SPATIALLY EXPLICIT MODELLING OF TOPOGRAPHIC AND CLIMATIC INFLUENCES ON VEGETATION DISTRIBUTION IN LABRADOR'S MEALY MOUNTAINS

SETH LOADER

Your file Votre référence
ISBN: 978-0-494-33430-0
Our file Notre référence
ISBN: 978-0-494-33430-0

SPATIALLY EXPLICIT MODELLING OF TOPOGRAPHIC AND CLIMATIC INFLUENCES ON VEGETATION DISTRIBUTION IN LABRADOR'S MEALY MOUNTAINS

BY

©SETH LOADER

A thesis submitted to the
School of Graduate Studies
in partial fulfilment of the
requirements for the degree of
Master of Science

Department of Geography
Memorial University of Newfoundland

March 2007

St John's Newfoundland and Labrador

ABSTRACT

Changing vegetation distributions due to climate change are a concern world wide. Understanding these potential changes is of special interest in the proposed Akamiuapishku (Mealy Mountains) national park (Labrador, Canada). The potential for change under future climate scenarios was investigated through spatially explicit statistical models. The models are based on sampling data from classified Quickbird high resolution satellite imagery. Topoclimatic variables were used to predict percentage cover by vegetation cover classes. Temperature was by far the most important predictor variable but other variables such as incident solar radiation and measures of slope and sheltering also proved to be useful predictors. The relationships between temperature and the percentage cover variables were nonlinear in most cases and so nonlinear parameter estimation was used to build the predictive equations. The predictions suggest there is great potential for an increase in abundance of coniferous forest in the study area given future climate scenarios used.

ACKNOWLEDGEMENTS

As a member of the Labrador Highlands Research Group I received a significant amount of support for my field work from the organization. I am indebted to those who created and control the project. I am also thankful to the people and organizations that help to support the LHRG and its activities; these include the Climate Change Impacts & Adaptation Program, Northern Environment Initiative, Inland Fish and Wildlife Division and Parks Canada. I also appreciate the opportunity to work in an area that is included within the traditional territory of the Innu Nation. My fieldwork would not have been a success or as enjoyable without the help of my field assistant, cook, boat pilot and brother; Thomas Loader. Dr Keith Lewis not only provided great support through managing the camp but also assisted with my fieldwork and has been a valuable resource in the writing stages of my thesis. My supervisor, Dr Alvin Simms, has provided great assistance throughout my masters program but also allowed considerable freedom to find my own path, which is greatly appreciated. My thanks go to Mariana Trindade for so many things, most noteworthy being her cheery mood when I was injured. Thanks go to Andrea Marich who provided many useful insights at the editing stages.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATIONS

LIST OF ABREVIATIONS

%γ: The nugget values expressed as a proportion of the total semivariance
γ: Semivariance

ANN: Artificial neural network

B: Beta coefficient

CART: Classification and regression tree

CBH: Circumference at breast height

CSH: Coniferous shrub and forest

DV: Dependent Variable

GAM: General additive model

GLM: General linear model

IV: Independent variable (predictors)

LGLM: linked general linear model

Moran's I: Moran's Index of spatial autocorrelation

MSA: Measurement of sampling adequacy

NE: North east

PCA: Principal components analysis

RMSD: The root mean squared differences

RMSE: Root mean squared error

S.E: Standard error

Sig: Significance level

SW: South west

Wald: The Wald Statistic

LIST OF COMMON PLANT NAMES USED

Common name	Latin name	Authority
Labrador tea (Northern)	*Rhododenron palustre*	L.
Bilberry	*Vaccinium uliginosum*	L.
Crowberry	*Empetrum nigrum*	L.
Bearberry (Alpine)	*Arctos alpina*	(L.) Spreng.
Sheep laurel	*Kalmia angustifolia*	L.
Sedges	*Carex spp.*	
Diapensia	*Diapensia lapponica*	L.
Moss spp.	*Rhacomitrium, and many other species*	
Black spruce	*Picea mariana*	(Mill) B.S.P.
White spruce	*Picea glauca*	(Moench) Voss
Balsam fir	*Abies balsamea*	(L.) Mill.
Larch	*Larix laricina*	(DuRoi) K. Koch
Alder (Mountain)	*Alnus crispa*	(Ait.) Pursh
Dwarf birch	*Betula glandulosa*	Michx.
Chuckley pear, Sasktoonberry	*Amelanchier spp.*	
Willow	*Salix spp.*	
Raspberry	*Rubus spp.*	
Moss Heather	*Harrimanella hypnoides*	(L.) Don

LIST OF APPENDICIES

1. INTRODUCTION

1.1. Nature and rationale

The literature discussing our planet's changing climate agrees that climate is changing at a faster rate than in the past (IPCC, 2001). There is, however, uncertainty about finer details such as how large the changes are going to be. Regardless of the details, changing conditions are likely to impact ecosystems because many of the processes that control ecosystems are affected by climatic conditions. Changing conditions are likely to have the most significant effects on high latitude ecosystems because these areas are predicted to experience the greatest climatic changes. Recent studies in the Canadian Arctic have shown that a dramatic rise in temperature around a decade ago caused displacement of tundra by forested areas (Danby and Hik, 2007). Figure 1.1 shows the results of a simulation from the Canadian Institute for Climate Studies (CICS). The example given is for mean summer temperatures but the trend of greater changes further north is similar in predictions for other conditions such as quantity of precipitation. Alpine areas such as the Alps (Pauli et al., 2007) and the Tibetan Plateau (Klein et al., 2004) have been shown to be particularly sensitive to climate change because climate is often the limiting factor with regard to vegetation distribution. The Mealy Mountains contain high elevation alpine areas which are likely to be sensitive to change. The alpine areas in the Mealy Mountains are also of particular interest because of their isolation, which increases the risk of local extinction of alpine species.

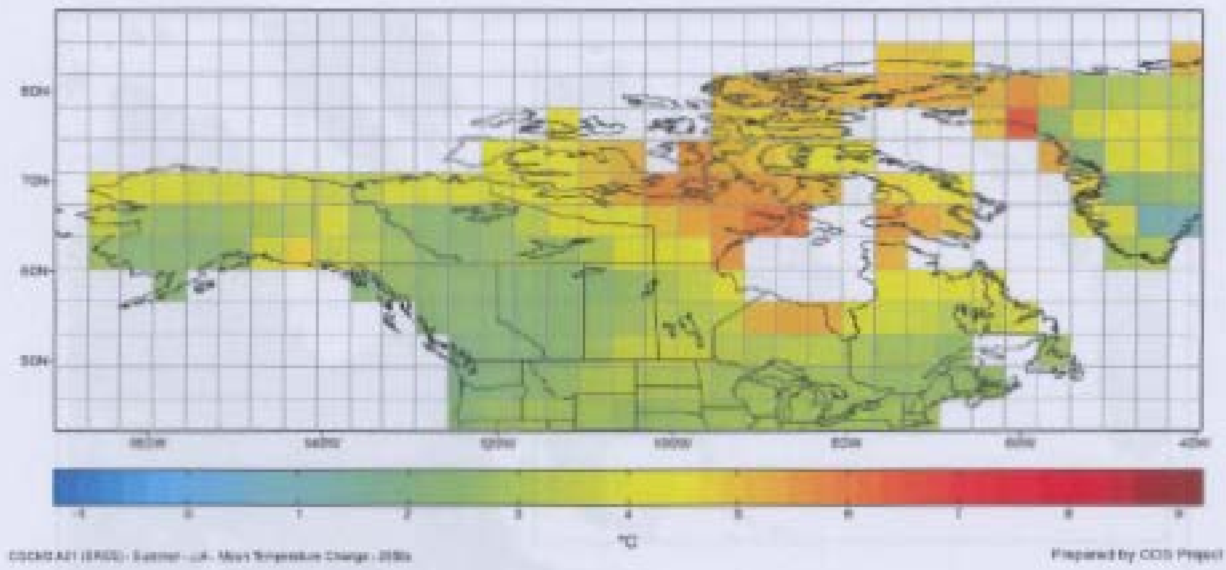

Figure 1.1 Predicted summer (JJA) mean temperature change - 2050s (Canadian Institute for Climate Studies, 2003).

Predictive modelling is one way to attempt to understand the link between an areas's climatic conditions and the vegetation that occurs there (Gottfried *et al.*, 1999, Guisan and Zimmermann, 2000). With this understanding further modelling may predict how distributions might be different in the future under different climatic conditions.

1.2. Study area

The study area is located inside the proposed Akamiuapishku (Mealy Mountains) national park. The Mealy Mountains are located south of Lake Melville in the south east of Labrador (Figure 1.2). The geographical extent of the study area ranges from 58.898° West and 53.634° North in the southeast corner to 58.777° West and 54.559° North in the northeast.

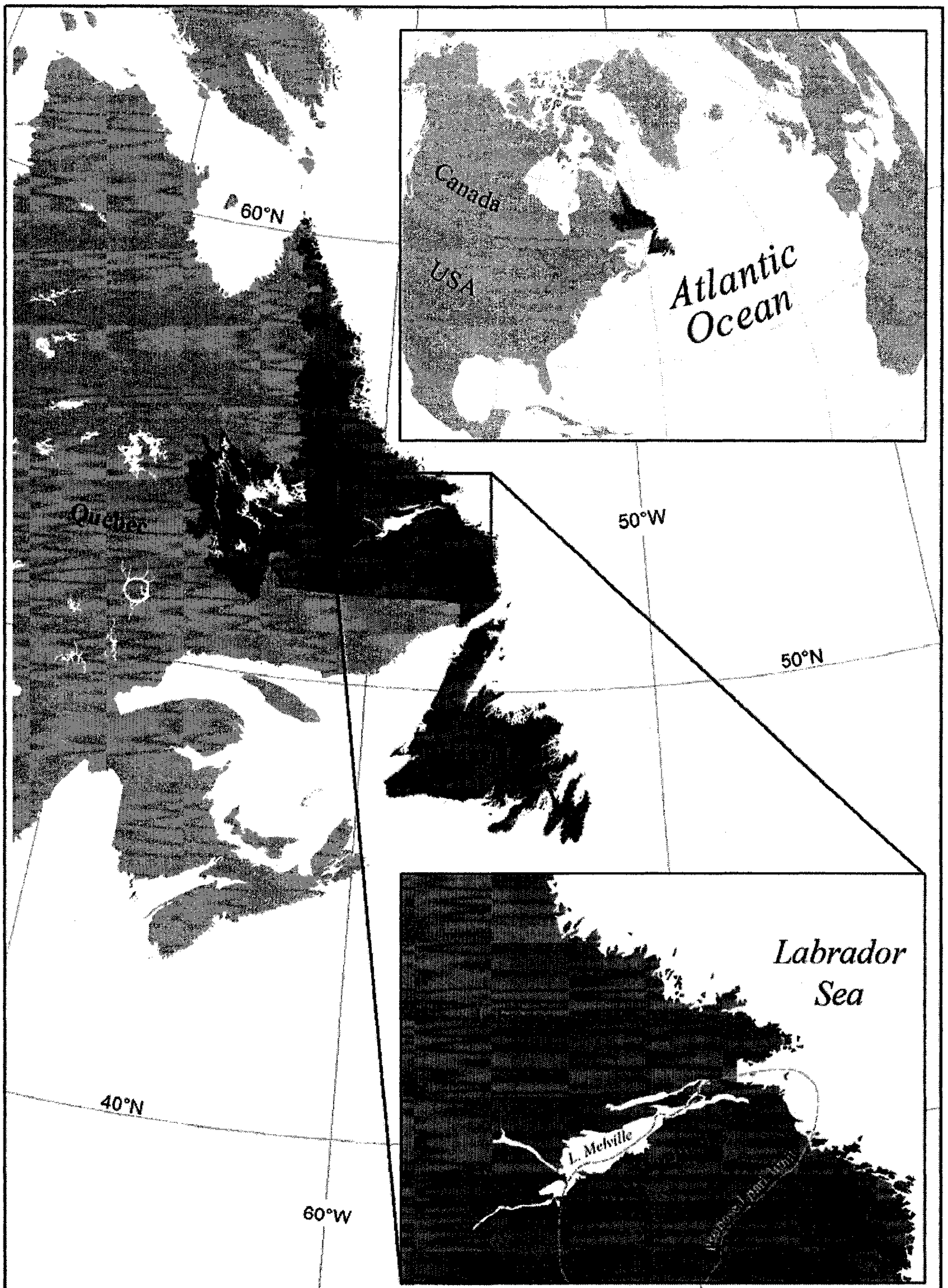

Figure 1.2 Study area location. Proposed park extent from Parks Canada (2005)

The summits of the Mealy Mountains do not generally reach sharp peaks but are rather more rounded and broad. Reaching a maximum height of 1100m (asl) the mountains are a relatively isolated area of high ground (Figure 1.3). The isolation of this region makes it particularly important for conservation because if species assemblages are displaced they may have nowhere to move to and will become locally extinct. The Mealy Mountains are composed mainly of the Canadian Shield "gneissic" rocks and have been affected by Quaternary glaciations and other tectonic and erosional forces (Gray and Lauriol, 1985, Jacobs *et al.*, 2005). The climate of the area is affected both by large scale continental circulation patterns and maritime effects of the Labrador Sea. In the winter, the Icelandic low, an area of low pressure south of Greenland, causes prevailing northwesterly winds. Cold arctic air is brought to the region when high pressure regions form in the north or northwest. A persistent low pressure area near Ungava Bay causes westerly winds in the summer (Keith, 2001). The close proximity of the Labrador Sea results in a damp maritime climate with large accumulations of winter snow some of which remains late into the summer (Jacobs *et al.*, 2005). Climate records (Environment Canada, 2007) show a warming trend in the region over the last 11 years, which is most significant for summer and fall temperatures. Climate modelling exercises predict warming in this region over the coming decades (Canadian Institute for Climate Studies, 2003). In the scenario used in this study taken from the Coupled Global Climate Model Two, with the results modified for the local situation, the mean summer temperature is predicted to rise from the current (1971-2000) 0.8°C to 1.8 °C in 2040-2069 and 3.0 °C in 2070-2099 (Jacobs, 2006)[1].

[1] Personal communication, Dr. John Jacobs , Department of Geography, Memorial University (2006)

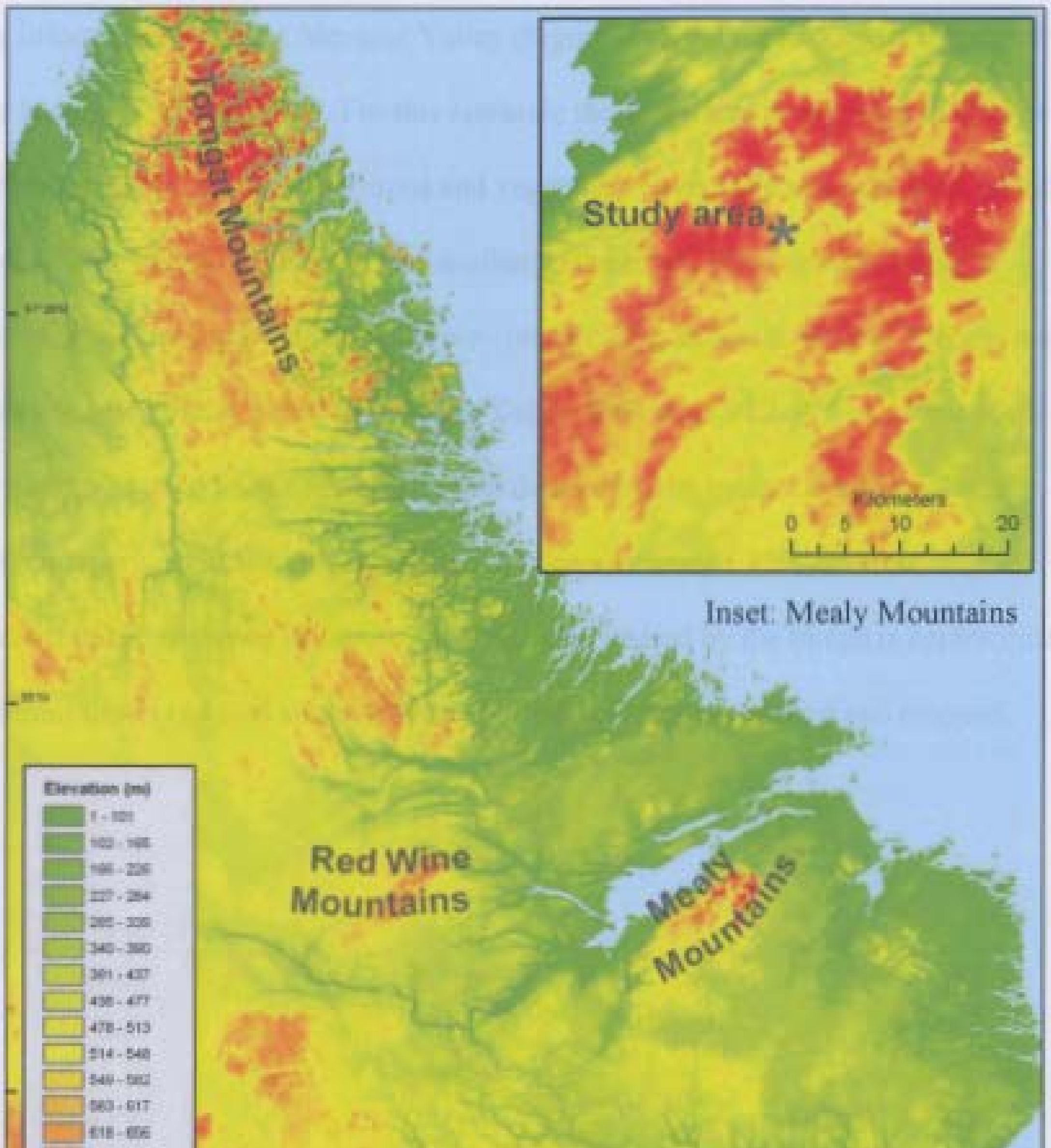

Figure 1.3 Mountainous areas and relief, Labrador. Derived from shuttle radar topography mission (SRTM) data (NASA *et al.*, 2002). The Mealy Mountains can be seen as an isolated area of higher ground. The study area is shown located in the centre of Mealy Mountains (inset).

The area of main interest to the Labrador Highlands Research Group (LHRG) is the valley informally known as Moraine Valley (Figure 1.4) and the summit at the top of the valley known as summit 1057. For this research, the study area is extended to include a wider range of topographic conditions and vegetation cover. The extended area includes two other valleys one to the north and another to the south. The northern valley is unnamed while the southern valley known (unofficially) as Swiss Valley. There is also a smaller unnamed valley between Moraine Valley and Swiss Valley. The extent of the study area to be used in this research is best defined by the extent of the Quickbird satellite imagery, from which vegetation cover was estimated. The extent of the Quickbird image is shown in Figure 1.4. The area covered by the image is approximately 63 km^2 but this is reduced slightly when the imagery is orthorectified and cropped.

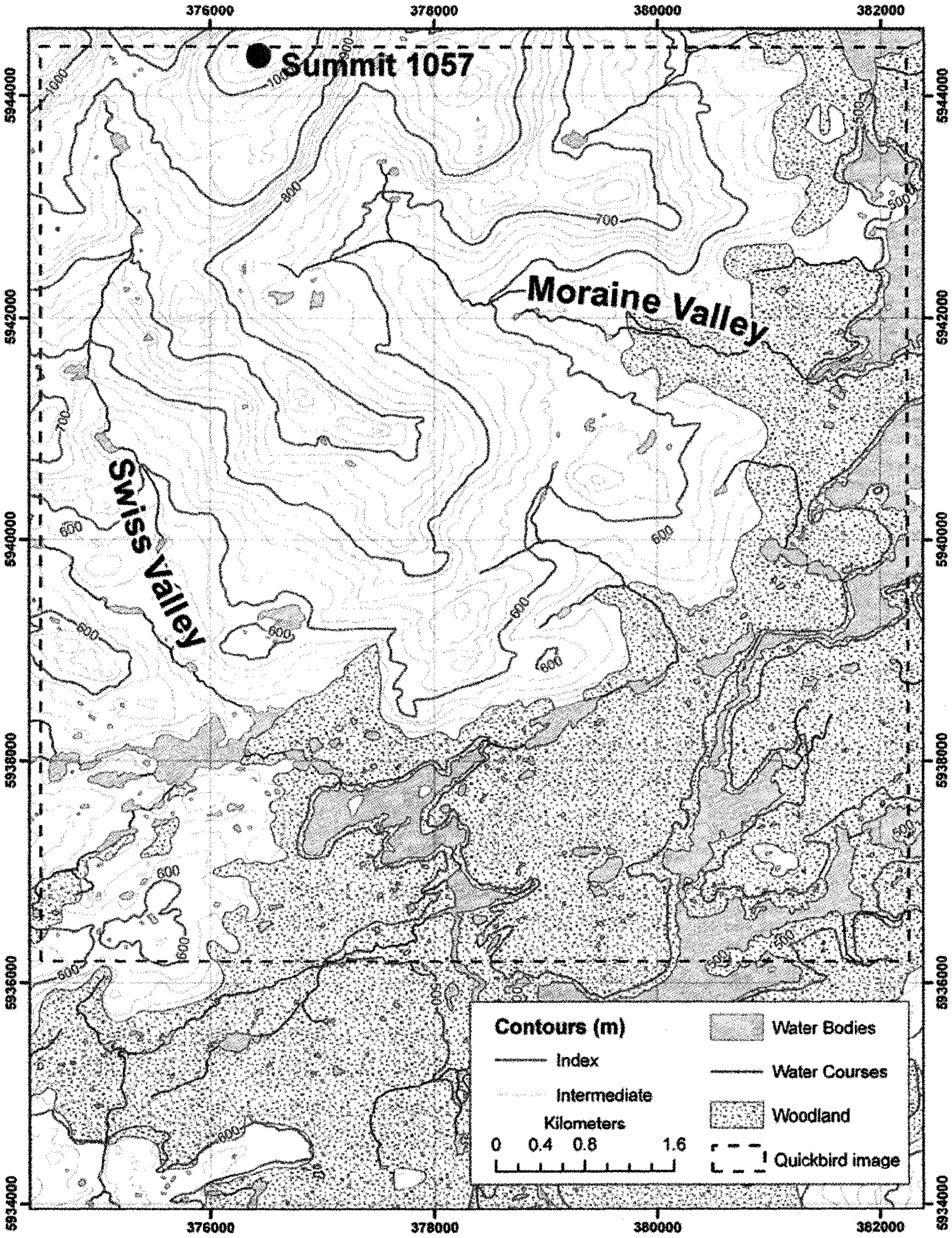

Figure 1.4 Study area extent; Derived from NTDB data (Natural Resources Canada, 2005). All place names are unofficial.

1.3. Purpose

Given that the study area falls inside the Mealy Mountains National Park (Parks Canada, 2005) knowledge about vegetation dynamics will aid future management. An understanding of how vegetation distributions might respond to different changes in conditions will allow better educated decisions in the future. This knowledge could be useful for operations such as park zoning, conservation prioritisation and future monitoring. Methodologies developed as part of this research may also be applied in other areas where similar issues are under investigation.

1.4. Theory

The common hypothesis about vegetation dynamics under a changing climate in northern hemisphere highland ecosystems is that species assemblages will migrate upwards in elevation and northward in latitude. Eventually the assemblages that currently occur at the highest elevations and northern most latitudes will be displaced (Pauli *et al.*, 2003, Pauli *et al.*, 2007). As shortage of suitable areas for these highland assemblages will force them to be out competed by species and assemblages better suited to the new conditions. This theory is commonly reduced to the level of tree-line shift, where vegetation is considered as either forested or not. Forests occur at lower elevations and more southerly latitudes, while at higher elevations and latitudes non forested areas prevail (Körner, 1998, Körner and Paulsen, 2004). The result is one northern latitudinal tree-line and many elevational tree-lines. The situation in the Mealy Mountains study area is a little different, mainly due to issues of scale. Tree-line rarely occurs as a distinct break between forest and tundra but rather occurs as a progression where trees become less frequent and more stressed (Körner and Paulsen, 2004). In mountainous areas that reach high elevations and have steep gradients the progression can occur in a relatively narrow range, when compared to the whole of the mountain. In the Mealy Mountains the progression occurs over a much larger proportion of the mountains range and is therefore not seen as a line. The lack of a distinct tree-line is shown in Figure 1.5, which looks down on a valley from a high ridge. Trees are generally seen lower down in the valley, closer to the lake, and the high ridge on the left side is free of trees. From extensive inspection of the study area it was seen that there is no clear elevation at which trees no longer occur and variation in the other types of vegetation do not follow discrete elevational gradients. Thus factors other than elevation, such as topographic sheltering and moisture retention have some impact on vegetation distribution (Körner, 1998). This is an issue of scale because on larger, steeper mountains these variations caused by other factors would be smaller relative to the overall distribution.

Figure 1.5 Topographic effects on altitudinal tree-line in the Mealy Mountains, Labrador. No clear elevation tree-line is seen as the distribution of trees appears very patchy and irregular.

1.5. Objectives

The primary objective of this research is to create a spatial model (spatially explicit) that can predict vegetation distributions for the current topoclimatic conditions and future conditions. This model will be calibrated for the study area but may be applicable to areas where conditions are similar. This objective can be broken down into the following five tasks:

1. Map the current distribution of vegetation types.
2. Build a database of current topoclimatic conditions.
3. Perform exploratory analysis to inform the model building process
4. Create a model that predicts current vegetation distribution based on the topoclimatic variables.
5. Apply the model to altered topoclimatic conditions that represent past or future scenarios.

1.6. Assumptions

The primary assumption in this research is that the distribution of vegetation types is at least partially dependent on climatic factors and their interaction with topographic conditions.

In this study the formulation of models was based on the current distribution of vegetation, which is assumed to be in equilibrium with the prevailing conditions though this is unlikely in reality (Bennet *et al.,* 1986, Ritchie and MacDonald, 1986, Malcolm *et al.,* 2002 and Midgley *et al.,* 2006). Rules and correlations about the distribution of vegetation cannot be created based on vegetation that is not in equilibrium with its environment because the vegetation is essentially occurring where it should not. This is a problem of inertia (resistance to change). When a type of vegetation is established in an area it may require a large change in conditions to cause it to be replaced (Körner, 1998). However the difference in initial conditions that would cause another type of vegetation to occur at that position may be small (Guisan and Zimmermann, 2000).

1.7. Context of research

This research is being conducted as part of the work of the Labrador Highlands Research Group (LHRG). The LHRG is a multidisciplinary group with students and faculty from the departments of Biology and Geography at Memorial University of Newfoundland and other associated institutes. The objectives of the group are: to better understand highland ecosystems in relation to their local climates, to determine how ecosystems developed in the past and to predict what will happen to alpine ecosystems under a future, perhaps very different, climate (Jacobs *et al.*, 2005).

This research will focus on the first and final aims, while information from other researchers in the project will be used as necessary. The primary information that will be required from the research group is climate data collected from the three climate stations in the study area (Jacobs, 2007).

2. LITERATURE

This section presents a review of the current knowledge regarding modelling techniques used to model vegetation distribution. There is a significant dichotomy in the approaches for the prediction of future vegetation distributions. The division occurs between equilibrium models and transient models (Guisan and Zimmermann, 2000). Transient models are commonly based on the processes that cause a particular distribution to be formed. Consequently transient models normally require extensive knowledge regarding processes such as seed dispersal and disturbance regimes that are crucial to the formation of these patterns. Transient models also tend to be considerably more complex than equilibrium models (Starfield and Chapin, 1996, Cousins *et al.,* 2003). Due to the significant increases in required data and model complexity transient models were not considered for use in this research and so modelling was restricted to equilibrium techniques. The second section of this review introduces some issues associated with scale. The third section introduces and discusses neighbourhood effects and spatial autocorrelation.

2.1. Equilibrium models

Equilibrium models can be constructed using a range of methodologies but are all based on a similar concept. The general concept of equilibrium modelling can be broken into three stages; calibration, altering input conditions, and making predictions (Figure 2.1). The model is calibrated to represent relationships between the environmental conditions and vegetation distributions. These relationships can be used with alternative conditions that represent possible future or past conditions.

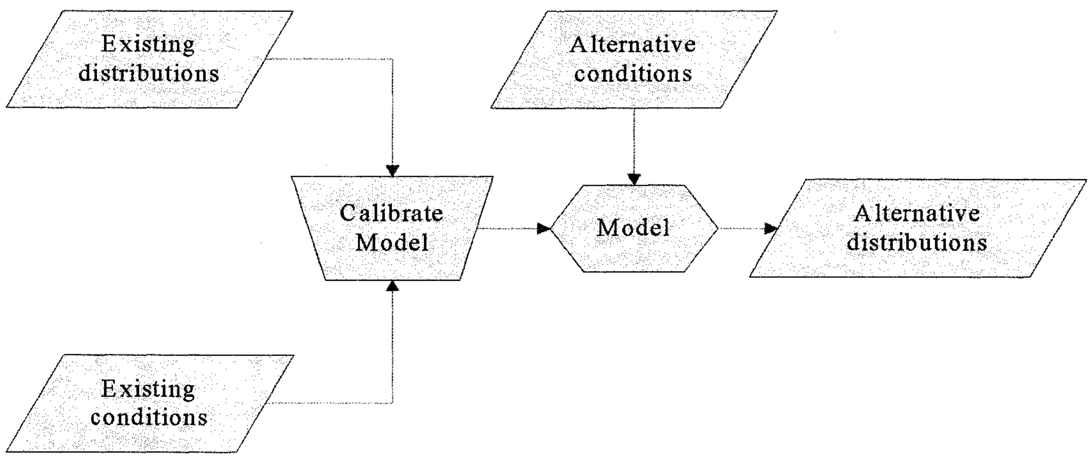

Figure 2.1 Conceptual overview of equilibrium models. Existing conditions and distributions are used to calibrate the model then the conditions are altered to make a prediction.

Equilibrium models are generally dependent upon the ecological theory of niches. Niches can be defined as either driven by the environmental requirements of the species or the impact the species has on the environment. In distribution modelling it is the environmental requirements of the species that are of interest. If it is possible to determine the environmental conditions that a species or assemblage requires to survive, then its distribution can be predicted based on those environmental conditions. There are two methods that may be used to quantify the niche of a species or assemblage (Guisan and Thuiller, 2005). The first is sampling of observed plant (or animal) distributions in relation to the variation in environmental conditions, known as *in situ* sampling. The second is *ex situ* sampling where different environmental conditions can be investigated individually, this requires controlled conditions such as those found on a laboratory. *In* *situ* sampling is preferred in this case for several reasons. First, *ex situ* sampling would be impractical especially as the focus here is on all assemblages present in the study area not just a single species. Second, *in situ* sampling has the advantages of ease of sampling and cost efficiency but has a large disadvantage. When the *in situ* distribution is sampled it is the realised not fundamental niche that is being assessed. The fundamental niche may not be realised because of factors such as competition and interactions (Guisan and Zimmermann, 20005). The result is that the niche is most likely underestimated.

Equilibrium models only predict potential distributions that an assemblage or species could achieve given certain conditions. There is no inclusion of movement from one distribution to another. The potential distribution is therefore the distribution that would be achieved once the vegetation has reached equilibrium with the environment. For these reasons equilibrium models are sometimes called static distribution models.

All of the stages in the conceptual model (Figure 2.1) will vary with each model but the stage that is of most importance is the method of calibration. The majority of calibration methods are statistical but other methods use artificial intelligence or machine learning methods (Cairns, 2001). Table 2.1 provides an overview of model construction methods used in the literature. This table is not in anyway exhaustive but does provide a framework for discussion of the variety of methods that are used. Some of these methods are discussed to provide an overview of what they involve and how they can be applied.

Table 2.1 An overview model construction methods, full details can be found in Guisan and Zimmermann, 2000.

