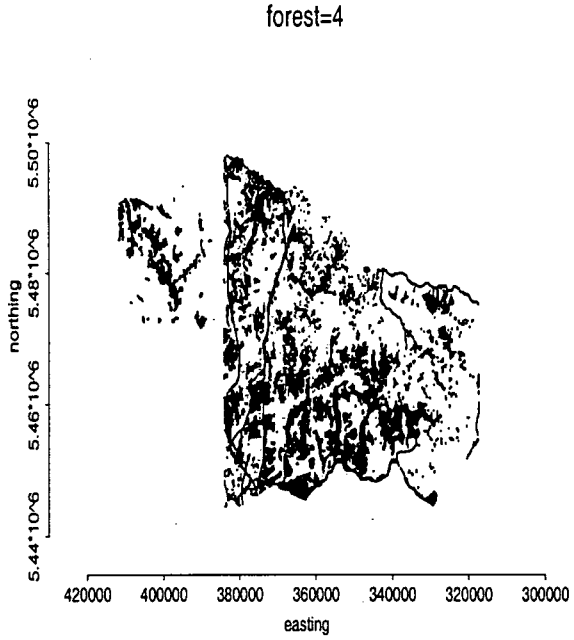

Figure 2.9: Forest area 4

forest class	proportion
1	0.0759
2	0.1964
3	0.5777
4	0.1500

Table 2.3: Proportion of area in different forest types

These are identified by a four digit code. The first digit in this code refers to the state of the "forest" variable, and therefore ranges from one to four. The physical meanings of these numbers have been outlined in the section describing that variable. The second digit in the code refers to the state of the "bec" variable. Again, digits have meanings similar to those described above. The last two digits refer to the distance from roads. The original coding for this variable had three states, labeled 10, 20, and 30, corresponding to 0-250m from road, 250-500m from road, and >500m from road. These were collapsed into two categories that use the labels 10 and 30, referring to <500m and >500m from the roads, respectively.

The following table presents the percentages of each of the 24 different types of habitat in the study area.

habitat classification	percentage	habitat classification	percentage
1110	1.21	3110	6.98
1130	2.31	3130	20.20
1210	0.95	3210	5.21
1230	2.40	3230	18.94
1310	0.14	3310	0.94
1330	0.58	3330	5.50
2110	2.28	4110	1.61
2130	6.89	4130	5.02
2210	2.02	4210	1.58
2230	6.47	4230	5.23
2310	0.32	4310	0.25
2330	1.65	4330	1.32

Table 2.4: Percentage of study area occupied by each habitat type

2.4 Bears

Two study zones were defined for the collection of the grizzly data. The first zone comprised 3000 km^2 of the Selkirks in northern Idaho and the second occupied 2700 km^2 of the British Columbian mountains. The entire area represents a contiguous block of land that sustains a large population of bears.

Within each study zone smaller trapping areas were selected. These areas were chosen because they were known to be desirable bear habitat (ie. lots of berries), and there existed signs of bear activity. Each was 100 km^2. Within these areas bears were trapped by Aldrich leg snares and immobilized with ketamine hydrochloride and xylazine hydrochloride. Immobilized bears were tattooed, ear-tagged, and fitted with activity sensing "drop-off" radio collars. Bears in Idaho were captured between May 1 and June 30 in the years 1985,1986, and 1987, and between August 8 and August 25 in 1989. Trapping in British Columbia occurred between May 25 and July 26 in 1988 and 1989. Basic measurements taken on the captured bears included sex, age, weight, and number of cubs.

The collared bears were then monitored by fixed wing radiotelemetry. Flights were conducted weekly during the non-denning periods (early April to early November) of the years 1985 to 1990.

These procedures resulted in data on 26 bears (14 female, 12 male). Because of the difficulties in locating bears, and the different times of capture and collaring, the number of observations on each bear varied, and observations were recorded at irregular times. The minimum number of observations on a single bear was five, and the maximum was 100. Three bears were collared in B.C. in the last year of the study, although most bears were observed for more than one season (often up to five years).

The following table presents the percentages of bear observations that occurred in each of the 24 different habitat types. This can be contrasted with table 4 to attempt to determine if the bears are using the habitats with the same frequency

that they occur.

habitat classification	percentage	habitat classification	percentage
1110	1.73	3110	5.52
1130	3.22	3130	18.18
1210	1.15	3210	6.67
1230	2.42	3230	19.22
1310	0.00	3310	1.74
1330	0.69	3330	7.83
2110	3.22	4110	2.07
2130	4.49	4130	6.56
2210	1.96	4210	2.07
2230	2.53	4230	5.98
2310	0.12	4310	0.46
2330	0.92	4330	1.27

Table 2.5: Frequency of bear sightings in each habitat type

Chapter 3

Issues and conventional methods

Studies of wildlife have generally been concerned with two major issues: resource utilization and location. Previous analyses have tended to view these two objectives as separate and as a consequence statistical methods have generally not focused on both simultaneously. There are a number of methods that attempt to describe how an animal (or group of animals) uses the surrounding habitat, and there are techniques that attempt to map the whereabouts of these animals. These methods are usually distinct and independent of one another.

Resource selection studies generally focus on the types of habitat that are available to the organisms, and how often each type is used by the organisms. The next section outlines some of the more prominent methods for this type of analysis.

Mapping of wildlife generally consists of simple techniques to describe where the animals have been observed. These methods are predominantly descriptive techniques that use past observations to define an area that is likely to contain an animal.

3.1 Resource selection techniques

There are a number of different techniques that have been proposed to analyze resource selection data. These techniques range from descriptive methods to those that can be used as hypothesis testing tools. They are predominantly based on the proportion of the observations that occur in areas with different resources. Alldredge and Ratti (1992) compare a number of techniques, we summarize their review.

Most of these techniques aim to test the relative frequencies of use of the different types of habitat. Often the statistical test is one of observed usage versus expected usage (often based on relative proportions of the different habitat types). These techniques use the data slightly differently than the present analysis does, but it is important to understand their strengths and weaknesses, as they may pertain directly to the analysis.

Analysis of the use of habitats generally consider how the observed usage differs from the available habitat. It is therefore necessary to collect data on both the animals, and the surrounding landscape. It is increasingly common to see exhaustive data about the surroundings, as they are easily obtained by remote sensing methods.

One common method of resource use analysis was presented by Neu et al. (1974). This method is easy to implement, and does not require many stringent assumptions. In fact, it is simply an application of the standard chi-square test of goodness of fit. This method compares the observed number of occurrences in each habitat to the number of occurrences the are expected if the habitat types are used in proportion to their availability.

To use this method data on both habitat availability and habitat usage is necessary. These data are generally expressed as proportions. The habitat is divided into distinct subgroups, and the proportion of the total area occupied by each subgroup is recorded. The proportion of animal occurrences in each habitat type is also recorded. The animal occurrences are calculated over all animals in the study, and thus there is an assumption that the habitat usage is equal for all animals. The

observed usage is then tested against the expected usage (the available proportions) using a χ^2 goodness of fit test.

