University of Alberta

A Comparison between Markov Random Field and Markov Chain Models with an Application to Pine Beetle Infestation

by

Dan Vlad Metes ©

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Master of Science
in
Statistics

Department of Mathematical and Statistical Sciences

Edmonton, Alberta
Fall 2008

Library and
Archives Canada

Published Heritage
Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et
Archives Canada

Direction du
Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 978-0-494-47394-8
Our file Notre référence
ISBN: 978-0-494-47394-8

ABSTRACT

Logistic models are a useful way of analyzing dichotomous data by including relevant covariates to explain the variability in binary responses. However, to account for spatial-temporal variability, markov random field and markov chain models can be used to improve the accuracy of the fit. For binary spatial-temporal data spatial dependence can be accounted for via neighboring information, while temporal dependence can be modeled via information from the previous year(s). The use of such models is central in many applications such as pine beetle infestation situations. The infestation is currently placing heavy burdens on the timber and forest industry in British Columbia and throughout North America. The aggressive use of preventive measures in infested areas is thus of utter importance and limiting the costs associated with such measures relies on the existence of good models. This paper examines such models, which can be used to tackle the spread of the infestation.

ACKNOWLEDGEMENTS

I would like to thank my supervisor, Dr. Subhash Lele, for the years of guidance and encouragement throughout my degree and for the useful suggestions and hints provided for the preparation of this thesis.

I would also like to thank Dr. Mark Lewis and Dr. Fangliang He who kindly agreed to be on my committee and who have been so very prompt in answering all my emails.

Also, a big thank you to my parents, who have offered me moral support throughout these last three years and to my friends who have helped me focus on my objectives.

Last, but not least, my appreciation extends to the numerous faculty members and staff of the Department of Mathematical and Statistical Sciences who have made a difference in my life.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

The pine beetle is one of the most devastating predators of pine forests throughout North America, causing millions of dollars in damages each year to the forestry and lumber industry. Whether we are talking about the Southern Pine Beetle (*Dendroctonus frontalis*), which decimates pine populations throughout the Southern United States and Central America (from Pennsylvania to Texas, from New Mexico to Honduras – Thatcher and Barry, 1982), or the Mountain Pine Beetle (*Dendroctonus pondoresae*) present in areas as far north as British Columbia and Alberta, and as far south as California, Arizona and even Northern Mexico, the spread and intensity of the infestation constitutes a heavy burden for the private forestry sector, as well as the government.

According to the BC Ministry of Forests and Range (2008 update), out of the 1.35 billion cubic meters of merchantable pine in British Columbia about half, or 710 million cubic meters of timber, has been affected by the mountain pine beetle (MPB). It is projected that 76 per cent of the entire pine volume will be destroyed by 2015. In the United States, low levels of MPB infestation were registered throughout the 90s, but the infestation has accelerated and has grown 10 fold from 1996 to 2002 (i.e. from a low of 21,570 ha in 1996 to a 209,465 ha in 2002). The utmost peak was registered in 1981 when almost 2 million hectares were infested throughout 11 states (Gibson, 2004). Although the infestation seems to come in cycles, the consequences of it are long-lasting and complex. Aside from the heavy losses endured by the forestry industry, the impact of the infestation can alter the forest ecosystem by changing the wildlife species composition of the habitat. The depleted pine forests are sometimes replaced by other tree species, or the terrain can be completely replaced by grass and shrubs which can lead to an increase in water yield following the infestation (Amman, McGregor, & Dolph 1990). Attacked and dead trees also constitute a fire hazard over time and can lead to forest fires unless removed.

For these and other reasons, it becomes imperative that appropriate prediction tools be created to assist with the monitoring and control of the affected areas. A number of factors have been considered when trying to explain the underlying processes that impact infestation. Some have suggested that climate variation is an important element in beetle dynamics and can act as a catalyst of the spread of infestation (Caroll 2006). Prolonged periods of extreme cold temperatures during the early stages of beetle development can greatly deter the expansion of the beetle population, by killing the over wintering beetle larvae (Regniere & Bentz, 2007). Warm temperatures as well as precipitation are also thought to impact infestation. In addition, the forest and tree characteristics are assumed to play an important role. Trees with large diameters seem to be the preferred choice of beetles (Geizler 1980), despite the fact that they are better equipped to produce resin and fend off the invasion. Tree vigor was found to be only marginally significant (Mitchell & Preisler, 1991) partly because the size of the invasion usually overwhelms the tree's defense mechanisms. Also, variables as tree age and proximity of trees in thinned respectively unthinned stands have been considered. Elevation and terrain factors have been used in explaining propagation of infestation, these elements being tied in with climatic variations.

In terms of modeling strategies used to predict outbreaks, logistic regression is a common feature of many analyses. Covariate information can easily be incorporated to account for the uncertainty in the response and the results are easily interpretable either as odds ratios or probabilities of infestation. Mitchell and Preisler (1991) use such a model taking into account tree characteristics, tree age, leaf area index, tree vigor and a distance measure that combines the proximity to adjacent trees as well as the diameter at breast height of the trees, from thinned and unthinned plots.

Although the inclusion of such variables can improve the quality of predictions, spatial correlation between adjacent sites cannot be fully accounted for in these models. Outbreaks for the Southern Pine Beetle in the United States present visible spatial patterns and due to the cyclical nature of the infestation temporal patterns also exist (Gumpertz 2000). Spatial dependence can be accounted for via the use of autoregressive models, or Markov Random Field (MRF) models, while the temporal dependence component can be accounted for by using Markov Chain (MC) models. The paper at hand makes use of logistic, MRF and MC models to compare the accuracy of prediction and to assess the significance of the covariates involved. Operating characteristic curves (ROCs) and their corresponding AUCs are assessed to establish the accuracy of the predictions and the results are compared with results from similar papers (Camino-Beck 2008; Zhu, Huang, & Wu, 2005).

Chapter 2 contains theoretical considerations for binary spatial and spatial-temporal models, together with corresponding parameter estimation methods, parameter selection procedures and accuracy assessment procedures. Chapters 3 and 4 make use of Chapter 2 models that are applied to two different data sets: a North Carolina data set in Chapter 3, and a British Columbia data set in Chapter 4. A detailed description of the data together with some references to the biology of the infestation will be provided with the analyses. Chapter 5 concludes the paper by offering suggestions as to how the models can be improved upon.

Chapter 2: Models for Binary Data

I. Various Models

a) Spatial Data – The Logistic Regression Model

a.1) General Description, Parameter Interpretation and Model Dependencies

Binary data arise in many fields and disciplines such as agriculture and forestry, computing science, demography, and ecology. Whether we look at the presence or absence of bird species throughout various counties in the United States, the HIV status of young females in the Sub-Saharan African region, or the pine beetle infestation status of forest populations in British Columbia or North Carolina, the binary nature of various response variables becomes apparent. Due to the wide range of applications in which binary data is involved, the need for statistical tools to model and analyze binary information has lead to the creation of particular classes of models that can handle the task.

One such class of models is the class of generalized linear models (GLM), and in particular the logistic regression models. Since binary data is a particular type of discrete data, linear regression analysis is not appropriate for modeling purposes. What is modeled, given that the response can only take the values 0 and 1, is the probability associated with these two values. Regression models for such responses are thus used to describe probabilities as functions of the response variable, and do not directly model the response. The usefulness of logistic regression lies in that the covariates it uses can be of any nature, that is continuous or categorical and, in that it uses less stringent assumptions than linear regression.

As a member of the GLM class, the logistic regression model is characterized by a systematic component (which includes the explanatory variables), a random component (the response variable) and a link function (in this case, the logit function) (Agresti 2000). A general form for such a model is given on the next page.

$$\eta_i = \log \frac{P(Y_i = 1 \mid \mathrm{X}_{i,1}, \ldots, \mathrm{X}_{i,k})}{P(Y_i = 0 \mid \mathrm{X}_{i,1}, \ldots, \mathrm{X}_{i,k})} = \beta_0 + \beta_1 * \mathrm{X}_{i,1} + \ldots + \beta_i * \mathrm{X}_{i,k}, \qquad i=1,\ldots, n$$

(1)

Equivalently,

$$p_i = \mathrm{P}(\mathrm{Y}_i = 1 \mid \mathrm{X}_{i,1}, \ldots, \mathrm{X}_{i,k}) = \frac{e^{\beta_0+\beta_1 * X_{i,1}+\ldots+\beta_k * X_{i,k}}}{1+e^{\beta_0+\beta_1 * X_{i,1}+\ldots+\beta_k * X_{i,k}}}, \quad i=1,\ldots,n \qquad (2)$$

The interpretation of the β_i parameter estimates is given in terms of the multiplicative effect on the log odds of infestation and the coefficient estimates are calculated via maximum likelihood. Unlike ordinary least squares linear regression, the logistic model does not assume homoskedasticity, normality, or a linear relationship between the dependent and independent variables. However, the observations ought to be independent and the covariates must be linearly related to the logit (i.e. log odds ratio) of the response.

Logistic regression is useful in that a wide range of covariates can be included in the model to account for the variability in the response variable. When it comes to beetle infestation, the variability in the infestation can be explained by a wide a range of factors such as: beetle biology, terrain information, climate variables and tree and forest characteristics. Despite the usefulness of logistic regression with modeling binary data, one of its major drawbacks when examining spatial binary data, is that it cannot take into account the spatial correlation present in the data. Although some of the spatial variability is accounted for by the covariates themselves, in many cases most of it remains unexplained and is, in effect, passed on to the residuals. As a consequence, the residuals are spatially correlated and the logistic model which uses the independence assumption is unsuitable. However, more complex models have been created to handle with this problem.

b) Spatial Data – The Markov Random Field Model

b.1) General Description

Markov Random Field Models (MRF), provide parametric ways of describing the spatial interactions between random variables that are spatially related. They are, in fact, an extension of the Markov Chain class in which the time component is replaced by its spatial analog and for which the Markovian property holds. To appropriately define a MRF one has to first introduce the concept of neighborhood. Following the notation of Cressie and Lele (1992) one can assume that the response variable, as a function of sites, is given by:

$$Y(S) = (Y(s_1), Y(s_2), \ldots, Y(s_n)), \qquad (3)$$

where $S = \{s_1, s_2, \ldots, s_n\}$, is the set of locations at which the response is recorded.

As Besag (1974) notes, the actual location of the sites is not important for modeling purposes. However, the spatial relationship between locations, which describes the neighboring structure of the data, does matter. Let us denote the neighborhood of the site s_i as N_i, where:

$$N_i = \{s_i \mid s_j \in S, j \neq i, s_i \,\&\, s_j \text{ neighbours}\} \qquad (4)$$

With this in mind, one can define the Markovian property for MRFs as:

$$Prob(Y(s_i) \mid Y(S \setminus \{s_i\})) = Prob(Y(s_i) \mid \{Y(s_j) : s_j \in N_i\}) \qquad (5)$$

That is, the probability of $Y(s_i)$ is conditional only on the information at neighboring locations of s_i. Assuming that the data have a joint probability of the response at all sites as given below,

$$Prob(Y(S)) = Prob((Y(s_1), Y(s_2), \ldots, Y(s_n)), \qquad (6)$$

one is interested in expressing it in terms of conditional probabilities, as given by the Markovian property. The Hammersley-Clifford theorem comes to help, providing the form that the joint probability density of a Markov random field must take. The density function is written as a sum of functions, where each function is expressed in terms of only those variables whose sites form a clique (a "clique" is every site or set of sites where each element is a neighbor of every other element in the set). It has also been shown (Besag, 1974; Cressie & Lele, 1992) that under sufficient conditional probability specifications Markov random fields can be obtained. MRF random fields have been derived from many members of the exponential family of conditional distributions, among which one can count beta, gamma and auto-logistic random fields.