Type of response variable	Probability distribution	Statistical approaches	Possible modeling technique	Examples of habitat modeling studies
Quantitative (continuous)	Gaussian	MULREG	WA, LS, LOWESS, GLM, GAM, Regression tree	Huntley et al., 1995; Heikkinen, 1996
		ORDIN	CANOCO	Hill, 1991; Gottfried et al., 1998; Guisan et al., 1999
	Poisson	MULREG	GLM, GAM	Vincent and Haworth, 1983; Guisan, 1997
	Negative binomial	MULREG	GLM, GAM	–
Semi-quantitative (ordinal)	Discretized continuous	MULREG	PO model, CR model	Guisan, 1997; Guisan et al., 1998; Guisan and Harrell, 2000; Guisan, in press2000
	True ordinal	MULREG	Stereotype model	–
Qualitative (categorical, nominal)	Multinomial	MULREG	Polychotomous logit regression	Davis and Goetz, 1990
		CLASSIF	Classification tree	Walker and Moore, 1988; Burke et al., 1989; Moore et al., 1991; Lees and Ritman, 1991
			MLC	Frank, 1988
			Rule-based class	Twery et al., 1991; Lenihan and Neilson, 1993; Li, 1995
		DISCR	DFA	Lowell, 1991
		ENV–ENV	Boxcar, Convex Hull, point-to-point metrics	Box, 1981; Busby, 1986; Carpenter et al., 1993; Tchebakova et al., 1993
	Binomial	MULREG	GLM, GAM, Regression tree	Nicholls, 1989; Austin et al., 1990, 1994; Yee and Mitchell, 1991; Lenihan, 1993; Brown, 1994; Van de Rijt et al., 1996; Guisan, 1997; Saetersdal and Birks, 1997; Franklin, 1998; Leathwick, 1998; Zimmermann and Kienast, 1999; Guisan et al., 1999; Guisan and Theurillat, 2000
		CLASSIF	Classification tree	Franklin, 1998; Franklin et al., 2000
		ENV–ENV	Boxcar, Convex Hull, point-to-point metrics	Busby, 1986; Busby 1991; Walker and Cocks, 1991; Shao and Halpin, 1995; Huntley et al., 1995
		BAYES	Bayes formula	Skidmore, 1989; Fischer, 1990; Aspinall, 1992; Brzeziecki et al., 1993

Adapted from Guisan and Zimmermann, 2000.

2.1.1. Multiple regressions

The multivariate form of the general linear model (GLM) forms the basis of multiple regressions in which the dependent variable is predicted as the sum of explanatory variables multiplied by their respective regression coefficients (Equation 2.1). The GLM has many weaknesses when used for vegetation distribution models, such as the limitation to linear relationships, but it is an important starting point from which many other methods are developed (Tabachnick and Fidell, 1996).

$$Y_i = A + \sum_{j=1}^{j=v} M_j \times X_{ij}$$

Equation 2.1. The general linear model

Where:
$Y_i =$ The dependent variable.
$X_j =$ The *j*th explanatory variable.
A = The intercept.
$M_j =$ The *j*th regression coefficient.
V= The number of explanatory variables.

(Shaw, 2003)

Different combinations of variable types (discrete, continuous and dichotomous) for the independent variables (IVs) and dependent variables (DVs) result in different analytical methods which fit the GLM. Cairns (2001) used a mixture of continuous and dichotomous IVs to predict continuous DVs. The continuous variables including slope and elevation, while fire and geomorphology were included as dichotomous IVs and the probability of a vegetation type occurring was predicted (continuous DV). This methodology can be thought of as analysis of covariance; the continuous predictor variables are the covariates and the discrete variables are the IVs.

Calef *et al.* (2005) also predicted the probability of the occurrence of a class but in the method they used (logistic regression) a threshold is used to divide the probability into two binary states. In their work they used a hierarchical structure to divide classes. The hierarchical structure works by first splitting the highest level of ecological difference, forested and non forest for example, then making further divisions of the classes below. Logistic regression is a generalized linear model and therefore can be used to deal with non linear relationships between IVs and DVs (Tabachnick and Fidell, 1996). Logistic regression also makes no assumptions about the distributions of the IVs further increasing is potential utility for vegetation distribution modelling.

Another extension to the GLM uses link functions to relate linearly combined IVs to the DV so that non-normal distributions can be modelled (Guisan *et al.*, 2002). This group of methods is confusingly called generalised linear models and so will not be abbreviated to GLMs here; instead the abbreviation LGLMs (linked general linear models) will be used to maintain the distinction. The main advantage of LGLMs is the ability to deal with DVs that have a wider range of distributions (Normal, Poisson, binomial), the IVs are, however, still combined linearly. LGLMs can also be extended to a more developed set of methods called general additive models (GAMs). The main aim of GAMs is to automate the process of creating the link functions that in LGLMs can be over reliant on the user. Miller and Franklin (2002) used LGLMs to model the distribution of vegetation assemblages in the Mojave Desert. They also used GAMs as a means of creating the link functions but used the simpler LGLMs in the actual modelling.

Miller and Franklin (2002) used a mixture of discrete and continuous environmental variables such as average summer precipitation and landform classes to model the distribution of four vegetation classes; a different model was used for each vegetation class.

2.1.2. Non parametric methods

Advances in the availability of computing power have aided development of a variety of methods that are not reliant on specific data distributions in the same way that the linear regression techniques described earlier are (O'Sullivan and Unwin, 2003). These methods are of increasing interest to those performing ecological modelling because many of the relationships in ecological studies are nonlinear, involving high order interactions that are not adequately represented in linear models (De'ath and Fabricicus, 2000). These methods come in various forms and with a host of different names such as artificial intelligence and machine learning. A full discussion of these methods is not relevant to this thesis but two methods of particular interest in vegetation distribution modelling will be reviewed briefly. The two methods to be covered here are artificial neural networks (ANNs) and classification and regression trees (CARTs).

A conceptual representation of an ANN is shown in Figure 2.2. ANNs get their name from their brain like structure and learning like behaviour (O'Sullivan and Unwin, 2003). The input layer contains the IVs, each IV is a node (neuron) and is represented by a single circle. The output layer contains the results, one node for each vegetation class, normally consisting of binary values, but values in a continuous range from zero to one can also be used (O'Sullivan and Unwin, 2003). Between the input and the output are a number of hidden layers, one is shown in Figure 2.2 but many more can be used. The hidden layers are where the actual processing occurs and they work by applying iteratively adjusted weights to the data. In vegetation distribution modelling applications ANNs are normally used in supervised mode, the model is trained using existing distribution data. Given a set of environmental variables as inputs and existing class distributions the model applies randomly created weights to the input in an attempt to predict the output. The initial random weights are iteratively adjusted until the model can predict the output (O'Sullivan and Unwin, 2003). Once the model has been trained it can be applied to different inputs to predict the distributions given those new conditions.

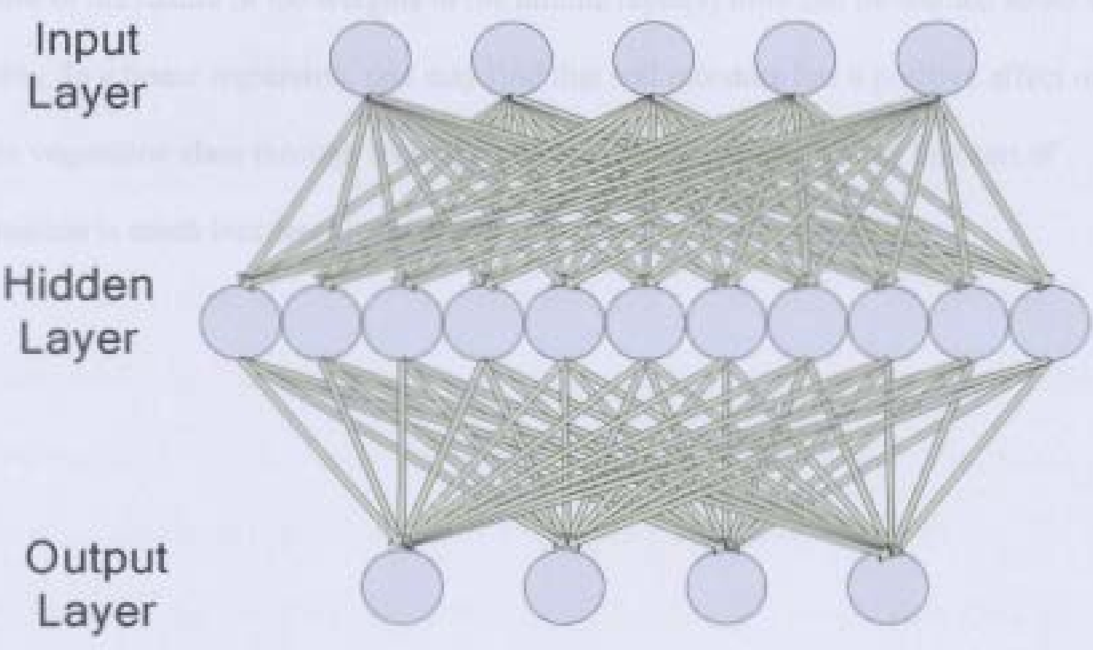

Figure 2.2 Conceptual representation of an ANN; Adapted from Cairns (2001). The nodes in the hidden layer(s) are iteratively adjusted in order to predict the output layer based on the nodes in the input layer.

ANNs have two common issues that restrict their usage: over-fitting and the black box effect. The internal structure of an ANN can be so complex that it predicts the training data very well but does not generalise so that the model could be used for other input data. The conceptual issue of over-fitting is illustrated in Figure 2.3, where two distributions can be divided by the two environmental conditions (x and y). The generalised model divides the majority of the points and finds the main trend, the over-fitted model correctly classifies all the points even those that do not fit the general trend. The over-fitted model therefore performs well when classifying the training data but is not as good for representing the main trend. This problem of over-fitting can be avoided by using an appropriate level of complexity in the hidden layer(s); this is a slightly subjective process that requires a certain amount of skill. The black box problem only matters when there is some desire to know what the trends are that divide populations. Because of the nature of the weights in the middle layer(s) little can be learned about their meaning. In a linear regression, one may find that soil moisture has a positive effect on a certain vegetation class through looking at the variable's weightings but this sort of information is much less clear in an ANN.

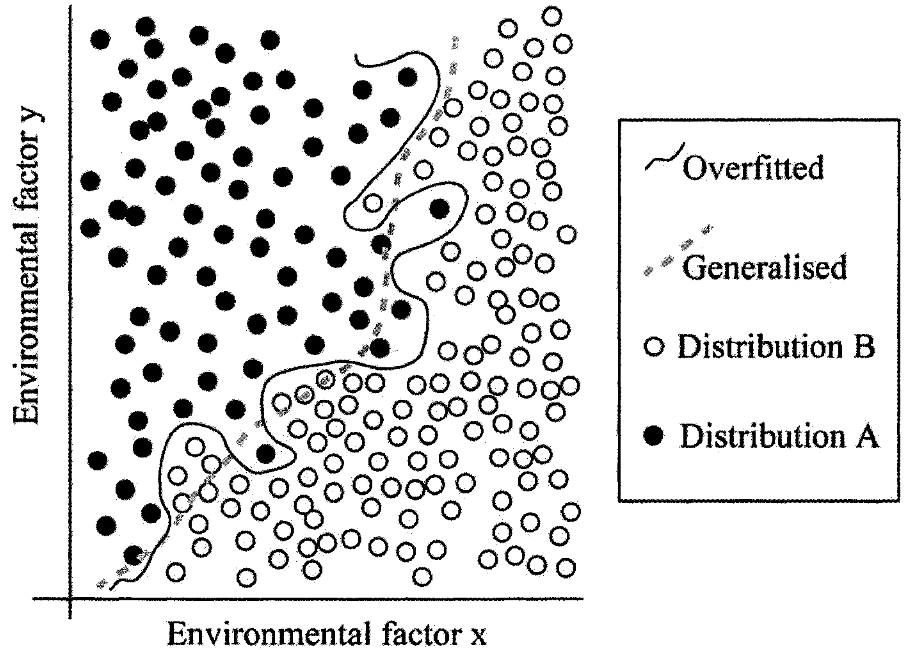

Figure 2.3 Over-fitting ANN models; Adapted from O'Sullivan and Unwin (2003). Two populations can be divided using two environmental factors (x and y). The over-fitted solution does not generalise the differences between the two distributions.

Classification and Regression Trees (CARTs) can use continuous, discrete or dichotomous IVs to predict either discrete DVs (classification trees) or continuous DVs (regression trees). CARTs work by dividing the data in a hierarchical manner, this is shown conceptually in Figure 2.4. The IVs (elevation, curvature, soil depth, air temperature, and soil moisture) are used to divide the data into new groups; the final groups are labelled as G6 to G11. Each division is based on one IV and aims to split the previous group into two groups that are different from each other but homogenous within themselves (De'ath and Fabricicus, 2000). The final groups (leaves) can be defined in two ways:

1. By the mean values of the observations that are put into the group.

2. By a frequency distribution of the classes the observations put into that group come from.

CARTs are especially advantageous because the divisions make simple rules that can be easily incorporated into other systems, for example the rules can be incorporated into expert systems.

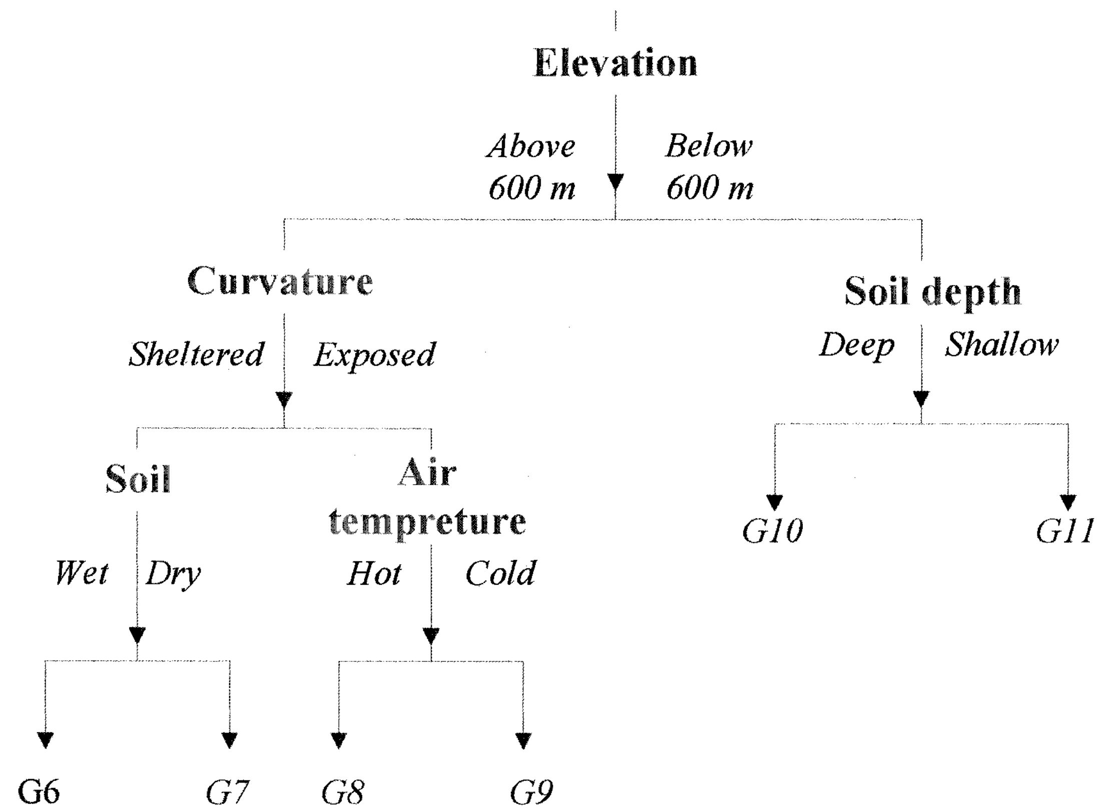

Figure 2.4 Classification and regression trees. Distributions are sequentially divided using environmental factors such as elevation.

Cairns (2001) performed a comparison between GLMs, ANNs and CARTs using these methods to classify vegetation into five cover classes. The classifications (predictions) were made using IVs such as topographic moisture potential, direct beam solar radiation and disturbance history. He found that each of the three methods had their respective advantages and that ANNs were generally the best predictor but they exhibited a higher degree of variability than the other methods.

2.2. Issues of scale

Scale is a critical issue in vegetation modelling. Working at the wrong scales can cause a wide range of problems including the complete failure of a model and the production of an essentially useless model. In modelling efforts abstraction is necessary, as there is no way to represent reality completely. Mismatching levels of abstraction have limited the progress of vegetation distribution modelling (O'Connor, 2002). The results of mismatching levels of abstraction could for example mean that the output from the model is too vague to be useful (not matching the model to its purpose).The planned level of precision could be too fine resulting in a model that can't make predictions with an accuracy. Therefore in performing this research scales were considered at an early stage so that a useful model could be produced. There are a number of scales or levels of abstraction and associated issues that must be considered:

1. **Spatial scales:** Model operation spatial resolution, input topographic data resolution, satellite data resolution, field sampling spatial resolution, neighbourhood size, and assemblage migration rates.

2. **Temporal scales:** Climate data period coverage (yearly averages, seasonal averages etc.), prediction temporal resolution, time differences between data collection, assemblage migration rates, and season changes in vegetation.

3. **Ecological scale:** Resolution of vegetation classification systems and resolution of ecological behaviour.

Most of the issues detailed above relate to the resolution of information in some scale. In this study resolution can be defined as:

"The smallest spacing between two displayed or processed elements; the smallest size of feature that can be mapped or sampled" (Burrough and McDonnell, 1998. p305).

It is important to consider that most of the scale issues are interrelated, often maximising the resolution on one scale will require a decrease in others. If a model is desired to make predictions at the ecological level of species it is likely to have to be more spatially and temporally coarse than a model predicting assemblages. Another consideration is that there is no one correct resolution for each scale but appropriate resolutions depending on research objectives (Morrison, 2002 p123).

2.3. Neighbourhood effects

W. R. Tobler (1970) stated his first law of geography "everything is related to everything else but, near things are more related than distant things". If this apparently obvious and relatively simple premise were false there would be little to study in the field of geography. The relationship Tobler (1970) stated can be described as positive spatial autocorrelation, negative spatial autocorrelation can also occur where similar values are found scattered and values close to each other are dissimilar.

Tobler's first law raises the important, but frequently ignored issue of suppressed variance. This issue can be illustrated through an example of an agricultural experiment (Griffith, 1987). The aim of the experiment was to test if fertilizing crops altered their yield. To investigate this effect crops were treated in patches some being fertilized and some not. The idea behind this treatment was that there would be some difference (or variance) between the yields in the treated and untreated areas. The resulting variance between the two treatments was not as great as was expected. This was found to be due to seepage of fertilizer from treated sections into untreated sections. The patches were therefore not as different as they should have been; the variance between them was suppressed by the seepage. The seepage effect is more widely known as a neighbourhood effect. A neighbourhood effect is a process that makes near observations more similar (or different) than is expected. Failure to account for any neighbourhood effects in an analysis results in the incorrect estimation of variance, as in the fertilizer study where the variance was underestimated.

It is not the spatial autocorrelation in a data set that causes issues and weakens analysis, rather it is neighbourhood effects that are unseen or unaccounted for. In the fertilizer example the actual level of spatial autocorrelation does not matter since it would have been largely controlled by the layout of the patch treatment. It is the suppression of variance due to seepage of fertilizer that weakens the analysis. If the researchers accounted for effect of seepage on the variance they could have estimated the real variance.

2.4. Summary of findings from the literature

The review of relevant literature confirms that objectives of this study are reasonable and that there are a range of established methods that can be used to achieve them. The methods used in this research will be limited based on the objectives, data and resources. Transient modelling techniques will not be used because they are beyond the scope of this project as are *ex situ* sampling methods. Consequently modelling will be based on *in situ* sampling using equilibrium modelling methods. Within these limits there is still a large range of possible methods that can deal with different data types. Parametric modelling techniques will be explored first and if required non-parametric methods will be investigated. In order to create a model that is useful various issues of scale must be considered and resolved, for example an appropriate spatial resolution must be chosen and a suitable level of ecological abstraction must be found. Neighbourhood effects must also be investigated as they could potentially lead to serious errors in model construction but also because they may be used to strengthen the model.

3. DATA

3.1. Field data

A vegetation survey was completed in the summer of 2005. This survey was designed to fulfill two purposes. First, vegetation distribution information was required for ground-truthing of satellite imagery. Second, the survey was designed to quantify present day variation in vegetation parameters such as tree height and density.

Sampling was carried out at two different levels of detail in order to capture the spatial variability of vegetation and record more detail at some of the sites. The detailed sampling (planned) was carried out at 80 of 201 potential predetermined sites, focusing on those sites within the area of planned satellite image acquisition. The less detailed sampling (ad-hoc) was carried out while travelling between the planned sites. The ad-hoc points were primarily designed to fulfill the satellite imagery ground-truthing objectives. The surveying was completed using a set of predetermined ground cover classes which were identified on the ground through visual inspection.

3.1.1. Ground cover classes

The ground cover class system was based on the system used in Natural Regions of Newfoundland and Labrador (Meades, 1990) and recommendations given by Dr. Luise Hermanutz (2005)[2] and Dr. John Jacobs (2005)[3]. With a total of 28 classes the system was designed to be finer than any system that will be used in modelling efforts. The finer

[2] Personal communication, Dr. Luise Hermanutz, Department of Biology, Memorial University of Newfoundland (2005).

[3] Personal communication, Dr. John Jacobs, Department of Geography, Memorial University of Newfoundland (2005).

classification systems allows for wider utility of the data and aggregation based on further analysis. The classes, summarised in Table 3.1, are intended to represent distinct plant communities that are based on species and form. These classes are intended to be easily identified on the ground and have significantly different environmental requirements. The system uses major and minor classes. The minor classes are subsections of the major classes, for example Heath is a major class and there a number of types of heath that form minor classes.

Table 3.1 Ground cover classes used in this study. The classes are identified by a major and minor class, for example sedge tundra (TUN_SED) is member of the tundra major class. The major classes are as follows: TUN = tundra, KHZ = Krummholz, DSH = deciduous shrub, CSH = coniferous shrub (low form conifers but not Krummholz), COP= coniferous open canopy, CCL= coniferous closed canopy, UND= understory, BRK= bare rock or soil, H20= Water, and FBG = fens and bogs.

No	Name	Major Class	Minor Class	Keys To identification	Photograph
1	Bilberry Heath (Tundra Heath)	TUN	HET	Labrador tea, Bilberry, Crowberry, Bearberry and Kalmia	
2	Sedge Tundra	TUN	SED	Sedges dominate with very few heath plants and dry soil not deep, wet peat.	

No	Name	Major Class	Minor Class	Keys To identification	Photograph
3	Low Heath (Alpine Heath)	TUN	ALP	Very low ground hugging forms of plants, cushion plants (Diapensia) and mosses.	
4	Moss Tundra	TUN	MOS	Clear dominance of Rhacomitrium	
5	Spruce Krummholz	KHZ	SP	Black spruce and White spruce occurring in krummholz form	
6	Balsam fir Krummholz	KHZ	BSF	Balsam fir in krummholz form	
7	Tamarack Krummholz	KHZ	TAM	Larch (Tamarack) in krummholz Form	
8	Dwarf birch shrub	DSH	DBIR	Dwarf birch in shrubby form	

No.	Name	Major Class	Minor Class	Keys To identification	Photograph
9	Alder shrub	DSH	ALD	Alder in shrubby form	
10	Mixed deciduous shrub	DSH	DBA	Dwarf birch and Alder (sometimes willow) in shrubby form	
11	Amalanchier shrub	DSH	AMA	Amalanchier in shrubby form	
12	Willow shrub	DSH	WIL	Willow in shrub form	
13	Spruce shrub	CSH	SP	White spruce and Black spruce in shrubby form	
14	Balsam fir Shrub	CSH	BSF	Balsam fir in shrubby form	
15	Larch shrub	CSH	TAM	Larch in shrubby form	
16	Open Canopy spruce	COP	SP	Erect spruce stems without a closed canopy	
17	Open Canopy Balsam fir	COP	BSF	Erect Balsam fir stems without a closed canopy	

No.	Name	Major Class	Minor Class	Keys To identification	Photograph
18	Open canopy Larch	COP	TAM	Erect larch stems without a closed canopy	
19	Closed canopy spruce	CCL	SP	Erect spruce stems very close or touching.	
20	Closed canopy Larch	CCL	TAM	Erect larch stems very close or touching.	
21	Closed canopy Balsam fir	CCL	BSF	Erect Balsam fir stems very close or touching	
22	Raspberry understory	UND	BBY	Understory with blackberry and grasses	
23	Bedrock	BRK	BED	Rock not broken into boulders	
24	Boulders	BRK	BOL	Rock broken into boulders	

No.	Name	Major Class	Minor Class	Keys To identification	Photograph
25	Gravel	BRK	GRV	Small fragments of rock	
26	Peat	BRK	PET	Bare peat with no vegetation	
27	Water	H20		Water	
28	Fens	FBG	FEN	Wet areas with deep peat and a dominance of sedges	

3.1.2. Planned sites

Survey site locations were pre-determined using stratified random points. These points were designed to be approximately one kilometre apart with no point occurring in mapped areas of water. The points were also checked for even representation of elevation classes and any areas with particularly dense or spare coverage had points removed or added to correct these problems. The points were located on the ground using GPS (global positioning system) receivers, using two receivers to keep track of any unusual inaccuracies and the general level accuracy achieved with receivers.

These planned sites were surveyed using a system of four transects radiating in cardinal directions from the centre point as shown in Figure 3.1 Each transect had four observation points at one metre spacing and the centre was also used as a recording point giving a total of 17 recording points for each sample site.

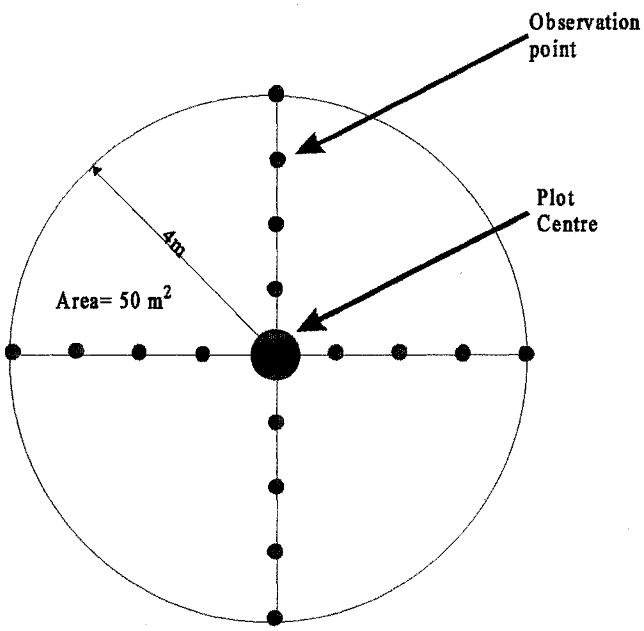

Figure 3.1 Plot sampling method. The plots included 17 observation points with four on each of the cardinal directions and one in the centre.

Further parameters, such as the number of erect stems, were also recorded for the site as a whole. Each of the 17 recording points at a site was classified into a major and minor class, discussed in 3.1.1. Where applicable, the under-storey type was recorded and the presence or absence of moss and lichen was noted. The major and minor class for each recording point was based on the class definitions and what plants occurred at the point and within 30cm of the point. Where more than one class was present the dominant class was chosen. If two or more classes appeared equally abundant the class that was most abundant in the surrounding metre was recorded. Counts of the number of upright tree stems and erect leaders showing high growth rates (long leaders) were recorded for the site. More detailed information about larger trees was recorded on a tree by tree basis; larger trees were defined as having a measurable (greater than four centimetres) circumference at breast height (CBH). Where larger trees were found in a plot the species, CBH and presence or absence of cones (or other reproductive structures) were all recorded for each tree.

3.1.3. Ad-hoc points

Ad-hoc points were chosen where good representations of classes were found with one class clearly dominating an area larger than a four metre radius. At ad-hoc point locations only the location and major and minor classes were recorded. Where possible the planned sites were visited via straight lines to ensure that the ad-hoc point collection was not biased towards more easily traversed terrain. Figure 3.2 illustrates the distribution of the ad-hoc data. Ad-hoc points are represented by the major class only as there are too many sub classes to summarise.

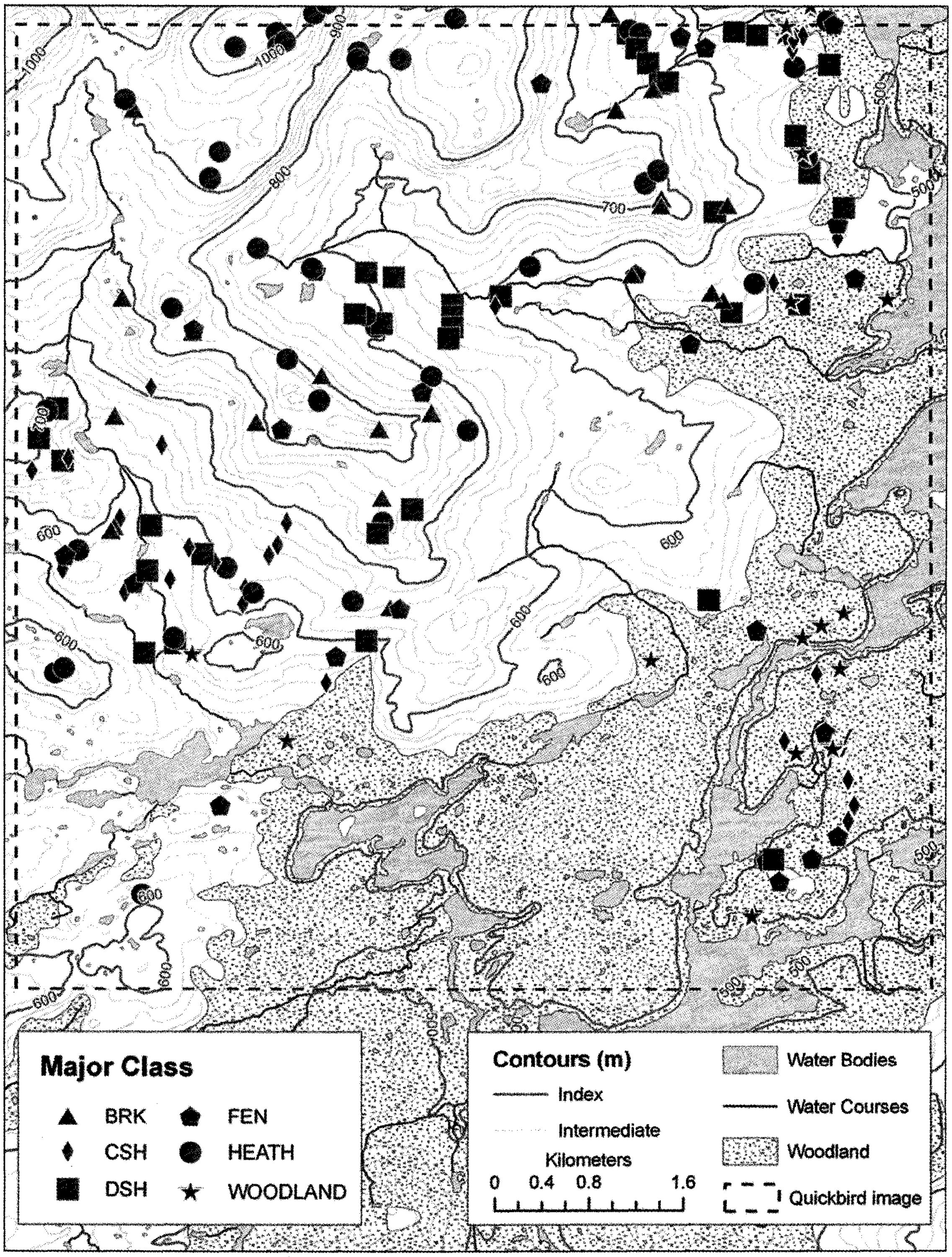

Figure 3.2 Ad-hoc point distribution; major classes are detailed in Table 3.1.