Obviously, this method has the same requirements as any other use of the χ^2 test. Specifically, the sample size needs to be large enough that the χ^2 approximation is valid, and there must be enough observations in all of the cells of the table.

Another common method was originally identified by Johnson (1980). This method compares ranks of habitat use to ranks of habitat availability for each animal. The rank ordering of habitat usage is computed for each individual animal. "The hypothesis tested is that the rank ordering of habitat use is the same as the rank ordering of habitat availability when averaged over all animals."

The third method is called the Friedman method, and it is an analysis of variance that is based on ranks for a randomized complete block design. It was initially proposed by Friedman in 1937 (Alldrege and Ratti, 1986) This method is applied by ranking the differences in use and availability for each animal. The animals are then considered as blocks and the habitats are considered treatments in an ANOVA. If a significant result is obtained Alldredge and Ratti suggest using Fisher's least significant difference procedure to determine which habitat is selected more frequently than the others.

The primary difference between this method and Johnson's is that the analysis in the latter is based on differences of ranks, while this one is based on ranks of differences.

The final method is called the Quade method (Quade, 1979). It is similar to the Friedman method, except that it allows for a different weighting of the animals. Those animals that show greater variability in the differences are weighted more heavily than the other animals.

3.1.1 Problems

While radio tracking has generally been a boon to wildlife researchers, there are some common problems that are often encountered in the analysis of data of this form. An excellent summary is provided in Aebischer et al. (1993) who outline four major problems that are commonly encountered in habitat utilization data.

Problem 1

In studies utilizing radio-collared animals it is easy to obtain very large data sets consisting of many observations on each of a group of animals. The result of this intensive sampling scheme are many records of location paired with the environmental covariates at that location. These records can then be used in statistical analysis. This is where the first common problem is encountered.

The problem occurs in the definition of the sampling unit, and the corresponding levels of sampling and sample size. If the study aims to make inferences about a population of animals, it is necessary to sample the animals, and count each individual animal as the sampling unit.

This is in contrast to considering the radio locations as the sample unit. The obvious result of considering each radio location as a sampling unit is that the apparent number of samples is much larger than it actually is. It is easily seen that taking 500 samples of one animal says very little about the population under study. This type of sampling says much about the behaviour of the single sampled animal, but it is still unknown if this animal follows the same behavioural patterns as the typical animal.

If inference on a population is desired, a good sampling plan would examine a large number of animals. An extreme sampling plan would entail a census of the animals in a populations, taking a single observationon each. Again, this is not necessarily the best plan, although it would give a better picture of the state of the population. This sampling plan would resemble a cross sectional study.

Viewing the radio location as the base sampling unit will lead to other problems. Radio locations from a single animal are likely to be correlated, especially locations that are separated by relatively short time periods. This problem would invalidate the use of standard techniques that require independent samples. This problem may not always be encountered in situations where the animal itself is considered the sampling unit (although it is possible, for example, in social or territorial animals).

This problem is one that should be considered prior to the collection of the data, as the number of observations of each animal can be decided at that time. If, as in this case, the data has already been collected the problem can be avoided by summarizing the observations on each animal and then analysing these aggregated data.

Problem 2

The most common measure of habitat usage is the proportion of observations in each of the different habitat types. This use of proportions leads to the second problem identified by Aebischer et al.. This is a problem for two reasons. Primarily, most formal statistical tests assume that groups are independent of each other, and this is clearly not the case with proportions, due to the unit sum constraint. The second reason is in terms of the biological implications, the occupation of one habitat preferentially necessarily decreases the proportion of observations in other habitats. This may result in a case where there is an apparent avoidance of a specific habitat. The converse is also possible (avoidance of a habitat interpreted as preference for another).

Problem 3

Another problem that may arise from analysing all of the data together, rather than by animal, is the loss of information on individual variation in animal behaviours.

There may be distinct differences in behaviour between different sub-groups of animals, and these differences may become obscured if each location is considered the basic sampling unit. This is especially true if there are different numbers of observations on different animals. It is much easier to notice differences in individual animals.

This third problem identified by Aebischer et al. pertains to the differential use of habitats by different groups of animals. Behaviour may be affected by an animals sex, age, size, or other factors. These factors will create sub-groups of animals that may use the habitat differently. If the analysis is carried out without accounting for these differences they may be obscured, and the "average" habitat use may actually be an artificial construct that doesn't accurately describe any of the animals.

Currently the analysis used to deal with group differences include chi-square tests and log-linear models. As both of these succumb to the problem of non-independence of proportions, Aebischer et al. suggest that a method similar to ANOVA would be appropriate. Such a method would compare the within group variation to the between group variation.

Problem 4

The last problem identified by Aebischer et al. is the arbitrary definition of habitat availability. In all methodologies for assessing habitat usage the animals' usage is compared to the available habitat, to determine if the usage is random or dependent upon the habitat types. Obviously it is necessary to assess what types of habitat are available.

At the largest geographical level, the animal is utilizing the resources within an arbitrarily defined study area. This area may be defined by convenience, and may include areas that no animal ever uses. The study area should be large enough to encompass all of the animals' movements. Information at this level is entirely

under the control of the researcher. A conscious decision to include or exclude areas (usually at the border of the study area) can be made, and this can affect the data on available habitat.

Within the total study area, it is also possible that some areas remain inaccessible to some animals, and therefore the true available habitat for those animals will be unknown. It may, therefore, be erroneous to use the proportions over the entire study area.

3.2 Mapping techniques

In general, there are few techniques that allow for the statistical analysis of animal location. With the advent of greater computing power and increased memory it has been easy to present all of the data in an easy to comprehend form. Modern GIS software allows for the display of all observations of an animal, in addition to the other variables in the area.

Often, the presentation of these data is augmented with an attempt to describe the home range of the animal. The home range of the animal is the area in which the animal spends most of its time. This is an imprecise definition, in that the amount of time spent in the home range is unclear. In some cases all observations are located in a small area, which is easy to define. More often though, there is an area with a high density of sightings, and surrounding areas with fewer sightings. In some cases there are even single observations at great distances from the core area. These situations lead to difficulties in defining the home range, and in interpreting it.

The most common method of identifying the home range is by finding the minimum area convex polygon (MCP) around the data (White & Garrott, 1990). That method describes an area that contains all of the observations yet has the smallest area, without resorting to concave boundaries. This method is desirable due to its simplicity. However, it is very sensitive to the presence of outliers. A

single point can greatly affect the size of the polygon, and subsequently the interpretations about the animals behavioural patterns. An outlier can greatly increase the size of the MCP, but the polygon will then have large areas that have few, or no, observations.