Although conditional models often provide a simpler and more intuitive understanding of reality, Besag (1974) argues that the specification of processes in some applications is more natural in terms of joint distributions rather than conditional ones and thus the motivation of deriving the joint probability from conditional ones. However, this is a contentious remark and the use of conditional probabilities can often be the preferred choice.

b.2) Neighboring Structure

Markov Random Fields' usefulness is driven from their ability to account for the spatial correlation present in the data by using the neighboring structure of the observations. Since this is an essential element of MRFs, the way in which neighbors are selected has a major impact on the effectiveness of the model. Depending on the structure of the map, the individual locations (i.e geographical areas, specific locations etc.) and the random variables used a number of spatial situations can be encountered. Some examples of such scenarios are:

- regular structures at specific sites with binary variables – such as presence and absence of tree infestation in a thinned pine plot (Note: the BC data analyzed in Chapter 4 fits this structure)
- regular structures of regions with continuous variables – such as orchard plots, where the total fruit yield is measured for the combined number of trees existing in each component of the plot
- irregular structures of regions with discrete variables – such as the number of species of birds present within each county of a particular state

In the context of binary data, one is mainly concerned with dichotomous outcomes on regular or irregular lattice structures. Assuming that the data was collected on a rectangular lattice structure, one could define the neighbors of a site in a number of ways. In a first order scheme on a rectangular structure, the neighbors of a site are defined as the four immediate sites adjacent to the given sight. A second order scheme, on the other hand, also takes into account the four diagonally adjacent neighbors of the site. Higher order schemes can be implemented, but first and second order schemes are the most frequently used in order to explain the spatial correlation present in the responses.

Irregular structures, on the other hand, require different methods of specifying the neighbors of a site. For geographical data, one could consider a county whose neighbors are the bordering counties. In other cases, neighbors could also be defined in terms of the proximity to the current location, or in terms of all sites located within a certain area. Unlike regular lattices, the sites on irregular structures can have varying numbers of neighbors and thus, require somewhat different models and ways of estimating the spatial correlation.

b.3) Auto-logistic Models and Model Dependencies

In his 1974 paper, Besag showed that the exponential family of conditional probabilities together with the neighboring structure can be used to generate Markov Random Fields. He named this class of models, which include gamma, poisson, exponential and normal MRFs as the auto model class.

For binary response data, the coined term corresponding to the auto model class is auto-logistic model and the expression for its conditional probability resembles the one for logistic regression. The major difference is that, in the case of auto-logistic models, the response variable rather than the covariates is directly involved in the model where they appear as individual observations. The full conditional probability of 2^{nd} order auto-logistic model is given on the next page:

$$\text{Prob}(\text{Y}(\text{s}_\text{i}) \mid \text{Y}(\text{s}_\text{j}), \text{s}_\text{j} \in N_i) = \frac{\exp\{Y(s_i)[\beta_0 + \sum_{1\le j\le n} \beta_{i,j} * Y(s_j)]\}}{1+\exp(\beta_0 + \sum_{1\le j\le n} \beta_{i,j} * Y(s_j))]} \qquad (7)$$

The capacity to model spatial dependence is one of the reasons why MRF models are widely used. As seen in section b.2), they can be applied to wide variety of spatial situations and offer a lot of flexibility in defining the neighboring structure of the sites. Also, the spatial dependence can be divided into orientations of dependence (i.e. into directional spatial components) and each of them can be estimated individually. For example, in a viticulture experiment, one could study whether the spatial pattern of infestation within neighboring orange trees located in the same row in an orchard is stronger than the spatial pattern of infestation amongst neighboring orange trees in adjacent rows. Similarly first and second order neighbor spatial effects could be compared.

One problem with the auto-logistic model defined in (7) is that it does not allow for useful covariates to be introduced in the model. However, the auto-logistic regressive model comes to help. With the specification of covariates, the auto-logistic regression model is able to better model the relationship between the binary response and the explanatory variables while also taking into account the spatial dependence of the responses. Greig, Porteous and Seheult (1989) described such a model whose conditional probability is as derived by Zhu, Hunag and Wu (2005) is given below:

$$f(Y(s_i) \mid Y(s_j), \underline{X}_{s_j,k}, s_j \in N_i) =$$

$$= \frac{\exp\{Y(s_i)[\sum_{k=1}^{p} \theta_k * X_{s_i,k} + \sum_{1\le j\le n} \beta_{i,j}(2Y(s_j)-1)]\}}{1+\exp\{\sum_{k=1}^{m} \theta_k * X_{s_i,k} + \sum_{1\le j\le n} \beta_{i,j}(2Y(s_j)-1)\}} \qquad (8)$$

The $\beta_{i,j}$'s in model (8) are the auto-regression coefficients with $\beta_{i,j} = \beta_{j,i} = \beta$, and $\beta = 0$ if s_i and s_j are not neighbors, while $\theta_1,\ldots,\theta_p$ are the p covariate coefficients.

The above model provides a better attempt at explaining the variability in the response values by using additional covariate information. However, a problem that arises is that covariates themselves can sometimes contain spatial correlation and it becomes difficult to distinguish between the spatial dependence accounted for by covariates and the one explained by neighboring dependence.

c) The Markov Chain Model for spatial-temporal data

c.1) General Description and Dependencies Modeled

In many situations spatial data are gathered throughout time, with observations being taken at each location on a regular basis. This is precisely the case in beetle infestation problems, where the infestation status together with covariate information is recorded yearly for each location within the area of interest. Due to the cyclical and cluster-like nature of the infestation, spatial and temporal dependence is induced in the response variable. Spatial-temporal Markov Chain models can be employed to capture both the spatial and temporal dependence among the observations.

Generally Markov Chains are processes where the current state of the process is dependent on a few previous states and are primarily utilized to capture time dependencies in the data. An extensive number of applications varying from weather forecasting, to gambling problems, to random walks can all be modeled as Markov Chain processes. Depending on the nature of the response measured at various times this processes can be discrete or continuous in nature.

In the case of binary data, Markov Chains can be used to model dependencies of current dichotomous response values on their immediate past values. In its most simplistic form a model that captures the time dependence from the previous time unit has the form:

$$\operatorname{logit}[P(Y_t = 1 \mid Y_{t-1})] = \beta_0 + \beta_1 * Y_{t-1} \quad (9)$$

However, in many situations covariate information is available and it can be included in the model via logistic regression. Assuming the covariates X_1, $X_2,\ldots,X_p$ are available, the one step Markov Chain model becomes:

$$\text{logit}[P(Y_t = 1 \mid Y_{t-1})] = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + ... + \beta_p * X_p + \alpha * Y_{t-1} \quad (10)$$

It often happens that temporal dependence is already present in the covariates themselves, especially when the covariates are time dependent. However, the temporal term in the above model captures the time unit to time unit variability present in the data that was not explained by the covariates included. Despite the fact that model (10), in its current form, cannot account for spatial dependence in the response variable, it is possible to incorporate neighboring information from previous or current time units to explain the spatial dependence. An example of such a spatial-temporal model is used by Zhu, Huang, and Wu (2005). The full conditional distribution of their model at location i and time t is given below:

$$p(Y_{i,t} \mid \{Y_{j,t} : (j,t) \in N_{i,t}) =$$

$$= \frac{\exp\{\sum_{k=0}^{p} \theta_k X_{k,i,t} Y_{i,t} + \sum_{i\sim j} \theta_{p+1} Y_{i,t}(2Y_{j,t} - 1) + \theta_{p+2} Y_{i,t}(2Y_{j,t-1} + 2Y_{i,t+1} - 2)\}}{1 + \exp\{\sum_{k=0}^{p} \theta_k X_{k,i,t} Y_{i,t} + \sum_{i\sim j} \theta_{p+1}(2Y_{j,t} - 1) + \theta_{p+2}(2Y_{j,t-1} + 2Y_{i,t+1} - 2)\}} \quad (11)$$

The neighborhood of location i at time t, considered in model (11) is defined as $N_{i,t} = \{(j,t) : j \sim i\} \cup \{(i,t-1),(i,t+1)\}$ and consists of all pairs (j,t) of neighboring locations of site i and time t, as well as the points corresponding to location i and previous time t-1, respectively future time t+1. $\theta_1,\ldots,\theta_p$ are the covariate coefficients, θ_{p+1} is the coefficient of the spatial covariate, while θ_{p+2} is the coefficient of the temporal covariate.

The above model contains covariate information, as well as spatial and temporal terms. The spatial component is based on neighboring information from the present, while the temporal component includes information from the current location recorded during the previous and immediately following one lag time units. The spatial-temporal MC model used later differs from this model, in that it only employs information from the previous time state, for both the temporal and the spatial covariate. The spatial covariate becomes thus a spatial-temporal covariate, but the nomenclature has been kept to emphasize the fact that neighboring information is used in its computation.

II. Inference Procedures

a) Inference Procedures for Logistic Regression

In the case of logistic regression, the estimation of the model parameters is done via maximum likelihood. Conditional on covariate information the responses at each location and time are considered to be independent of each other and the joint probability is expressed as a product of conditional distributions. The log likelihood is computed together with the score equations and the MLEs that optimize the likelihood function are found in an iterative fashion. That is the gradient and the log likelihood are evaluated at the current estimates and the numerical algorithm iteratively improves the parameter estimates until the gradient is sufficiently close to zero. Various numerical approximation methods (such Newton-Raphson) can be employed to compute the MLEs since there are no closed form expressions for these estimates. The covariance matrix is the inverse of the matrix of second derivatives (i.e inverse Fisher information matrix) and its computation gives the standard errors of the parameters.

b) Inference Procedures for Markov Random Field Models (and for the spatial-temporal MC Model)

Upon deciding on a MRF model to be fit to the data at hand, the question of interest becomes how to estimate the model parameters. A large number of academic articles exist in the literature regarding parameter estimation for MRFs. Among the proposed methods of estimation one can count pseudo-likelihood (PL), generalized PL and Monte Carlo Maximum Likelihood (MCML).

When dealing with ordinary logistic regression, maximum likelihood estimation is the method of choice. However, in the case of auto-logistic models where spatial correlation exists among the observations, the independence assumption which is necessary for conducting maximum likelihood estimation is invalid, and no closed expression for the likelihood function can be specified. One way to approximate the likelihood function is by using MCML. The method uses Monte Carlo re-sampling and it provides consistent and asymptotically normal estimates. However, the method is computationally inefficient. Pseudo-likelihood (Besag, 1975) and coding (Besag, 1974a) were the common estimation methods before MCML was introduced and before the technological advances that have made computations easier.

In order to examine the way PL works let us assume that we are dealing with a model with conditional probability defined in terms of neighboring locations. Provided that once we condition on the values at neighboring locations the observations are independent, the PL is but the product of full conditional likelihood functions.

$$\mathrm{PL}(\delta, \beta \mid \mathrm{Y(S)}) = \prod_{s_i \in S} \frac{\exp\{Y(s_i)[\beta_0 + \sum_{s_j \in N_i} \beta * Y(s_j)]\}}{1 + \exp(\beta_0 + \sum_{s_j \in N_i} \beta * Y(s_j))} \qquad (12)$$

where β=0 if s_i,s_j are not neighbors.