3.1.4. Ground control features

Differential GPS (DGPS) was used to digitize the outlines of features such as ponds and gravel areas. Ground control point (GCP) collection was concentrated in the proposed Quickbird image acquisition area. The coverage was not as dense or as uniform as would be ideal, but the data proved to work well for the orthorectification of the Quickbird imagery. Figure 3.3 below shows the distribution of collected features in relation to the topography and the Quickbird image.

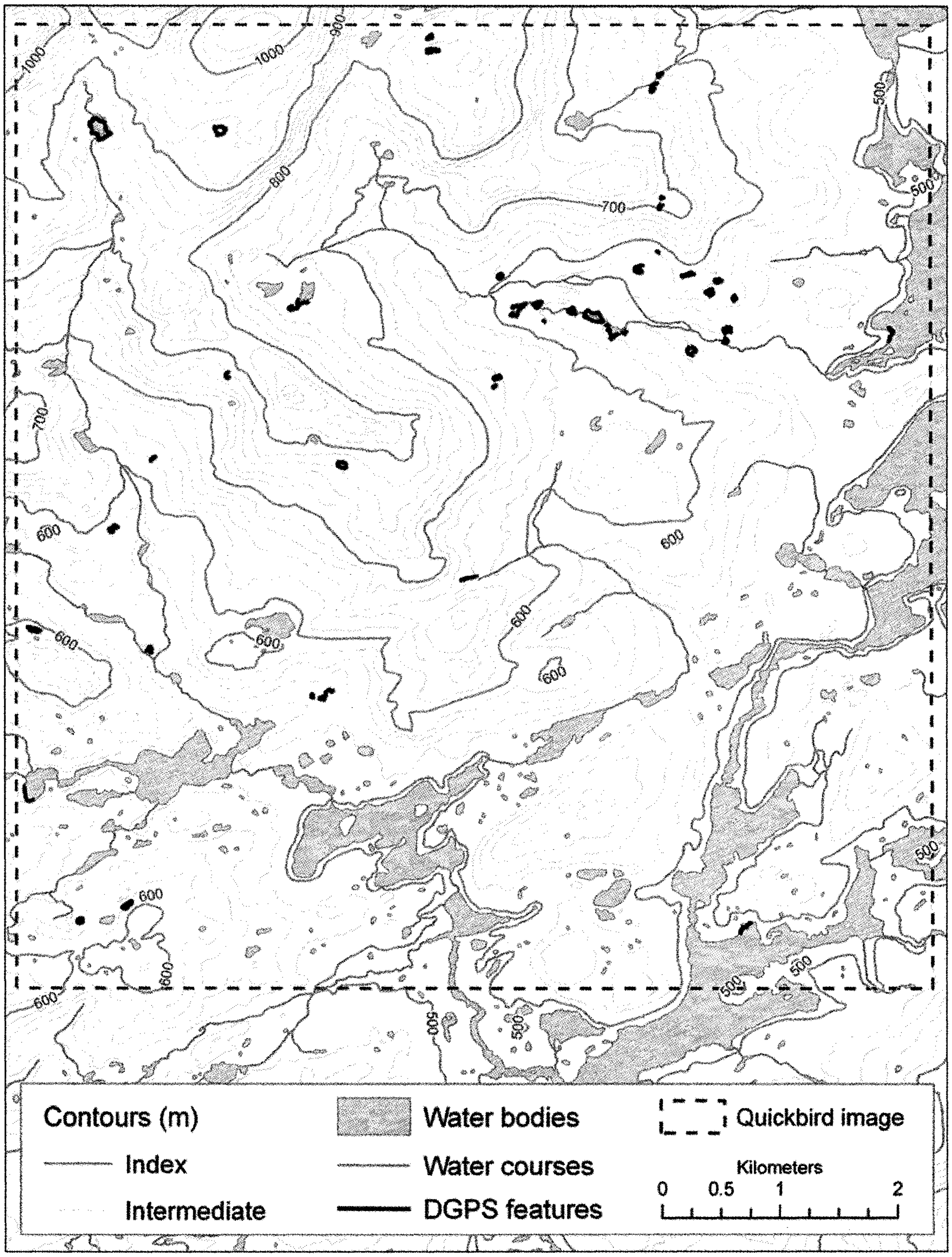

Figure 3.3 Ground control point collection. The ground control points are a bit sparser in the southeast than would be desired.

3.2. Satellite data

A Quickbird image covering the study area was acquired on the sixth of September 2005 at 3:04 pm. LHRG purchased the ortho ready kit bundle product from this acquisition. This product includes four multi-spectral bands, three visible and one infrared, at a resolution near to 2.4m and a panchromatic band with a pixel resolution close to 60cm. The acquisition included some cloud over the northwest corner of the study area, which excludes data from some of the higher areas. Both the panchromatic and multispectral images were orthorectified using the DGPS features collected in the field season as ground control (PCI Geomatics, undated). The orthorectified multispectral images were then used to create a classified vegetation map. The classification was performed using a maximum likelihood algorithm with signature data collected in the field. Polygons were digitized around ground truth points using the imagery as a basis to constrain the polygons to areas where vegetation was identical. Extra polygons were also digitized for areas where ground cover was clear, such as the rock and water classes. The resulting classification was then aggregated based on spectral and ecological differences between classes. The aggregation of classes is shown in Table 3.2. The most notable aggregation is the grouping of shrubby form and full form trees coniferous trees into the CSH class, this is necessary because the spectral signatures are very similar. This grouping is unfortunate since there would be value in being able to differentiate the two classes. The two classes are, however, not that distinct in reality as there is more a progression from shrubs up to full form trees rather than two clearly discrete classes.

Table 3.2 Aggregation of vegetation classes.

Initial Class (Major_Minor)	Aggregated class	Description
brk_bed	Rock	Bare ground or rock
brk_bol		
brk_grv		
brk_pet		
h20	Water	Water
fbg_fen	Fens	Fens and bogs
tun_het	Heath	Heath communities
tun_alp		
tun_sed		
tun_mos		
bbry		
khz_sp	CSH	Coniferous shrub and forest
khz_bsf		
khz_tam		
csh_sp		
csh_bsf		
csh_tam		
cop_sp		
cop_bsf		
cop_tam		
ccl_sp		
ccl_tam		
ccl_bsf		
dsh_dba	DSH	Deciduous shrub
dsh_dbir		
dsh_ald		
dsh_ama		
dsh_wil		

The accuracy of the resulting classified image is detailed in Table 3.3, where the classification accuracy is presented in terms of pixels. The overall percentage of cells correctly classified is 98.6% when water is included and 81.3% when water is excluded. Most individual classes are also well classified, with the exception of fens, which show a large amount of omission error (92%) with many pixels being miss-classified into the CSH class. A further problem with the fens class is the very small proportion of the study area occupied by the class. Consequently the fens class was not considered in the main analysis.

Table 3.3 Vegetation classification accuracy. Accuracy is presented in pixels and summarized in percentages; the classes are as detailed in Table 3.2.

		Ground truth							
		Rock	CSH	DSH	Fens	Heath	Water	Producers Accuracy (%)	Omission (%)
Classification	Rock	964	0	0	0	32	2	96.6	3.41
	CSH	0	1361	55	3	34	1	93.6	6.40
	DSH	0	182	275	0	12	0	58.6	41.4
	Fens	5	118	0	16	11	47	8.12	91.9
	Heath	69	147	26	0	411	0	62.9	37.1
	Water	5	2	0	2	3	48997	99.9	0.0245
	Users accuracy (%)	92.4	75.2	77.2	76.2	81.7	99.9		
	Commission (%)	7.57	24.8	22.8	23.8	18.3	0.102		

3.3. Topoclimatic Variables

The topoclimatic variables were designed to represent the interaction of climate and weather with the topography of the land. None of the conditions were sampled extensively in the field so surrogate variables were used. These surrogates estimate the spatial variability in conditions such as wind exposure using a digital elevation model (DEM) and knowledge about the local climate. A large number of topoclimatic variables were produced so that the important variables could be identified. Two main types of topoclimatic variables were used. The first type is purely topographic measurements such as slope and aspect that are not modified with knowledge of climatic conditions. The second type is variables that are produced by combining the topographic variables with information about climate. For example aspect can be modified to produce a variable that represents the degree to which an area is exposed to winds from a certain direction.

The topoclimatic variables were integrated with the vector based point and area data sets discussed later in this chapter. The topoclimatic variables were integrated with the point-based data sets by extracting the values from each topographic variable raster layer to each point. To allow investigation of the spatial scales of interactions the values from smoothed topoclimatic variables were also integrated. The topoclimatic variables (with the exception of elevation) were smoothed using circular median filters with varying diameters. Filters with diameters of 100m, 500m and 1000m were used to produce a total of four versions of each topoclimatic variable. The topoclimatic data were integrated with the area-based data through the use of the zonal statistics tool in ArcView 3.2. The zonal statistics tool produces statistics in the form of a table for input summary zones. In this case the input zones are the buffers and the statistic of interest is the mean value of the topoclimatic variable. The zonal statistics tool produces a separate summary table for each topoclimatic variable; the summary variables of interest were then integrated using a unique identifier to build a complete summary data set.

3.3.1. Elevation

Elevation is related to many topoclimatic conditions but the one of primary importance is air temperature. In general air temperature decreases with increasing elevation, though inversions do occur and have been observed in the study area (Dr. John Jacobs 2005[4]) they are not considered to be the general climatic situation. Temperature has well documented effects on vegetation and especially the vegetation dynamics in areas where a transition from forest to tundra occurs (Arsenault and Payette,1992, Hobbie and Chapin 1998, Körner1998, Körner and Paulsen, 2004). For the initial investigations, elevation was taken directly from the digital elevation model. In later analyses, temperatures were estimated from elevation using information from the climate stations in the study area. The elevation range for the area surrounding the study area is zero (sea level at Lake Melville) to 1188m but within the study area the elevation ranges from 477m to 1046m.

3.3.2. Slope

A slope variable was calculated from the DEM. Slope is related to a number of conditions such as exposure and soil moisture holding potential. The slope variable is measured in degrees (°), but does not have issues associated with circular variables as the potential range is between zero degrees and 90°.

3.3.3. Aspect

Aspect, defined as the direction the maximum slope faces, is a problematic circular measure. When an average is taken of two values near north (ten degrees and 350° for example) the result would be a south facing aspect (180°). To avoid this problem two

[4] Personal communication, Dr. John Jacobs , Department of Geography, Memorial University of Newfoundland (2005)

alternative variables were made from an aspect variable, these layers were created by taking the sine or cosine of aspect. The cosine of aspect works as a measure of northness, values close to north are near one, values close to south approach minus one and values near east or west move towards zero. The sine of aspect produces a measure of eastness with similar behaviour to northness (Calef *et al.,* 2005).

3.3.4. Curvature

Curvature was calculated from a DEM using ArcGIS 9.1. There are a number of different forms of curvature, these include profile and planiform curvature but the form of interest here is basic curvature. Basic curvature is a measure of how concave or convex a surface is; in the layers produced using ArcGIS 9.1 convex surfaces have positive values and concave surfaces have negative values. Curvature was also calculated from a smoothed version of the DEM. This was done because curvature over a wider area is probably more important than local curvature. The units of curvature are based on the amount of deformation of a surface over a unit of distance, in the output from spatial analyst the units are one over 100 elevation units (meters in this case).

3.3.5. Solar radiation

Annual solar radiation was estimated using an ArcView 3.2 script (McCune and Keon, 2002). The values created by this script are based on the work of Buffo *et al.,* (1972) where incident solar radiation was measured in a wide range of topographic situations. These measurements recorded by Buffo *et al.,* (1972) are stored in tabulated form and are therefore of little direct use. The script predicts solar radiation using equations derived from regressions of the tabled data from Buffo *et al.,* (1972). The output of the script is the estimated annual incident solar radiation in MJ cm^{-2} yr^{-2}. The values in the study area range from 0.47 to 0.918 MJ cm^{-2} yr^{-2}.

3.3.6. Topographic moisture index

A topographic moisture index was calculated based on one used by Cairns (2001). The index was calculated by taking the natural logarithm of the ratio between the upslope area draining through a location and the slope at the location (Equation 3.1). This is based on the theory that the larger the areas that drain into a site the wetter it will be, but steeper slopes will drain more quickly and therefore hold less moisture. Soil moisture could be an important factor as without sufficient moisture potential changes due to increased temperature may not be possible (Black and Bliss, 1980).

$$\ln(A/S)$$

Equation 3.1 Topographic moisture potential

Where:

Ln = The natural logarithm

A = Upslope area draining into a location in square meters

S = Slope in degrees.

This index was created using the Model Builder in ArcGIS 9.1. In the model, represented by a flow chart in Figure 3.4 the inverse of the result is taken so that potentially wetter areas have higher values. The model uses two tools from ArcToolbox 9.1 to estimate the area draining into a location (flow accumulation). The first of these tools processes the DEM to remove pits and the second calculates the flow accumulation. The range of values produced by the model for the study area and surrounding landscape is between -4.4 and 14.4.

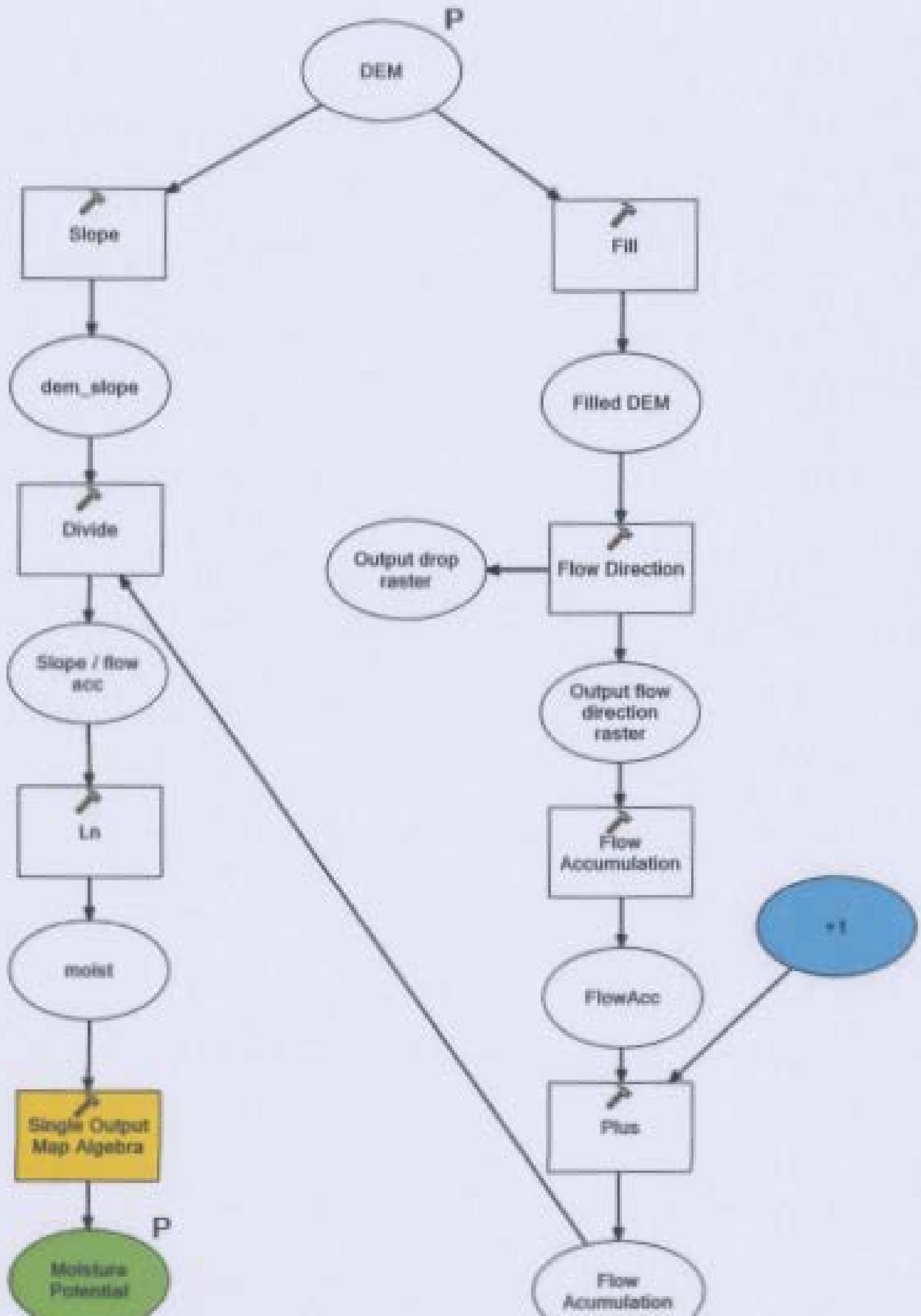

Figure 3.4 Topographic soil moisture index model. The model builder in ArcMap 9.1 was used to calculate this index which is based on water accumulation and slope.

3.3.7. Wind exposure

Wind can cause limitations to developing trees through desiccation and defoliation of trees that grow beyond the protection of surrounding vegetation. Two types of wind exposure variables were created. The first is based solely on local aspect and wind direction. The second, called Angle to sheltering topography, is intended to represent how areas may be sheltered by higher areas upwind (section 3.3.8). The first set of wind variables were calculated in a similar way to the measurement of northness or eastness discussed earlier. The difference is that rather than being a measure of how exposed to the north or east a slope is, the variables are changed to represent exposure to specific wind directions. Two wind directions are used based on the climate normals for the closest weather station (Goose Bay A) (Environment Canada 2002, Dr. John Jacobs 2005[5]). The wind directions used were 257.4°, based on an average of winds coming from the west and southwest, and 22.5° which represents the north easterly winds that dominate in the spring.

3.3.8. Angle to sheltering topography

High ground upwind of an area provides shelter from the wind. The steeper the angle to a sheltering feature the greater the amount of shelter. This variable was calculated using hillshading. Hillshades were performed using the desired wind direction as the sun direction. A number of hillshades were performed at incrementally decreasing sun azimuths between 85° and 15°. The area of shadow was then taken from each hillshade so that areas in shade were coded one and areas not in shade were coded zero. A new layer

---

[5] Personal communication, Dr. John Jacobs , Department of Geography, Memorial University of Newfoundland (2005)

was then created by adding the values from areas of shade. The greater the number in the new layer the greater the angle is to the sheltering feature and so the more sheltered the spot is. The process was repeated three times, once for the dominant wind direction, once for 20° north of the dominant wind direction and once for 20° south. The three resulting variables were then combined using a weighted average where the dominant wind direction had the highest weight (50%). The use of multiple wind directions is intended to account for variation in the dominant wind. Because of the large number of operations required to build this variable the model builder in ArcGIS 9.1 was used to perform required processing, this also allows the index to be created using different inputs such as alternate wind directions. The angle to shelter variable was calculated for the two wind directions identified previously (257.4° and 22.5°).

3.3.9. Snow potential index

Snow has a number of influences on vegetation including providing shelter from harsh winds in the winter and blocking out sunlight in the spring. To include and assess the importance of snow cover an index of snow potential (SPI) was created. The snow potential index used is modified from Cairns (2001). The main modification is that elevation is not included to prevent issues of covariance. This index can be represented by the following equation.

$$SPI = -\left(\left(\frac{C - C_{\min}}{C_{\max} - C_{\min}}\right) \times \left(\frac{(\cos(A - W)) + 1}{2}\right)\right)$$

Equation 3.2 Snow potential index

Where:

C = A curvature layer

A = An aspect layer

W = The dominant wind direction

The index is essentially the inverse of curvature multiplied by shelter from the dominant wind with the two variables converted so that their ranges are constrained between zero and one. Constraining the ranges of the variables ensures that they both have equal influence on the final index value. Although the potential range of the SPI is between zero and one in the study area and surrounding landscape the highest value observed was 0.72.

3.4. Randomly located samples from classified imagery

Both of the previously discussed point-based data sets are limited to relatively small sample sizes (80 and 203). The ad-hoc points, with 203 sites, has a reasonable sample size but given the large number of independent variables a larger sample is desirable (Tabachnick and Fidell, 1996). The classified Quickbird image presents the possibility for a much larger number of samples though classification is limited to the aggregated classes in the classified image. Another disadvantage of using the classified imagery is the error that is added into the analysis due to errors in the classification. To facilitate analysis outside of a raster based environment the classified image was sampled using random points. A large number of sample points were randomly located with a condition that limited the minimum distance between points. This threshold of the minimum distance reduces the issue of clustering. 100m was used as the minimum separation distance; this distance was chosen as a trade off between having too few samples and too much overlap between buffers. Classes from the vegetation classification were then extracted to the points. The classified image does not differentiate between full size trees and shrubby form trees. However the difference between the two classes can be visually detected in the higher resolution panchromatic satellite imagery. The points in the forest or coniferous shrub class were then visually inspected and assigned to either the forest class or coniferous shrub class. The resulting data set is summarised in Table 3.4. The data set contains a total of 1387 sample points; as with the ad-hoc data the heath class dominates.

Table 3.4 Cover class frequencies in the random point-based data set.

Vegetation Class	Frequency	Percent
CSH	312	22.5
DSH	325	23.4
FEN	96	6.9
HEATH	455	32.8
ROCK	64	4.6
WOODLAND	135	9.7
Total	1387	100

3.5. Area-based data

Field observations and preliminary investigations suggested that the high degree of local variability in vegetation type would cause the point-based data to be of little use. The point-based data records what occurs in a very small area, but what occurs at that location is likely to be as much a result of stochastic processes as the influence of topoclimatic variables. To overcome this potential issue an area-based data set was developed. The aim of the area-based data set is to summarise the vegetation that occurs in a larger area, and reduce the influence of random processes. The area-based data set was created by buffering randomly located points. The points were randomly located with a condition so that points did not occur closer than a specified threshold. To facilitate an investigation of the relationship between spatial scales and predictability, buffers were created with different diameters. Every random point generated was buffered to produce circles with diameters starting at 50m and increasing in 50m increments to the largest buffer of 500m. The vegetation cover, from the classified image, and the topoclimatic variables were then summarised for each buffer. The summary for the vegetation cover provides the number of cells of each vegetation type found in the buffer which can be used to calculate percentage cover. The summaries of the topoclimatic conditions produce a mean value of the variable within each buffer. The summaries were then integrated and simplified.

3.6. Vector land cover data

There is little land cover data available at a suitable resolution for this project, the East–Central Labrador Ecological Land Inventory covers the study area but is based on 1:125,000 maps and surveys dating back to the 1970s (GeoGratis, 2003). The National Topographic Data Base (NTDB) provides information at the 1:50,000 scale for the whole of the study area but the data is very generalised (spatially and ecologically) and comparison with satellite imagery suggested the data would not suit this project (Geomatics Canada, 1997).

4. EXPLORATORY ANALYSIS AND DATA PREPARATION

Model creation cannot logically start without some idea of how the data behaves in statistical analysis. Without knowledge of the behaviour of the data, important decisions such as what modelling methods to use cannot be made (Sibley, 1987). Exploratory analysis is also required to discover if there is any predictability that can be modelled or if vegetation distribution is apparently random. Analysis should also not be conducted on data that have not been prepared to resolve issues that may cause incorrect results. There are eight main questions that must be addressed in the exploratory analysis before modelling can begin (Tabachnick and Fidell, 1996):

1. Are all the values within the acceptable ranges of the chosen variables?
2. Are there any univariate or multivariate outliers in the data?
3. Are there any correlations between vegetation distribution and topoclimatic variables?
4. If correlations do exist what is their nature (linear or nonlinear)?
5. What is the spatial nature of the correlations, are all relationships strongest at the same scale?
6. Are there significant correlations between the topoclimatic variables (IVs)?
7. Do conditions such as heteroscedasticity exist?
8. Are there neighbourhood effects and if so what is their nature?

The data sets of primary interest for exploratory analysis are the area-based data set and the randomly generated point samples. These data sets were checked for the above issues.

The majority of data sets in this study have a large number of cases, which is desirable for multivariate analysis but does have one major draw back. With large numbers of cases, the confirmatory style of analysis, performing statistical test to validate hypotheses, is limited as most tests will appear statistically significant regardless of their actual importance (Tabachnick and Fidell, 1996). The use of hypothesis testing is also complicated by issues of spatial autocorrelation in the data sets (Griffith, 1987). These issues increase the importance of exploratory analysis as assessments of significance will be more subjective and should be based on a sound understanding of the data.

4.1. Screening for errors

Because the data sets of interest were created through automated processes and did not require any manual input the chances of data errors are reduced. The data sets created using manual input (ad-hoc and planned) were screened for errors and corrected at an earlier stage as correctness was required for operations such as image classification. All variables were checked to ensure that the values of each case fitted within the range of the variable. No cases were found with values beyond the limits of the variable. In both random point-based and buffer based data sets there were a small number of cases with no data in some variables, the number of cases were less than five percent of each data set. These missing values are caused by flat areas which result in a division by zero (0° slope) in some variables. These cases were removed from the data set as they represented only a small proportion of the sample. In the point data set the "no data" cases accounted for less than one percent of the initial data, after their removal there were 1180 cases left. Because the creation of the area-based data set involves an averaging process there was only one point with division by zero problems created by a slope of zero degrees.

4.2. Outliers

All data sets were checked for both univariate and multivariate outliers using methods in Tabachnick and Fidell (1996). Univariate outliers were defined as cases with Z scores above three or below negative 3. There were not a large number of univariate outliers in either data set (less than 10%) and because the values represent true environmental conditions that will have to be predicted for in the later analysis the outliers were retained.

Multivariate outliers were identified using Mahalanobis distance as described in Tabachnick and Fidell (1996). In the area data, the effect of each variable in creating multivariate outliers was assessed by calculating Mahalanobis distance with each variable removed. None of the variables appeared to have a particularly high impact on Mahalanobis distance so the measurement calculated using all variables was used to define multivariate outliers. In both data sets, cases identified as multivariate outliers were removed. The complete removal of multivariate outliers is not of great concern since they account for only two percent of the point data set and just below four percent of the area-based data.

4.3. Spatial nature of correlations

Before the correlations were investigated in detail, it was necessary to assess how they changed with different spatial scales and choose a spatial scale at which to work. The number of different buffer sizes in the area-based data makes a provision for this investigation. Curve fitting was performed for each variable at each buffer size for every vegetation class. In order to fit the full range of curves available some variables were altered to remove negative values, this was done by adding constant value to the values of variables with negative numbers.

The aim of the curve fitting is to fit a relationship between the percentage cover of a cover class and an independent variable. The curve fitting process generates a large amount of information as there were ten buffer sizes, 12 variables, four ground cover classes, and 11 different potential curve types; the number of potential curves that results is 5280. The number can be greatly reduced as only the best fitting curve for each relationship is of interest. Additionally many of the relationships cannot be fitted to a curve. The R^2 value for the curve estimations was used to assess how strong the relationships were and find those that were of interest. To investigate the differences between buffer sizes the relationships between the cover classes and elevation was chosen as it is the main variable that shows strong relationships. The change in R^2 for predictions of cover classes using elevation and curve fitting at the different buffer sizes was graphed (Figure 4.1). For each relationship the curve type that fitted the relationship best throughout the range of buffer sizes was selected. The resulting graph summarizes how the relationships change with increasing buffer size. The most dominant trend is the increase in the strength of relationships with increasing buffer size. This trend most likely occurs because in small areas vegetation is quite random but over a broader area it is more controlled by topoclimatic factors. Another useful trend is that the shapes of relationships generally stay the same over the different buffer sizes. This means that if the parameters of a curve created for a 500m buffer was used to predict percentage cover in a smaller buffer they would predict reasonably well. This aspect of the data was tested by assessing the differences in the predictions for 200m buffers with curve parameters from both the 200m data and 500m data. This investigation showed little difference between the predictions and so further affirms the hypothesis that relationships do not change significantly within the range of the buffer sizes. There is a noticeable decline in the rate of increase in the strength of relationships with increasing buffer size. This decline in the rate of increase suggests that there is an optimum size of area with which to work; this is the point where most of the increases in the strength of relationship have occurred. 200m was chosen as the optimum size area and much of the analysis in the following section is carried out using only the data from the 200m buffers. The 250 or 300m buffer sizes could also be reasonably justified, however it is desirable to minimise the buffer size due to the small size of the study area.

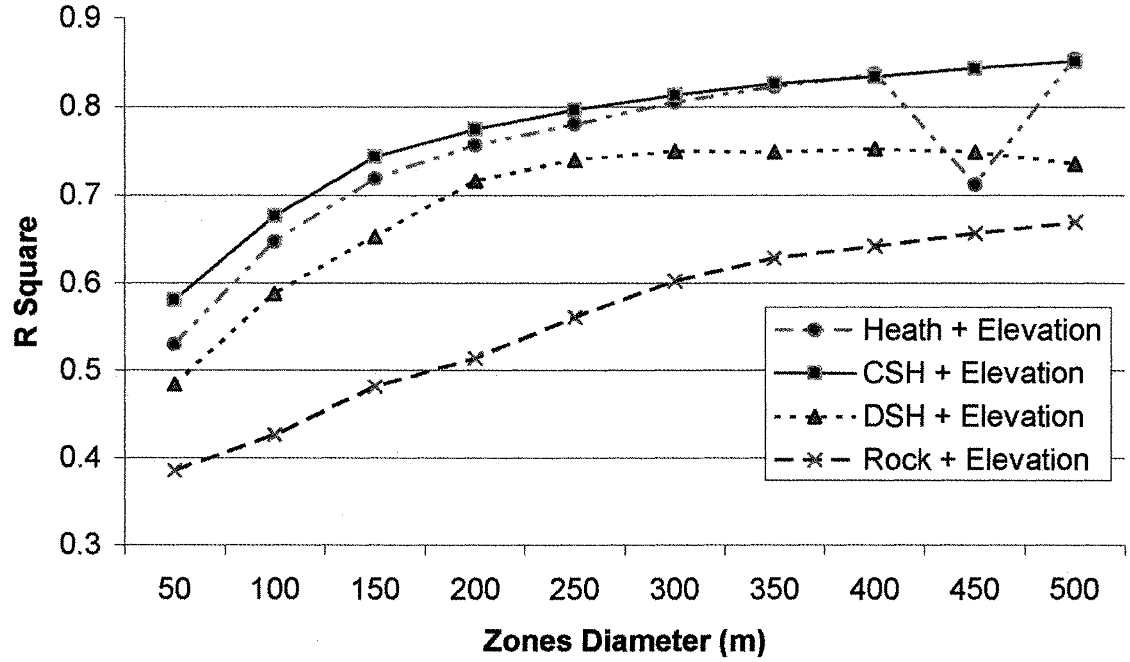

Figure 4.1 Spatial changes in relationships between IVs and elevation. The increase in the strength of relationships starts to level off around 200m.

4.4. Correlations between independent variables

As many of the independent variables (IVs) are produced using similar methods there is a high probability that multicolinearity, variables with high bivariate correlations with other variables, will occur. This is not of great concern at the exploratory stage but in further analysis knowledge of how the variables are interrelated will be important to avoid the issue of multicolinearity. Since the IVs in the two data sets are the same the interrelationships are likely to be similar in both the point and area data sets. The 200m buffer subset of the area data set was chosen to assess interrelationships and any issues found are assumed to be similar in the point data. The simplest way to assess correlations that might be an issue is through a bivariate correlation matrix (Appendix I). The downside to this method is that the matrix is very large due to the high number of inputs.

From investigation of the correlation matrix a number of conclusions can be drawn:

1. Many of the topoclimatic variables are highly correlated but few relationships (six) have bivariate correlations above 0.9 (Appendix I).

2. Variables that include some measure of aspect, such as the measures of wind exposure, are all interrelated.

3. Topoclimatic variables generally have high correlations with the topographic variables used to derive them.

4. Slope and elevation have a high positive correlation. This is due to the shape of the landscape, steeper slopes are only found at higher elevations.

Because slope and elevation were highly correlated a new variable was created by normalising slope with elevation. This variable is referred to as normalised slope and produces a measure of slope that is independent from elevation so that the influence of slope may still be investigated.