For this reason, outlying points are often excluded from the analysis. Exclusion is often an arbitrary procedure. The data with outliers removed can then be used to define a core area, and the outliers can be included in an "expanded" range definition. Interpretations of these different areas will depend on the nature of the study.

Another problem with the MCP method is that it identifies only a single contiguous block of land, and does not identify areas within this home range that receive more visits than others. For example, an animal may spend time in three distinct patches of land, and only use the land between them during travel. The MCP method would identify the three patches and the area between them as the home range.

A second, and very useful method is the Kernel method of determining a home range (Worton, 1989). This method uses bivariate density estimation techniques to identify areas of high probability. The home range is delimited by isopleths containing a predefined probability. Common values are 50%, 75%, and 95%. The standard reference for density estimation is Silverman (1986), and the methods described in it are applicable to the home range estimation problem. This method shall be elaborated on later in this thesis.

3.3 Autologistic model

A desirable method of analysing data of wildlife locations should incorporate information from both of the above types of analysis. Knowledge of characteristics of the surrounding landscape can aid in the mapping of the animals. The mapping techniques presented above rely exclusively on the locations of the animals, either

at a specific moment or over a given time period.

These methods are generally inadequate in that they do not take into account the state of the many other variables that are present in the same area. In a simple case, if an animal is observed around the perimeter of a lake, it makes little sense to claim that the area the animal inhabits is that contained by the observations. Similarly, there is no reason to exclude an area of valuable resources simply because an animal was not observed there at the time of sampling. A superior method would account for both the presence of the animal, and the environmental characteristics in the study area.

The autologistic model is one method that attempts to account for both the presence of covariates and the spatial correlation of the wildlife. This is done by explicitly modeling the presence of wildlife in adjacent areas. Although this method was first proposed by Besag over twenty years ago (Besag, 1974) it has only recently seen much use (Wu & Huffer, 1997; Augustin et. al. 1996, 1998)

A typical logistic regression relates a binary response variable to a number of covariates. This can easily be used in a wildlife mapping situation, by examining the covariates at locations where the animals are present and at locations where the animals are absent. This standard method results in a model of the form:

$$\Pr(Y_i = 1|\mathbf{X} = x) = \frac{\exp(\beta_0 + \sum_{k=1}^{m} \beta_k x_{ki})}{1 + \exp(\beta_0 + \sum_{k=1}^{m} \beta_k x_{ki})}$$

where x_{ki} is the k^{th} covariate for the i^{th} record. In the case of mapping wildlife, it is necessary to discretize the area under study, as viewing the area as a collection of infinitely many points leads to severe technical challenges. The discrete cells (AKA quadrats) in the study area would then be sampled, and the states of both the response and covariates recorded. Then, the model would be fit with the data from the $i = 1, \ldots, n$ sampled cells.

This model is deficient in that it does not acknowledge the effects of sociality (ie, autocorrelation) in the animals under study. It simply relates the presence or absence of the animal to a number of other variables. It is a very real possibility

that the presence of an animal nearby will affect the probability of finding another in cell i. Depending on the nature of the animals, a nearby individual could increase the probability (in the case of social animals) or decrease it (in the case of territorial animals). Further, it is possible that the presence of an animal in one location is independent of the state of adjacent cells.

This deficiency is addressed by the addition of an "autocovariate" to the logistic model (hence the name "autologistic model"). This model is almost identical to the preceding one, except for the addition of this new term.

$$\Pr(Y_i = 1|\mathbf{X} = x) = \frac{\exp(\beta_0 + \sum_{k=1}^{m-1} \beta_k x_{ki} + \sum \delta_{ij} y_j)}{1 + \exp(\beta_0 + \sum_{k=1}^{m-1} \beta_k x_{ki} + \sum \delta_{ij} y_j)} \quad \forall \ j \neq i$$

where $\sum \delta_{ij} y_j$ is the autocovariate term for cell i. This is simply a function of the presence/absence of the animal at the other cells in the study area. It can be expressed as $\beta_m x_m$, and the model can then be fit as before.

The choice of the autocovariate term is rather arbitrary, as a number of different weights can be used, and the size of the neighbourhood to be used is also flexible. In their implementation, Augustin et al. suggest using weights that are inversely proportional to the Euclidean distance between cells.

The addition of this term results in a model that can account for the tendencies of the animals to aggregate or vice versa. This leads to a more useful model in that it incorporates the knowledge that the animals do have certain behavioural preferences for where they locate themselves. For example, the mere presence of food may not be enough to convince an animal to inhabit that specific location, whereas the presence of food *and* another animal will influence the animal to move to that location.

The introduction of the autocovariate term in the model also introduces a major problem. Since this term is constructed from information in neighbouring cells it is necessary to have data from all of the neighbouring cells. Thus, it seems that the only implementations of this technique would occur when the entire area had been sampled, in which case a model is not necessary. It has been proposed

that in situations where there is not data for every cell simulation methods can be used to "fill-in" the gaps in the data. The method suggested by Augustin et al. is to use a Gibbs sampler to estimate presence/absence at unsampled squares. The standard EM algorithm is also a candidate for this estimation.

This model was assessed relative to the standard logistic model, and found to perform better for predicting the spatial distribution of the wildlife (Augustin et al., 1996). Thus, it seems as though this model has promise for the type of analysis that is desired for the present study.

Unfortunately, this model suffers from the same problem that plagues the analysis of the grizzly bear data. Specifically, this method requires that the data be collected from random locations, where presence or absence of the study organism is recorded at the randomly sampled locations. In the present study, we do not sampled locations and recorded binary responses, but instead we have "location" as the response.

A modification of this technique that would allow for this particular sampling scheme would be useful. Such a modification would allow for precise maps of bear locations, and provide information on how the bears are using their habitats.

Chapter 4

A Conditional Approach

4.1 Background

The model underlying the methodology used in this section begins with an assessment of the environment given the presence of a bear. The sampling regime has given us a vector (or "cluster") of environmental characteristics that have essentially been chosen (preferred) by the bear. The bear itself has a trajectory (unknown to the observers) that it follows indefinitely. At certain times the bear was located, and the state of the covariates was recorded. This differs from standard sampling techniques, where the presence of animals or a count of the number of animals is recorded along with the state of the covariates. In this case, the researcher deliberately seeks and finds a specific bear, so that at the sampling time there is always a bear present. This leaves us with the challenging situation of having no observations without a bear present.

In this case, the state of the covariates is known for all locations in the study area, due to remote sensing and satellite technology. Thus, when one views the map, the locations of the bears are isolated points in a large grid of covariates. The number of locations without a bear greatly outnumbers those with. It might naively seem as that standard analysis could be used. However, the areas with no

bear recorded are not necessarily "bear-free". They were simply free of any of the bear(s) being tracked, at the time of tracking. It might well be that a different bear is at the location with a zero recorded, or that a tracked bear was in that area at a different time. Thus, the "zeros" in any recorded location are not necessarily true zeros. In some cases, of course, it would seem likely that the zero's for some areas are actual zeros (ie. no bears would be sighted in the middle of a lake).