Thus, MPL estimators are the parameter values that maximize PL(δ, β | Y(S)). Although in the case of large data sets, the consistency and asymptotic normality of these estimators holds, because Y(s_i) and {Y(s_j), $s_j \in N_i$ } are not independent, the pseudo-likelihood is not the true likelihood function, except in the simple case of independence between observations. For this and other reasons MCML is sometimes a better method of estimation, especially when the spatial correlation is significant.

Another problem one is faced with when using PL, is that the standard errors of the estimates do not have a closed form. Numerous re-sampling techniques of estimating standard errors such as parametric bootstrap exist throughout the literature, but they rely heavily on generating data under the updated parametric models and with the underlying dependence structure of the MRF. This is not always an easy task. One way to overcome this difficulty is to estimate the standard errors via jackknife estimation. Lele (1991) describes a way to come up with proper estimates and standard errors using a jackknife procedure. In the usual jackknife, one observation is deleted from the data at a time, and the remaining observations are used to obtain the needed estimates. Yet, removing observations from serially correlated datasets causes problems with the jackknife. To overcome this problem a component of the pseudo likelihood is deleted instead and the formulae from below are used to estimate the model coefficients and their standard errors. This method will be used in the applied sections of chapters 3 and 4 for both the MRF model and the spatial-temporal MC model.

$$\mathrm{JK}\delta_n = \delta_n - \frac{n-1}{n} * \sum_j (\delta_{n,-j} - \delta_n), \quad B_{nj} = \delta_{n,-j} - \delta_n$$

$$\text{s.e.}(\mathrm{JK}\delta_n) = (n-1) * \sum_{i=1}^{n} \sum_{j \in N(i)} (B_{ni} - \overline{B_n})(B_{ni} - \overline{B_n}) \quad (13)$$

where δ_n are the original pseudo likelihood estimates, $\delta_{n,-j}$ are the estimates from removing the j-th component of the pseudo likelihood, the Bs are the biases and the N(i) is the neighborhood of location i.

III. Model Selection and Adequacy

a) Model Selection

Model selection is important in making sure that only the relevant covariates are kept in the model. Knowing which factors are relevant in explaining the variability in the response is an important question for researchers. In general, a candidate model may generate an improved likelihood while at the same time leading to an over fitted system. If model selection is solely based on ML (or least squares), the higher the number of variables present in the model the better the fit. However, the superiority of various models is ranked according to goodness of fit as well as other criteria: the principle of parsimony, complexity of the model parameters etc. The AIC or Akaike Information Criterion is a method of selecting the model with the best likelihood while penalizing for model complexity according to the formula:

$$\text{AIC} = -2 * \log L(\theta_1,\ldots, \theta_k| \text{ data}) + 2 * k \quad (14)$$

where $\theta_1,\ldots, \theta_k$ are the MLE parameter estimates.

While this is a useful way of selecting the appropriate covariates in a logistic model, the ML cannot be computed when using a MRF (or a spatial-temporal MC model). However, one can repeatedly apply the Jackknife method of estimation described in section II.b), to come up with consistent estimates of the parameter standard error. The p-values can be computed thereafter and the method of backward substitution can be used to remove covariates with large p-values from the model, until all covariates left are significant at the significance level of choice (usually $\alpha = .05$). This method will be applied for the MRF and spatial-temporal MC models in chapters 3 and 4.

b) Model Adequacy

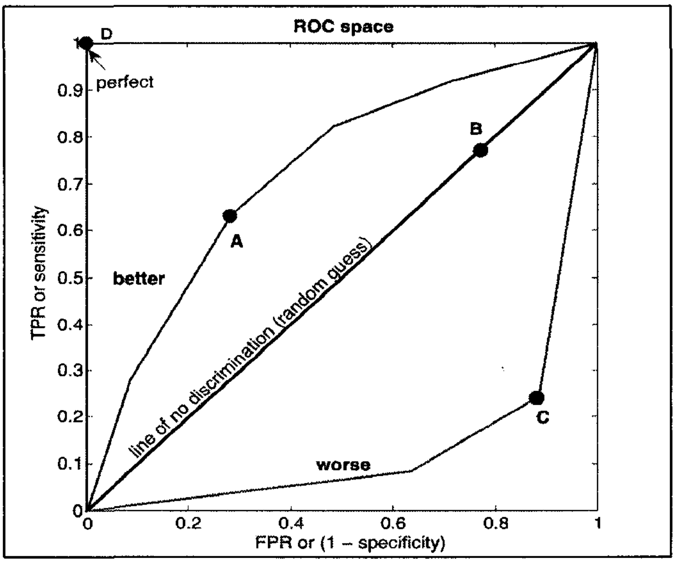
While each model has advantages and limitations, some models are better than others. To decide which models are better, one has to establish what qualifies as better and come up with a common measure of assessing it. The receiver operating curve (ROC) is a measure used to determine the adequacy of the fit when using a binary model. The ROC is a graphical portrayal of the tradeoffs between sensitivity (Sn) and specificity (Sp) (where sensitivity is the probability of detecting the true positives and specificity is the probability of detecting the true negatives). The formulae for sensitivity and specificity are given below:

$$\text{Sn} = \frac{\#\, \textit{true} \text{ positives}}{\# \text{ true positives} + \# \text{ false negatives}}$$

$$\text{Sp} = \frac{\#\, \textit{true} \text{ negatives}}{\# \text{ true negatives} + \ \# \text{false positives}} \quad (15)$$

In its traditional form the ROC consists of plotting sensitivity (Sn) on the Y axis and 1-specificity (1-Sp) on the X axis, and thus, is a measure that relates the rate of true positives and the one of false positives. The curve itself cannot be deemed good or bad, but rather the diagnostic test associated with the model can be evaluated. The plot given below displays a number of such ROC curves:

Figure 1 – ROC Curves and Cutoff Values



A cutoff value is a value between 0 and 1 on the ROC, which allows us to classify the infestation status of a location, based on the probability predicted (by the model) for that location. For example, a cutoff value of .1 means that based on it, areas with predicted values higher than .1 would be classified as infested, whereas areas with associated probabilities less than .1 would be seen as non-infested. One has to remember that each such cutoff value has an associated pairing of sensitivity/specificity which tells us how well the model is fairing in terms of its fit and based on this testing cutoff value. What one wants to see is a model under which few mistakes are made in detecting infestation or lack of it, across a relatively wide range of cutoff values. That is, one wants an ROC, which climbs quickly towards the top left corner of the plot with a large area under the curve. The diagonal line in the Figure 1, corresponds to a random test, where any cutoff value (i.e. any point on the curve) has equal Sn and 1-Sp. That is, tests based on cutoff values on this curve fare the same as if we were to decide the infestation status of a particular location by tossing a coin. Points A and D represent cutoff values on ROCs that fare much better than the random test (D more so than A). The ROCs corresponding to these points, are also characterized by areas under the curve (AUCs) that are closer to 1 than the .5 AUC value of the random test.

The values for the AUC range from 0 to 1, with 1 corresponding to a perfect test. Values between .75 and .9 are considered good, while AUCs between .9 and 1 are seen as excellent. However, these qualify as rough guidelines rather than the absolute standard.

Although the AUC is one of the most commonly used measures when comparing different ROC curves, the Youden Index (Youden, 1950) is frequently used in practice and it is computed as a function of sensitivity and specificity. The Youden Index (i.e.{Sn(cutoff) + Sp(cutoff) -1}) is a way of summarizing the test accuracy into a single value and the cutoff value that maximizes it can be an indicator of the accuracy of the predictions. This optimal cutoff is provided on all the ROC plots that appear in the following sections and the different ROC plots can be compared with one another in terms of their Youden Index values corresponding to the same cutoff value. However, the Youden Index is not the only criteria of choosing the optimal cutoff value, other measures such as efficiency and misclassification-cost existing throughout the literature (Greiner, 2000).

Although each of the models used with the data throughout this paper can be assessed in terms of ROCs, one has to be careful about their interpretation. While logistic the regression model and the spatio-temporal MC model can be used to make predictions for the immediate future, the MRF model which depends on conditioning on current response values does not have this capability. This has to be taken into account when comparing the various models by this measure.

Chapter 3: The North Carolina Data and Analysis

I. Putting things in Context

a) The Biology of the Southern Pine Beetle and the Nature of the Infestation

Dendoctronus frontalis, or the southern pine beetle (SPB), is one of the most devastating pests of pine forests in southern United States. According to Meeker, Wayne, Foltz, and Fasulo (2008) more than sixteen southern states are or have been affected by SPB from 1960 until now. It is estimated that over a span of thirty years the SPB has caused more than $900 million of damage to pine forests in the US. The infestations come in cycles which occur every six to twelve years and last for a few years. During endemic periods (low levels of infestation) the attacks are limited to weakened trees that cannot produce a large amount of resin. However, when infestations reach their peak the large number of beetles can overwhelm a tree's defense mechanisms even for the healthiest of trees.

The adult female beetles bore through the bark of pine trees into the phloem and release pheromones which attract large numbers of other pine beetles, both male and female. While the tree is somewhat capable to protect itself by producing resin, the overwhelming number of beetles that attack it during epidemics makes it impossible for the tree to repel the invasion. Not long after the original attacks, the blue-stain fungus carried by the beetles blocks the water conducting tissues of the tree eventually leading to the tree's death. Once a tree is overcrowded with pine beetles new pheromones are emitted by the beetles, causing the attacks to switch off to adjacent trees. Mating takes place soon after the initial attacks and eggs are laid in S-shaped galleries (Thatcher, 1980). The maturation cycle lasts 26 to 54 days, depending on the season, and several generations of beetles are born within each year. The reproductive cycle begins early in the spring when, after a period of inactivity, the beetles emerge from the bark. They then continue their mating and dispersal well into the summer months and early fall. According to Meeker, Wayne, Foltz, and Fasulo (2008) during high infestation years the beetles can increase their population ten-fold.

b) The NC data

The North Carolina SPB data set has been looked at before by Zhu, Huang, and Wu (2005) and contains infestation as well as covariate data gathered for each of the 100 counties of North Carolina over a period of thirty seven years (from 1960 to 1996). Among the covariates thought to impact infestation, one counts climate variables (precipitation and temperature), geographical variables (elevation, county area), as well as soil and forest characteristics such as hydric and xeric proportion, saw volume, and size of the natural forest.

While some of these variables, such as temperature, are more directly related to the spread of the infestation in terms of their impact on the reproductive cycle of the beetles, others, such as elevation and county area, might indirectly influence the infestation process by their association with climate or forest characteristics. According to Thatcher and Barry (1982) persistent temperature drops of -18C or less during winter, or long periods in excess of 35C during spring and summer can kill large numbers of broods in the Gulf States. On the other hand, low precipitation rates (during spring and summer) can have a weakening effect on the tree population (later in the year) and can thus lead to increased and more successful beetle attacks. The larger county area and higher elevation can be associated with a greater presence of pine tree forests in the region, while saw volume is a measure of the vigor of the tree and could negatively impact the infestation. While one can come up with various explanations as to why each of these variables could impact infestation, statistical methods can be applied to decide upon their significance.