It is also useful to assess how many variables can actually be used. Because of the high levels of interrelationships clearly not all variables can be used, otherwise problems with multicolinearity will arise (Tabachnick and Fidell, 1996). The number of dimensions that the data actually varies on can be assessed through principal components analysis (PCA). The number and make up of components produced by PCA can help decide how many variables should be used. As with the bivariate correlations PCA was performed only on the area-based data. PCA was performed in SPSS 13 using the factor module with principal component extraction, Varimax rotation and Kaiser Normalisation. PCA was carried out on standardised scores of a restricted set of topoclimatic variables. Variables were not included if they were in some way directly replicated. For variables that had, for example, an exposure and shelter variant only one variant was included. PCA was repeated a number of times in order to find a set of inputs that performed as required. The Kaiser-Meyer measure of sampling adequacy was used to confirm that the variables were related; a value of 0.843 in the final PCA indicated that a high proportion of the variance in the data is shared. Variables were assessed for fitness within the group by looking at the measurement of sampling adequacy (MSA). The majority of variables fitted well within the group, with MSA values greater than 0.5. Curvature and topographic moisture index had very low values and were therefore excluded from the final PCA. Curvature and the topographic moisture index both represent very different factors to the other variables so it is not surprising that they do not fit with the other variables. The rotated component matrix from the final PCA (Table 4.1) was interpreted to assess the number and nature of reliable variables in the data. Two stable components are present in the final rotated matrix. The first component contains variables that have a strong aspect element and the second contains slope. From the PCA and inspection of the correlation matrix it appears that there are only three strong variables in the data; aspect, slope (which is related to elevation) and curvature (which was not included in the PCA). This does not mean that only those variables can be used but that it would, for example, be unwise to try to make a model that included two variables with dominant aspect elements.

Table 4.1 Rotated component matrix used to investigate how many dimensions the data is variable on.

	Component	
Variable	1	2
SW angle exposure	-.886	.157
SW wind exposure	-.960	.248
Snow potential index	.957	-.253
NE angle exposure	.772	-.420
NE wind exposure	.855	-.504
Solar Radiation	-.369	.793
Northness	.365	-.875
Eastness	.973	.058
Elevation	-.155	.474
Slope	.096	.647

4.5. Correlations between the topoclimatic variables and vegetation variables

In order for any modelling methods to be worthwhile there must be relationships between the topoclimatic variables (IVs) and the vegetation variables (DVs). Curve fitting, where a variety of functions such as exponential and logarithmic curves are fitted to correlations between an IV and a DV was used to achieve this task. Curve fitting also allow an investigation of the shape of the relationships, be they linear, or nonlinear. The investigation of correlations between topoclimatic variables was completed somewhat simultaneously with investigations of the spatial nature of correlations, which is described in the previous section. These two issues are described separately for purposes of clarity.

4.6. Curve Fitting

Curve fitting was performed on the area-based data with all but the 200m buffer size excluded. The first variable that was investigated was elevation as it is already known to have the strongest relationships with most of the cover classes. The relationship between elevation and percentage cover by heath is represented reasonably well by a linear function (Figure 4.2) with an R^2 value of 0.630. The relationship is more accurately modelled with a quadratic function. The quadratic function also matches the nature of the distribution, which increases from the lower elevations upwards and the decreases again higher up. A further benefit of a quadratic function is that it would not result in percentage coverage predictions above 100%.

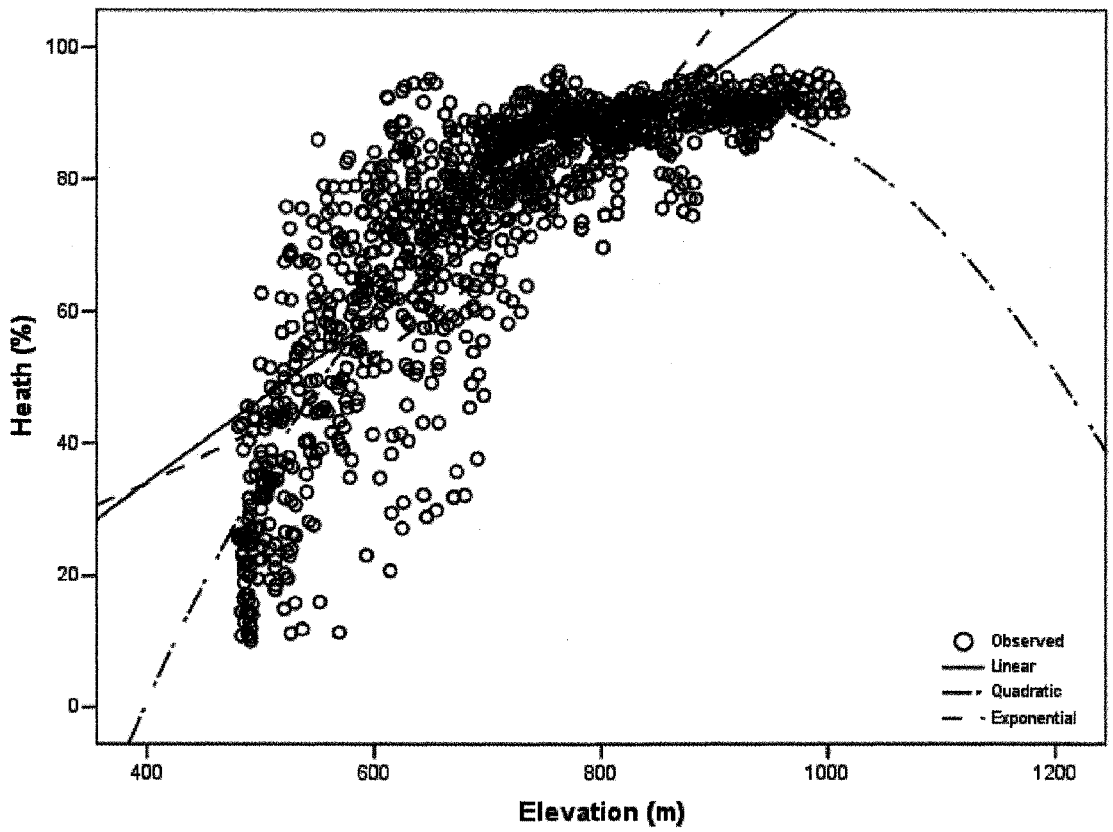

Figure 4.2 The relationship between percentage cover by heath and elevation. The quadratic model provides a good fit for the distribution.

Elevation has a strong correlation with the percentage cover of coniferous shrub. Cover by CSH appears to decreases exponentially with increasing elevation (Figure 4.3). It is worth noting that this relationship is slightly heteroscedastic with more variation at lower elevations.

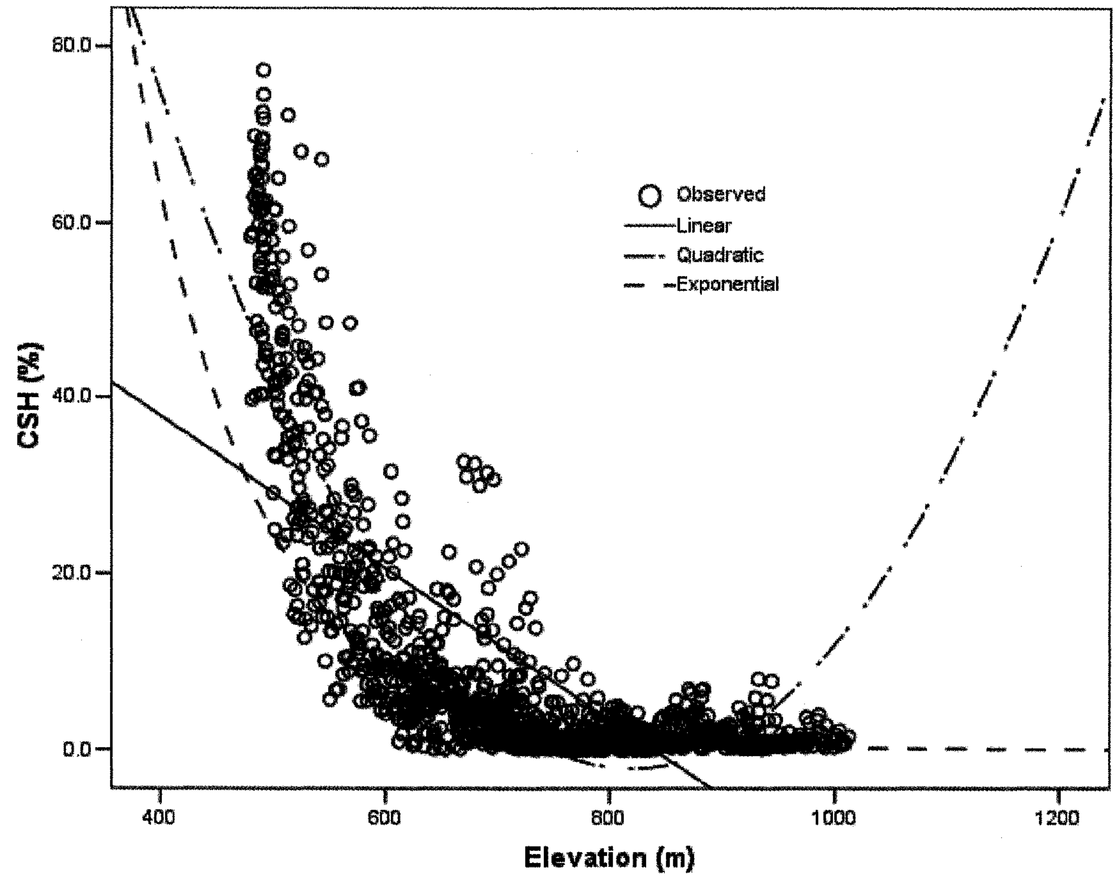

Figure 4.3 The relationship with coniferous shrub and elevation. The quadratic and exponential models both provide good fits but the exponential model is more logical for the cover class

The relationship between elevation and percentage cover by the rock class is modelled well by a linear function ($R^2 = 0.71$). This correlation is also heteroscedastic (Figure 4.4).

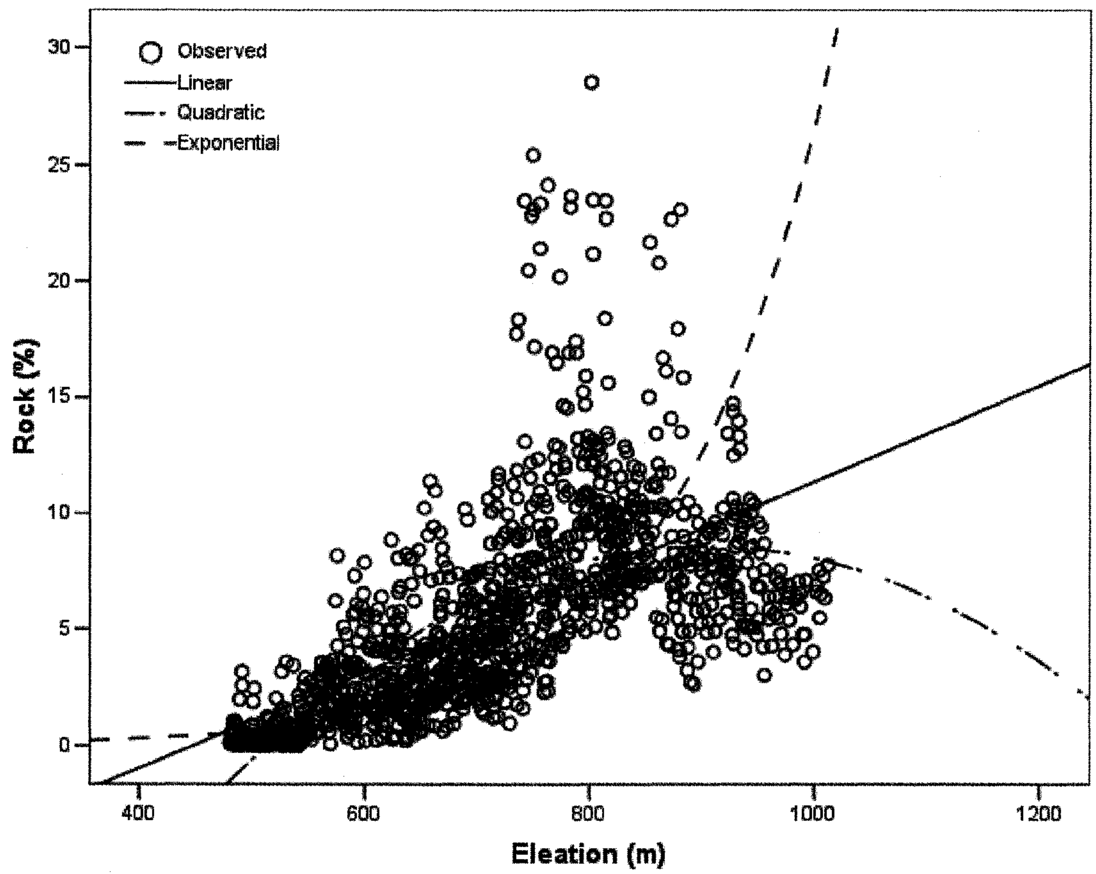

Figure 4.4 The relationship between rock and elevation. A linear model explains a reasonable amount of variance in the distribution ($R^2 = 0.71$) but does not perform well at the higher elevations.

Correlations with other variables are not as strong as those with elevation but some are still of interest. The correlation between fens and NE wind exposure is not clear at the 200m buffer size but at 500m a useful correlation appears. As with many of the other correlations the relationship is heteroscedastic. The measure of alignment with northern aspects (northness) has a strong correlation with percentage cover by fens, which is modelled well with an exponential function. This relationship is only clear with the 500m buffer size. Fens also correlate with solar radiation at the 500m buffer size, an exponential functions fits this relationship with an R^2 value of 0.452. The correlation is negative and heteroscedastic with more variance at lower solar radiation values.

4.7. Differences between groups

The randomly sampled point-based data only contain the vegetation class at a point as opposed to the percentage cover by each cover class. From analysis that has already been conducted it is expected that results from categorical data will be of limited use due to the high degree of local variability. The big advantage of point-based data is that trees are distinguished from shrubs. The point-based data can therefore be used to investigate if trees can be separated from shrubs using the topoclimatic data. Two main types of analysis were performed on the point-based data. The first technique was an investigation of values of Wilks' lambda from discriminant functions analysis. Wilks' lambda is a measure of the proportion of the total variance in discriminant scores not explained by differences among the groups. Wilks' lambda values close to one indicate very little of the variance is explained by differences between groups (Tabachnick and Fidell, 1996). The second technique used was examination of box and whisker plots, which can be used as the graphical equivalent of a difference of means test. These plots were used to determine if there were significant differences between the distributions of topoclimatic variables for each cover class.

4.7.1. Wilks' Lambda: discriminant function analysis

The values of Wilks' Lambda for each of the variables and at each of the spatial filter sizes are shown in Table 4.2. This investigation helped to confirm that the point-based data are less useful than the area-based data. All of the values in the table are close to one indicating that the topoclimatic variables have little power to divide the data into the cover classes. Even elevation which has been shown to have some predicting power previously has a Wilks' Lambda score near to one. The increasing size of filters for the topoclimatic variables does generally decrease the values of Wilks' Lambda, meaning that the variables become more useful, but they are still not powerful enough. Because of the low values of Wilks' Lambda discriminant function analysis was not taken any further.

Table 4.2 Wilks' Lambda scores. The majority of the scores are close to one indicating the variables have little power to divide the classes.

Variable	Raw Points	100m filter	500m filter	1km filter
Elevation	0.709	-	-	-
Eastness	0.985	0.981	0.984	0.988
Northness	0.928	0.927	0.922	0.896
Curvature	0.997	0.976	0.979	0.954
Slope	0.973	0.944	0.839	0.766
Solar radiation	0.941	0.942	0.955	0.968
Angle to summit exposure NE	0.978	0.977	0.974	0.964
Wind aspect exposure NE	0.966	0.961	0.967	0.953
Snow Potential	0.981	0.98	0.976	0.974
Angle to summit exposure SW	0.98	0.98	0.988	0.989
Wind aspect exposure SW	0.982	0.982	0.976	0.974
Moisture Index	0.988	0.961	0.903	0.835

4.7.2. Box and whisker plots – differences in distributions

Inspection of box and whisker plots reconfirmed the previous findings, that there is little potential to divide the point data into the categories using topoclimatic variables. The plot for elevation (Figure 4.5) illustrates the sort of distributions that are desirable. Even with elevation (the strongest predictor) the distributions are not ideal. It would be preferable if there were no overlaps between the distributions of the values for each vegetation class. The figure shows that elevation could be used to divide the data into two or three classes. The first classes would be at the lower elevations, Woodland, CSH, DSH, and fens. Woodland could possibly be separated from the other classes at lower elevations. Rock and heath could be aggregated as a group at higher elevations. In multivariate analysis other variables could be used to refine these divisions.

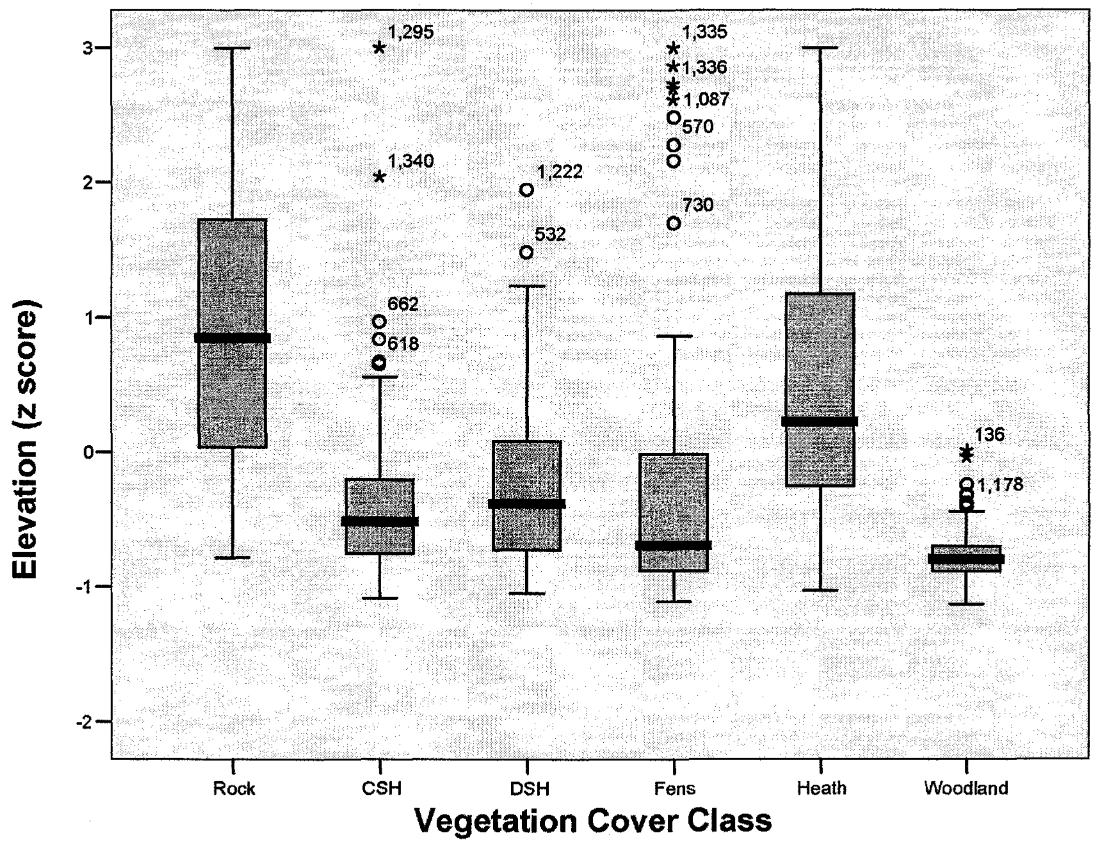

Figure 4.5 Box and whisker plot for elevation and cover classes. There are some differences between the distributions of the vegetation classes with respect to elevation but they are limited and likely to be of little use.

The results with the other variables indicate less useful relationships than were found with elevation. Taking the example of eastness (using the 100m filter results) the distribution of values for each class overlap considerably and there is very little difference in the median values (Figure 4.6). Most of the other variables show similar distributions. Some showed a certain amount of discriminating power in separating a small number of classes but nothing of particular interest was found; however this does not mean that useful multivariate relationships do not exist.

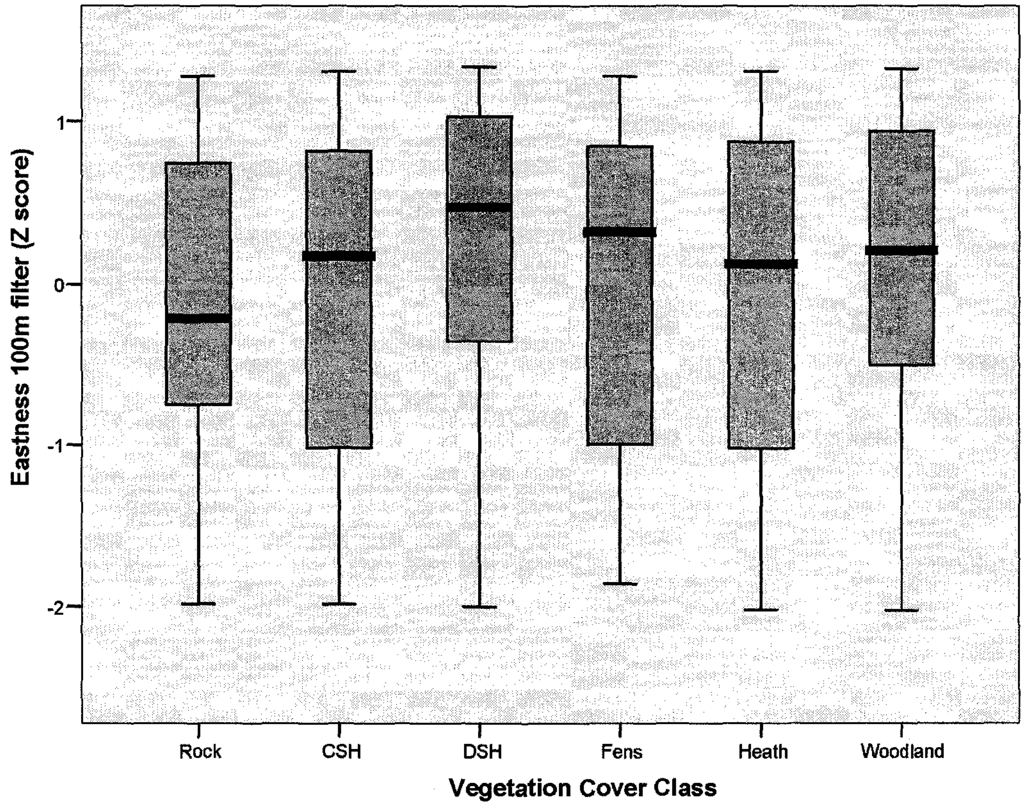

Figure 4.6 Box and whisker plot for Eastness and cover classes. There is little difference between the distributions of the vegetation classes with respect to eastness.

4.8. Spatial autocorrelation - DVs

Four methods were used to investigate spatial autocorrelation in the percentage cover by cover classes. All methods used the 200m buffer area data. Two of the methods were completed using tools in ArcToolbox and two other methods used the geostatistics analysis program GS+ from gamma design software. The first of the two methods, completed using ArcToolbox, used Moran's Index of spatial autocorrelation (Moran's I) to assess how similarity in values is related to separation between their locations. The second method also used Moran's I but in a local sense so that spatial variation in how similar or different close events are could be investigated. The analysis performed in GS+ allows a more detailed investigation of how the degree of similarity between cases changes with separation distance.

Moran's I was calculated using ArcToolbox, with inverse distance as the type of relationship and Euclidean distance as the distance method. The search radius used was one Kilometre, based on prior knowledge of the spatial relationships, and Morans's I was calculated separately for each cover class. The results from this investigation are shown in Table 4.3. All of the results show significant clustering, meaning cases that are close are likely to have similar values. Comparing the values of Moran's I for each cover class provides some useful information about the different spatial behaviour of the classes. Fens and DSH both have comparatively low levels of clustering which indicates that there is less spatial similarity in the values for this class. On the ground this translates to fens and DSH being more patchy and fragmented than the other classes. The heath and rock classes have much higher values indicating more homogenous and extensive areas with similar levels of these cover classes.

Table 4.3 Moran's I scores by cover class for one kilometre distance. The fens and DSH classes appear to be more heterogeneous than the heath, CSH and rock classes

Cover class	Moran's I	Z Score
Heath	0.65	35.1
Fens	0.33	17.7
DSH	0.37	20
CSH	0.54	29
Rock	0.7	37

The calculation of local Moran's I was performed using the Anselin Local Moran's I tool in ArcToolbox 9.1. The values for local autocorrelation were saved as a set of features for each cover class. The values of local spatial autocorrelation at each point were then mapped to investigate spatial differences in the amount of autocorrelation. Figure 4.7 shows the patterns for the heath class: they are similar to those found with the other cover classes. The most interesting pattern, which is shown in all maps, is an area with lower spatial autocorrelation at the middle elevations and two more autocorrelated areas at the lower and higher elevations. The area at the middle elevations is therefore more patchy and variable and could be thought of as a transition zone.

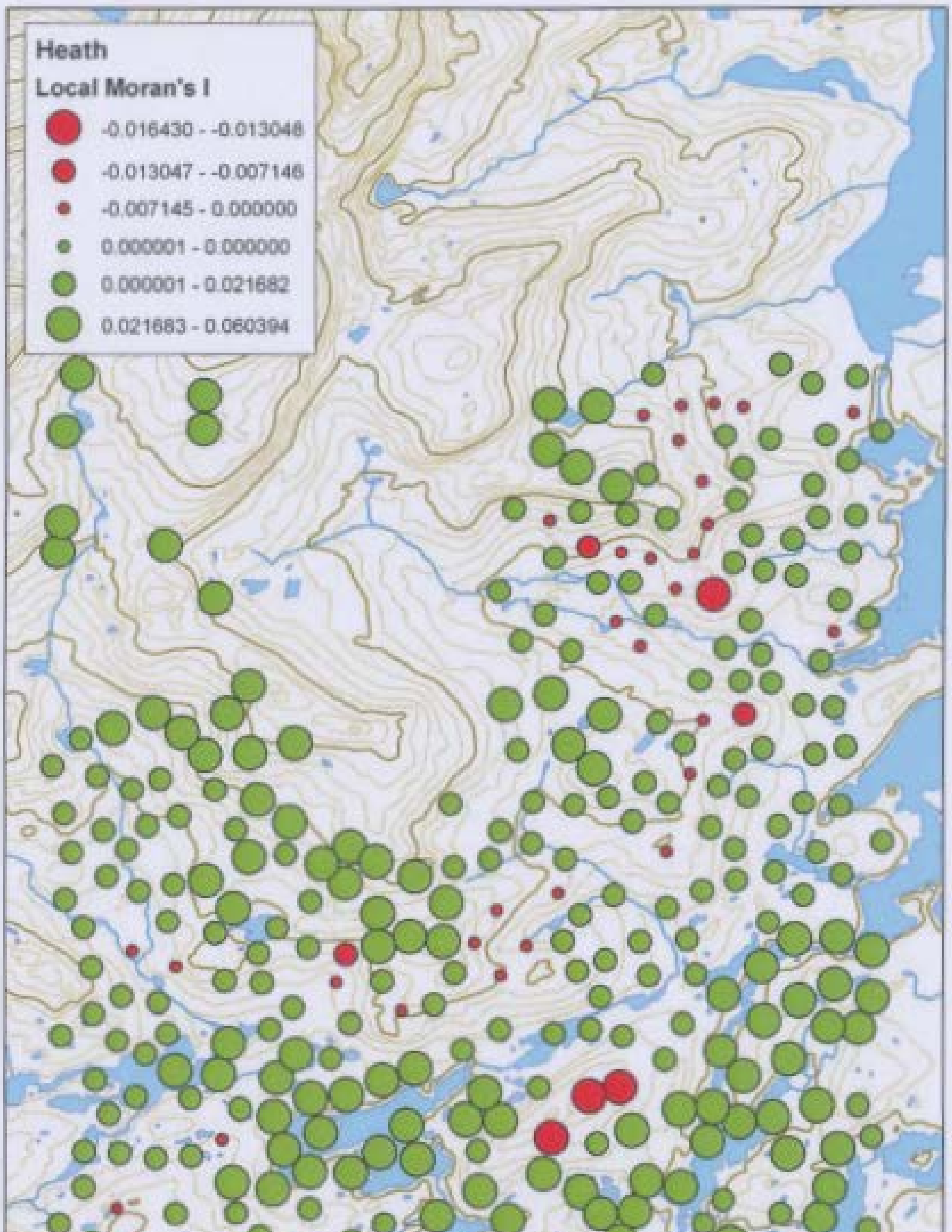

Figure 4.7 Variation in local spatial autocorrelation for the heath cover class. Positive autocorrelation is shown in green and negative autocorrelation is shown in red. Most of the sites are autocorrelated but there are patches that are negatively autocorrelated.

The geostatistical analysis performed in GS+ used both variograms and correlograms for each of the cover classes. There are two main aspects of the plots that were of use in gaining information about the nature of the data. The nugget, the predicted amount of semivariance with no separation between pairs, provides some information about how locally random distributions are. The results from this analysis should provide similar conclusions as the calculation of global Moran's I performed earlier. The second source of information is the shapes of both the variograms and correlograms, which show how the similarity between cases changes with increasing separation distance. To investigate the size of the nugget for each class and how it varies between classes the nugget values are expressed both as the semivariance (γ) and as a proportion of the total semivariance (%γ) (Table 4.4). The results from investigation of this table lead to similar conclusions as the global measure of Moran's I. The nugget is largest for the two classes that had the lower Moran's I values, these are the patchy classes. Conversely the nugget (random error) is lowest for the heath, rock and CSH classes that have high values for global Moran's I. These results therefore help to confirm the earlier findings.

Table 4.4 Local randomness in distributions. Nugget values are expressed both as the semivariance (γ) and as a proportion of the total semivariance. Again the fens and DSH classes appear more heterogeneous.

Cover Class	Nugget (γ)	Nugget (%γ)
Heath	54	10.4
Fens	15.1	26.5
DSH	129.2	32.9
CSH	36	6.3
Rock	15.1	6.5

Inspection of the shapes of the variograms and correlograms reveals a number of interesting trends. First, from the correlograms (Figure 4.8 and Figure 4.9) it appears that pairs of values become unrelated (zero autocorrelation) at around 3km (Figure 4.8 as an example). In some cases there is a notable difference between the anisotropic correlograms at 90° and at 135° (Figure 4.9). In the anisotropic correlograms at 90° the autocorrelation tends to decrease at a slower rate than at 135°. A possible explanation for this is that 90° approximately aligns perpendicular with most of the valleys (Figure 1.4) and is therefore perpendicular to the increase in elevation. However 135° aligns with the valleys and so there is a greater change in elevation between pairs of values than at 90°.