As outlined previously, the autologistic model is a potential method for assessing the probabilities of finding a bear present. Both the effect of other variables and the effect of spatial correlation are accounted for by this method. Again, however, this model is not suited to the data at hand. This method requires a random sample of cells in the grid, and cannot be used with the selection biased, non-random sample of bear locations we have.

Due to the insurmountable problems inherent in the data, standard methods could not be applied. Instead, we develop a new approach that treats the data as conditional on the presence of a bear, rather than treating the bear as the response variable of interest. Thus, the environmental data becomes the response, conditional on the presence of a certain bear (ie. $Y = y|\Theta = \theta_i$). The autologistic model is then used merely as a guide for constructing the new method. The resulting new method has a number of desirable qualities. The most important characteristics of the autologistic model are its capacity to use other variables to aid in the mapping of an animal's presence propensity, and the ability to use other observations of the animals to define a region that the animals frequent.

The primary goal of the analysis is to produce a map of probabilities that accurately reflect both individual bears' location and habitat preference, and the overall preferences of the entire bear community. Following this, it would be useful to provide a method for testing whether certain habitats are used more frequently than chance would predict, and whether there are differences in habitat use and general locational choices in different subgroups of the bear population (ie. differences due

to sex, age, and so on).

Consider a response variable that consists of a number K of mutually exclusive and exhaustive categories. These different categories represent different habitat types and other environmental variables. In this case the different categories are created by the juxtaposition of different variables. Given, say, two binary environmental variables, the response variable would be created by examining all four possible combinations of the variables. This method seems to work well for a few variables, each with a few possible outcomes. It is easily seen that as the number of variables increases, and as the number of possible outcomes of each increases, the number of distinct realizations for Y increases rapidly. For this reason we limit the number of environmental covariates in our study. It then becomes feasible to treat this response variable as a multinomial random variable with K distinct outcomes,Y_k, and a vector of K corresponding probabilities, $\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_K)$.

There are two options for interpreting the data at this stage. The data could be grouped over all bears, providing a summary of the habitat usage by the entire population of bears, or the data could be analysed by the individual bear. In this case, $\boldsymbol{\theta}_i$ would represent bear i's vector of habitat preference probabilities. That is, the proportion of time spent in each different type of habitat.

In this data set the value of Y is known for all locations in the study area. When the bears are sampled, we can assess the state of Y at the locations in which the bears were found. We can interpret this situation as if the bears are "sampling" the surrounding landscape. One can therefore not construe this as a random sample of habitats. It can however, be treated as an observation of Y conditional on the bear. If the times at which the bears are sampled is randomly chosen, we end up with a random sample of each bear's trajectory.

The sequence of observations conditional on the bear would be correlated unless they were well separated in time. However, for simplicity, we assume independence as an approximation. For each bear the observation at one point in time

is correlated with those at other times, especially those at subsequent time periods. This need not be a problem in principle, as the basic sampling unit should be considered an individual bear. We are interested in reconstructing the trajectory of the bear, and so it doesn't really matter if the observations on each bear individually are correlated. A more serious problem with correlated observations will arise when there are correlations between bears. If the presence of one bear precludes that of another bear significant technical problems will arise. This will have implications in hypothesis testing and comparisons between bears.

It may well be possible to identify this correlation, but development of methodology for doing so will be left for future work

Thus, we may summarize our data by assuming that each observation in the data can be expressed as follows.

$$Y_{ij} = y_{ij}|\Theta = \theta_i, T = t_j \qquad i = 1,\ldots,n; \quad j = 1,\ldots,T_i$$

where θ_i denotes the bear observed (from a total of n bears), and t_j represents the time at which the bear was sampled. We are presently allowing for different numbers of observations on each bear.

The eventual goal is to use data of this form to assess the probability of finding a bear in a given habitat (ie. $P(\theta|y)$) essentially by application of Bayes rule.

4.2 Single Bear Analysis

The method of analysis begins with a Bayesian approach that will allow for an assessment of the probabilities of finding bears given the surrounding characteristics. Application of Bayesian methods to data of the form $y|\theta$ can lead to two different quantities, the posterior density $p(\theta|y)$ or the posterior predictive distribution $p(\tilde{y}|y)$, where y denotes the observed values of the random variable, θ denotes the

parameters of the distribution of Y, and $\tilde{y}$ represents unobserved data.

$$p(\theta|y) = \frac{p(\theta)p(y|\theta)}{\int p(\theta)p(y|\theta)d\theta}$$

or,

$$p(\tilde{y}|y) = \int p(\tilde{y}|\theta, y)p(\theta|y)d\theta = \int p(\tilde{y}|\theta)p(\theta|y)d\theta$$

In both of the above equations, the integration should be replaced by summation over theta in the case of discrete random variables.

For the data in hand, we have defined

$$y|\theta \sim \text{Multinomial}(\boldsymbol{\theta}),$$

where $\boldsymbol{\theta}$ is a vector with components $\theta_1, \ldots, \theta_K$. Each θ_i represents the probability that $Y = y_i$.

We take the prior distribution of $\boldsymbol{\theta}$ to be Dirichlet, with parameters $(\boldsymbol{\alpha})$. The vector of parameters is again a K-vector. Each α_i represents a subjective weight on the importance of the different habitat types. If it is unknown to what degree each habitat is utilized, it is possible to take $\alpha_i = \alpha_j$ for all i and j. It is also possible to take $\alpha_i = 1$ for all i to give a non-informative (uniform) prior. Alternatively, it is possible to adopt an empirical Bayes approach by estimating $\boldsymbol{\alpha} = \hat{\boldsymbol{\alpha}}$.

The Dirichlet distribution is a conjugate prior for a multinomial, and therefore the posterior distribution of the $\theta|y$ is also Dirichlet. The updated parameters of this distribution are $(\boldsymbol{\alpha} + \mathbf{y})$, where $\mathbf{y}$ is the vector of counts in each of the K categories.

Thus, estimates of the probabilities of selection by a bear can be estimated for each of the K different habitat types. These estimates are calculated easily from the posterior, and are of the form:

$$\hat{\theta}_i = \frac{\alpha_i + y_i}{\sum_{j=1}^{K}(\alpha_j + y_j)}$$

This is simply the posterior expectation $E(\theta_i|y)$. The posterior variance of each θ is given by

$$V(\theta|y) = \frac{\hat{\theta}_i(1-\hat{\theta}_i)}{1+\sum_{j=1}^{K}(\alpha_j+y_j)}$$

The posterior covariance between pairs of θ's is also easily calculated. Its general form is as follows:

$$C(\theta_i,\theta_j) = \frac{\hat{\theta}_i(\hat{\theta}_j)}{1+\sum_{j=1}^{K}(\alpha_j+y_j)}$$

Note that these estimates are very similar to the naive, and intuitive, estimate of $\hat{\theta}_i = y_i/\sum y_j$. If the prior distribution were taken to be the non-informative (uniform, $\boldsymbol{\alpha} = \mathbf{1}$) then the posterior estimates are identical to the naive form.