II. Applying the Models to the Data

a) The Logistic Model

In order to get a clearer picture of the importance of the covariates considered above, a logistic regression model was fit to the data. The model used appears below:

$$\eta_{it} = \text{logit}(p_{it}) = \log \frac{P(Y_i = 1 \mid X_{i,1}, \ldots, X_{i,13})}{P(Y_i = 0 \mid X_{i,1}, \ldots, X_{i,13})} = \beta_0 + \beta_1 * X_{i,1} + \ldots + \beta_{14} * X_{i,14},$$

$$i=1,\ldots, 100, t=1,\ldots,36 \text{ (i.e from 1961 to 1996)} \quad (16)$$

Where:

X_1 = Saw Volume (volume of straight section of a certain length from stump height to top) (m^3) = SAW

X_2 = Proportion of land area classified as XERIC = XERIC

X_3 = Proportion of land area classified as Hydric = HYDRIC

X_4 = Mean Daily Maximum Fall Temperature (F) = MAXTF

X_5 = Mean Fall Precipitation (cm) = PRCPF

X_6 = Mean Daily Maximum Winter Temperature (F) = MAXTW

X_7 = Mean Winter Precipitation (cm) = PRCPW

X_8 = Mean Daily Maximum Spring Temperature (F) = MAXTSP

X_9 = Mean Spring Precipitation (cm) = PRCPSP

X_{10} = Mean Daily Maximum Summer Temperature (F) = MAXTSU

X_{11} = Mean Summer Precipitation (cm) = PRCPSU

X_{12} = Ln Elevation (m) = LNELEV

X_{13} = Size of Natural Forest (thousand ha) = NATFOR

X_{14} = Land Size Area (hundreds of thousands of acres)

Table 1- Coefficient Table of the Full Logistic Model

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.292e+01  3.781e+00  -3.418  0.00063 ***
SAW         -6.428e-03  1.531e-03  -4.198 2.69e-05 ***
XERIC       -5.352e-03  2.305e-03  -2.322  0.02025 *
HYDRIC      -5.282e-03  3.769e-03  -1.401  0.16115
MAXTF        1.215e-01  1.963e-01   0.619  0.53614
PRCPF        3.017e-01  1.997e-01   1.511  0.13076
MAXTW       -2.772e-01  1.373e-01  -2.020  0.04343 *
PRCPW        7.764e-01  2.400e-01   3.235  0.00122 **
MAXTSU      -1.018e-02  1.673e-01  -0.061  0.95146
PRCPSU      -5.400e-01  1.297e-01  -4.165 3.12e-05 ***
MAXTSP       2.517e-01  1.318e-01   1.910  0.05612 .
PRCPSP      -1.110e-01  3.338e-01  -0.332  0.73952
LNELEV      -1.493e-01  6.535e-02  -2.285  0.02234 *
NATFOR       7.769e-04  3.145e-03   0.247  0.80491
ACRES        1.437e-01  6.058e-02   2.372  0.01767 *
    Null deviance: 3276.0  on 3599  degrees of freedom
Residual deviance: 3132.4  on 3585  degrees of freedom
AIC: 3162.4
```

Only a few of the covariates in the above table (Table 1) are significant at a level of significance of 0.05; namely, the saw volume, the xeric proportion, the logarithm of the natural elevation, the county area, as well as some of the climate variables (the winter and summer mean precipitation rates, respectively, the mean daily maximum winter temperature). The mean daily maximum spring temperature appears to be borderline significant at 0.05. In order to eliminate the variables that appear unimportant, AIC selection was used. The coefficient table for the reduced logistic model is given below (Table 2).

Table 2 - Backward AIC selection Coefficients of the Reduced Logistic Model

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.027e+01  1.737e+00  -5.913 3.37e-09 ***
SAW         -6.951e-03  1.447e-03  -4.805 1.55e-06 ***
XERIC       -5.755e-03  1.894e-03  -3.039 0.002370 **
MAXTW       -2.730e-01  7.411e-02  -3.683 0.000230 ***
PRCPW        8.828e-01  1.137e-01   7.765 8.17e-15 ***
PRCPSU      -5.557e-01  1.046e-01  -5.313 1.08e-07 ***
MAXTSP       3.293e-01  6.500e-02   5.066 4.05e-07 ***
LNELEV      -1.761e-01  5.662e-02  -3.111 0.001867 **
ACRES        1.305e-01  5.270e-02   2.477 0.013256 *
    Null deviance: 3276.0  on 3599  degrees of freedom
Residual deviance: 3137.2  on 3591  degrees of freedom
AIC: 3155.2
```

Other than the variables that were significant in the full model, the mean daily maximum temperature during spring is now significant at 0.05. Also, the AIC has been reduced from 3162.4 to 3155.2. The reduction in the AIC is not a surprise since the variables that were removed from the model did not have a significant contribution to the fit of the model and thus, did not help reduce the likelihood substantially. Note that the AIC is based on the likelihood (a measure of goodness of fit), but also penalizes for the increase in complexity of the model caused by the increase in the number of covariates used.

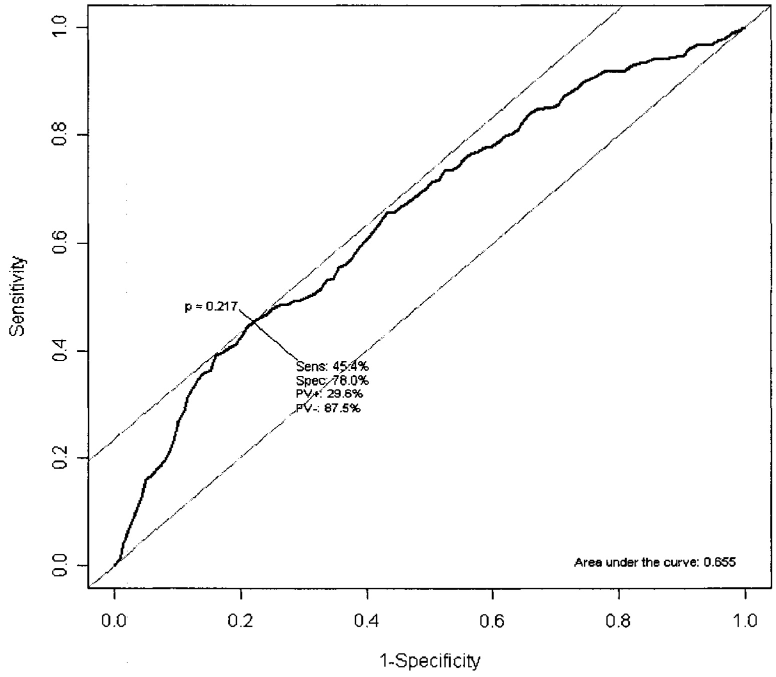
The interpretation of the coefficients provided in Table 2, is tied in with the log odds of infestation. For example, an increase of one degree F in the mean daily spring maximum temperature (MAXTSP) while everything else is kept constant, translates into an increase in the odds of infestation of $\exp\{0.3293\}$ or 1.3899. While this supports the belief that low spring temperatures can have a negative impact on the survival of the beetle, the sign of some of the other variables seems to contradict such hypotheses. According to the literature, one would expect that MAXTW would have a positive coefficient since sustained harsh winter conditions are thought to decimate the beetle population. However, the results from the logistic regression (i.e the negative sign of MAXTW) seem to indicate the opposite. This might be the result of confounding between MAXTW and some of the other covariates in the model such as LNELEV or MAXTSP.

Also, being able to assess what the sign of a covariate should be beforehand is not always straightforward. SAW volume can be seen as an expression of tree vigor and higher values of it could be associated with a higher capacity of the pine tree to repel the invasion. On the other hand, during periods of high infestation the beetles are able to attack and overwhelm large healthy trees, which are often a preferred target. That would suggest that the coefficient of SAW volume could be negatively associated with infestation rates. This seeming contradiction also holds for other variables. While higher elevations are associated with lower temperatures and a harsher climate, the presence of pine forests might be higher at higher elevations (i.e. presence of mountains).

Nevertheless, both SAW volume and the log of elevation were found to have negative coefficients, higher values for these variables leading to lower odds of infestation.

One cause of concern regarding the reduced logistic model is the fact that the reduction in the residual sum of squares between the null (overall mean) model and the current model is not considerable, which might indicate a certain lack of fit of the model and thus, lack of predictive power. To better assess the accuracy of the fit of the reduced logistic model the ROC of the reduced model was plotted (Figure 2) and is given below:

Figure 2 – ROC curve for Reduced Logistic Model

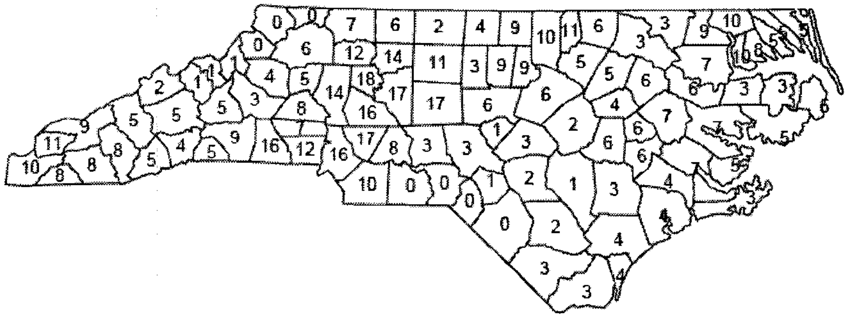


As the above plot shows, the area under the ROC is .655. Ideally, a perfect test is one that reaches a specificity and sensitivity of one and has an area below the curve equal to one. The worst case scenario corresponds to a test with area .5, where specificity and sensitivity are 50% (i.e. a test with a predictive capability no better that that of tossing an unbiased coin). Although the guidelines for what represents a good ROC area are somewhat subjective a value of .75 to .90 is considered good, while values above .9 are seen as excellent. Since the reduced logistic model has an ROC with area .66 it appears that the model is not a very good candidate when it comes to correctly identifying the pine beetle infestation status of the various counties in North Carolina. Also, the cutoff with the highest sensitivity/specificity pairing is at .217, where the sensitivity reaches 45.4% and specificity is 78.0%. While sensitivity and specificity vary quite a bit within the range of the cutoff value, the values of highest pairing for the current ROC are a sign that the predictive ability of the model is not great. However, this can be improved upon by considering the spatial dependence of the response among the counties.

b) The MRF Spatial Model

In order to improve the adequacy of the logistic model one has to take into account the presence of spatial dependence among the response values of different locations. However, to justify the inclusion of a spatial covariate, one should first examine the spatial dependence between the infestation statuses of the neighboring counties of North Carolina.

Figure 3 – NC County Map – No. of Years of Southern Pine Beetle Infestation (during 1960-1996)



The numbers in the above plot describe the number of times (in years) a county has been infested by southern pine beetle between 1960 and 1996. The graph shows that counties with higher infestation numbers seem to be surrounded by similar counties. This is an indication of spatial dependence between adjacent counties, due to factors such as the geography of the place, forest and soil characteristics etc., which tend to be more alike for adjacent counties. As one might expect some of this dependence may as well be captured by the covariates used. However, the unexplained dependence could be accounted for by using the neighboring information.