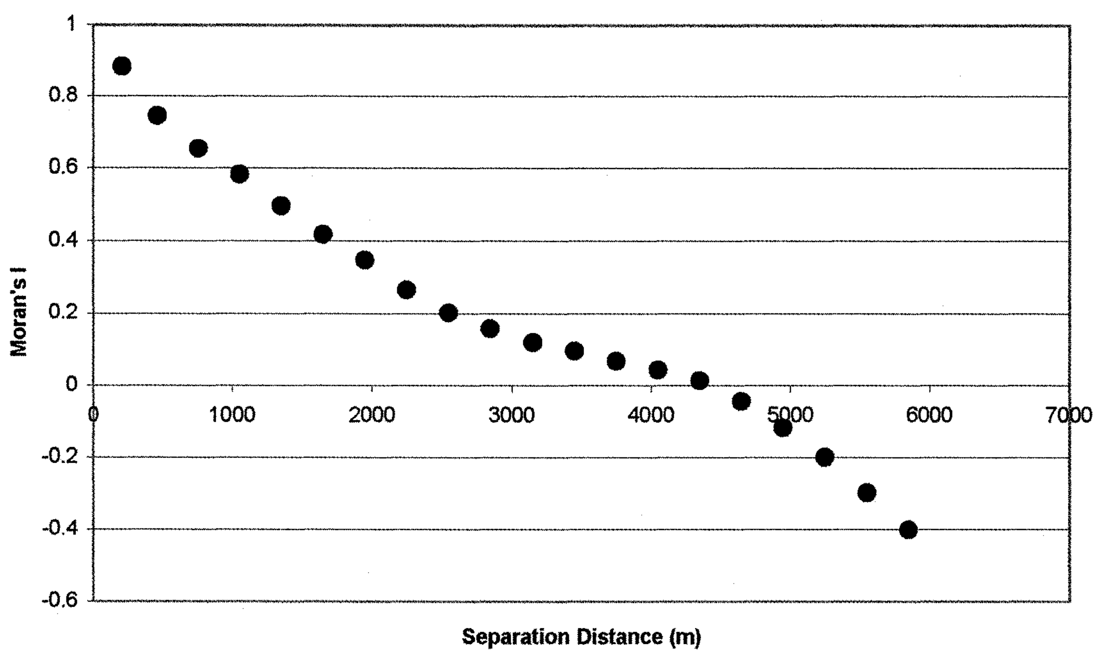

Figure 4.8 Isotropic correlogram for the CSH cover class (non-stratified 200m buffer data). Values become unrelated around 4.5 km.

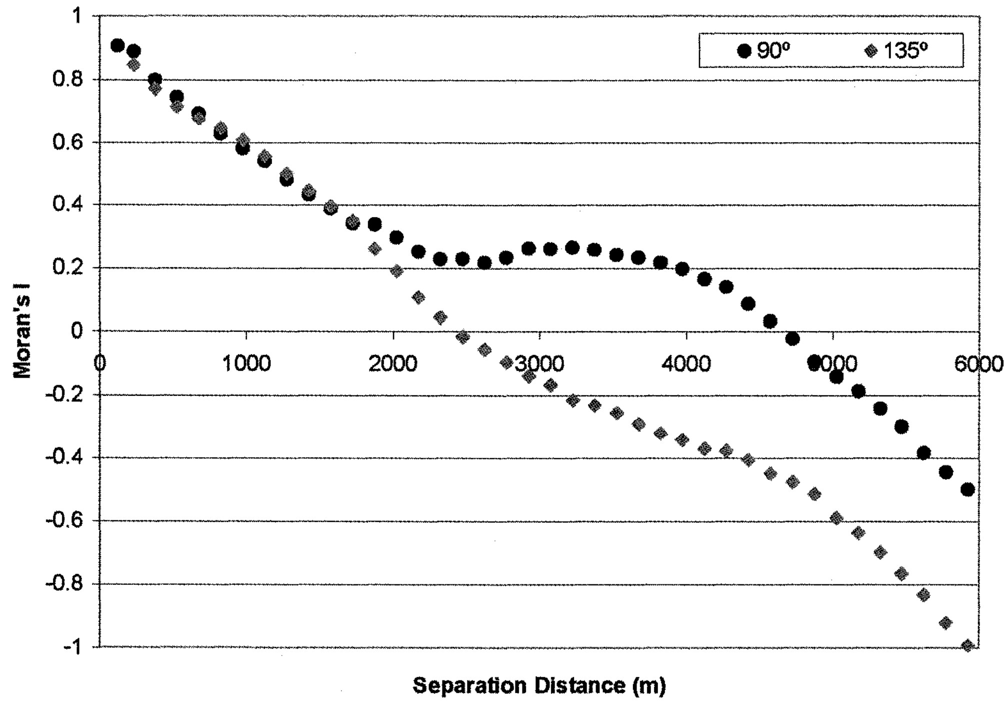

Figure 4.9 Correlograms for CSH cover class at 90 ° and 135 ° (non-stratified 200m buffer data). The autocorrelation tends to decrease at a slower rate in the correlogram at 90° than at 135°.

4.9. Summary of important findings from exploratory analysis

The large sample sizes used in this analysis increase the importance of the exploratory analysis. The significance of a test becomes misleading as the sample size increases. Subjective assessment of significance is therefore more important and these assessments are better educated with information from exploratory analysis. Spatial autocorrelation and neighbourhood effects also complicate the issue of statistical testing. From the exploratory analysis it was decided that the main data to be used for model calibration were the 200m buffer data. The buffer data were chosen because point-based information is too random and unpredictable, while the summaries within a buffer follow trends that can be generalised and used for prediction. The 200m size was chosen because it optimises the trade off between predictability and spatial resolution. The only strong bivariate relationships useful for predicting cover class included elevation and many of these relationships were nonlinear. Other variables may be of more use in multivariate situations where interactions can occur.

5. METHODS

This section details the analysis and modelling that was performed using the data discussed in section three. The exploratory work in section two was used to aid decision making regarding the selection of techniques used in this section. Some additional data sets were also generated for use in this section based on discoveries in the exploratory analysis. This section includes details of the processes used to develop models and make predictions using those models. In the final part of this section the sensitivity of the models to some of the subjective decisions made throughout the process is examined.

5.1.1. Additional data sets

Three additional data sets were created to aid the investigation of a number of potential problems. The first additional data set was created using Voronoi polygons and the second is very similar to the main buffers, except that the cases were stratified by elevation. Both of these data sets were processed and cleaned in the same way as the main data set (the non-stratified buffers). Climate data were also required to enable the models to be built based on temperature rather than elevation, which is required to facilitate predictions based on alternative climate regimes.

5.1.2. Voronoi zones

The first of these additional data sets was designed to allow an investigation of the effect of the overlap between the circular buffers. A vector based regular grid could have been used for this purpose but it would not have produced a random sample and the area of cloud would have caused significant problems. Voronoi polygons, also know as

Thiessen or Dirichlet polygons (Burrough, 1998) were used to avoid potential issues of a regular grid. The Voronoi polygons were created from a distribution of random points that were designed to result in Voronoi polygons with areas similar to the 200m radius buffers. The mean area of the Voronoi polygons is within 0.25% of the 200m buffers area. The variation in the size of the Voronoi zones is significant with a coefficient of variation close to 30%. After data cleaning there were 412 cases in the Voronoi polygons data set, which is significantly fewer than the main circular buffers data set (n= 1702). The spatial distribution of the Voronoi polygons is shown in relation to elevation in Figure 5.1. As with the non stratified buffers the distribution of cases is significantly skewed in relation to elevation because there is more ground at lower elevations (Skewness = 1.437).

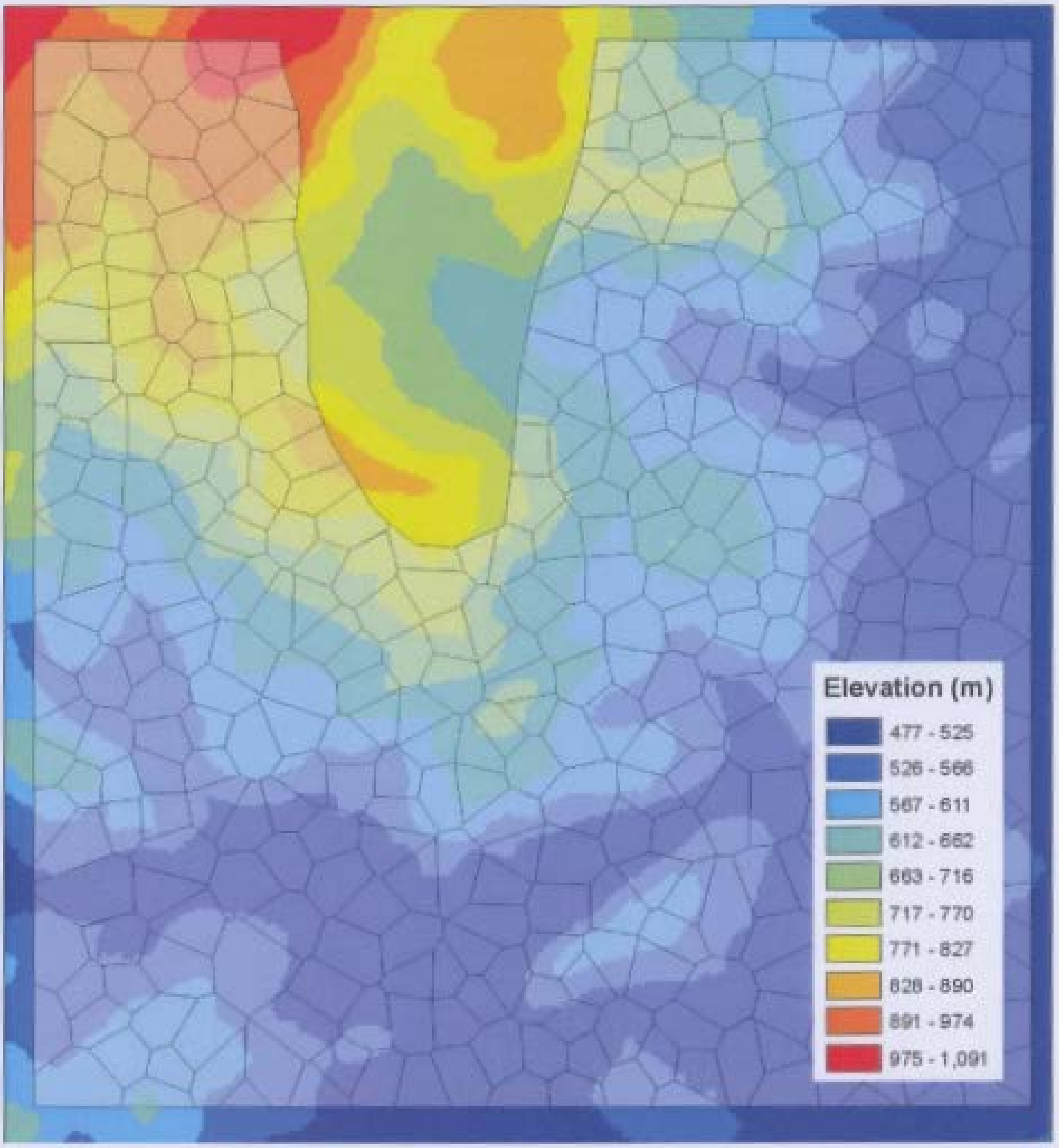

Figure 5.1. Voronoi polygons constructed to fit the study area and exclude the area of cloud.

5.1.3. Stratified buffers

The purpose of the stratified buffers data set is to enable analysis of how the skewness of elevation values affects the data and modelling process. Because elevation is the prime predictor it is worthwhile to investigate what effect the skewness of values has on model construction and robustness. The stratified buffers were created in much the same way as the main buffers data set except an attempt was made to make the distribution of elevation values as close to normal as possible in contrast to the skewed distribution of an unstratified sample. This stratification by elevation was performed by classifying the range of elevation values into discrete classes and calculating how many cases should occur in each class to achieve a normal distribution. When the points were randomly located the stratification scheme was used to determine the number of points located in each elevation class. The resulting data set has 1166 cases and a very low level of skewness of elevation values (skewness = 0.117) compared to the non-stratified buffers (skewness = 1.549). The spatial distribution of the stratified buffers does of course suffer as a result of the stratification (Figure 5.2). Because of the skewness in the landscape (more ground at lower elevations) and the area of cloud that occupies an area of higher ground the distribution is very dense at high elevations and very sparse lower down.

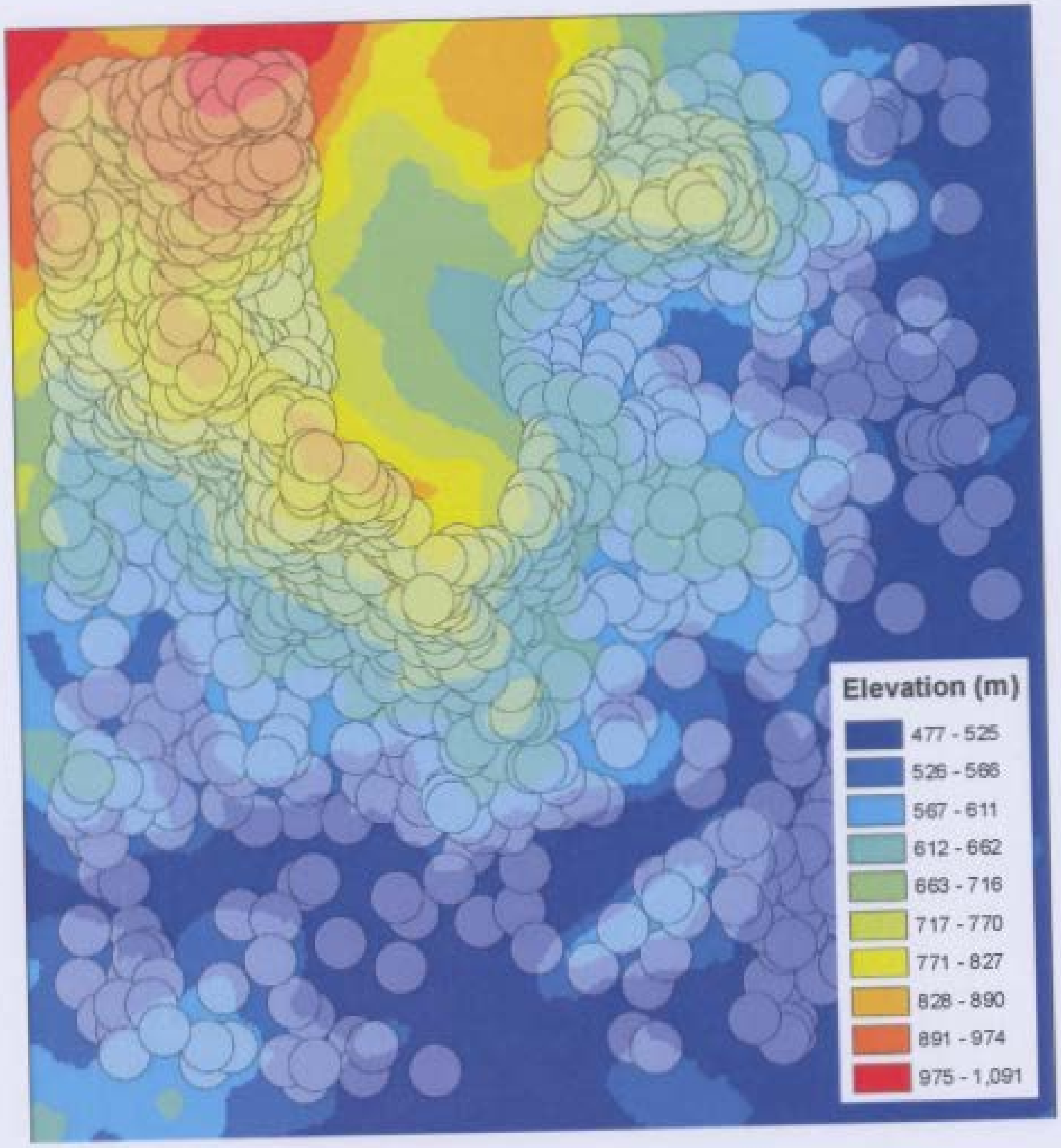

Figure 5.2. Stratified buffers spatial distribution. The distribution is much denser at higher elevations (northwest) because there is less ground at higher elevations.

5.1.4. Climate data

The generation of suitable climate information was completed by Jacobs (2006)[6] and is summarized here to illustrate the nature of the data. To make predictions of how the vegetation classes might respond to climate change the models need to be calibrated with current climate variables. Future versions of the same variables need to be available to make the future predictions. Temperature is the only climate variable that had a long enough history of records on the site to make links with other sources of climate information. The mean summer (April to October) temperature is used as it is generally considered to have the largest impact on vegetation growth (Sirois, 2000, Kirdyanov *et al.*, 2003). Temperature has been recorded at two climate stations in the study area since January 2001. The climate stations are set up at different elevations to allow the calculation of lapse rate which is used to predict the temperature at any given elevation in the study area. The records from the climate stations are useful but not extensive enough to be equivalent to the normals predicted by general circulation models which were used for future scenarios. Current normals (1971 – 2000) were generated using the short history of records from the two stations in the study area and more extensive data from climate stations in Goose Bay (Goose Bay A) and Cartwright. A statistical model was built to represent the relationships between the study area climate stations and the Goose Bay and Cartwright stations. This model was then used to generate synthetic normals for the study area. The future time periods, 2010-2039, 2040-2069, and 2070-2099 were based on published scenarios (grid based data) from the Canadian Centre for Modelling

---

[6] Personal communication, Dr. John Jacobs , Department of Geography, Memorial University of Newfoundland (2006)

and Analysis (2006) (CGCM2 model). The results of this process are summarized in Table 5.1.

Table 5.1. Estimated current normal and future scenario normal summer temperature for the upper climate station in the study area.

Period	Mean April to October temperature (°C)
1971-2000	0.8
2010-2039	1.6
2040-2069	1.8
2070-2099	3.0

Using the information from the downscaling analysis detailed above current and future temperatures were estimated based on elevation. This operation was performed using the lapse rate (rate of change with elevation) calculated using the difference between the two stations in the study area. A simple linear equation was constructed to predict temperature based on elevation, this model was then used with the predicted future temperatures to estimate the distribution of future temperatures over the study area. This method assumes the lapse rate remains the same in the future.

The predicted future temperatures should not be viewed as what the future temperature will be or even the best possible estimate; they should rather be seen as one of many possible scenarios. The impact of the estimated error in this scenario was investigated through sensitivity analysis once the predictions of future vegetation cover were complete.

5.2. Model building

In this section the processes used to develop the models to predict current and future vegetation cover are described. Before the processes are described in detail an overview of the methods used is provided. The model building will then be discussed in detail one cover class at a time.

5.2.1. Overview of the model building process

All three of the data sets (non-stratified, stratified and Voronoi) were treated for outliers and out of range values before this stage of the process. The issue of skewed dependent variables, however, remains in all three data sets and was addressed at this stage to allow a full investigation of the impact of this skewness. Table 5.2 details how skewed each of the dependent variables (DVs) is and the effect of square root and log base ten transforms. Additionally this table highlights the relationship between skewness in the DVs and the stratification of the sampling scheme. Because the vegetation classes have a different distribution their response to a sampling scheme stratified by elevation varies. Stratification increases the skewness in the CSH class DV but decreases skewness in the Rock class DV. This occurs because there is more rock at higher elevations and more CSH at lower elevations. The effect of transforming the skewed DVs was investigated by repeating the modelling process with the transforms applied and comparing the power and robustness of the resulting models with the models based on untransformed (skewed) data.

Table 5.2. Transforms for skewed dependent variables. The response to a sampling scheme stratified by elevation varies because the vegetation classes have different distributions with regard to elevation.

Variable	Skewness		
	Stratified buffers	Non-stratified buffers	Voronoi zones
CSH %	1.635	0.279	0.307
CSH log transform	-0.514	-1.285	-1.543
CSH square root transform	0.917	-0.287	-0.279
DSH %	1.457	0.843	0.945
DSH log transform	-0.683	-1.980	-1.725
DSH square root transform	0.396	-0.215	-0.195
Heath %	-0.806	0.238	0.178
Heath log transform	-1.846	-0.723	-0.983
Heath square root transform	-1.249	-0.172	-0.277
Rock %	0.482	1.571	1.514
Rock log transform	-1.801	-0.901	-0.758
Rock square root transform	-0.362	0.580	0.603

All of the models were built and tested using the same basic methodology outlined in Figure 5.3. The process starts with one of the four DVs being investigated and was repeated for each of the classes. Skewness transforms were then applied and if they resulted in a significant reduction in skewness the rest of the process was completed with the best transformed data set and the original data. In the next stage the DVs were tested for linearity with temperature. This was only carried out for temperature because preliminary analysis suggested that the only significant nonlinearity was between DVs and temperature. Where significant nonlinearity between the DV and temperature was found, curve estimation was used to correct the temperature independent variable (IV) to be linearly related to the DV. This correction of the temperature IV is only used at this preliminary stage, if the correction was found useful nonlinear modelling was used in later stages.

The next set of processes aimed to find those variables that had a significant relationship with the DV. At first, all of the variables were used in a non-sequential linear regression. The squared semipartial correlations were then used to order the variables entry into a series of non-sequential linear regressions where one more variable was entered at each stage. The change in coefficient of correlation (R^2) at each of these stages was then attributed to the addition of the variable added at that stage and used as a measure of the usefulness of that IV. The R^2s were then used to choose which IVs were included in the further stages of the modelling. In the next stage modelling was repeated with one of each of the IVs that were found to be useful omitted in each repetition to further test their importance. At this stage nonlinear parameter estimation was used if there was a nonlinear relationship between the DV and temperature.

Once the important predictors and the form of the model were identified the models were tested using a number of methods to determine their robustness. In some cases it was not clear how many predictors should be used, so two or three models were put through the testing process. The first test was bootstrapping, the sample data sets were randomly split into two sets, 70% was used for training the model and the remaining 30% was used to test the model. The difference between the R^2 for the test and training data was then used to interpret how well the model generalised the trends in the data. The other two tests were based on the model residuals. The residuals were tested for correlations with the IV and spatial clustering, both of which would indicate a model was not robust. In the final stage all the possible models for predicting a cover class were compared and the model that was found to have the best performance and robustness was chosen and used in later stages to make predictions.

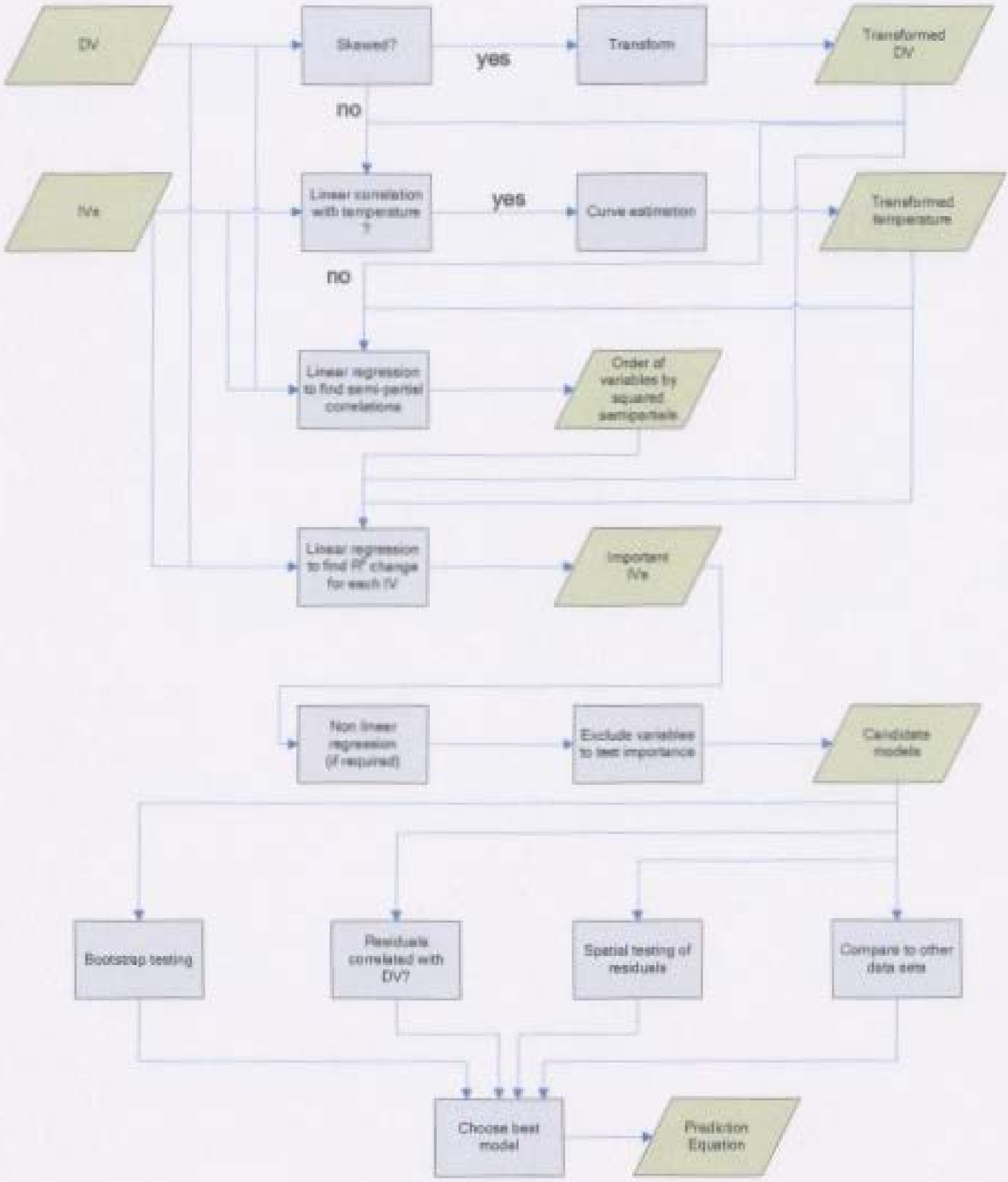

Figure 5.3 Overview of the model building process.

5.2.2. The Heath model

The level of skewness of the DV in the non-stratified and Voronoi data sets is very low (skewness = 0.178), whereas in the stratified data the skewness is much greater (skewness = -0.806) but the absolute value is still below one which would be considered significantly skewed. No further investigation of skewness was performed for this cover class as the DVs do not appear to be significantly skewed.

The percentage cover by heath shows a clear nonlinear relationship with temperature (Figure 5.4)

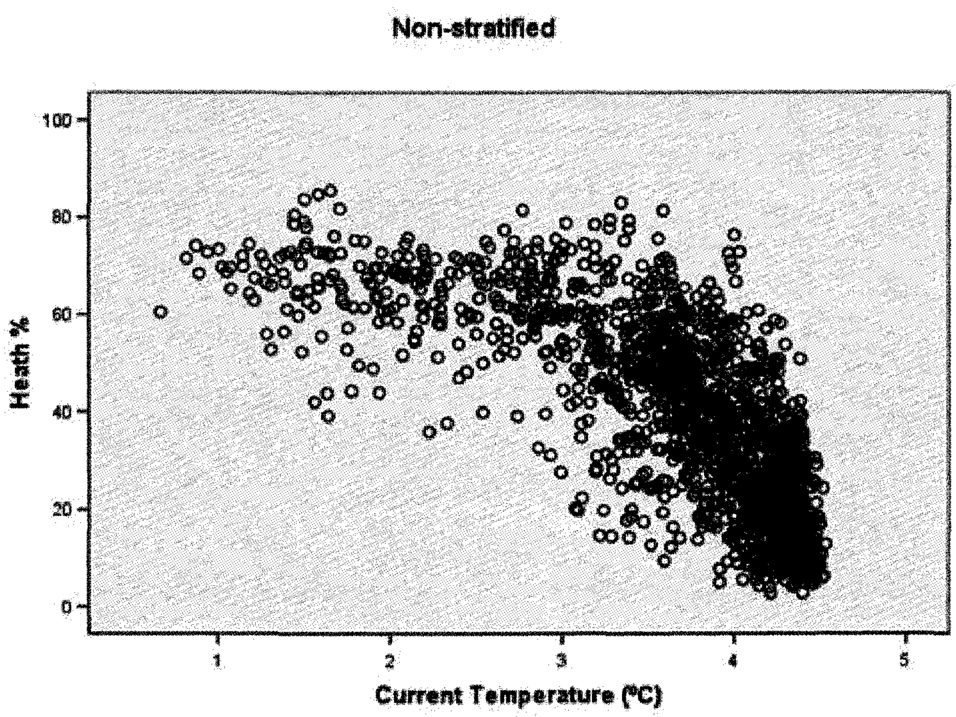

Figure 5.4 Relationship between temperature and heath percentage cover in the non-stratified data. This relationship is clearly nonlinear.

This nonlinear relationship can be represented well by a quadratic function, which accounts for over 63% of the variance in the heath variable (Table 5.3). The quadratic function has the advantage that it matches the nature of the distribution in the real world.

An exponential function for example would imply that the cover by heath kept on increasing with increasing elevation (lower temperatures) when in reality cooler temperatures limit growth.

Table 5.3 Model fits for the nonlinear relationship between heath and temperature. A quadratic function fits this relationship well accounting for over 63% of the variance in the heath variable.

	Model R^2		
Model	Stratified	Non-stratified	Voronoi
Linear	.543	.549	.564
Quadratic	.648	.632	.670
Exponential	.480	.422	.428

With the newly transformed temperature variable created the squared semi partial correlations were calculated (Table 5.4) and then used to create the order that variables were entered into the sequence of linear regressions.

Table 5.4 Squared semi partial correlations from the initial Heath model.

	Squared semi partial correlation		
Variable	Stratified	Non- stratified	Voronoi
Transformed temperature	0.266	0.378	0.307
Normalized slope	0.008	0.033	0.011
NE angle exposure	0.002	0.007	0.001
Moisture potential index	0.001	0.007	0.006
Solar radiation	0.012	0.002	0.000
SW angle exposure	0.000	0.001	0.006
Northness	0.004	0.000	0.001
Curvature	-	0.000	-
Smoothed curvature	-	0.000	0.001
Snow potential index	0.000	0.000	0.002
Eastness		0.000	0.001
SW wind exposure	0.000	-	-

From the R^2 changes (Table 5.5) temperature is clearly a strong predictor while a few other variables (normalized slope, NE angle exposure, solar radiation, snow potential index) can explain a small amount of the variance. The variables that explain small amounts of variance do not appear to be significant at this stage, however their full power was tested in the next stage where nonlinear parameter estimation was used and the influence of each variable was tested by removing it from the model.

Table 5.5 R^2 changes for heath models. Temperature is strong predictor while some other variables explain a small amount of variance.

	R^2 Change		
Variable	Stratified	Non-stratified	Voronoi
Transformed temperature	0.648	0.632	0.564
Normalized slope	0.018	0.033	0.003
NE angle exposure	0.000	0.029	-
Moisture potential index	0.002	0.015	0.004
Eastness, Snow potential index	-	0.002	-
Solar radiation	0.017	0.002	0.001
Northness	0.003	0.001	0.000
Curvature	-	0.000	-
SW angle exposure	-	0.000	-
SW wind exposure	0.005	-	-
Snow potential index	0.000	-	0.021
Eastness	-	-	0.003
SW angle exposure, Smooth curvature	0.000	-	0.001

The nonlinear modelling is summarised separately for each data set due to the differences in which variables statistically contribute to the explained variance. For the non-stratified buffers data set the addition of three linear predictors, normalized slope, NE angle exposure and moisture potential index were tested and the increase in R2 was used to examine the contribution of each linear predictor. The resulting models are presented in

Table 5.6. An R2 increase of 0.087 can be gained by adding all three linear predictors, though this is probably not the best model since the added power of the model is probably not justified by the increase in complexity. The model that includes NE angle exposure and Normalized slope adds 0.069 to the quadratic function and probably achieves the best balance between prediction power and simplicity. This model was then used in testing of robustness. Residuals from the model have a weak correlation (R2 = 0.29%) with the DV, which is a slight concern but not a strong indication of weakness in the model. The bootstrapping test also indicates the model is robust with a R2 difference of only 0.03 between the test and training data. This suggests the model generalizes well and does not over fit to the training data.

Table 5.6 Influence of predictors in nonlinear modelling using the non-stratified buffers data set. The model that includes NE angle exposure and Normalized slope adds 6.9% to the quadratic function and probably achieves the best balance between prediction power and simplicity.