The form of the posterior predictive density is slightly more complicated. Because θ is an uncertain nuisance parameter (for the purpose of prediction) it must be integrated out:

$$\begin{aligned}
p(\tilde{y}|y) &= \int p(\tilde{y}|\theta,y)p(\theta|y)d\theta \\
&= \int p(\tilde{y}|\theta)p(\theta|y)d\theta \\
&= \int \theta_1^{\tilde{y}_1}\theta_2^{\tilde{y}_2}\cdots\theta_K^{\tilde{y}_K}\frac{\Gamma(\alpha_1+y_1+\cdots+\alpha_K+y_K)}{\Gamma(\alpha_1+y_1)\cdots\Gamma(\alpha_K+y_K)} \\
&\qquad \times \theta_1^{\alpha_1+y_1}\theta_2^{\alpha_2+y_2}\cdots(1-\sum_{j=1}^{K-1}\theta_j)^{(\alpha_K+y_K)}d\theta \\
&= \frac{\Gamma(\alpha_1+y_1+\tilde{y}_1)\Gamma(\alpha_2+y_2+\tilde{y}_2)\cdots\Gamma(\alpha_K+y_K+\tilde{y}_K)}{\Gamma(\alpha_1+y_1+\tilde{y}_1+\alpha_2+y_2+\tilde{y}_2+\cdots+\alpha_K+y_K+\tilde{y}_K)} \\
&\qquad \times \frac{\Gamma(\alpha_1+y_1+\alpha_2+y_2+\cdots+\alpha_K+y_K)}{\Gamma(\alpha_1+y_1)\Gamma(\alpha_2+y_2)\cdots\Gamma(\alpha_K+y_K)}
\end{aligned}$$

To implement this method it is necessary only to define a vector of habitats in which it is possible for the animal to live. An attempt to determine the relative importance of each habitat type may be made as well, in order to assign values to the vector $\boldsymbol{\alpha}$. Following this, a simple count of the number of occurrences in each habitat type is recorded. This is enough information to proceed with the calculation of both of the posterior distributions.

These two distributions give us an assessment of how each of the different habitat types compares to the others, with respect to the prediction of bear behaviour. They don't, however, tell us which type of habitat the bears actually prefer in "absolute terms". It is entirely plausible that the most desirable habitat occurs less frequently than the others in the surrounding area, and as a consequence the bears spend relatively little time in that desirable area (as one would predictively expect).

After these distributions are found, it is possible to construct a map of the posterior probabilities of habitat selection by a bear. The area can be divided into distinct regions based on the type of habitat present, and then the probability of occurrence in each area can be displayed. It should be noted that the probabilities calculated are for selection of each habitat type. They do not identify the probability of selection of specific area within the habitat type. That is, if the entire study area had, say, four distinct areas with the same type of habitat, they would all have the same probability of selection, regardless of if a bear was found only in one of those areas.

In the production of the map it is also necessary to scale each probability to account for the amount of each habitat type to get a predictive distribution for habitat type and location. It is possible to discretize the entire area into small cells, and then assign a probability to each cell based on the habitat type present there, and the number of other cells that also contain the same habitat type. Thus, if the predicted probability of bear selection in a habitat was 0.2, and there were 20 cells that contained that type of habitat, the map would identify the probability of occurrence in each cell as 0.01. Of course, this assumes that the probability of occurrence in a cell is equal for all cells in a region. This is not necessarily true, and shall be accounted for later. The end result of this is that the sum of the probabilities over the entire map area equals one.

4.3 Multiple Bears

The astute reader will notice that the preceding discussion is based on probability vectors $\boldsymbol{\theta}$. These probabilities have been ambiguously defined as the vector of probabilities associated with the multinomial random variable Y. This vector of probabilities may be derived for a single bear, or it may represent the probabilities calculated from all observations on all bears.

There are a few issues that should be reiterated at this point. It is not advisable to pool all of the data for all of the bears, especially if there are different numbers of observations for each bear. Pooling the data may obscure differences between subgroups of the bears, and it may also give undue attention to those bears that happened to be sampled more frequently than the others. These issues have been discussed previously in the section about problems in this type of analysis.

Since it is inadvisable to pool the data, it should be examined on the individual bear level. Thus, for each bear i, there exists a vector of probabilities ($\boldsymbol{\theta}_i$) that describe the habitat preference of that animal. These are to be estimated individually by the method outlined above. It is very likely that these probabilities are not equal for all bears, and it is conceivable that there are large discrepancies between some of them.

One can speculate about why these vectors might differ between bears. Perhaps they reflect some innate bear preference (thus, that differences represent bear-to-bear variability), although, it seems more likely that the differences would reflect the availability of certain habitats as well. For example, if a bear lives in an area that is relatively devoid of a "desirable" habitat type, then that bear's vector of probabilities will have a lower probability associated with that habitat. It is also conceivable that there are different "definitions" of "desirable" habitats in different groups of bears. A female with two cubs might require a food rich area more than a single bear does, and may then opt to spend more time in areas with more food, even if they are sub-par in terms of other variables. There is also the possibility of

interactions among the bears. The use of certain habitats might be constrained to those that are present, which may be influenced by the presence of other animals.

Returning to the main objective of this thesis, it is now possible to produce a map that accounts for the different preferences of individual bears. The procedure for mapping outlined above could be used, although it would now require that the probabilities associated with an individual bear were attached to cells in the regions frequented by that bear. The obvious difficulty is the definition of regions for each bear, and how to map the probabilities in regions that do not clearly fall into the territory of any specific bear.

There are three methods that may work in this situation. We need a method that will allow us to assign a bear to each cell, or a weighted "average" of bears, so that we can then use the appropriate vector of probabilities in our map.

Firstly, it is possible to use some sort of clustering algorithm over the spatial area to group coordinates according the bear that most frequently inhabits that area. Standard cluster methods such as K-means, or something more complex that allows for non-linear cluster boundaries will all provide a means to assign each map cell to a single bear (the cluster). From there, we can create the map of probabilities, although it will be "discrete" in the sense that the clustering will produce distinct areas for each bear. In this "discrete" map it may be possible to have identical habitats that are adjacent, yet have very different probabilities associated with them.

Since it is quite likely that certain bear ranges overlap, and the bears actually share resources, it might be better to go with a regression type approach that will allow for "averaging" in areas frequented by more than one bear. To do this we propose a multinomial logistic regression model with the n bears as the response variable, and the X and Y coordinates as the predictor variables. This will result in a set of probabilities of occupation for each bear in any given cell. From here we can map the overall probabilities (for all bears) by weighting the different $\boldsymbol{\theta}_i$'s with the probability of encountering the bear associated with it.