We can thus introduce a new spatial variable that combines the neighboring structure of the NC counties and the infestation status of counties within each year. That is one can count the number of infested neighbors that a county has within a particular year and assign that value to the spatial covariate, for that same year. The spatial covariate is thus time dependent. The MRF model given below contains such a covariate together with covariates already included in the logistic model, and in effect, adds a new layer of spatial complexity to the logistic model:

$$\log \frac{P(Y_{i,t} = 1 \mid \underline{X}_i, spatialCov_{i,t})}{P(Y_{i,t} = 0 \mid \underline{X}_i, spatialCov_{i,t})} = \quad (17)$$

$$= \beta_0 + \beta_1 * X_{i,1} + \ldots + \beta_{14} * X_{i,14} + \beta_{15} * spatialCov_{i,t}$$

Where:

$spatialCov_{i,t} = \sum_{j \in N_i} a_{j,i} * Y_{j,t}$ and N_i is the neighborhood of county i.

The $a_{i,j}$ takes a value of 1 if (i,j) are neighbors and 0 otherwise. Since the NC state map has the structure of an irregular lattice, different counties have different numbers of neighbors. Also, by definition, two counties are considered neighbors of each other if they share a common border. On average, the counties cover a large enough area that second order neighbors need not be considered. The table of coefficients of the MRF model (17) appears on the following page.

Table 3 – Jackknife Coefficients and St. Errors for the MRF Model

	Estimate	Std. Error	z value	Pr(>|z|)		Bias
(Intercept)	-1.023e+01	6.959	-1.471	1.412e-01		2.82081
spatialCov	1.476e+00	0.100	14.731	4.023e-49	***	0.05298
SAW	-7.362e-03	0.00830	-0.887	3.749e-01		0.00103
XERIC	7.969e-04	0.00766	0.104	9.172e-01		-0.00082
HYDRIC	1.887e-04	0.00839	0.0225	9.821e-01		0.00021
PRCPF	-2.668e-01	0.925	-0.289	7.730e-01		-0.00884
MAXTW	3.410e-01	0.436	0.781	4.345e-01		-0.01633
PRCPW	9.344e-01	0.928	1.007	3.138e-01		0.27060
MAXTSU	1.736e-01	0.544	0.319	7.496e-01		-0.08613
PRCPSU	-3.858e-01	0.301	-1.280	2.006e-01		-0.07581
MAXTSP	-2.062e-01	0.324	-0.637	5.240e-01		-0.00211
PRCPSP	2.748e-01	1.767	0.156	8.764e-01		-0.23512
LNELEV	-1.294e-01	0.144	-0.901	3.674e-01		0.00656
NATFOR	-9.805e-03	0.0125	-0.782	4.341e-01		-0.00063
ACRES	-5.539e-02	0.371	-0.149	8.814e-01		-0.07792

While the parameter estimates for the MRF model can be computed by pseudo-likelihood (calculated as the product of independent conditional distributions at each location) and are asymptotically normal (Guyon, 1986), the standard errors provided by the glm function in R are not consistent. The jackknife estimating equation method described in section II. b) of chapter 2 was used to compute consistent standard error estimates and the results are displayed in Table 4. The table also contains the parameter bias, computed as the difference between the parameter estimates found by glm and the jackknife estimates. According to the p-values displayed in Table 3, the only significant covariate at α=0.05 is the spatial covariate, spatialCov. Backward selection together with jackknife estimation can be used sequentially to eliminate the uninformative covariates, until all variables left in the model are significant at 0.05. The coefficient table of the reduced model is given below.

Table 4 – Jackknife Coefficients and St. Errors for the Reduced MRF Model

	Estimate	Std. Error	z value	Pr(>|z|)		Bias
(Intercept)	-9.317	3.714	-2.508	1.213e-02	*	1.556e-01
spatialCov	1.488	0.0907	16.415	1.487e-60	***	1.876e-02
SAW	-0.00845	0.00282	-2.992	2.771e-03	**	-9.502e-05
MAXTW	0.118	0.0578	2.042	4.111e-02	*	-1.229e-03
PRCPW	0.862	0.2164	3.981	6.853e-05	***	2.383e-02
PRCPSU	-0.444	0.233	-1.907	5.646e-02	.	-3.265e-02
LNELEV	-0.216	0.0855	-2.529	1.143e-02	*	-1.180e-02

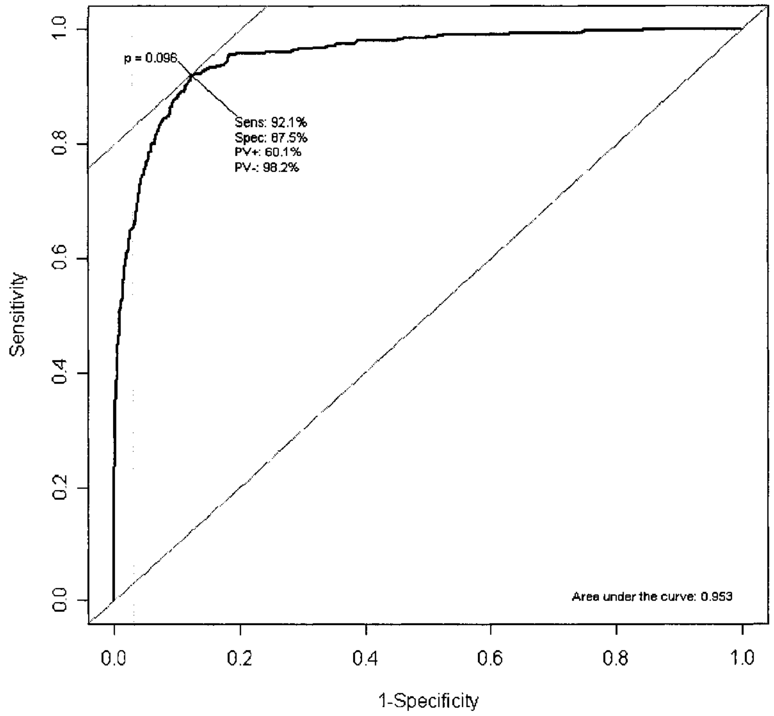
Unlike the full MRF model, the reduced MRF model contains a few significant covariates other than spatialCov. While the spatial covariate is by far the most significant variable (i.e. its p-value ~ 0), the log of natural elevation, the SAW volume, as well as the mean maximum temperature during winter, and the winter and summer precipitation rates are significant or marginally significant at 0.05; One can also see that each and every single one of these covariates were also significant in the reduced logistic model.

The bias of each of the parameter estimates is also very small, ranging from 1.1% for MAXTW to about 7.3% for PRCPSU. This is important because it is a reflection of the parameter stability of the model.

In terms of the model coefficients, the spatial covariate has a positive coefficient. That is the more infested neighbors one country has the greater the odds of infestation. In terms of actual values, an additional infested neighbor for a particular county (provided everything else stays the same) increases the odds of infestation by a factor of exp{1.488}, or 4.428. While a positive relationship was to be expected between odds of infestation and infested neighbors, the strength of the relationship is noticeable. It is also interesting to note that the sign of MAXTW is now positive which from a biological standpoint makes a lot more sense. It is possible that the removal of MAXTSP from the model has lead to the change in the sign of MAXTW, since periods of sustained cold temperatures during winter and/or spring can have a devastating impact on the spread of the beetle population.

In order to assess the usefulness of the MRF model the ROC curve was plotted and it appears on the next page.

Figure 4 – ROC Curve for Reduced MRF Model



Judging from the above plot there was a substantial increase in the AUC. While the AUC for the reduced logistic was 0.655, the AUC for the reduced MRF model is .953. At a cutoff level of .096, the pairing of sensitivity/specificity is 92.1%/87.5% (compared to 45.4%/78% at a cutoff of .211 for the logistic model), which is proof of a better fit of the new model. However, one has to be careful when comparing the AUCs of the two models, because the AUC of the reduced MRF model is based on conditioning on response values from neighboring locations from the current year. Thus, although the fit of the model is superior and the predictions based on it are excellent, one cannot think of the ROC as extremely informative when it comes to predicting things in the future because these predictions are reliant on future response values.

c) The Spatial-Temporal MC Model

The MRF model from the previous section (II.b) is superior to its reduced logistic counterpart from part (II.a) both in terms of goodness of fit as well as dependencies modeled. However, the NC data contains information gathered from all locations throughout the years and is thus a spatial-temporal data set. In order to take into account both spatial and temporal dependencies, a spatial-temporal MC model can be fit to the data. The model considered is given below:

$$\log \frac{P(Y_{i,t}=1 \mid \underline{X}_i, Y_{i,t-1}, spatialCov_{i,t}, timeCov_{i,t})}{P(Y_{i,t}=0 \mid \underline{X}_i, Y_{i,t-1}, spatialCov_{i,t}, timeCov_{i,t})} = \qquad (18)$$

$$= \beta_0 + \beta_1 * X_{i,1} + \ldots + \beta_{14} * X_{i,14} + \beta_{15} * spatialCov_{i,t} + \beta_{16} * timeCov_{i,t}$$

Where:

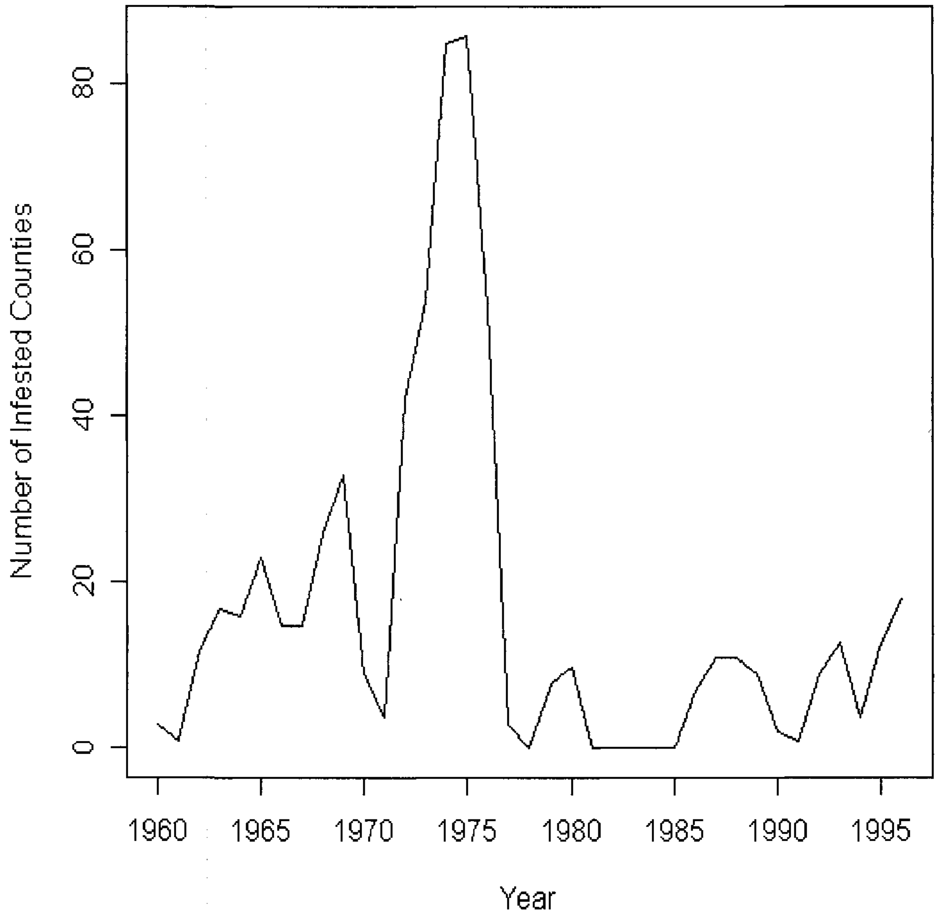
$$spatialCov_{i,t} = \sum_{j=1}^{n} a_{j,i} * Y_{j,t-1}$$

$$timeCov_{i,t} = \begin{cases} 1 & ,if \ldots Y_{i,t-1} = 1 \\ 0 & ,if \ldots Y_{i,t-1} = 0 \end{cases}$$

The above model makes use of information from the previous year and takes into account the infestation status of both neighboring locations and the current location. The spatial covariate is more appropriately thought of as a spatial temporal covariate and it basically amounts to describing the number of infested counties (adjacent to the county of interest) from the previous year. The temporal covariate has no spatial information due to the fact that it captures the infestation status of the location of interest from the previous year. One would expect both covariates to play an important role in explaining the variability of infestation since infestation occurs in cycles and previous year information can offer a clue as to where the infestation could spread. The spatial dependence of the infestation was depicted in Figure 3, while the temporal dependence can be seen by looking at the number of infested counties during each year.