Predictors	R^2	R^2 increase
Current temperature (quadratic)	.632	0%
Current temperature (quadratic) Normalized slope (linear) NE angle exposure (linear)	.701	6.9%
Current temperature (quadratic) Normalized slope (linear) Moisture potential index (linear)	.675	4.3%
Current temperature (quadratic) NE angle exposure (linear) Moisture potential index (linear)	.655	2.3%
Current temperature (quadratic) Normalized slope (linear)	.668	3.6%
Current temperature (quadratic) Moisture potential index (linear)	.638	0.6%
Current temperature (quadratic) NE angle exposure (linear)	.651	1.9%
Current temperature (quadratic) NE angle exposure (linear) Normalized slope (linear)	.701	6.9%
Current temperature (quadratic) NE angle exposure (linear) Normalized slope (linear) Moisture potential index (linear)	.719	8.7%

In the nonlinear model testing for the stratified data two linear predictors were tested, NE angle exposure and Normalized slope. The R^2 increase due to the linear predictors is not large (Solar radiation = 1.7% and Normalized slope 1.6%) and the most significant gain is achieved through adding both variables (4.2%) (Table 5.7). Because of the questionable value of adding both variables further testing was performed on the basic model (only the quadratic function) and an extended model (solar radiation and normalized slope included as linear predictors).

When the model residuals were tested for correlation with the IV, both models had R^2 values slightly higher than the non-stratified zones model. Both models also performed well in the bootstrapping test, which showed little difference between the two models.

Table 5.7 Influence of predictors in nonlinear modelling using the stratified buffers data set. The R^2 increase due to the linear predictors is not large.

Predictors	R^2	R^2 increase
Current temperature (quadratic)	.648	0
Current temperature (quadratic) Solar radiation (linear) Normalized slope (linear)	.690	4.2%
Current temperature (quadratic) Solar radiation (linear)	.665	1.7%
Current temperature (quadratic) Normalized slope (linear)	.664	1.6%

The nonlinear modelling for the Voronoi data results in models with similar power and robustness. The only important difference with the Voronoi zones is that none of the linear predictors were found to add enough explanation of variance to warrant their inclusion. When comparing models generated from all three data sets the differences are not large and choosing the best model is not straight forward. It appears that the non-stratified zones model is the best and the inclusion of linear predictors may be warranted but the effects should be monitored.

5.2.3. The CSH model

As with the heath cover class the CSH DV is not significantly skewed in the non-stratified or Voronoi data sets but is more skewed in the stratified data (skewness = 1.64). Square root and $\log^{10}$ transforms were both tested but neither resulted in better models and so the following discussion includes untransformed models only.

The relationship between CSH and temperature is nonlinear in all three data sets, the shape of this relationship is illustrated using the non-stratified data as an example (Figure 5.5). A quadratic function provides the closest fit though the difference between the quadratic and exponential functions is only significant in the stratified data set. A quadratic function may perform best for predicting current conditions but could give very unreal predictions for future conditions. Using the quadratic function the percentage cover by CSH would begin to rise with temperatures decreasing below two degrees Celsius and this does not match the real world behaviour of the cover class. The exponential function provides a good fit that matches the real world behaviour of the cover class and so was used in the rest of the model construction process.

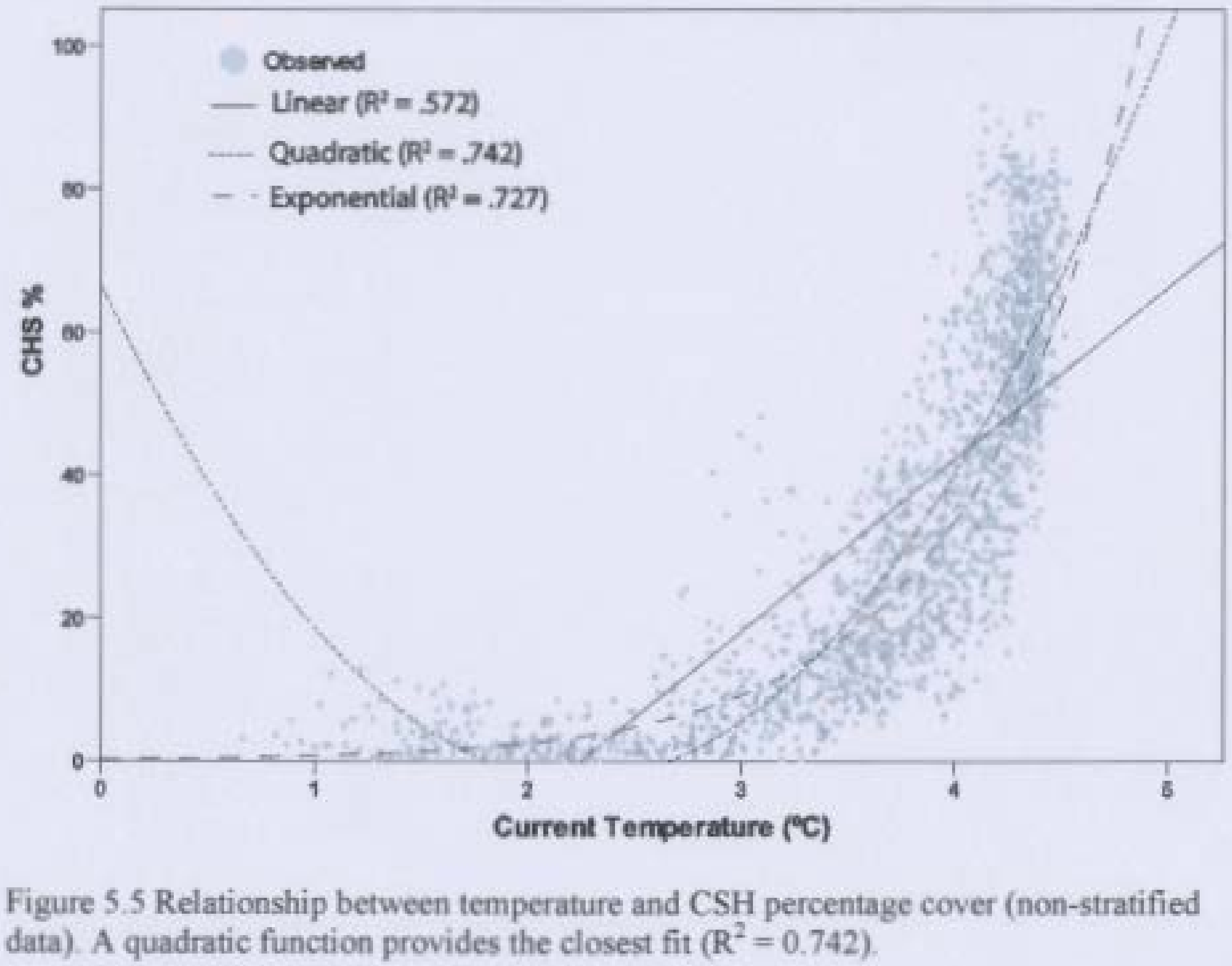

Figure 5.5 Relationship between temperature and CSH percentage cover (non-stratified data). A quadratic function provides the closest fit ($R^2 = 0.742$).

Once the squared semipartial correlations for prediction of the CSH DV were calculated and used to create the order of variables entry into the series of linear regressions the R^2 change attributed to each variable could then be investigated to find those potentially useful variables. The prediction power of the transformed temperature variable is very high (Table 5.8) but only one other variable, solar radiation, shows significant potential as a predictor. Though solar radiation does not add a large amount of explanation it is significant for all three data sets, which suggests there is a real relationship with the DV and not just an artefact of the regression model.

Table 5.8 R^2 changes for CSH models. Temperature is a very useful predictor but the only other variable that has potential to be of use is solar radiation.

	R^2 Change		
Variable(s)	Stratified	Non-stratified	Voronoi
Transformed Temperature	0.849	0.756	0.785
Solar Radiation	0.043	0.080	0.060
Smoothed curvature	0.003	0.001	0.000
Moisture Potential Index, NE angle exposure	0.002	-	-
Normalized Slope	0.001	0.003	0.000
SW angle exposure	0.001	0.014	0.009
Eastness	0.001	-	0.000
Snow potential index	0.001	0.000	-
Northness	0.000	-	-
Moisture Potential Index	-	0.000	0.004
NE angle exposure	-	0.000	0.004
Northness, Snow potential index	-	-	0.000
Northness, Eastness	-	0.000	-

The importance of solar radiation was then tested in nonlinear modelling. Because there is only one linear predictor under investigation only two nonlinear models were required, with and without solar radiation included. This stage of testing produces very similar results to the linear modelling with transformed temperature variables, this is not a surprise as both methods result in essentially the same type of model. In the non-stratified data sets the added explanation provided by solar radiation appears to be significant and justified. In the stratified data set the variables' inclusion is questionable due to its limited addition to the level of prediction. Further testing of the models was performed with solar radiation included.

Testing of the residuals from the three models suggested all models were robust with very low correlations ($R^2 < 0.17$) between the DV and the residuals. The maximum variation between the parameter estimates from each model are shown in Table 5.9 where the difference is represented as a percentage of the parameter average. All but parameter b2 show variation above 50%, this is mainly due to the difference between the stratified and non-stratified data sets. The variation between the non-stratified zones and Voronoi zones are less. This level of variation means that at the prediction stage it is worth investigating the variability between the models to check how much the difference in the parameters created by the alternative sampling schemes effects the final prediction.

Table 5.9 Parameter differences from the CSH models. where the model takes the form CSH = b0 + (b1 * (Exp (b2 * [TEMP]))) + (b3 * [SOLAR]).

Parameter	Max difference
b0	60.1%
b1	54.2%
b2	7.5%
b3	60.6%

The model derived from the stratified data appears to be the best model as all are robust and the stratified data model explains the most variance. However this model may not be ideal due to the relationship between the sampling scheme and the distribution of the cover class. The stratified data have a lower proportion of samples at lower elevations, which is where values for the cover class are higher. This means that the model could be appearing to perform better just by predicting lower values. To check that the increased performance was the result of a better model and not an artefact of the modelling process the model was tested against a new independent non-stratified data set. A non-stratified data set is used to better represent the actual prediction the model will be making later on. When the stratified model was compared to the validation data the R^2 was slightly lower, reduced from 0.898 to 0.826, which is less than the R^2 for the non-stratified model both from training and testing with the same validation data set. The differences are not great but they do suggest that the increased performance of the model made when using the stratified data comes at the cost of applicability to its end use of making predictors for the study area. Both of the models show a tendency to underestimate higher values leading to weak correlations between the model residuals (when tested against the validation data) and the DV. The mean value predicted by the stratified model is also slightly lower (1.5% lower) but this difference is not of real significance. It therefore appears that there are no real differences between the models and any perceived gains in using the stratified sample data are lost when the model is applied to the actual study area. It is therefore suggested that the non-stratified model is the best choice since the sample data are not weakened by the large amount of overlapping between buffers seen in the stratified data and the models power is not overestimated. The impact of the choice of sampling scheme will be investigated later by comparing the resulting distributions from both models.

5.2.4. The rock cover class model

Rock is included as a cover class because areas of rock can both increase and decrease in the same way that areas of vegetation can. Rock coverage can decrease due to colonisation by vegetation or increase due to death of vegetation and soil erosion. The resulting model from the building process did not explain enough variance in the distribution of the rock cover class to warrant further investigations. The poor performance of the models in predicting the rock cover class is illustrated in Table 5.10. The most noticeable weakness is the poor performance of current temperature, which is a strong predictor for other cover classes (Heath and CSH) but for this class doesn't account for more than 50% of the variance.

Table 5.10 Model performances for the rock cover class. Temperature, which is normally a strong predictor explains less than 50% of the variance while a few other variables show a small amount of predictive potential.

	R^2 Change		
Variable	Stratified	Non-stratified	Voronoi
Current temperature (°C)	0.465	0.483	0.485
Snow potential index	0.042	0.007	0.006
Northness	0.014	0.000	0.004
Eastness	0.014	0.015	0.039
Normalised Slope	0.010	0.024	0.003
NE angle exposure	0.004	0.001	0.000
Moisture Potential Index	0.003	0.044	0.010
SW angle exposure, Solar Radiation	0.003	0.001	0.003
Smoothed curvature	0.002	0.004	0.037

Some of the problems with the data for this cover class are discussed to illustrate why an effective model was not developed. One of the biggest issues with the rock cover class is the high degree of skewness, the levels of skewness in the original data sets and after transforms were applied is detailed in Table 5.2. Three transforms, Square root, inverse and $\log^{10}$ were tested and removal of any cases with values below five percent was also tested. The stratification of the sampling and some of the transforms improve the level of skewness but the modelling process still failed to explain enough of the variance to make any of the models worthwhile.

Table 5.11 Skewness and transforms for the rock cover class.

		Stratified	Non-stratified	Voronoi
Original data	All	.482	1.57	1.51
	>5%	.591	1.18	.997
Log^{10} transform	All	-1.80	-.901	-.758
	>5%	-.528	.135	.081
Inverse transform	All	17.8	8.85	5.68
	>5%	1.53	.656	.687
Square root transform	All	-.362	.580	.603
	>5%	.006	.609	.518

A further problem with the rock cover class is that the relationship with the DV and the strongest predictor, temperature, is not easily modeled. This relationship is illustrated in Figure 5.6 using the stratified data set as an example. The relationship appears approximately linear for cases warmer than two degrees Celsius but cases colder than that appear to have a very different relationship. The change between the two sections is too sharp to be modeled by an exponential or quadratic function and there are not enough cases below two degrees Celsius to build two separate relationships. An R^2 of .643 can be achieved by modelling the cases above two degrees Celsius with a linear function and representing cases below .643 with the mean value for those cases. This piecewise model is adequate for modelling the current distribution but does not allow for predictions using alternate temperatures and so is not sufficient for the purposes of this study. The rock cover class is not of prime interest to this study so more complex methods of dealing with this difficult relationship are not warranted.

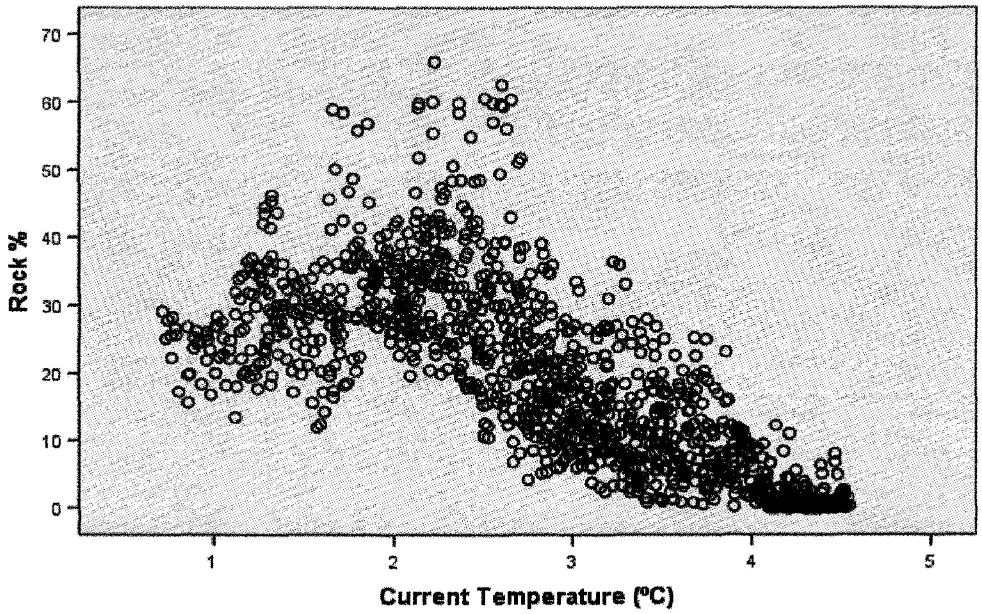

Figure 5.6 The relationship between the rock cover class and current temperature (stratified data). The relationship appears approximately linear for cases warmer than two degrees Celsius but cases colder than that appear to have a very different relationship.

5.2.5. <u>The DSH Model</u>

With untransformed data the modelling process did not result in any models that performed well enough to be developed any further. When a $\log^{10}$ transform was used to correct the high degree of skewness in the DSH DV (stratified data set,) a viable model was developed. A quadratic function using temperature as a predictor accounted for a large amount of the variance in the $\log^{10}$ transformed DSH DV and a few other IVs also added to the model. The set of nonlinear models is shown in Table 5.12; on its own temperature accounts for 73.5% of the variance and an extra 5.2% can be gained by adding northness and normalized slope as linear predictors. The model also appears robust, though there is a slight correlation (R^2= 0.275) between the DV and the residuals. Because only one usable DSH model was developed no comparisons between the resulting models from the different sample strategies were completed.

Table 5.12 Nonlinear models for the DSH DV using the stratified sample data. On its own temperature accounts for 73.5% of the variance and an extra 5.2% can be gained by adding northness and normalized slope as linear predictors.

Model	R^2	R^2 increase
Current temperature (quadratic)	.735	0
Current temperature (quadratic) + normalized slope (linear)	.770	3.5%
Current temperature (quadratic) + Northness (linear)	.766	3.1%
Current temperature (quadratic) + Northness (linear) + normalized slope (linear)	.787	5.2%

5.3. Making predictions

This section discusses how the regression models developed in the previous section are used to make spatially continuous predictions for future time periods with different mean summer temperatures.

The desired output format of the predictions is a spatially continuous estimate of future percentage cover by each cover class. The models were developed using point-based data, the points represented the centre of zones used to generate the information, but the models can be applied to other data types. The data format chosen for the predictions was raster, because the data structure provides the spatial continuity desired and the predictions can be made easily using the raster based inputs (topoclimatic variables). The raster calculator in ArcMap was chosen to make the raster based predictions because of it's simplicity for scripting and ease of integration with the existing data. Before the predictions could be completed the resolution of the outputs had to be chosen. The resolution had to be consistent with the model building process and the projects goals. Since the models were built using data derived from 200m radius circles it could be argued that the resolution should be 354.4m, which is the equivalent area for a square pixel. This resolution is however very coarse compared to the size of the study area and would hide a certain amount of spatial variation in the resulting predictions. The option preferred in this study is to treat the predictions in a similar way to the points used to build the models; that is that the value at the point is a summary for the area surrounding it. Therefore a finer resolution can be used which allows better visualisation of the spatial trends in the predictions. A 20m cell size was chosen based on the resolution of the original data and the extent of the study area (approx $63km^2$). To ensure the inputs to the predictions were consistent with the model calibration, the topoclimatic layers were smoothed using a 200m radius mean filter. The result of the mean filter is that the value for each becomes equivalent to the mean value at the points in the data used to build the model. The final step before making the predictions was to generate the temperature layers in raster format to use in the predictions; this was done using the simple linear models of the relationship between temperature and elevation discussed earlier. The simple linear equations were applied to a digital elevation model in raster calculator.

The predictions were calculated in ArcMap's raster calculator using a script to allow the process to be repeated with modified inputs. The processes used in the script are detailed in Figure 5.7. The process was repeated for each time period and for each cover class as well as for different version of the models (basic and extended). The initial stage in the process is the application of the nonlinear model to the raster input DVs, this produces a prediction but the values are not limited to the possible values as the equation can predict values below zero percent and over 100%. In the following steps two masks are used to restrict the data values to the zero to 100 range. After the corrected predictions were created, the scripts performed calculations to compare the distributions in different years and to calculate predicted change.

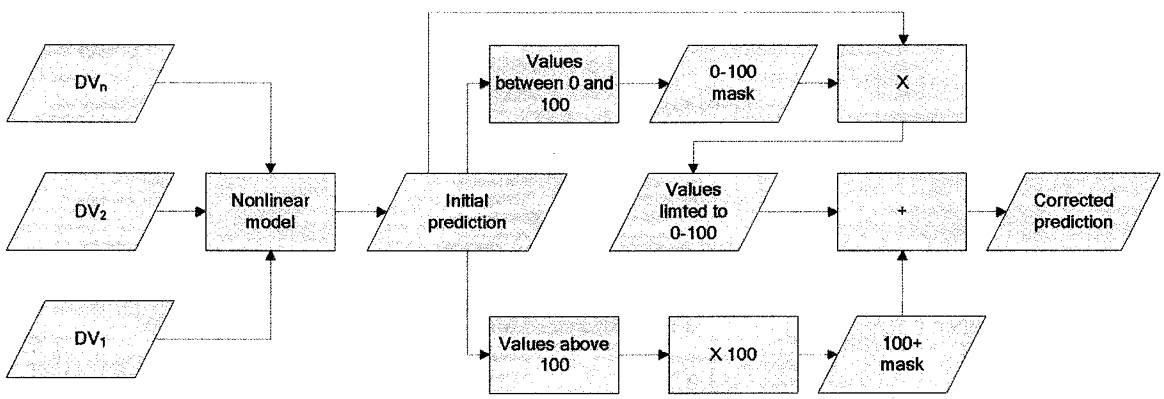

Figure 5.7 Steps used in predictions scripting.

When calculating the predicted changes there are a number of possible ways of generating the baseline (present) data. The options considered here are using the prediction based on the current temperature or an interpolation of point-based data. The interpolated baseline percentage cover raster layers were created from the non-stratified buffers data set because the sample coverage density is more even. Block kriging was used to perform the interpolation and the output was then tested against a validation data set. The performance of the interpolation in comparison to the predictive models is summarized in which also serves as validation for the predictive models when applied to the actual study area as a whole. The validation is based on a new and independent data set (from the training data) containing 892 cases created in the same way as the non-stratified data set. The effect of smoothing the input topoclimatic variables on the predictive models is also tested through running the predictions with smoothed inputs and with unsmoothed inputs. Interpolation is clearly the best method of generating the baseline data as the validation R^2 is above 90% for all three cover classes and is consistently higher than validation R^2 for the predictive models. One of the most noticeable features of the validation table (Table 5.13) is the very poor performance of the DSH predictive model, the maximum R^2 reached is 0.325 which is not good enough to warrant use of the models'. The difference between the performance for the training data and the validation is probably due to an overestimation of the models' performance in the building stage. The overestimation of the models' performance is likely due to the manipulation of the data, stratified sampling, $\log^{10}$ transform and the quadratic function. From the residual statistics (Table 5.13) it can be seen that the mean residuals are further from zero than they should be, the negative numbers indicate the models tend to under predict. The estimated power of the model is therefore likely an artefact of the process used to create it and the model should not be use to make any predictions. The DSH model will therefore not be used any further. The use of smoothed or non smoothed inputs shows only a slight influence on the performance of the predictive models, the validation R^2 are mostly lower by around two percent which indicates no real consequence. The lack of difference between the two treatments is likely due to the high level of spatial autocorrelation in the landscape, the mean at a location for the surrounding 200m is not substantially different from the value for that location. The CSH and Heath models perform well with validation R^2 high enough to use the models for prediction using future temperature scenarios. The drop in validation R^2 when compared to the training R^2 is small for all models, all differences were less than two percent, this reaffirms that the models are robust and not over-fitted.

Table 5.13 Comparison of potential methods for creating baseline percentage cover data and validation of predictive model performance. Interpolation is clearly the best method of generating the baseline data. The DSH model performs very badly when tested against the validation data set while the CSH and Heath models perform well with validation R^2 high enough to use the models for prediction.

			Validation R^2	Residual Statistics			
				Std. Deviation	Minimum	Maximum	Mean
CSH	Interpolation		0.925	6.56	-42.8	25.6	-1.15
	Smoothed inputs	Basic model	.722	11.9	-43.1	34.0	.377
		Extended model	.730	9.36	-37.0	26.4	.0549
	Non Smoothed inputs	Basic model	0.716	12.2	-44.5	45.4	.609
		Extended model	0.666	10.9	-34.0	42.3	.795
Heath	Interpolation		0.927	5.36	-23.15	16.2	-.1615
	Smoothed inputs	Basic model	0.617	12.0	-41.3	39.8	.0113
		Extended model	0.685	10.9	-32.4	32	-.579
	Non Smoothed inputs	Basic model	0.619	11.9	-39.7	40.4	-.0724
		Extended model	0.646	11.7	-37.5	34.7	-.652
DSH	Interpolation		0.979	2.14	-9.6	13.33	.112
	Smoothed inputs	Basic model	0.112	13.6	-62.1	19.9	-3.26
		Extended model	0.325	12.5	-41.2	67.9	-1.99
	Non Smoothed inputs	Basic model	0.117	13.6	-62.0	19.9	-3.26
		Extended model	0.219	17.8	-43.95	93.0	.342

All of the models were also tested at this stage to see if the residuals were spatially autocorrelated. The residuals from all of the models showed some autocorrelation with Moran's I scores (using a threshold distance of one kilometre) around 0.4 for the CSH and heath models, the DSH models were higher at around 0.5. The differences between the model variants for each cover class were very low, below 0.06, suggesting the addition of the linear predictors has little effect on the spatial clustering of residuals. The level of spatial autocorrelation in the residuals is acceptable since the values show some clustering but not a strong tendency towards clustering. A certain amount of clustering is to be expected as the models generalize the relationships in the data and should not fit to areas of unusual conditions.

5.4. Sensitivity analysis

5.4.1. Basic and extended models

Up to this point some issues regarding which form of models to use remain partially unresolved. Because of the small differences between models choices between basic or extended models and stratified or non-stratified data are not clear cut. For this reason sensitivity analysis was performed to assess what influence these decisions have on the end results. To assess the difference between the basic and extended models, variables were constructed that represented the difference in predicted changes. In order to maintain the direction of the differences the predicted change value for the basic model was simply subtracted from the extended model, as opposed to using squared differences. These difference variables were then classified based on standard deviations (using the variable with the greatest range of values) to allow a comparison of the relative areas over which the different degrees of differences occur. The results of the analysis show very little difference between the results from the basic and extended models. For the CSH cover class the standard deviation classes were based on the 2010-2039 time period, the resulting distribution of area within each of the classes is shown in Figure 5.8. The limits on the first standard deviation are very small (-3.26 to 2.11) showing that for most of the data there is very little difference between the predicted change from the basic or extended models. The distributions also become more peaked in the later time period as the area within the first classes (-3.26 to 2.11) increases, this means that not only are the differences very small but over time they become less important. The trends for the heath cover class are very similar though the range covered by the first standard deviation class is slightly smaller (-2.78 to 1.28).

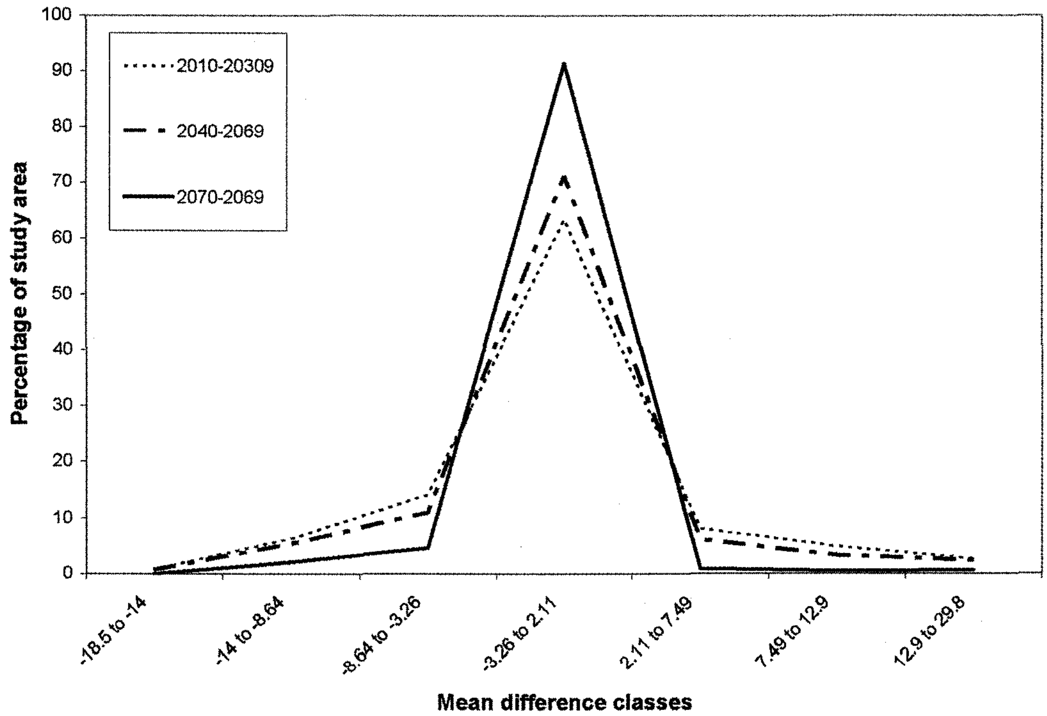

Figure 5.8 Area occupied by mean difference classes for the CSH cover class. For most of the data there is very little difference between the predicted change from the basic or extended models. These changes decrease in the later time periods.

5.4.2. Sampling schemes

The sensitivity to selection of the sampling scheme, stratified or not, was investigated through the use of the standard deviation of the differences between predictions from equivalent models made with different sample data (stratified and non-stratified). The values in Table 5.14 clearly show that the selection of the sampling scheme has very little influence on the results and their interpretation. As with the differences between basic and extended models the sensitivity decreases in the later time periods. For the majority of the study area (± one standard deviation of differences) in all time periods the differences between the models was less than the error in the models.

Table 5.14 Standard deviation of differences between predicted change based on stratified and non-stratified data (units are percentage cover). The selection of the sampling scheme has very little influence on the results and their interpretation.

Time period	Heath	CSH
Present	3.69	2.68
2010-2039	3.60	2.42
2040-2069	3.45	1.22
2070-2099	1.68	3.41

5.4.3. Climate Scenarios

A further source of great uncertainty in the predictions is the error in the future climate scenarios. The scenarios provided for this study do not include any estimation of the level of uncertainty so hypothetical uncertainty scenarios were used instead. The sensitivity to three levels of uncertainty, five percent, 30% and 50% was tested by analyzing the level of differences this uncertainty would cause. To perform this analysis the predictions were repeated with the temperature increased and decreased by the levels of uncertainty. The difference between the predictions with increased temperatures and those with decreased temperature were then calculated to provide the range of values under each uncertainty scenario. This method was used because of its simplicity. Because the models are nonlinear a single range cannot be computed as the levels of difference vary across the site. The root mean squared differences (RMSD) between values generated by this analysis are detailed in Table 5.15. The five percent error scenario has very little effect on the predictions, all the RMSDs are below six percent which is small relative to the levels of change that occur in the predictions. In the 20% error scenario the range generated is still low for the heath cover class but much higher for CSH. When the level of uncertainty in the 20% error scenario is combined with the other main error sources in the data the change in percentage cover has to be over 21% (2010-2039) to be significant, this makes the change for nearly a quarter of the study area non-significant (within the error of the model). The 50% error scenario obviously has a large impact for example the change in cover percentage for nearly 40% of the study area for the 2010 to 2039 time period is no longer significant with that level of error in the temperature estimates.

Table 5.15 Root mean squared differences for the range of values generated by each error scenario. The 50% error scenario has a very large effect on the predictions and at this level of uncertainty the results of the models are of little value.

		Error scenarios		
		5%	20%	50%
Heath	2010-2039	2.18	8.45	19.62
	2040-2069	2.47	9.34	18.17
	2070-2099	2.92	10.61	22.33
CSH	2010-2039	4.55	16.74	35.55
	2040-2069	4.90	17.54	36.84
	2070-2099	5.71	19.16	37.19

5.5. Neighbourhood effects

The problem of neighbourhood effects in regression analysis is best explained through the requirement for random independent samples by most regression models. Depending on the sampling scheme geographic data sets can be random but the requirement for independence will not be fulfilled if neighbourhood effects are present (Odland, 1988). If neighbourhood effects are present each case will have some dependence on near cases and the assumption of independence will be violated.

In this study each sample is not independent since processes such as seed spread and insect infestation mean that what grows in one location will have an influence on what grows nearby. Additionally the fact that the 200m buffer zones included some overlap will mean that sites within 400m of one another will have some interdependence due to shared ground.

The issues that need to be resolved in this study are not whether there are any neighbourhood effects but rather if they are important in the models being used and if so how to improve the models to account for them. The first basic test for problems with missed neighbourhood effects is to check the residuals for spatial autocorrelation (Odland, 1988). If the residuals are spatially autocorrelated then the model is not performing properly, this could be due to a missing variable or incorrect functional form of the model. This basic test was performed for the two main models in this analysis. The two models predict percentage cover for the CSH and heath classes. The residuals were calculated using a validation data set that was sampled separately from the data used to build the models. Each model has two forms, a basic and extended form. All of the four models were tested to see if the residuals were autocorrelated using global Moran's I as a measure of spatial autocorrelation. The Moran's I scores were all low and the scores for the extended models were below 0.37 (Table 5.16).