The last option involves estimating the range of each bear, by the use of bivariate density estimation techniques, in order to find the probability of encountering a given bear in an area. This would then be used as a weight for the θ's as above. This would also require some form of rescaling because the area under each of the n density estimates would equal one, and as a consequence, the total area under these ranges on the map would equal n. In essence, each cell in the grid would contain n probabilities, each representing the chance of finding a specific bear in that cell. This option has the benefit of being easy to display.

4.3.1 Bayesian posterior predictive distribution

It is possible to assume that all bears have some intrinsic preference for each habitat and that the proportions of time spent in each are random variables. From this, we can construct a posterior density as follows.

$$p(\tilde{y}|\mathbf{y}^{(1)},\ldots,\mathbf{y}^{(n)}) = \int p(\tilde{y}|\boldsymbol{\theta}_0)p(\boldsymbol{\theta}_0|\mathbf{y}^{(1)},\ldots,\mathbf{y}^{(n)})d\boldsymbol{\theta}_0$$

$$p(\boldsymbol{\theta}_0|\mathbf{y}^{(1)},\ldots,\mathbf{y}^{(n)}) = \int p(\boldsymbol{\theta}_0,\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_n|\mathbf{y}^{(1)},\ldots,\mathbf{y}^{(n)})d\boldsymbol{\theta}_1\cdots d\boldsymbol{\theta}_n$$

$$\propto \int \left[\prod_{i=1}^{n} p(\mathbf{y}^{(i)}|\boldsymbol{\theta}_i)\right] \pi(\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_n|\boldsymbol{\theta}_0)d\boldsymbol{\theta}_1\cdots d\boldsymbol{\theta}_n$$

we take $\boldsymbol{\theta}_i \sim^{ind}$ Dirichlet($\boldsymbol{\alpha}$), so that

$$\pi(\boldsymbol{\theta}_0,\ldots\boldsymbol{\theta}_n) = \int \left[\prod_{i=0}^{n} \pi(\boldsymbol{\theta}_i|\boldsymbol{\alpha})\right] \pi(\boldsymbol{\alpha})d\boldsymbol{\alpha}$$

Furthermore, if $\boldsymbol{\alpha}$ can be reasonably well estimated from the data we may adopt the empirical Bayes approximation

$$\pi(\boldsymbol{\theta}_0,\ldots\boldsymbol{\theta}_n) \simeq \prod_{i=0}^{n} \pi(\boldsymbol{\theta}_i|\hat{\boldsymbol{\alpha}})$$

We wish to calculate $p(\tilde{y}|\mathbf{y}^{(1)},\ldots,\mathbf{y}^{(n)})$, and for this we need $p(\tilde{y}|\boldsymbol{\theta}_0)$ and $p(\boldsymbol{\theta}_0|\mathbf{y}^{(1)},\ldots,\mathbf{y}^{(n)})$. $p(\tilde{y}|\boldsymbol{\theta}_0)$ has been defined as a multinomial($\boldsymbol{\theta}$), so no further work is necessary on this piece of the equation.

$p(\boldsymbol{\theta}_0|\mathbf{y}^{(1)},\ldots,\mathbf{y}^{(n)})$ is defined above, and for it we need to identify two more probabilities: $p(\mathbf{y}^{(i)}|\boldsymbol{\theta}_i)$ and $\pi(\boldsymbol{\theta}_0,\ldots,\boldsymbol{\theta}_n)$. $\mathbf{y}^{(i)}$ is the realized counts in each habitat by bear i, and $\boldsymbol{\theta}_i$ is the probabilities of occupation for bear i. These probabilities can be considered a random effect, they are one possible outcome of a number of viable bear occupation preferences.

Again, $p(\mathbf{y}^{(i)}|\boldsymbol{\theta}_i)$ is multinomial, conditional on the specific probabilities $\boldsymbol{\theta}_i$.

Thus, the only quantity needed before the computation can begin is $\hat{\boldsymbol{\alpha}}$. This can be obtained by means of marginal likelihood.

$$\begin{aligned} L(\boldsymbol{\alpha}|\mathbf{y}^{(1)},\ldots,\mathbf{y}^{(n)}) &\propto f(\mathbf{y}^{(1)},\ldots,\mathbf{y}^{(n)}|\boldsymbol{\alpha}) \\ &= \int \textstyle\prod_{i=1}^{n} f(\mathbf{y}^{(i)}|\boldsymbol{\theta}_i)\pi(\boldsymbol{\theta}_i|\boldsymbol{\alpha})d\boldsymbol{\theta}_i \end{aligned}$$

We take all $\boldsymbol{\theta}_i$ to be identically and independently distributed as Dirichlet $(\boldsymbol{\alpha})$ random variables.

Thus,

$$f(\mathbf{y}^{(i)}|\boldsymbol{\theta}_i) = \frac{T_i!}{y_1^{(i)}!y_2^{(i)}!\cdots y_K^{(i)}!}\theta_{i1}^{y_1^{(i)}}\theta_{i2}^{y_2^{(i)}}\cdots\theta_{iK}^{y_K^{(i)}}$$

where T_i represents the number of observations on the i^{th} bear, and the y's represent the counts in each of the K habitat types for the i^{th} bear. And,

$$\pi(\boldsymbol{\theta}_i|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_1+\alpha_2+\cdots+\alpha_K)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\cdots\Gamma(\alpha_K)}\theta_{i1}^{\alpha_1}\theta_{i2}^{\alpha_2}\cdots(1-\sum_{j=1}^{K-1}\theta_{ij})^{\alpha_K}$$

Using these pmf's we can calculate the likelihood of $\boldsymbol{\alpha}$ as

$$\int\prod_{i=1}^{n}\left[\frac{T_i!}{y_1^{(i)}!y_2^{(i)}!\cdots y_K^{(i)}!}\theta_{i1}^{y_1^{(i)}}\theta_{i2}^{y_2^{(i)}}\cdots\theta_{iK}^{y_K^{(i)}}\frac{\Gamma(\alpha_1+\alpha_2+\cdots+\alpha_K)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\cdots\Gamma(\alpha_K)}\theta_{i1}^{\alpha_1}\theta_{i2}^{\alpha_2}\cdots(1-\sum_{j=1}^{K-1}\theta_{ij})^{\alpha_K}\right]d\boldsymbol{\theta}_i$$

which can be cleaned up slightly and expressed as

$$\int\left[\frac{\Gamma(\alpha_1+\cdots+\alpha_K)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)}\right]^n\prod_{i=1}^{n}\left[\frac{T_i!}{y_1^{(i)}!\cdots y_K^{(i)}!}\theta_{i1}^{\alpha_1+y_1^{(i)}}\cdots(1-\sum_{j=1}^{K-1}\theta_{ij})^{\alpha_K+y_K^{(i)}}\right]d\boldsymbol{\theta}_i \quad (4.1)$$

Integrating out the $n\times K$ parameters θ_{ij} leads us to the following form for the likelihood.

$$f(\mathbf{y}^{(1)},\ldots,\mathbf{y}^{(n)}|\boldsymbol{\alpha}) =$$

$$\left[\frac{\Gamma(\alpha_1+\cdots+\alpha_K)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)}\right]^n \prod_{i=1}^{n}\left[\frac{T_i!}{y_1^{(i)}!\cdots y_K^{(i)}!}\times\frac{\Gamma(\alpha_1+y_1^{(i)})\cdots\Gamma(\alpha_K+y_K^{(i)})}{\Gamma(\alpha_1+y_1^{(i)}+\cdots+\alpha_K+y_K^{(i)})}\right] \tag{4.2}$$

It is this likelihood that would be used for the estimation of $\hat{\boldsymbol{\alpha}}$ by numerical methods.