As Figure 5 indicates, the infestation seems to go through cycles that span 2-4 years. The highest peak seems to have occurred during 1974-1975, when more than 80 out of the 100 counties presented infestation problems. However, the infestation seemed to die off soon afterwards, no infestations being recorded during 1981 to 1985.

Figure 5 – Number of Infested Counties in NC during each Year (1960-1996)



The coefficients of the spatial-temporal reduced MC model are given on the next page. As with the MRF model, the glm function in R was used to obtain initial pseudo likelihood parameter estimates for the MC model; jackknifing of the estimating equations was performed afterwards and backward substitution was applied after each fit to eliminate the uninformative covariates from the model.

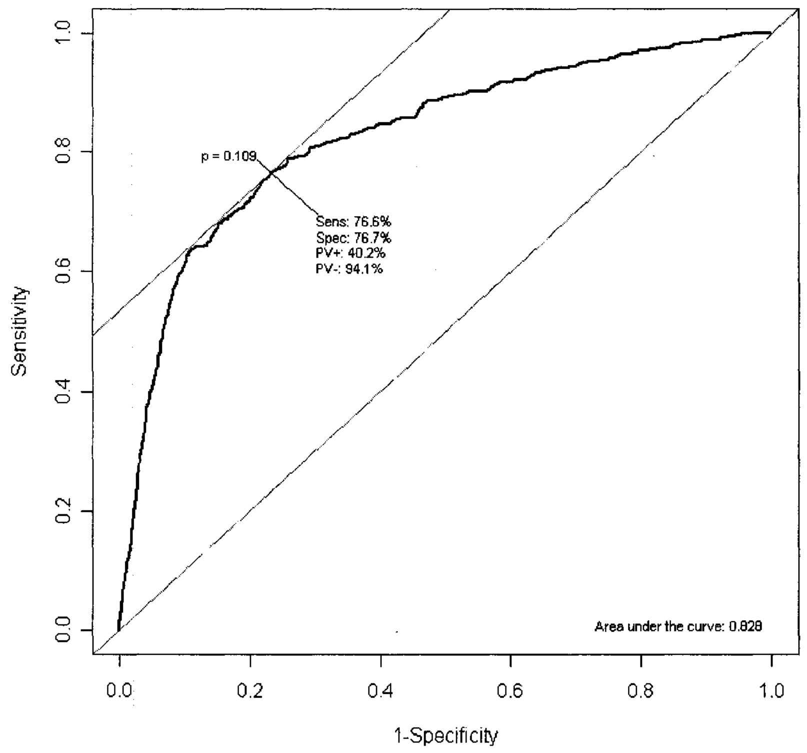
Table 5 – Jackknife Coefficients and St. Errors for the Reduced MC Model

	Estimate	Std. Error	z value	Pr(>|z|)		Bias
(Intercept)	-6.515	2.559	-2.546	1.089e-02	*	-2.279e-02
spatialCov	0.361	0.0416	8.679	3.999e-18	***	4.186e-03
timeCov	1.471	0.147	10.038	1.042e-23	***	-2.273e-02
SAW	-0.00399	0.00154	-2.592	9.542e-03	**	-5.177e-05
PRCPW	0.520	0.149	3.486	4.901e-04	***	1.718e-02
PRCPSU	-0.459	0.154	-2.987	2.817e-03	**	-2.797e-02
MAXTSP	0.0731	0.0316	2.313	2.074e-02	*	1.752e-03
LNELEV	-0.111	0.0588	-1.893	5.838e-02	.	-6.527e-03

The two most significant covariates in the Table 5 are the spatial-temporal and the temporal covariates: spatialCov and timeCov. The sign of both variables is positive indicating that a higher value for each of these covariates increases the log odds of infestation. As such, a county that has been infested during the previous year has odds of infestation $\exp\{1.471\} = 4.354$ times higher than if the country hadn't been infested. Also, having an additional infested neighbor increases the odds of infestation of a county by a factor of $\exp\{0.361\} = 1.435$ (provided everything else is kept constant). As for the remaining covariates they were also present in the reduced MRF model with the exception of MAXTSP which has replaced MAXTW (as noted before the two covariates are similar in terms of the impact they have on the beetle population). Last but not least, the biases of the parameter estimates are quite small ranging from 0.3% to a 6%, while the standard errors are consistent.

In terms of the model fit and the accuracy of the predictions, a close examination of the ROC function (Figure 6) reveals that the model is quite useful. The AUC is .828, while the pairing of sensitivity/specificity (76.6%/76.7%) reaches its peak at a cutoff of .109. While the AUC is lower than it was for the reduced MRF model, the model at hand allows for predictions to be made for the immediate future (i.e the next year) since only past information is used when fitting the model. This is an important feature that the MRF model did not have.

Figure 6 - ROC for the Reduced MC Model



d) A brief look at the at the Zhu, Huang and Wu Model

Zhu, Huang, and Wu (2005) have also analyzed the NC data set, using a spatial-temporal auto-logistic model. The model includes covariates as well as spatial and temporal components and its formula was given in chapter 2 (formula (11)). Unlike the models employed in the previous sections (i.e parts b) and c)) the spatial covariate is based on information from neighbors located within a radius of 30 miles of the current location. Different specifications for the spatial and temporal terms are also used.

While the spatial-temporal MC model in part c) counts the number of infested neighbors from the previous year for a certain location, the Zhu, Huang, Wu model considers neighboring infestation information from the present when computing the spatial covariate, and both past and future infestation values from the current location for the temporal covariate. The formulae for the spatial and temporal covariates at location i and time t are as follows:

$$spatialCov_{i,t} = \sum\nolimits_{i \sim j} Y_{i,t}(2Y_{j,t} - 1)$$

$$timeCov_{i,t} = Y_{i,t}(2Y_{j,t-1} + 2Y_{i,t+1} - 2)] \quad (19)$$

As Zhu, Huang and Wu explain, the temporal covariate is related to the mean difference between consecutive time points at the same site with same values and those with opposite values. The infestation status at location i and time t, is multiplied by 1(s) if the location was infested previously (at time t-1) or is infested the year after (at time t+1) and is multiplied by -1(s) if no infestation was present. Similarly, the spatial covariate, is a sum of ones or negative ones depending on whether the neighboring locations of location i at time t are infested or not. In addition, Zhu, Huang, and Wu make use of two interaction terms (SAW* *MAXTW, SAW*MAXTSU) and as well as transformed covariates (SAW, HYDRIC, XERIC and NATFOR are employed with their square root values).

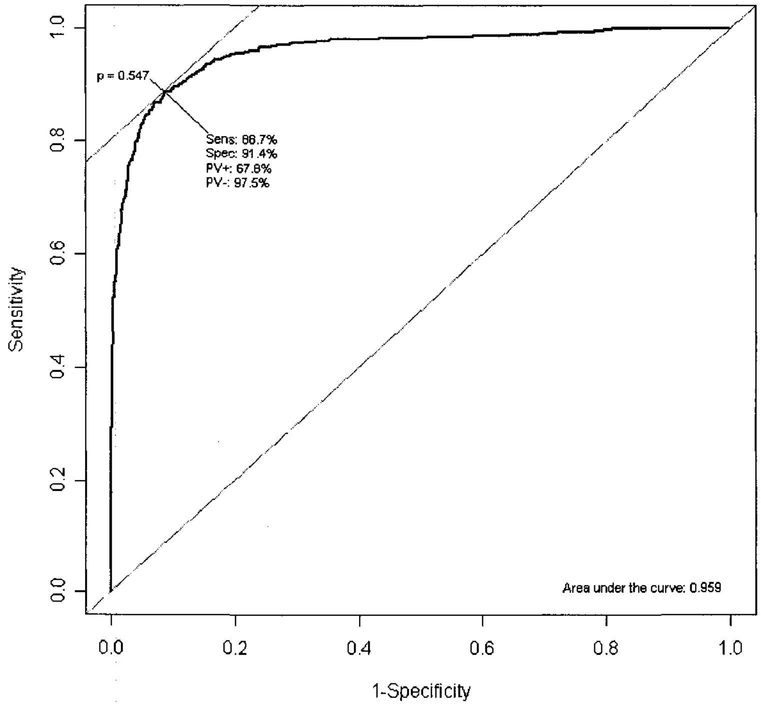
Estimation is done by pseudo likelihood, while consistent errors are computed via parametric bootstrap. The model reduction is achieved via backward substitution and optimal predictions are obtained by MCMC for the period between 1991 and 1996. The model estimates and standard errors are given in the table below.

Table 6 – Coefficients and St. Errors for the Zhu, Huang Wu Reduced Model

	Estimate	Std. Error
(Intercept)	-23.848	8.093
spatialCov	.807	.088
timeCov	.813	.121
Sqrt(SAW)	1.318	.772
Sqrt(HYDRIC)	-.068	.056
Sqrt(Xeric)	.040	.033
MAXTF	-.249	.153
PRCPF	.666	.250
MAXTSU	.515	.193
Sqrt(SAW)* MAXTSU	-.015	.009

Histograms of the predicted number of counties for each of the six years between 1991 and 1996 were obtained and, based on the mode and rounded means of these distributions, the prediction error rates were computed for each year. The yearly reported errors, calculated as the proportion of counties for which the infestation status was incorrectly predicted based on the mode of the Gibbs distributions were: .05, .15, .19, .06, .17 and .24 (for 1991 to 1996). Based on the coefficients given in Table 6 and a neighboring structure where neighbors are counties with centers located within 30 miles of each other, the ROC was plotted and is given below:

Figure 7 – ROC curve for the Zhu, Huang Wu Reduced Model



At a cutoff probability of infestation of .54 the sensitivity/specificity pairing is 88.7%/91.4%, which is indicative of an excellent fit. The ROC is comparable to the one of the MRF model from section b) and is superior to the one of the MC model in part c). Given the relatively good prediction rates as well as the temporal and spatial dependencies it captures, model (11) is thus a good candidate for predicting the MPB infest status for the NC data.

III. Summary of Models and Results

The four models considered above are options one is presented with when analyzing the NC data. The logistic model offers simplicity in modeling and is easily interpretable. Its parameters estimates and associated standard errors can be estimated by standard methods as glm in R. The model allows for covariates to be included and it therefore links the variability of different factors with the variability in the response. The logistic model offers thus a bridge to understanding the underlying factors that impact the MPB infestation. However, the fit of the model is not particularly striking (AUC .655) and the model fails to account for various dependencies in the response.