Table 5.16 Moran's I for residuals from the main models using a threshold distance of one kilometre. The values all show some autocorrelation in the residuals but not enough to cause concern.

Model form	CSH	Heath
Basic	0.354	0.418
Extended	0.366	0.369

A further way of examining the autocorrelation of residuals is through correlograms, which allow the visualisation of how the level of autocorrelation changes depending on separation distances. Figure 5.9 and Figure 5.10 show correlograms of the residuals from the extended models for CSH and heath respectively. With both models the level of autocorrelation drops to near zero by 1125m to 1500m separation distance. In the CSH model the drop is quicker and the decrease in autocorrelation flattens around 400m. The drop to near zero by 400m is interesting since this is the minimum separation distance where the 200m buffer zones no longer overlap. It is possible that autocorrelation of residuals within that 400m is at least partly due to the samples sharing some of the same data. This suggests the overlap of the buffers should be further investigated. This investigation was done through the use of the Voronoi zones data set. To test if performance of the models was over estimated the R^2 values from the non stratified buffer zones were compared to equivalent models made using the Voronoi data. In all cases the difference in R^2 was less than five percent indicating the performance of the models was not overestimated and the overlap issue is not serious. The parameters from the models were also compared to test if the models found similar relationships. In most cases the parameters were close to within ten percent of each other with the exception of one parameter in the CSH extended model, which showed a difference of 31.5%. This test is less relevant to overlapping causing problems with neighbourhood effects but it does reinforce the idea that the overlapping zones do not generate problems.

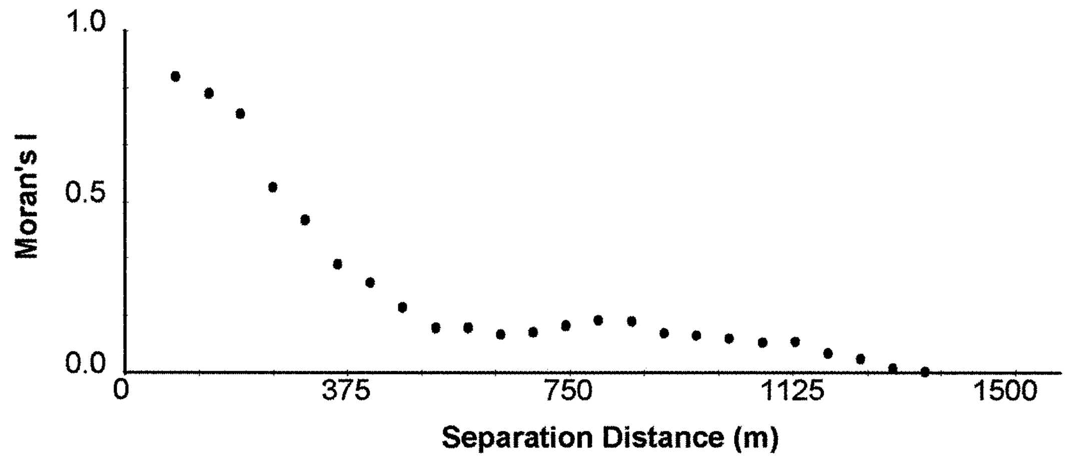

Figure 5.9 CSH extended model residuals correlogram.

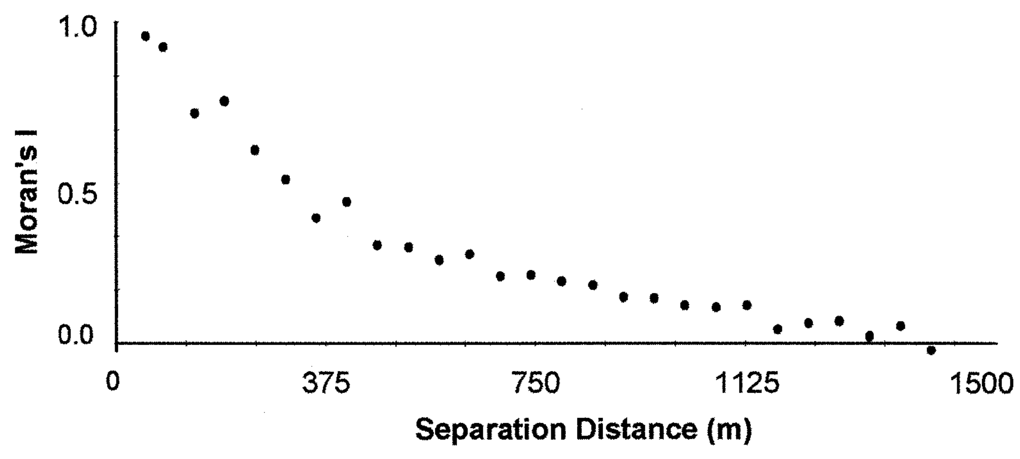

Figure 5.10 Heath extended model residuals correlogram.

The second test for problems with neighbourhood effects was based on comparing the level of autocorrelation in the dependent variables and the predictors. The predictors are topographic variables and since the landscape does not change rapidly it is clear that the IVs will be spatially autocorrelated. It is also logical that if the predictors are autocorrelated the dependent variables should be also. The question that needs to be resolved therefore is not "are the DVs spatially autocorrelated ?" but "are there significant neighbourhood effects that make the DVs more spatially autocorrelated than the IV? ". A simple way to answer this question is to compare the level of spatial autocorrelation in the DVs with the predicted values. The predicted values can be used because they are simply a multivariate nonlinear combination of landscape variables found to be good predictors of percentage cover. Correlograms provide a clear and detailed way to visualise the relationship between spatial autocorrelation in the IVs and DVs. Figure 5.11 and Figure 5.12 compare the spatial autocorrelation in the CSH and heath percentage cover variables with the predictors. The figures show there is very little difference between the spatial autocorrelation in the DVs and predictors. Bivariate correlations for the relationships are all above 0.95 indicating there is no real difference in the level of spatial autocorrelation in the DVs and IVs. From this analysis it can be concluded that there are no neighbourhood effects making the DVs more spatially autocorrelated than the predictors.

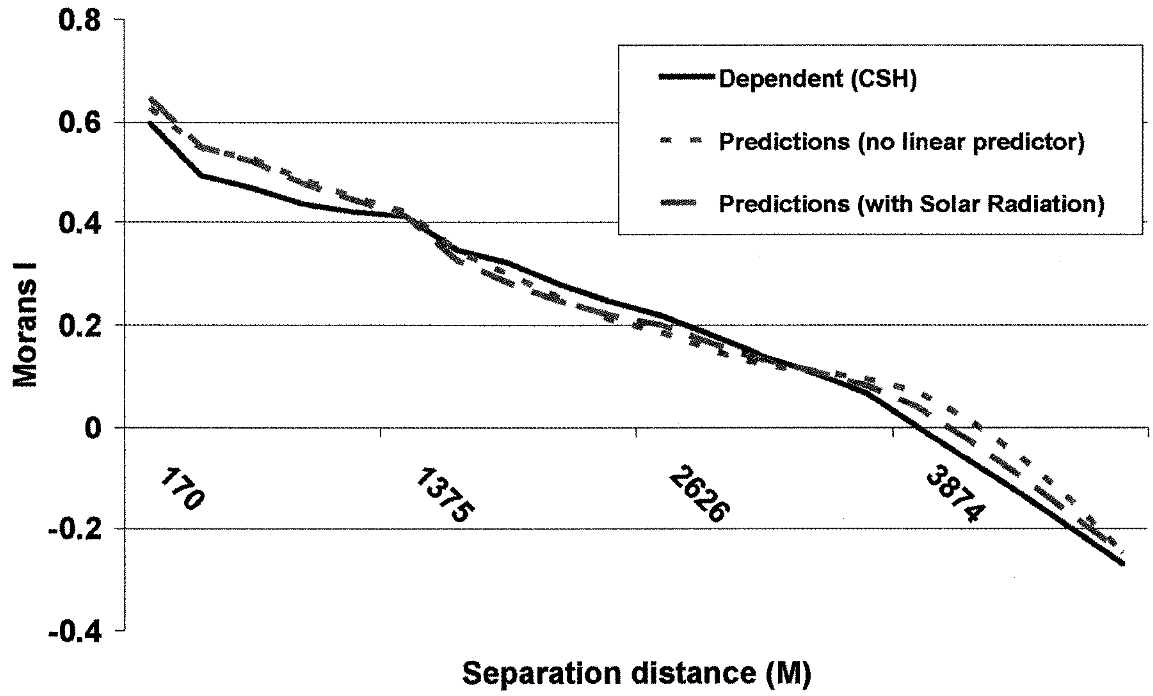

Figure 5.11 Spatial autocorrelation in the CSH DV compared to the predictors. There is very little difference between the autocorrelation in the dependent variable and the predictors.

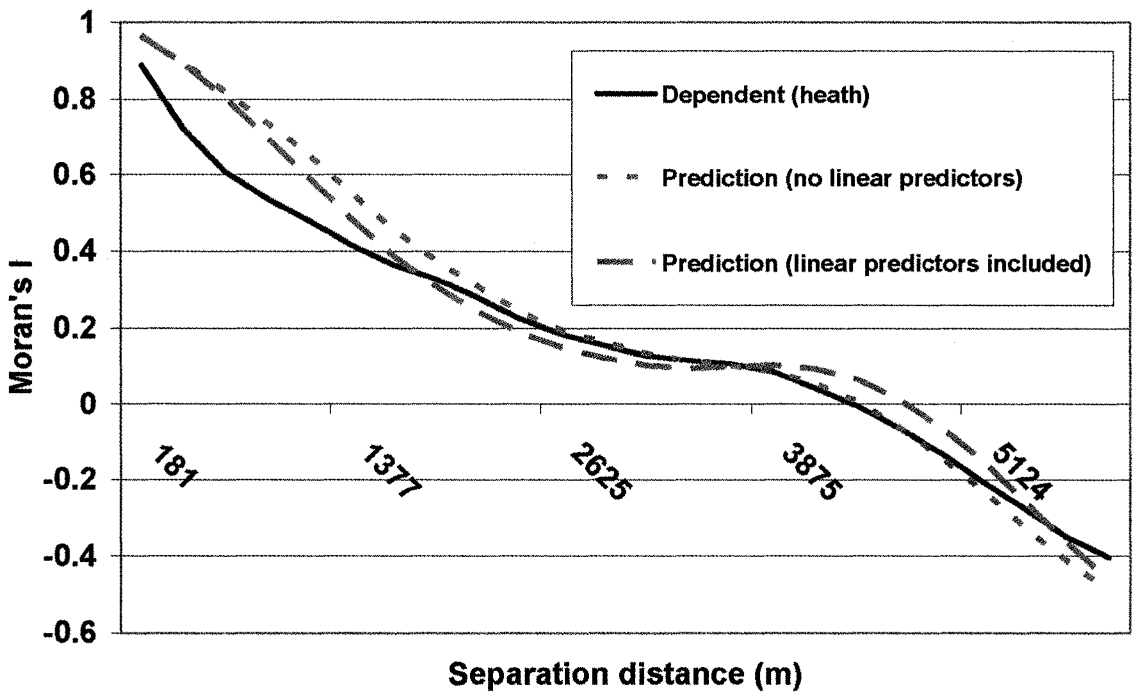

Figure 5.12 Spatial autocorrelation in the heath DV compared to the predictors. There is very little difference between the autocorrelation in the dependent variable and the predictors.

Though the data, both IVs and DVs, in this analysis are autocorrelated no evidence was found for significant neighbourhood effects. This study suggests no further treatment of neighbourhood effects or spatial autocorrelation is required.

5.6. Final models

The final models that were used for predictions are for the two cover classes that are of primary importance, CSH and Heath. The models developed from the non-stratified data were used in both cases and the inputs were smoothed. The extended versions of the models were used in both cases as the validation R^2s indicate the linear predictors make a justifiable addition to the models. The resulting nonlinear models are detailed in Equation 5.1 (CSH) and Equation 5.2 (Heath) where they are shown as raster calculator expressions, the parameter values are shortened to three significant figures for display only. In the next chapter the resulting predictions from these models will be investigated.

CSH = 165 + (0.0540 * (Exp (1.60 * [TEMP]))) + (-200 * [SOLAR])

Equation 5.1 CSH predictive model.

Heath = 38.2 + (44.4 * [TEMP]) + (-10.2 * Pow([TEMP], 2)) + (-1.01 * [SLOPE_E]) + (-191 * [NEANGEX])

Equation 5.2 Heath predictive model.

Where:

[TEMP] = Smoothed temperature

[SOLAR] = Smoothed solar radiation

[SLOPE_E] = Smoothed normalized slope

[NEANGEX] = Smoothed NE angle exposure

5.7. Summary of Methods

Models were successfully built for two of the cover classes, CSH and heath. The other cover classes could not be modelled reliably by the methods used here. The two models developed use nonlinear functions to include the relationship with elevation and one or more linear predictors to explain further variance. The models trained with data sampled randomly and not stratified by elevation appear the most robust when tested against the study area as a whole. The sensitivity analysis suggests that although the models are affected by choices in their calibration and errors in climate scenarios these differences are not likely to affect conclusions derived from the resulting modelled predictions of change.

6. RESULTS AND DISCUSSION

The previous chapter discussed how the model predictions of vegetation change were created. In this section the predictions will be presented and interpreted. The results for the CSH and Heath cover classes are the only ones examined because they are the only classes for which useful models could be constructed. These two classes were both classified from the satellite imagery with a good level of accuracy. The consumer's accuracy was 81.7% for the Heath class and 75.2% for the CSH class. There was some misclassification between the two classes seen in a cross tabulation between the ground truth and classified image, 8.12% of the CSH class was misclassified as Heath and 6.76% of the CSH was misclassified as Heath.

Figure 6.1 to Figure 6.3 represent the predicted change in percentage cover (not percentage change) in each of the three time periods for the CSH cover class, and Figure 6.4 to Figure 6.6 contain equivalent predictions for the heath cover class. In the figures values within the root mean squared error (RMSE) are shown in grey as these values do not represent significant change as the predicted changes are less than the standard error in the model. The RMSE used is based on a combination of the RMSEs from the predictive models and the interpolations according to error propagation formulas from Eastman (1993). The resulting RMSE for CSH is ±11.52% and ±12.15% for the heath cover class, the errors were calculated using the validation data set discussed previously. Due to the nature of the predictive formulas the CSH cover class can only increase in abundance with the increased temperature scenarios. Reduced abundances are possible, but only due to differences between the baseline data and the training data. The heath class can both increase and decrease due to the quadratic function used to model it. In the study area, increases only occur in small areas at high elevations. It is therefore possible for both classes to increase in the same cell. This does occur at high elevations which is consistent with colonization by both heath and CSH of previously sparsely vegetated (rocky) areas. The predictions show that the areas most susceptible to change are the transition areas at the middle elevations. At lower elevations CSH is quite stable and heath will not increase, in the upper elevation CSH only starts to increase dramatically in the final time period (2070-2099) when heath also shows large declines. The sensitive areas are therefore the middle elevations and this is particularly so because these are the areas where there is already a significant abundance of CSH required for the spread of the species included in that cover class. The middle elevations are also where there is the greatest uncertainty in the predictive models because this is where the model fits are the weakest (Figure 5.4 and Figure 5.5).

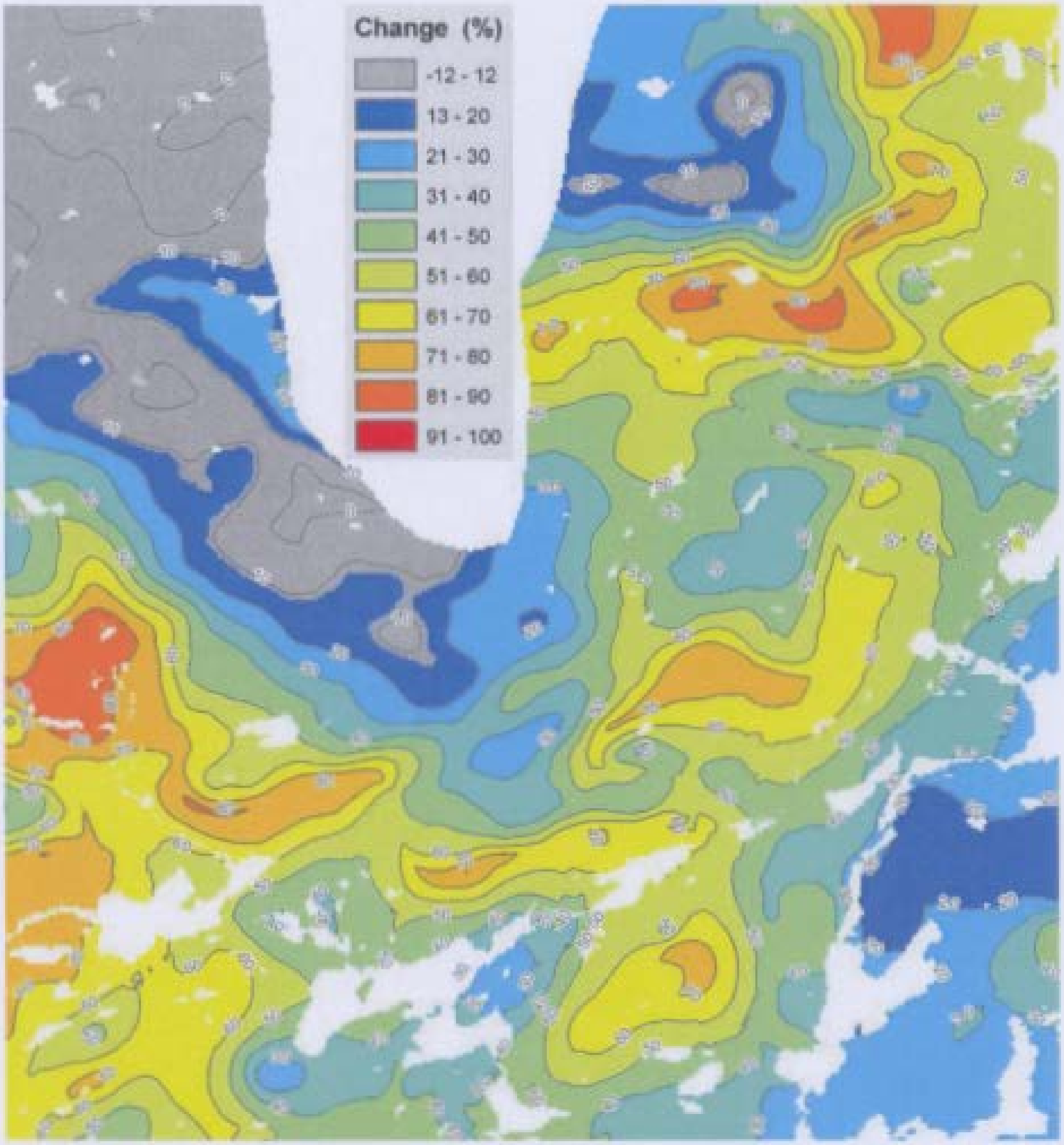

Figure 6.1 CSH predicted change in percent cover for the 2010 to 2039 time period. Grey areas represent non-significant change that is with the error of the model

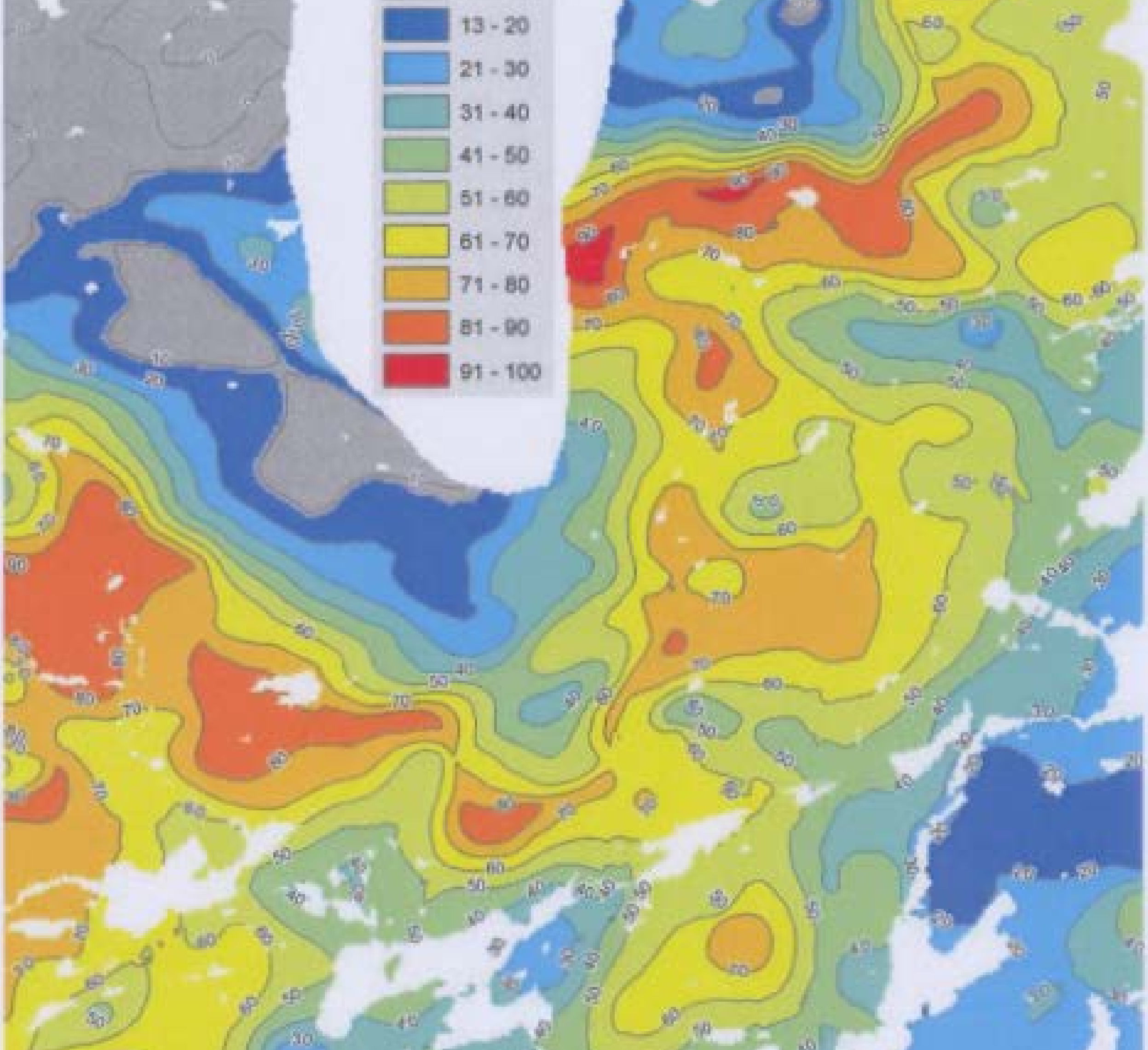

Figure 6.2 CSH predicted change in percent cover for the 2040 to 2069 time period. Grey areas represent non-significant change that is with the error of the model

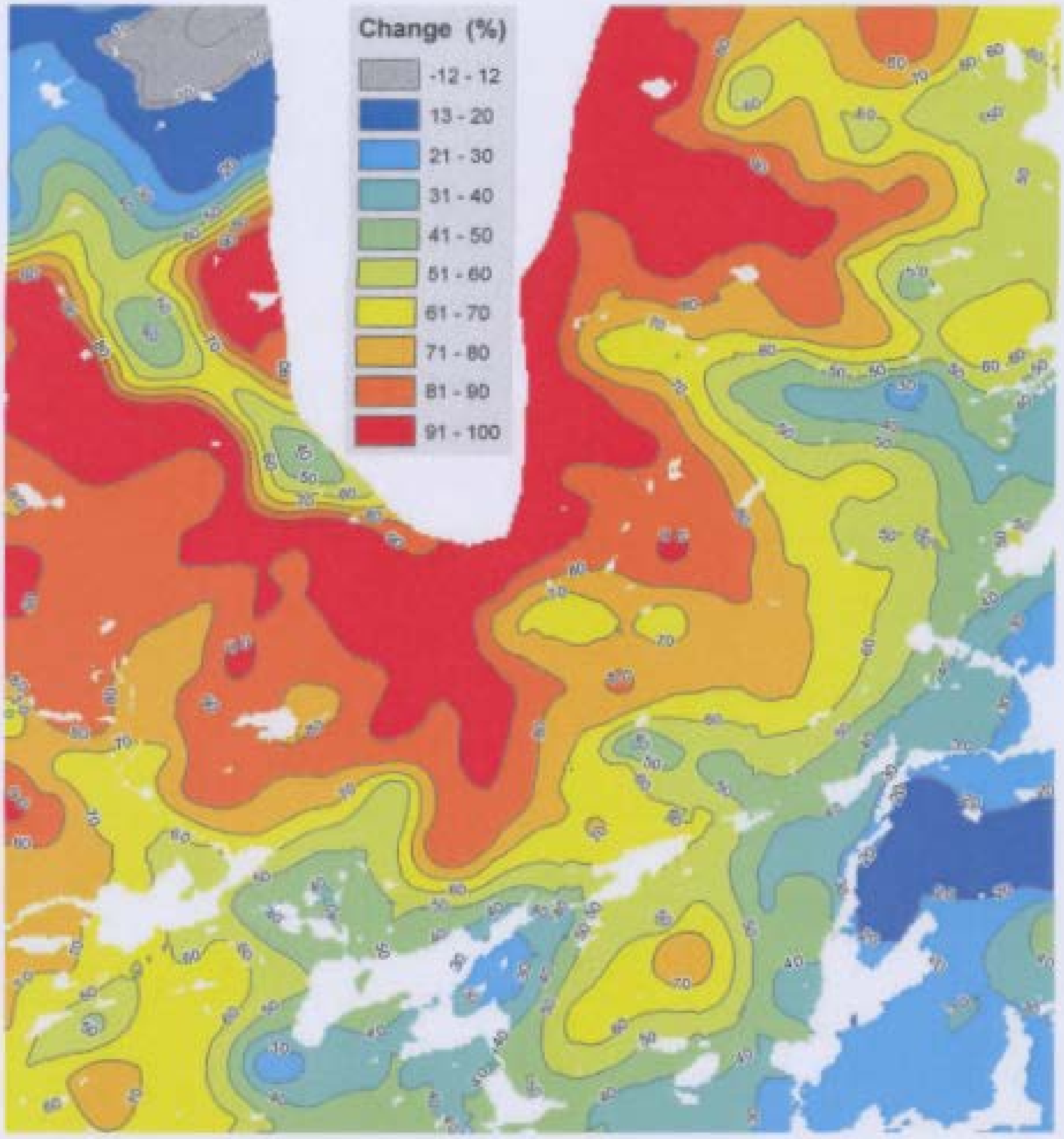

Figure 6.3 CSH predicted change in percent cover for the 2070 to 2099 time period. Grey areas represent non-significant change that is with the error of the model

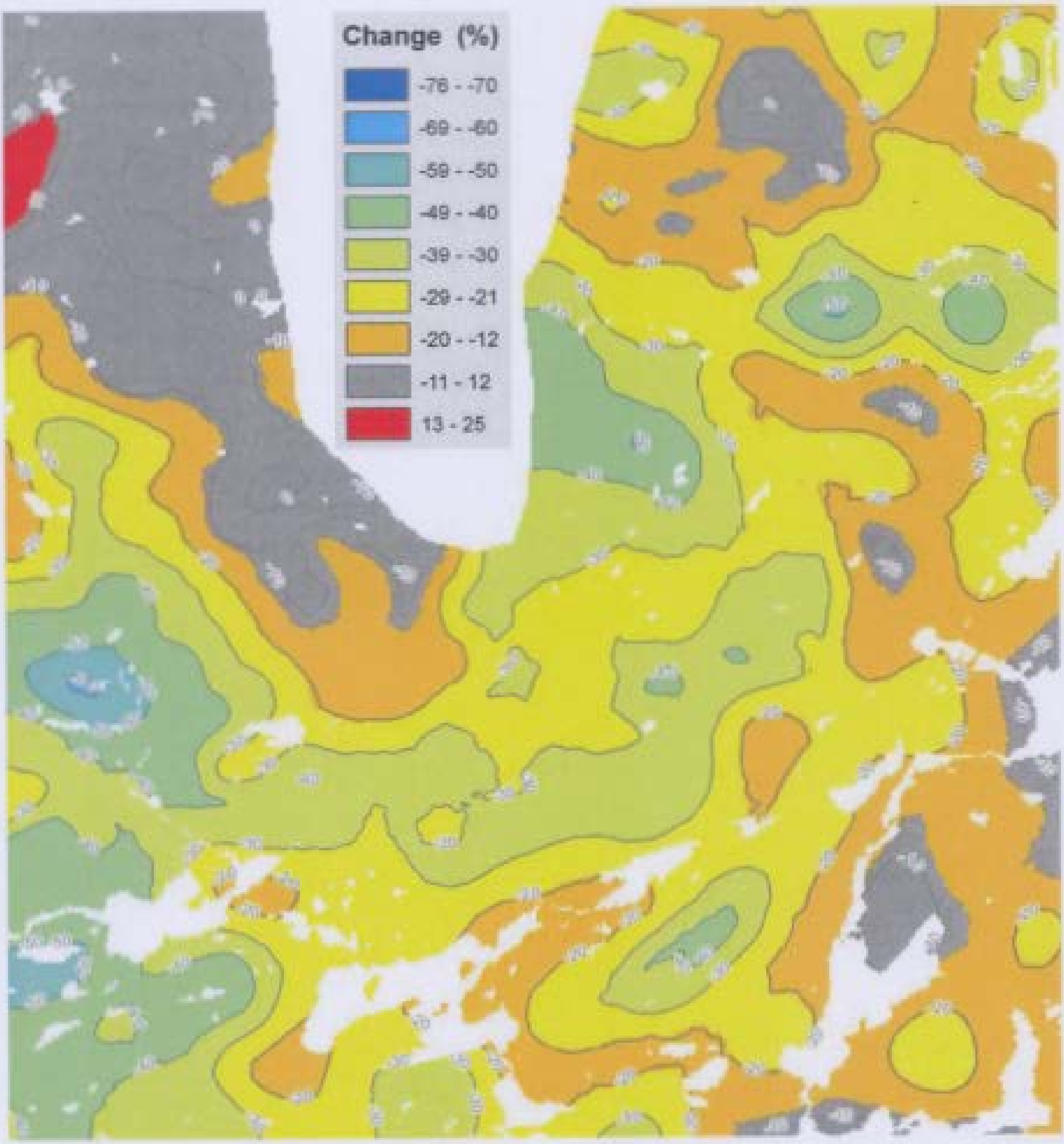

Figure 6.4 Heath predicted change in percent cover for the 2010 to 2039 time period. Grey areas represent non-significant change that is with the error of the model