Since the likelihood of $\boldsymbol{\alpha}$ is exceedingly difficult to work with, its estimator was constructed from the estimates of the probabilities $\boldsymbol{\theta}$, which were obtained from the observed $\mathbf{y}$. The expected value of $\theta_i|\alpha_i$ is $E(\theta_i) = \alpha_i/\sum_{j=1}^{k}\alpha_j$. Standard method of moments estimation techniques allow for an estimate of α_i by equating $E(\theta_i)$ with it's sample analogue, $\bar{\theta}$. In this specific case we do not have observed values of $\boldsymbol{\theta}$, but we can estimate them from the observed $\mathbf{y}$. Thus, our estimate (and point prior) of $\boldsymbol{\alpha}$ will be found as follows:

$$\hat{\alpha}_i = (\sum_{j=1}^{k}\alpha_j)\frac{1}{n}\sum_{l=1}^{n}\hat{\theta}_{il}$$

where $\hat{\theta}_{il}$ is estimated as $\frac{1}{T_l}\sum_{j=1}^{T_l} y_{ij}$.

4.4 Density Estimation

The second step of the analysis is to take the locations where the bears were found and estimate the general geographic preference of the bears. This has the effect of distinguishing between areas that have similar geographical variables, but different numbers of bear sightings. The previous step of the analysis estimates the probability of a habitat being selected by a bear , and this one complements that by calculating probabilities of locating a bear, based on the locations that the bears have been observed at. This makes intuitive sense, because the probability of an observation should be large if there were many observations in nearby locations.

This is a simple application of standard kernel smoothing methods. A simple bivariate density estimation algorithm from Silverman (1986) was used on the bear

locations, with the easterly and northerly UTM coordinates as the bivariate vector associated with the bear observations. The density estimation was carried out in Splus using a slight modification of the function `kde2D()`, which was written by Guy Nason and Martin Maechler, and posted on the statlib website maintained at Carnegie Melon University.

There is a problem here, in that the density estimates are based on all recorded sightings of the bears. Thus, if there is a difference in the intensity of which individual bears are sampled, the estimate will be biased towards those bears with greater numbers of observations. This is another reason to analyse each bear individually.

There is also the potential to introduce bias by recording many observations from easy-to-reach locations, and few observations from the more difficult-to-reach locations. This problem is directly attributable to the method used for collection of radio locations, and may manifest itself differently for different methods. For example, if the animals are tracked from an airplane (as in this study) the entire area may be sampled relatively easily, with perhaps the exception of some canyons and/or particularly nasty mountain spires. This method would still be preferable to tracking from a truck or car, which would be constrained by the system of roads built into the area. This would limit the efficacy of the tracking regime, although it may be possible to access some different regions with all terrain vehicles. Of course, these may come associated with their own problems. It is well known that the noise from ATV's can affect the behaviour of the animals that are being tracked.

The questions of how to resolve these problems remain. The second problem is more easily dealt with, although it may require a great deal of work on the part of those collecting the data. The problem about unequal sampling effort can be skirted easily by planning a study in which each of the animals is tracked an equal number of times. This is the easy way to deal with the problem, and the density can then be estimated for all of the radio locations.

Also, there is still the issue of the bears locations being correlated. More specifically, these density estimation algorithms assume that we have a random sample of points. This is certainly not the case, as the observations from one bear are correlated with each other. This should not be a problem if the bear is sampled randomly in time, because the point of the estimation is to estimate where the bear spends most of its time, and the individual points are a random sample of the locations that the bear frequents. More importantly is the correlation between the locations belonging to different bears. In some cases this won't be a problem, such as in bears whose ranges overlap, (ie. cubless females), although it may be quite problematic in cases where the animals are territorial.

4.5 Results of the bear study

As an illustration of the applicability of this method, we present a map of the probability of encountering a bear in the study region described above. Figure 4.1 shows an estimate of the density of occurrences for the bear in question. Table 4.1 gives the posterior probability of selection for each habitat type that was calculated using the aforementioned Bayesian approach.

habitat classification	percentage	habitat classification	percentage
1110	1.72	3110	5.52
1130	3.19	3130	18.16
1210	1.11	3210	6.63
1230	2.45	3230	19.26
1310	0.00	3310	1.72
1330	0.73	3330	7.85
2110	3.19	4110	2.08
2130	4.53	4130	6.51
2210	1.96	4210	2.08
2230	2.57	4230	6.01
2310	0.12	4310	0.49
2330	0.87	4330	1.23

Table 4.1: Posterior probabilities of selection (Expressed as percentages)