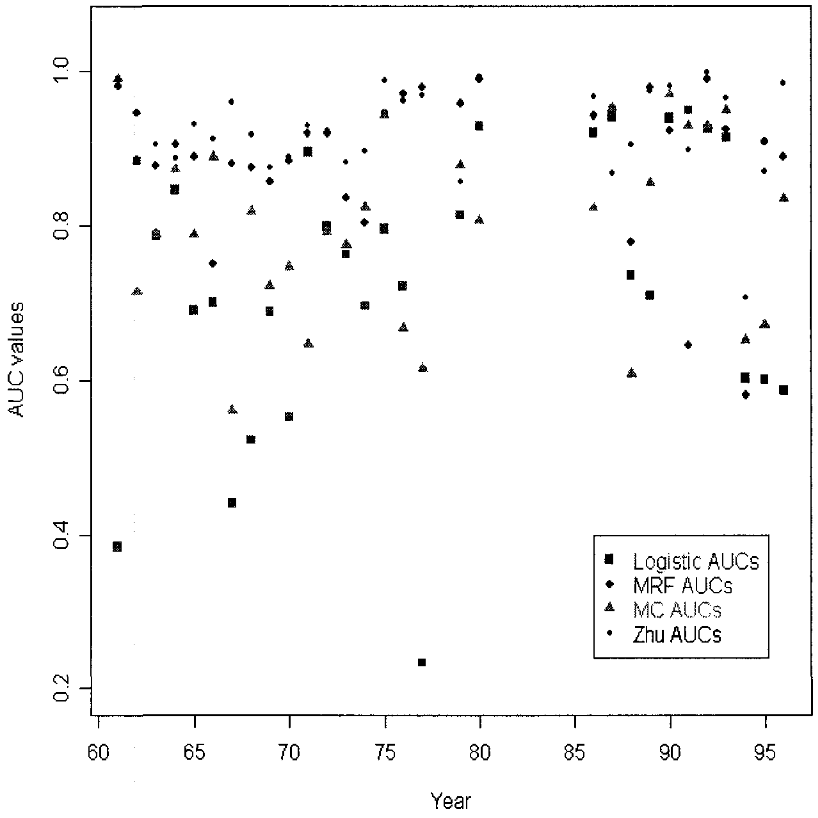
The MRF model is far better than its logistic counterpart. It allows for the spatial dependence among responses to be modeled, thus greatly improving the accuracy of the predictions (AUC .954). However, the AUC is a conditional AUC in the sense that in order for future predictions to be made, the response values at the neighboring locations from the year of prediction have to be known. Also, the increased model complexity comes at the expense of estimation (the standard errors have to be computed by a jackknifing procedure to ensure their consistency).

The MC model adds a new perspective to the analysis, allowing for temporal (as well as spatial) trends to be captured. This is often useful when data is gathered over time. Parameter estimates are obtained similarly to the MRF model. The AUC is significantly lower (.828) than the ROC of the MRF model, but large enough for the model to be considered good. Also, since only values from the previous year are used, predictions for the immediate future can be easily obtained; this is an important advantage of this model.

Last but not least, the Zhu, Huang, Wu model is a mélange between the MRF and MC models in that the spatial covariate depends on neighboring information from the current year, while the temporal covariate include past and current information. Both spatial and temporal dependencies can be captured with it, and the ROC has an excellent AUC (.959). However, this model uses a slightly different definition for the neighboring structure which makes the comparisons with the other models somewhat problematic. Also, the reliance on future values when the temporal covariate is computed is counterproductive and speculative.

The plot displayed below (Figure 8) shows the AUC values for most years during the study period (years with no infestation are excluded) for each of the models used. As expected the MRF and Zhu models have better AUC values for most years but their AUC values are conditional on current or future responses. The MC model performs worse than the aforementioned models but its predictions are still reasonable.

Figure 8 - AUC values per Year for the logistic (squares), MRF (diamonds), MC (triangles) and Zhu (circles) Models



Chapter 4: The British Columbia Data and Analysis

I. Putting things in Context

a) The Biology of the Mountain Pine Beetle and the Nature of the Infestation

Over the last decade or so, British Columbia has been confronted with a rapidly expanding Mountain Pine Beetle (MPB) infestation problem. According to the British Columbia Mountain Pine Beetle action plan for 2006-2011, the province is currently experiencing the largest MPB epidemic ever recorded in North America. The Ministry of Forests and Range estimates that at the current rate of spread more than 80% of the merchandisable pine resources in the central and southern interior parts of the province will be compromised by 2013. While in the short term some communities may benefit from the increased harvesting of trees (before their decay), the long term prospects look grim as the pine forest populations die off. The crisis is amplified by the fact that in some regions of the province pine forests make up to 50% of the harvestable timber. It is thus imperative that aggressive measures based on accurate predictions are taken, in order to reduce the extent of the infestation to controllable levels.

The MPB (Dendoctronus ponderosae) has a lifespan of about a year. The biology of the MPB is similar to the one of the SPB. The adult female beetles bore through the bark of pine trees where they lay eggs that turn into larvae and feed off the tree throughout the winter and spring months. During the initial attacks the female beetle release pheromones which attract large numbers of other pine beetles overwhelming the tree's defense mechanisms. The blue-stain fungi carried by the beetles block the water nutriment conducting tissues of the tree leading to the tree's death. Eventually, the larvae develop into adult beetle and emerge from under the bark during mid summer and only to begin their reproductive cycle.

b) The BC data

The original data were provided by Dr. Thomas de Camino-Beck and Dr. Mark A. Lewis (University of Alberta, the Department of Mathematical and Statistical Sciences) and were modified from their original form. The original data set contained the province-wide outbreak surveys, climatic data and topographic maps data for a large number of areas in British Columbia gathered over 40 years (between 1963 and 2003, except for 1996 and 1997). All maps were standardized to 100 ha pixel resolution raster maps, Albers (NAD83) projection. Canada Environment provided the Digital Elevation Map (DEM), and the terrain covariates, DEM, slope and aspect were computed using ArcGIS and rescaled to 100 ha pixel raster maps.

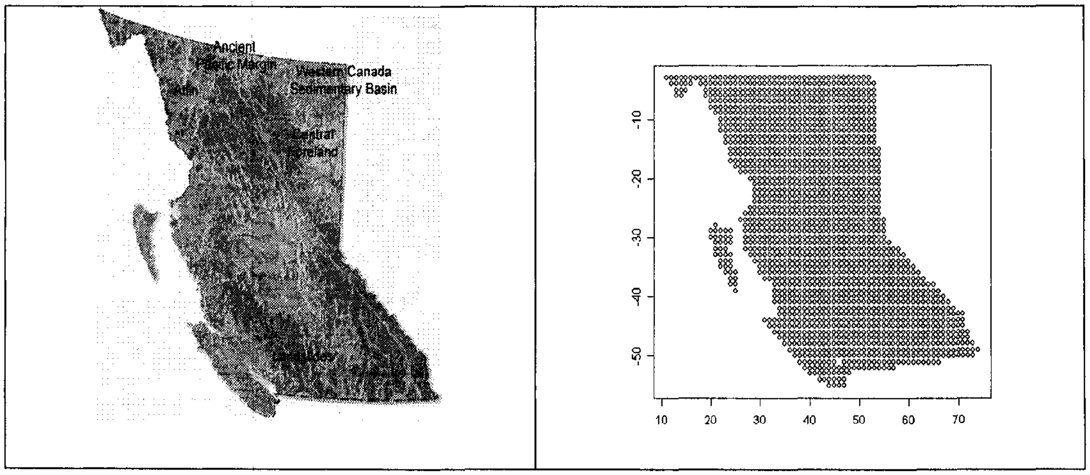
Outbreak regions presenting red top trees (in a 1:50000 map) were surveyed from the air. These surveys were digitized to vector maps, which were then converted to raster maps by superimposing a grid of 100 ha pixel units on top of the vector maps. Infestation status severity codes of 1, 2 or 3, which depict the degree of new infestations (1 for 1-10% stand killed, 2 for 11-29% of stand killed, and 3 for more than 30% of stand killed) were assigned to the 100 ha pixel units. However, the modified data contains only binary information, where only the presence of new infestations and lack thereof are represented. If an area presented any new infestations (1, 2 or 3) it was assigned a value of 1, while areas with no new infestations were given a value of 0.

In addition, the original data set was given for a grid of 1371 by 1844 one hundred hectare (1km by 1km) areas, whereas the modified data has the form of a regular grid with 55 by 74 units, with each unit corresponding to a 25km by 25km area. The map of locations considered, together with the BC map, is given on the following page (Figure 9). The reduction in scale was necessary to reduce the computational difficulties in handling the data, but the data could nonetheless be analyzed in its primary form using the very same models.

Covariate information corresponding to the new units was computed by averaging over the covariate values for the original 1km by 1km units. Among the variables considered the data set included: the tree age class, the pine coverage (representing tree and forest characteristics), the slope, aspect and elevation (as related to the terrain properties), as well as climate information such as minimum temperature and degree days (calculated as the cumulative sum of days with temperatures higher than 5.5C). While, averaging over subunits does not radically alter the interpretation of the covariates, it can have an impact on the magnitude of the model coefficients. This is because the larger variability within the smaller subunits (e.g. substantial different climate or terrain conditions within small mountainous regions) is lost when averaging is applied.

With regards to the infestation status, if any of the original units were infested, the new larger unit was deemed to be infested. However, useful information is lost by increasing the scale of the locations examined. The new grid of locations at which the infestation status is to be modeled has 4070 cells (i.e. 55*74), many of which correspond to ocean locations as well as Alaska and Alberta land areas. Of the 4070 such unit areas 1706 correspond to BC land mass areas and only these observations are used when various models are fit.

Figure 9 - BC Map and Unit Area Approximation to BC Map
(BC map as found at: http://gsc.nrcan.gc.ca/cogmaps/prov/bc_e.php)



Due to incomplete information during 1996 and 1997, the data set used was constrained to the period between 1964 and 1996. This was done in part because the Markov models rely on information from previous year(s) and missing information would create problems in terms of modeling. However, the amount of information gathered over 33 years is more than enough to assess the importance of different covariates and modeling the infestation status.

Before proceeding with the data analysis, one has to recognize the limitations of coding the infestation status in binary terms. While, binary responses can be modeled to determine the newly infested areas, they cannot be used to answer questions regarding the degree of infestation affecting the respective areas. This is an important aspect to reflect on when deciding which models should be considered, and a multinomial model might be preferred instead. Also, when modeling infestation one should be primarily focused on areas where pine forests exist since the MPBs rely heavily on them for their survival. One could inspect the map of pine forests in BC and select only those locations where pine forests exist. However, the current model contains the pine coverage covariate, which allows for pine forest presence (or lack thereof) to play a role in predicting infestations.