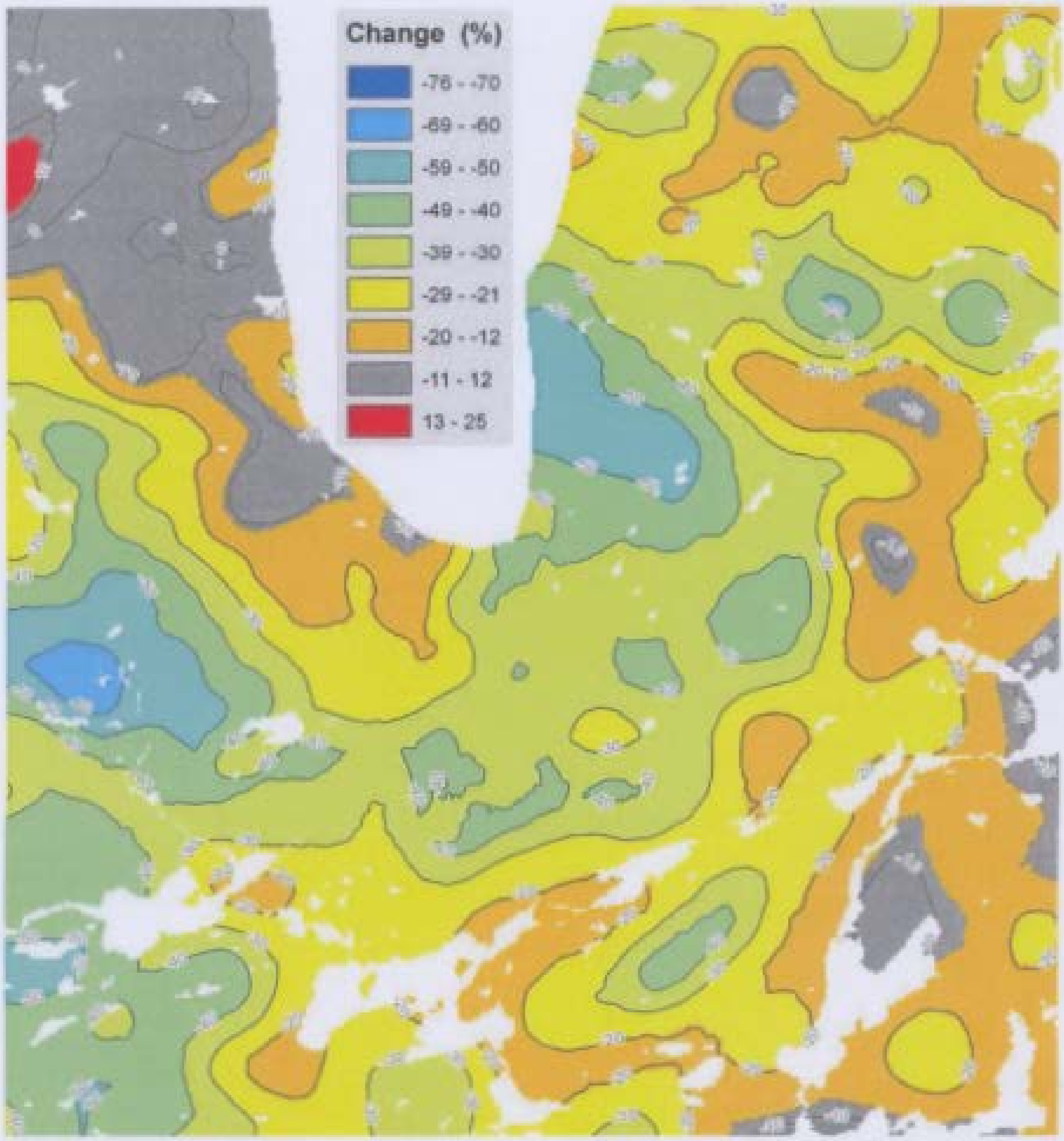

Figure 6.5 Heath predicted change in percent cover for the 2040 to 2069 time period. Grey areas represent non-significant change that is with the error of the model

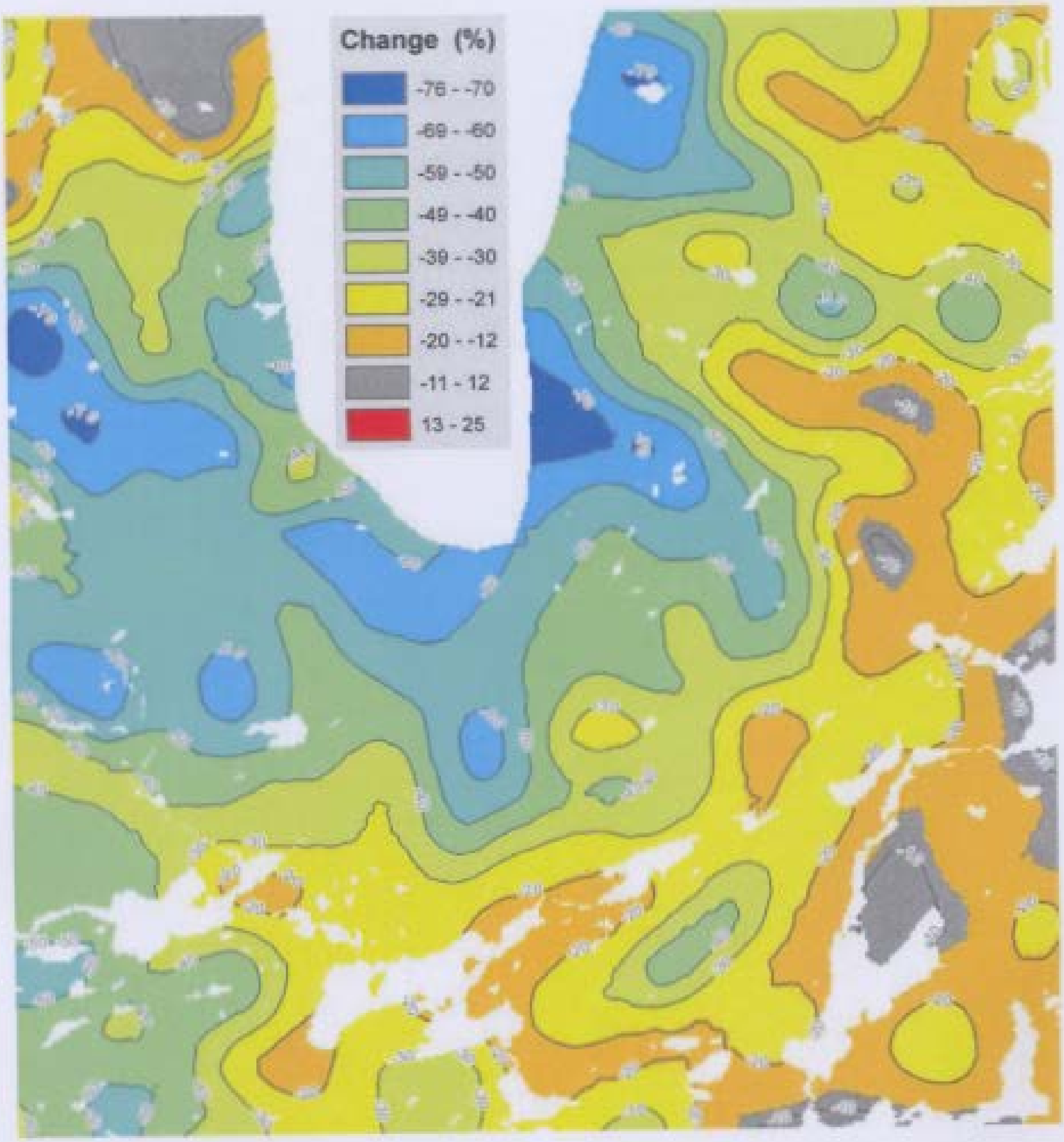

Figure 6.6 Heath predicted change in percent cover for the 2070 to 2099 time period. Grey areas represent non-significant change that is with the error of the model

The predictions can be summarized using charts to compare the proportion of the study area occupied by different abundance classes for each of the two cover classes. In the two charts that follow (Figure 6.7and Figure 6.8) the pixels from each cover class are assigned to abundance classes depending on the percentage cover in the 200m radius around that pixel. These abundance classes are then plotted against the proportion of the study area occupied by pixels in that class. The total of all the bars for CSH equals 100% because 100% of the study area has between zero and 100% coverage by CSH. For the present time period (Figure 6.7) both cover classes have a reasonable spread of areas occupied by the range of cover classes. Low abundances of CSH are more common than high abundances but the spread of heath abundance is relatively even. In the predictions for the 2070-2099 time period very high abundances of CSH (90% to 100%) occur in nearly 90% of the study area, conversely most of the study area has very low abundances of heath. The equivalent graph for the 2040-2069 time period (Appendix II) shows a very similar pattern to the 2070-2099 period though there are some areas containing higher abundances of heath. The difference between the two figures shows a very dramatic change from a reasonably mixed landscape to a landscape dominated by coniferous forest and shrub (CSH).

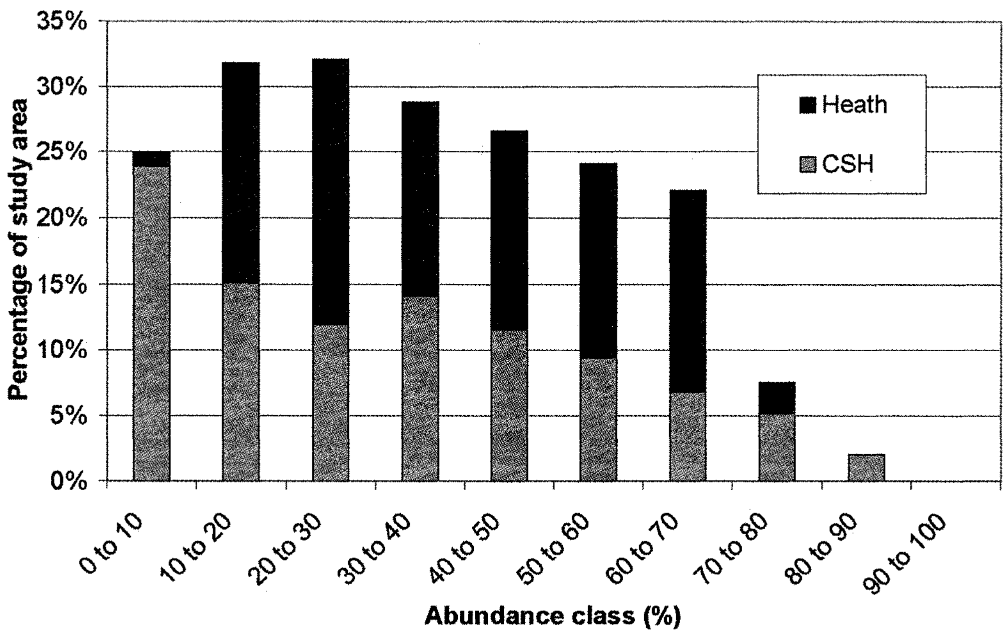

Figure 6.7 Percentage of the study area occupied by CSH and Heath abundance classes for the present.

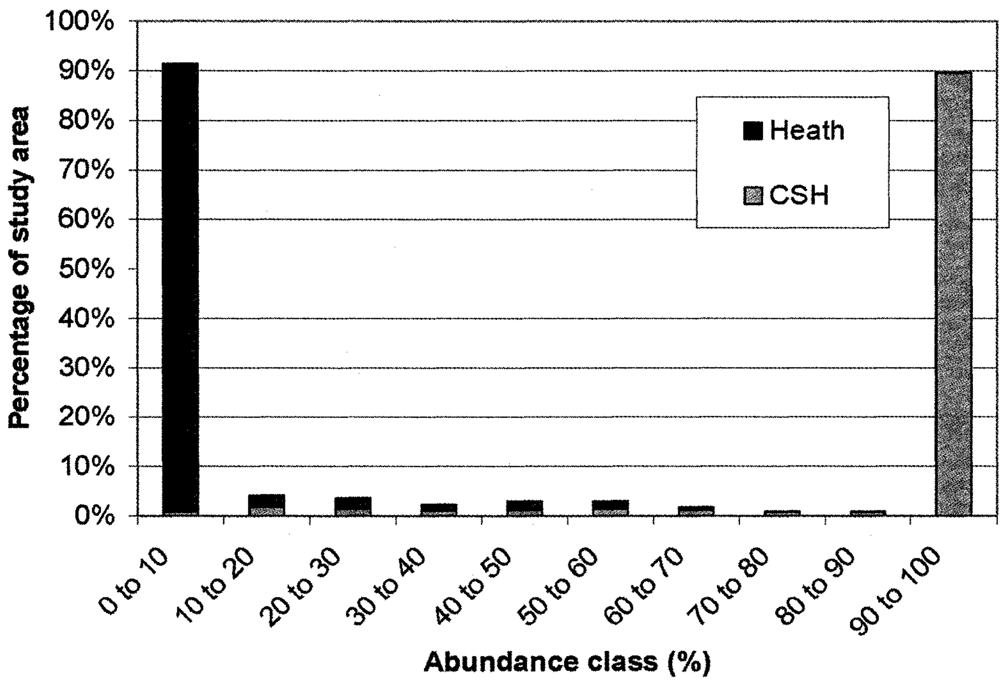

Figure 6.8 Percentage of the study area occupied by CSH and Heath abundance classes for the 2070-2099 time period.

These predictions include no restrictions on the speed at which vegetation can change and so should be interpreted as what the potential coverage might be under the future climate scenarios if the trees and shrubs can increase in abundance at a rate fast enough to maintain equilibrium with the changing climate. Otherwise the realized coverage will be less (Guisan and Zimmermann, 2000, Midgley *et al.* 2006).

7. CONCLUSION

The primary objective of this research was to create a spatially explicit model that could predict vegetation distributions for the current topoclimatic conditions and future conditions. These objectives were broken down into five tasks that had to be completed to meet the primary objective:

1. Map the current distribution of vegetation types.
2. Build a database of current topoclimatic conditions.
3. Perform exploratory analysis to inform the model building process
4. Create a model that predicts current vegetation distribution based on the topoclimatic variables.
5. Apply the model to altered topoclimatic conditions that represent past or future scenarios.

These tasks were completed successfully and predictions have been made that can be of use in understanding and managing the future vegetation cover of the Mealy Mountains under different growing-season temperature scenarios.

7.1. Vegetation mapping

The mapping of vegetation types was completed using orthorectified Quickbird imagery and ground truth data collected as part of the summer 2005 field season. Earlier work with Landsat imagery and field observation suggested that high resolution imagery was important because the landscape is too mixed for the resolution of Landsat pixels. The heterogeneous landscape results in mixed signatures and poor classification. The classification of the Quickbird imagery performed well and produced a raster based vegetation map that was suitable for model building in later stages. The vegetation map is limited in that the form of coniferous trees cannot be separated; full trees, shrubs and krummholz remain grouped. There was also some misclassification between the two main classes (Heath and CSH) which could have limited the performance of the models.

7.2. Topoclimatic Variables

The set of topoclimatic variables was constructed for the study area-based on the range of variables used by other researchers working towards similar aims. These variables are mostly surrogates since actual variations across the site were not available. The range of variables included measures of sheltering, exposure to solar radiation, potential for snow accumulation and potential for soil moisture. Results from the exploratory analysis suggested that the slope variable was too highly correlated with elevation to be used in its raw form. This correlation occurs because of the form of the landscape, slopes are steeper at higher elevations. Potential issues with collinearity were avoided through normalization of the slope variable by elevation. The resulting normalised slope variable was a useful predictor in one of the two final models. In the final models only a small number of the topoclimatic variables were useful in predicting vegetation distributions. Temperature, derived solely from elevation, was a strong predictor for most of the vegetation classes and could have been used without any other topoclimatic variables to make predictions with a reasonable amount of accuracy.

7.3. Exploratory analysis

Exploratory analysis provided a great deal of information for the model building process. The most important conclusions from the exploratory analysis were concerned with the spatial scales appropriate for modelling. As was previously hypothesized, the landscape is too heterogeneous for models to be built on the information at single locations or very small areas. What occurs at a specific location, in a metre radius for example, is quite random but when larger areas are examined trends can be seen. From the investigation using a range of different sized areas it appears that the optimum sized area to work with is around 0.126km2 (a circle with a radius of 200m). The data sets used for model building were therefore based on 200m radius circles and the values used were average values of the topoclimatic variables and percentage cover by each class in the circular zone. The zone based data (opposed to point-based) contain continuous values for each cover class (percentage cover) and so allow a broader set of modelling techniques than the discrete classes in the point-based data.

The exploratory analysis also showed the presence of a transition zone in the landscape. This zone occurs at the middle elevations and the vegetation coverage is more mixed and heterogeneous than at the higher and lower elevations. This area is at the greatest risk to change and is also the hardest to make predictions for because it is more randomized.

There are only three true variables in the data, aspect, elevation (temperature) and curvature. There are a number of topoclimatic variables that are too highly correlated with each other to use together and these relate to a common one of the three true variables. In the final models there are no variables used from common true variables. The heath model only uses temperature, which is essentially elevation and solar radiation which is highly related to aspect. The CSH model uses temperature and NE angle exposure which is an aspect variable. The CSH model also uses slope which in its raw form is highly related to elevation but in its normalised form collinearity is not an issue.

7.4. Model building

The model building and validation process resulted in two models that predict percentage cover for their respective cover class, heath and CSH, with a good level of accuracy (R2 > 0.685). The DSH model appeared to work well but later stages of validation showed that it did not generalise well when applied to the actual study area. The apparent performance of the DSH model was assumed to be an artefact of the modelling process, caused by the combination of stratified sampling, log transform and non linear modelling. The successful models were quite basic using a nonlinear function for the temperature variable and one or two linear predictors that added a small but useful amount of prediction power.

The use of stratified sampling schemes for model development showed no real gains over the non-stratified sampling. At the training stages some models trained with stratified data appeared slightly more powerful, CSH for example, these gains however did not hold up to full validation. The models based on non-stratified data performed best for the study area as a whole. The implication of the choice of sampling scheme was also investigated through sensitivity analysis. The interpretation of the results from models built with either data set would be identical or at least very similar, so the choice is almost of no consequence.

The overall finding from the model building process was that nonlinear modelling performed well but any modifications such as altered sampling schemes and data transforms did not add enough to the power of the models to warrant their use. As has held true throughout the analysis, temperature is the only strong predictor for the distribution of the cover classes. Other topoclimatic variables were less important than expected. The minor influence of the other topoclimatic variables could be due to scale effects. At the scales used here they may not appear to be of great importance but at other scales they could be more important. The topoclimatic variables that were used in the two final models were solar radiation, NE angle exposure and normalised slope. Both NE angle exposure and solar radiation show some aspect dependency. In the heath model the DV is negatively related to normalised slope indicating heath is less abundant on the steep areas.

The inclusion of the variables in the models should not be confused with cause and effect, the temperature variable is a strong predictor but temperature is not necessarily the cause of the distribution. The distributions could be caused by variables not mentioned in this project that are highly correlated with the variables found to be useful. Another possibility is that temperature does have an actual influence on the distributions but other variables, highly correlated with elevation such as precipitation, also play some role. This issue is important because when the temperature variable is altered to make predictions it is altered based solely on temperature change, therefore assuming that temperature is the only cause of the correlations with elevation.

The predictions of change in the future are best seen as representing how the potential for change varies across the site under a given climate scenario. The predicted future distributions may not be achievable within the time period of the temperature change as plant assemblages take time to spread and develop (Bennet *et al.,* 1986, Ritchie and

MacDonald 1986, Malcolm *et al.* 2002, Midgley *et al.* 2006). The change should also not be seen as what the potential change will be in the future but what the potential for change would be given the climate scenario used. It is clear that the accuracy of the prediction will be heavily influenced by differences between the actual future climate and the scenario. From the sensitivity analysis it appears that with a level of difference of 30% in the temperature prediction there are still large areas of significant potential for change even at 50% there is some significant potential for change. The real value of the predictions is not in showing specific changes in single locations but rather in showing where the greatest potential for change is and how great that potential is.

7.5. Predictions and implications

The predictions suggest large potential for change in the middle elevations where the vegetation is currently quite heterogeneous. If the predictions are realized and the potential for change is achieved the landscape will be clearly dominated by coniferous forest and shrub after the 2040-2069 time period. The prediction for this time period indicates about 65% of the study area will have only ten percent or less local coverage (within a 200m radius zone) by the heath class. Conversely over 65% of the study area will be covered by a vegetation type consisting of 90 to 100% coniferous forest and shrub.

The findings of this study agree with the consensus of upward shifts of altitudinal tree-lines (Parmesan, 2006). Though tree-line is not included explicitly in this study the results could be modified to allow analysis in terms of tree-line. Tree-line could be defined by a threshold percentage cover beyond which each pixel would be classed as forest. The resulting boundary between forest and nonforest could then be used to create a tree-line and predicted future tree-lines. This analysis was not completed because the choice of threshold would be arbitrary and the forest and no forest land cover classes dichotomy does not suit the patchy homogenous nature of vegetation found in the study area.

Assuming that the climate scenarios used in this analysis or changes to a similar degree are unavoidable how can the findings of this study be of use? The key finding that can be of use in protecting these areas is that it is the middle elevations that are at the greatest risk. The middle elevations should therefore be afforded the greatest available protection.

8. REFERENCES

Arseneault, D. & Payette, S. (1992). A Postfire Shift From Lichen-Spruce to Lichen-Tundra Vegetation at Tree Line. *Ecology* 73 1067-1081.

Bennett, K. D., Lawton, J. H., Gibbs, A.,Williamson, M. H., Holdgate, M. W. & Hamilton, W. D., *et al.*. (1986). The Rate of Spread and Population in Increase of Forest Trees During the Postglacial [and Discussion]. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 314(1167), 523-531.

Buffo, J., Fritschen, L. & Murphey, J. (1972). *Direct solar radiation on various slopes from 0 to 60 degress north latitude.* (USDA Forest Service Reseach Papers). Portland, OR:.

Burrough, P. & McDonnell, R. (1998). *Principles of Geographical Information Systems.*New York: Oxford University Press.

Cairns, D. (2001). A comparison of methods for predicting vegetation type. *Plant Ecology* 156 3-18.

Calef, M., McGuire, A., Rupp, T. & Shugart, H. (2005). Analysis of vegetation distribution in Interior Alaska and sensitivity to climate change using a logistic regression approach. *Journal of Biogeography* 32(5), 863-878.

Cousins, S., Lavorel, S. & Davies, I. (2003). Modelling the effects of landscape pattern and grazing regimes on the persistence of plant species with high conservation value in grasslands in south-eastern Sweden. *Landscape Ecology* 18(3), 315-332.

Canadian Institute for Climate Studies. (2003). *Canadian climate impacts scenarios.* Retrieved 01/30/2006, 2006, from http://www.cics.uvic.ca/scenarios/data/select.cgi#map.

Danby, R. & Hik, D. (2007). Variability, contingency and rapid change in recent subarctic alpine tree line dynamics. *Journal of Ecology* 95(2), 352-363.

De'ath,G. & Fabricicus, K. (2000). Classification and Regression Trees: A Powerful Yet Simple Technique for Ecological Data Analysis. *Ecology* 81 3178-3192.

Eastman, J. (1993). *Idrisi Version 4.1 Update Manual.*Worcester, Massachusetts: Clark University.

Environment Canada.(2003). *Canadian Climate Scenarios*, from http://www.cccma.bc.ec.gc.ca/data/cgcm2/cgcm2.shtml.

Environment Canada. (2007). *Climate Trends and Variations Bulletin,* from http://www.smc-msc.ec.gc.ca/ccrm/bulletin/archive_e.cfm.

GeoGratis. (2003). *East-Central Labrador Ecological Land Survey, Canada: Data Product Summary,* from http://geodiscover.cgdi.ca/gdp/search?action=entrySummary&entryType=productCollection&entryId=6033&entryLang=en&displayHeader=true.

Geomatics Canada. (1997). *Standards and Specifications of the National Topographic Data Base*, 2006, from http://www.cits.rncan.gc.ca/fich_ext/1/text/products/ntdb/stdntdb3.pdf.

Gottfried, M., Pauli, H., Reiter, K. & Grabherr, G. (1999). A fine-scaled predictive model for changes in species distribution patterns of high mountain plants induced by climate warming. *Diversity and Distributions* 5 241-251.

Gray, J. & Lauriol,B. (1985). Dynamics of the Late Wisconsin Ice Sheet in the Ungava Peninsula Interpreted from Geomorphological Evidence. *Arctic and Alpine Research* 17(3), 289-310.

Griffith, D. (1987). *Spatial Autocorrelation: A Primer.*Washington DC: Association of American Geographers.

Griggs, R. (1938). Timberlines in the Northern Rocky Mountains. *Ecology* 19(4), 548-564.

Guisan, A., Edwards,T.C. & Hastie, T. (2002). Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecological Modelling* 157 89-100.

Guisan, A. & Zimmermann, N. (2000). Predictive habitat distribution models in ecology. *Ecological Modelling* 135 147-186.

Hobbie, S. & Chapin III, F. (1998). An experimental test of limits to tree establishment in Arctic tundra. *Journal of Ecology* 86 449-461.

IPCC (Intergovernmental Panel on Climate Change) (2001). *Climate Change 2001: The Scientific Basis. (eds. Houghton, J.T.; Ding, Y.; Griggs, D.J.; Noguer, M.; van der Linden, P.J.; Dai, X.; Maskell, K.; Johnson, C.A.) Cambridge University Press, 881pp.* Cambridge University Press.

Jacobs, J., Hermanutz, L., Bell, T. & Simms, A. (2005). *Annual Report of research 2004, Labrador Highlands Research Group.* Retrieved January 11th, 2007, from http://www.mun.ca/geog/lhrg/Annual_Report_2004.pdf.

Jacobs, J. (2007). Climate Studies in the Mealy Mountains. *Labrador Highlands Research Group, Report of Research 2006* (pp. 11-4 Memorial University, Dept of Geography, St. John's, NL.

Keith, T. (2001). *A Natural History and Resource Inventory of the Proposed Mealy Mountains (Akamiuapishku) National Park Study Area, Labrador.* (Park Establishment Branch, Parks Canada.

Kirdyanov, A., Hughes, M. V., E., Schweingruber, F. & Silkin, P. (2003). The importance of early summer temperature and date of snow melt for tree growth in the Siberian Subarctic. *Trees - Structure and Function* 17(1).

Klein, J., Harte, J. & Zhao, X. (2004). Experimental warming causes large and rapid species loss, dampened by simulated grazing, on the Tibetan Plateau. *Ecology Letters* 7(12), 1170-1179.

Körner, C. (1998). A re-assessment of high elevation treeline positions and their explanation. *Oecologia* 115(4), 445-459.

Körner, C. & Paulsen, J. (2004). A world-wide study of high altitude treeline temperatures. 31(5), 713-732.

Malcolm, J., Markham, A., Neilson, R. & Garaci, M. (2002). Estimated migration rates under scenarios of global climate change. *Journal of Biogeography* 29(7), 835-849.

McCune, B. & Keon, D. (2002). Equations for potential annual direct incident radiation and heat load. *Journal of Vegetation Science* 13 603-606.

Meades, S. (1990). *Natural regions of Newfoundland and Labrador.* St. John's, NF: Protected Areas Association.

Midgley, G. F., Hughes, G. O., Thuiller, W. & Rebelo, A. G. (2006). Migration rate limitations on climate change-induced range shifts in Cape Proteaceae. *Diversity and Distributions* 12(5), 555-562.

Miller, J. & Franklin, J. (2002). Modelling the distribution of four vegetation alliances using generalized linear models and classification trees with spatial dependence. *Ecological Modelling* 157 227-247.

Morrison, P. (2002). Introduction to Part 2. In Scott, J., Heglund, P. and Morrison, M. (Eds.), *Predicting Species Occurrences: Issues of Accuracy and Scale* (Island Press, USA.

NASA, NIMA, DLR & ASI. (2002). *Shuttle Radar Topography Mission (SRTM) Spatial Metadata Dataset*, from http://www.metrokc.gov/gis/sdc/raster/elevation/ShuttleRadarTopographyMissionSpatialMetadata.htm.

O'Connor, R. (2002). Introduction. In Scott, J., Heglund, P., Morrison, M., Haufler, J., Raphael, M., Wall, W. and Samson, F. (Eds.), *Predicting Species Occurrences: Issues of Accuracy and Scale* (Island Press, USA.

Odland, J. (1988). *Spatial Autocorrelation*.Newbury Park: SAGE Publications.

O'Sullivan, D. & Unwin, D. (2003). *Geographic Information Analysis.* Hoboken, New Jersey: John Wiley and Sons, Inc.

PCI Geomatics. (undated). *Orthorectifying QuickBird Data.* Retrieved 01/17/2006, 2006, from http://www.pcigeomatics.com/services/support_center/quickguides/QuickBird.pdf.

Parmesan, C. (2006). Ecological and Evolutionary Responses to Recent Climate Change. *The Annual Review of Ecology, Evolution, and Systematics* 37 637-669.

Parks Canada. (2005). *Meayly Mountains Akamiuapishk National Park Feasibility Study.* Newsletter #1 Summer 2005

Pauli, H., Gottfried, M. & Grabherr, G. (1996). Effects of climate change on mountain ecosystems -- Upward shifting of alpine plants. *World Resource Review* 8(3), 382-390.

Pauli, H., Gottfried, M., Dirnböck,T., Dullinger,S. & Grabherr, G. (2003). Assessing the Long-Term Dynamics of Endemic Plants at Summit Habitats;. *Ecological Studies* 167 195-207.

Pauli, H., Gottfried, M., Reiter, K., Klettner, C. & Grabherr, G. (2007). Signals of range expansions and contractions of vascular plants in the high Alps: observations (1994-2004) at the GLORIA* master site Schrankogel, Tyrol, Austria. *Global Change Biology* 13 147-156.

Ritchie, J. & MacDonald, G. (1986). The Patterns of Post-Glacial Spread of White Spruce. *Journal of Biogeography* 13(6), 527-540.

Sirois, L. (2000). Spatiotemporal variation in black spruce cone and seed crops along a boreal forest – tree line transect. *Canadian Journal of Forest Research* 30 900-909.

Shaw, P. (2003). *Multivariate Statistics for the Environmental Sciences.* New York: Arnold.

Sibley, D. (1987). *Spatial Applications of Exploratory Data Analysis.* (Vol. 49). Norwich: Geobooks.

Starfield,A.M. & Chapin III, F. (1996). Model of Transient Changes in Arctic and Boreal Vegetation in Response to Climate and Land Use Change. *Ecological Applications* 6(3), 842-864.

Tabachnick, B. & Fidell, L. (1996). *Using Multivariate Statistics.* (3 ed.). New York: Harper & Row Publishers.

Tobler, W. R. (1970). A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography* 46(Supplement: Proceedings. International Geographical Union. Commission on Quantitative Methods), 234-240.

Appendix I

Pearson's Bivariate correlation matrix for the 200m buffer data created using a non-stratified sampling scheme.

	Slope	Normalised Slope	SW wind exposure	Snow potential index	Smoothed curvature	NE angle exposure	NE wind exposure	Moisture Potential Index	Solar Radiation	Northness	Eastness	Elevation
Slope	1.000	**0.924**	0.249	-0.257	0.191	-0.276	-0.328	-0.757	0.108	-0.369	-0.126	-0.503
Normalised Slope	**0.924**	1.000	0.170	-0.176	0.136	-0.223	-0.241	-0.710	0.049	-0.296	-0.067	-0.160
SW wind exposure	0.249	0.170	1.000	**-0.998**	0.014	-0.756	**-0.942**	-0.124	0.472	-0.531	**-0.927**	-0.276
Snow potential index	-0.257	-0.176	**-0.998**	1.000	-0.064	0.744	**0.941**	0.154	-0.467	0.529	**0.926**	0.282
Smoothed curvature	0.191	0.136	0.014	-0.064	1.000	0.188	-0.004	-0.592	-0.071	0.018	-0.025	-0.166
NE angle exposure	-0.276	-0.223	-0.756	0.744	0.188	1.000	0.796	0.030	-0.600	0.612	0.608	0.214
NE wind exposure	-0.328	-0.241	**-0.942**	**0.941**	-0.004	0.796	1.000	0.175	-0.690	0.784	0.748	0.320
Moisture Potential Index	-0.757	-0.710	-0.124	0.154	-0.592	0.030	0.175	1.000	0.013	0.212	0.051	0.401
Solar Radiation	0.108	0.049	0.472	-0.467	-0.071	-0.600	-0.690	0.013	1.000	-0.873	-0.161	-0.135
Northness	-0.369	-0.296	-0.531	0.529	0.018	0.612	0.784	0.212	-0.873	1.000	0.174	0.300
Eastness	-0.126	-0.067	**-0.927**	**0.926**	-0.025	0.608	0.748	0.051	-0.161	0.174	1.000	0.188
Elevation	-0.503	-0.160	-0.276	0.282	-0.166	0.214	0.320	0.401	-0.135	0.300	0.188	1.000

Appendix II

Predicted percentage of the study area occupied by CSH and Heath abundance classes for the 2040-2069 time period