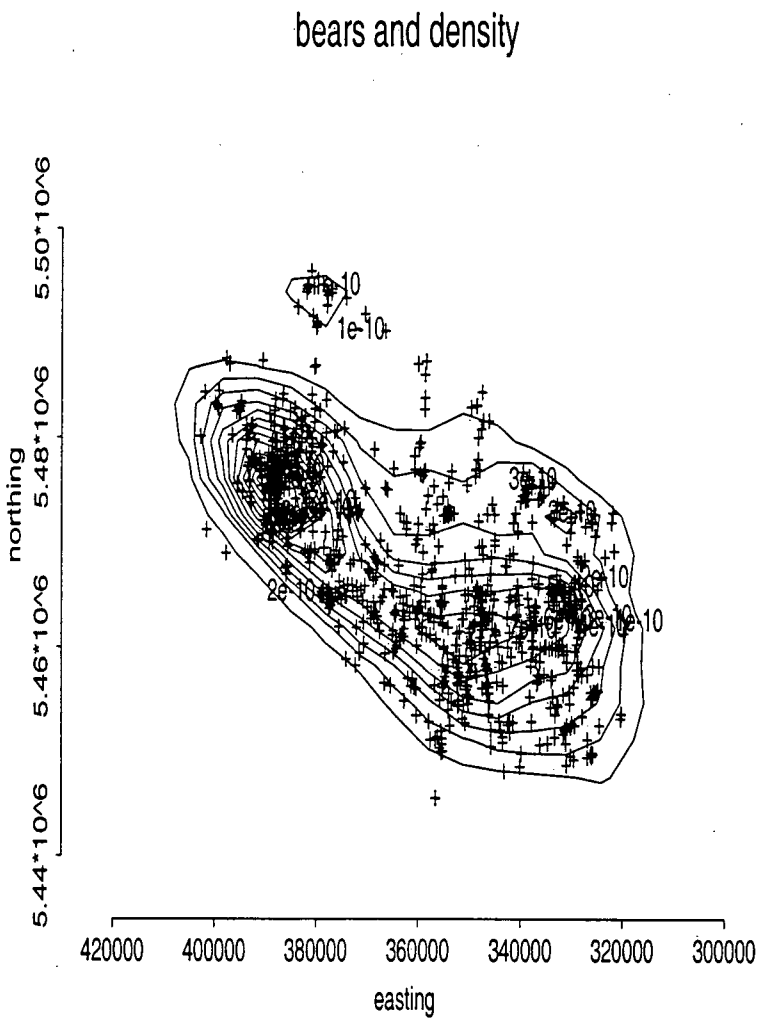

Figure 4.1: Density estimate of bear locations

Chapter 5

Conclusion

In this thesis we have attempted to present and summarize the major problems associated with data collected from radio-tracked animals. Identification of these problems is paramount, as they cannot be avoided or rectified by any simple modification of the sampling plan. Methods of analysis must treat the data as is, and as such, cannot rely on overly simplistic assumptions.

Presently, there are a number of methods that allow for an assessment of how animals use their environment. We have identified the more common of these techniques, and shown how they can be used with radio-tracking data. We have also identified some of the major problems associated with these methods.

Identifying the spatial distribution of the radio-tracked animals is also of importance to researchers. Again, the more common methods of identifying the areas inhabited by a specific animal have been presented.

The primary aim of this thesis was to identify a technique that would allow for an assessment of the habitats that a given animal preferred. It was also desired that the technique allow for the construction of a map showing the probabilities of occupation by an animal, given the past locations and habitat preferences of the animals in that area.

Our work has indicated that the Bayesian approach outlined in this thesis is

such a method. It has a number of features that make it suitable for use in radio-tracking studies, including ease of use and the ability to be used with either a single animal or multiple animals.

Further work on this method is certainly required, particularly with regard to the extensibility of the new method. How can it be enhanced, how can the precision of the estimates be improved, and how can differences in subgroups of animals be accounted for in the analysis? More detailed comparison of the new method with the older ones is also required.

A GIS implementation of the Bayesian approach is also a necessity, as this is the format for the vast majority of data from radio-tracking studies. For spatial data, GIS software is a more powerful tool than Splus, especially in tasks such as determining the intersection of different habitat types. In addition, the display capabilities of GIS software are far superior than those found in Splus.

Generally, however, the new method has shown promise, and with slight enhancements it could become part of the wildlife biologist's standard toolkit.

Bibliography

[1] Aebischer, N.J., Robertson, P.A. and Kenward, R.E. (1993), "Compositional analysis of habitat use from animal radio-tracking data", *Ecology*, 74, 1313-1325.

[2] Alldredge, J.R. and Ratti, J.T. (1986) "Comparison of some statistical techniques for analysis of resource selection", *Journal of Wildlife Management*, 50, 157-165.

[3] Alldredge, J.R. and Ratti, J.T. (1992) "Further comparison of some statistical techniques for analysis of resource selection", *Journal of Wildlife Management*, 56, 1-9.

[4] Anderson, D.E., and Rongstad, O.J. (1989) "Home-range estimates of Red-tailed Hawks based on random and systematic relocations", *Journal of Wildlife Management*, 53, 802-807.

[5] Augustin, N.H., Mugglestone, M.A., and Buckland, S.T. (1996) "An autologistic model for the spatial distribution of wildlife", *Journal of Applied Ecology*, 33, 339-347.

[6] Augustin, N.H., Mugglestone, M.A., and Buckland, S.T. (1998) "The role of simulation in modelling spatially correlated data", *Environmetrics*, 9, 175-196.

[7] Bernardo, J.M. and Smith, A.F.M. (1994) *Bayesian Theory*, New York: John Wiley & Sons.

[8] Besag, J. (1972) "Nearest-neighbour systems and the auto-logistic model for binary data", *Journal of the Royal Statistical Society, Series B*, 34, 75-83.

[9] Dickey, J.M., Jiang, J-M. and Kadane, J.B. (1987) "Bayesian methods for censored categorical data", *Journal of the American Statistical Association*, 82, 773-781.

[10] Goovaerts, P. (1997) *Geostatistics for Natural Resources Evaluation*, New York: Oxford University Press.

[11] Johnson, D.H. (1980) "The comparison of usage and availability measurements for evaluating resource preference" *Ecology*, 61, 65-71.

[12] Krebs, C.J. (1989) *Ecological Methodology*, New York: Harper Collins.

[13] Manly, B.F.J., McDonald, L.L, and Thomas, D.L. (1993) *Resource selection by animals: statistical design and analysis for field studies*, Chapman and Hall, New York, NY.

[14] Neu,C.W, Byers, C.R., and Peek, J.M. (1974) "A technique for analysis of utilization-availability data" *The Journal of Wildlife Management*, 38, 541-545.

[15] Osborne, P.E. and Tigar, B.J. (1992) "Interpreting bird atlas data using logistic models: an example from Lesotho, Southern Africa", *Journal of Applied Ecology*, 29, 55-62.

[16] Quade, D. (1979) "Using weighted rankings in the analysis of complete blocks with additive block effects" *Journal of the American Statistical Association*, 74, 680-683.

[17] Silverman, B.W. (1986) *Density Estimation for Statistics and Data Analysis*, Chapman and Hall: London.

[18] Tucker, K, Rushton, S.P., Sanderson, R.A., Martin, E.B., and Blaiklock, J. (1997) "Modelling bird distributions - a combined GIS and Bayesian rule-based approach" *Landscape Ecology*, 12(2), 77-93.

[19] Welch, D., Staines, B. W., and Catt, D. C. (1990), "Habitat usage by red (Cervus elaphus) and roe (Capreolus capreolus) deer in a Scottish Sitka spruce plantation." *Journal of Zoology* , 221, 453-76.

[20] White, G.C. and Garrott, R.A. (1990) *Analysis of wildlife radio-tracking data.*, Academic Press, San Diego.

[21] Wielgus, R.B., (1993) *Causes and consequences of sexual habitat segregation in grizzly bears*, Ph.D. Thesis, University of British Columbia.

[22] Worton, B.J., (1989) "Kernel methods for estimating the utilization distribution in home-range studies.", *Ecology*, 70,164-168.

[23] Wu, H. and Huffer, F.W. (1997) "Modelling the distribution of plant species using the autologistic regression model", *Environmental and Ecological Statistics*, 4, 49-64.