II. Applying the Models to the Data

a) The Logistic Model

As with the NC data, in order to assess the usefulness of the covariates considered a logistic model was fit. The model used is given below.

$$\eta_{it} = \text{logit}(p_{it}) = \log \frac{P(Y_i = 1 \mid X_{i,t,1}, \ldots, X_{i,7})}{P(Y_i = 0 \mid X_{i,t,1}, \ldots, X_{i,7})} = \beta_0 + \beta_1 * X_{i,t,1} + \ldots + \beta_7 * X_{i,7},$$

$$i=1,\ldots, 1706,\ t=1,\ldots,33 \text{ (i.e from 1964 to 1996)} \qquad (20)$$

Where:

X_1 = Degree Days = DD (cumulative sum of days with temperature >5.5C)

X_2 = Minimum Temperature = Min (C)

X_3 = Age class = Age (forest average age measured in decades)

X_4 = Pine Coverage = Pine $\qquad$ X_5 = Digital Elevation = DEM (m)

X_6 = Slope $\qquad$ X_7 = Aspect (degrees)

Some of the variables in the above model are just spatially dependent (Age, Pine, Slope, DEM and Aspect), while others depend both on location and the year the observations were recorded (DD and Min). AIC selection was used to eliminate the insignificant covariates and the reduced logistic model coefficients are given in the table from below.

Table 7 - Coefficient Table of the Reduced Logistic Model

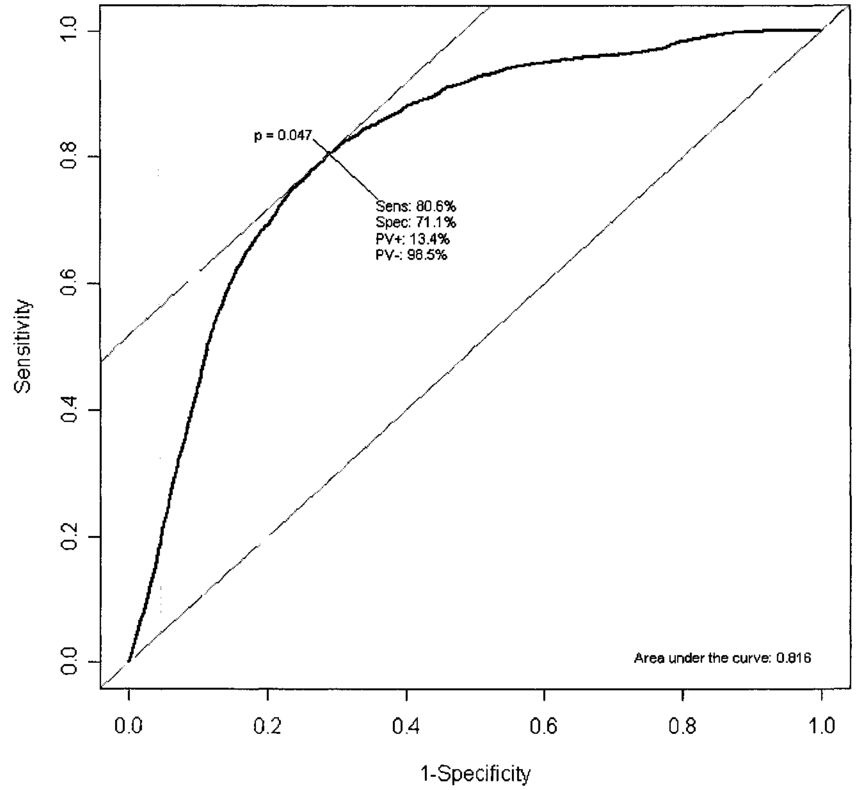
```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.688e+00  2.594e-01 -29.637  < 2e-16 ***
dd           2.835e-03  1.081e-04  26.237  < 2e-16 ***
min          9.192e-03  3.813e-03   2.411  0.01591 *
age          4.941e-02  1.536e-02   3.216  0.00130 **
pine         7.589e-04  4.541e-05  16.711  < 2e-16 ***
dem          4.011e-04  6.938e-05   5.781 7.42e-09 ***
slope        7.079e-02  7.749e-03   9.135  < 2e-16 ***
aspect      -3.356e-03  9.306e-04  -3.607  0.00031 ***
    Null deviance: 23181  on 56297  degrees of freedom
Residual deviance: 19416  on 56290  degrees of freedom
AIC: 19432
```

All the model covariates are significant at a level of significance of α=0.05. Degree days, pine coverage and slope appear to be the most significant factors in the above model. Also, aspect is the only covariate with a negative coefficient. Given these results an additional unit increase in the degree days corresponds to an increase of exp{.00284}=1.0028 in the odds of infestation, provided everything else is kept constant. Similarly, one unit (i.e. one decade) increase in the age class classification of the pine trees translates into an increase of exp{.0494}= 1.051 in the odds of infestation.

The sign of many of these coefficients has intuitive appeal. For example, more days with temperatures higher than 5.5C, translates into a more favorable overall environment for the beetle and thus, higher survival rates for the beetles; also the larger the pine coverage, the greater the probability of infestation. Minimum temperature is thought to have a detrimental effect on the life cycle of the MPB, which is exactly the case judging by the positive coefficient of the variable min in the above table. That is a decrease in the minimum temperature is followed by a decrease in the log odds of infestation. However, the sign of other variables, such as age class, might not be easily guessed. While adult trees might be a preferred choice of habitat for MPBs, they are also better at defending themselves by producing resin. Thus, an increase in age class might not straightforwardly be thought to lead to an increase in the infestation rate. Nevertheless, in the above model, age class positively impacts infestation.

Having looked at the significance of the model covariates, one can examine the fit of the model in terms of its predictive accuracy. The ROC curve for the logistic model is given below (Figure 10).

Figure 10 – ROC Curve for Simple Logistic Model



The highest pairing of sensitivity (80.6%) and specificity (71.1%) at a cutoff of .047 indicates a relatively good fit of the model. The AUC (.816) reconfirms the usefulness of the logistic regression. Although other improvements and complexities can be added, given its simplicity the model seems to behave quite well. However, one is to expect that some degree of spatial dependence exists between adjacent areas on the BC map and thus, one could try to account for such dependencies by introducing a spatial variable that taking into account neighboring information.

b) The MRF Spatial Model

Unlike the North Carolina data set, where the neighboring information was captured by using a matrix that describes the relationship between each of the 100 counties of the state (i.e. weather two counties border each other or not), the BC data has the shape of a regular lattice and the neighbors of a particular location can be described in a more systematic way. That is, one can consider as neighbors of a spatial unit the four immediately adjacent locations next to it (ie. to the north, south, east and west). The neighboring structure can be further extended to include the immediate diagonal neighbors or the neighbors located within a certain distance, but the spatial covariate defined in the current MRF model is based on the first order neighbors.

However, not all locations have four neighbors. Sites located on the edge of the BC land map, whose neighboring locations are part of the ocean, Alberta and Alaska land areas, or fall outside the considered grid, have fewer neighbors. While the focus of this paper is on using BC land covariate information and infestation status, the inclusion of AB information from locations adjacent to BC, could prove useful and should be investigated later on. That is because the two provinces share a mountainous border populated with pine forests and infestation easily can spread from one province to the other. Ocean locations, on the other hand, do not contain relevant information when it comes to modeling infestation. The MRF and MC models used in this paper discounted information from areas outside BC when computing the spatial covariate, counting only infested neighbors within the BC map. As such, sites located in corners or on lateral edges of the BC contour could have at most one, two, or three infested BC neighbors.

The MRF model for the BC data appears below:

$$\log \frac{P(Y_{i,t}=1\,|\,\underline{X}_i, spatialCov_{i,t})}{P(Y_{i,t}=0\,|\,\underline{X}_i, spatialCov_{i,t})} =$$

$$= \beta_0 + \beta_1 * X_{i,t,1} + \ldots + \beta_7 * X_{i,7} + \beta_8 * spatialCov_{i,t} \quad (21)$$

Where:

$spatialCov_{i,t} = \sum_{j \in N_i} a_{j,i} * Y_{j,t}$ and N_i is the neighborhood of county i.

The spatial covariate can only take as value a number from 0 to 4, and it denotes the number of neighbors (for a maximum of four) of location i that are infested at time t. One expects that the more infested neighbors a location has, the higher the odds of infestation at that location. The coefficient table for the above model is given below (Table 8). The parameter estimates for the MRF model were computed by pseudo-likelihood via the glm function and are asymptotically normal; consistent standard errors were obtained via jackknife.

Table 8 – Jackknife Coefficients and St. Errors for the MRF Model

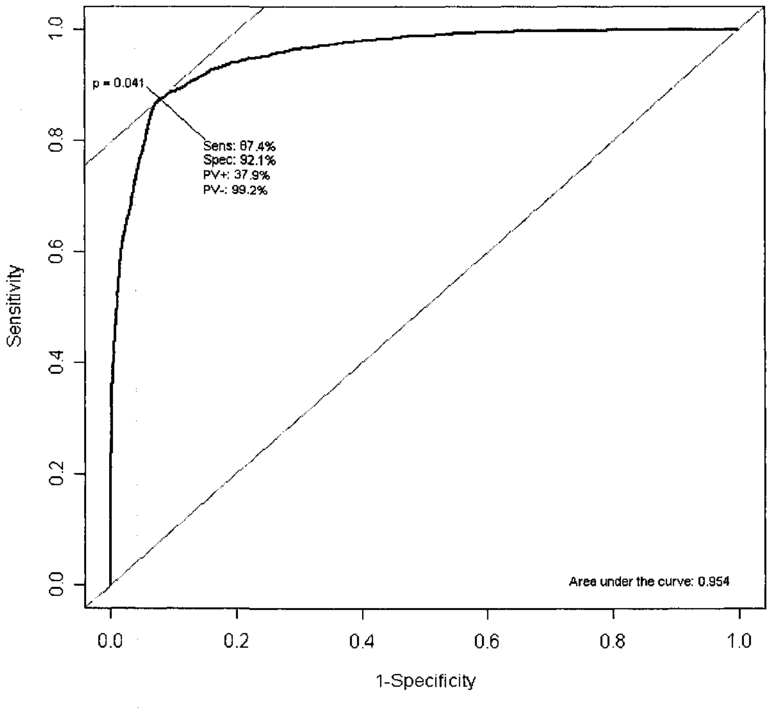
```
              Estimate   Std. Error  z value  Pr(>|z|)            Bias
(Intercept)    -7.308    3.886e-01   -18.806  6.723e-79 *** -1.108e-01
dd              0.00206  1.435e-04    14.358  9.556e-47 *** -2.315e-04
min             0.00598  1.787e-03     3.347  8.176e-04 *** -1.391e-02
age             0.0747   1.264e-02     5.911  3.398e-09 *** -3.192e-03
pine            0.000305 3.201e-05     9.529  1.588e-21 ***  6.674e-05
dem            -0.000339 6.832e-05    -4.968  6.757e-07 ***  1.191e-05
slope           0.0963   4.801e-03    20.069  1.379e-89 ***  1.039e-03
aspect         -0.00247  7.031e-04    -3.515  4.399e-04 ***  4.821e-04
spatialCov      2.0381   3.071e-02    66.368  0.000e+00 *** -1.350e-01
```

Of all covariates fitted, the spatial covariate (spatialCov) is the most significant. This is not all that surprising considering that the infestation seems to occur in clusters. As expected, the spatial covariate has a positive coefficient, an increase in the number of infested neighbors leading to an increase in the log odds of infestation. To be more precise, each additional infested neighbor of a particular location increases the odds of infestation of that location by a factor of exp{2.038}, or 7.675. In addition, the other covariates are also significant at 0.05 and only the elevation (dem) and the aspect have a negative impact on the log odds of infestation.

The estimation bias for the jackknife coefficients is small to moderate ranging from 1% to 22%, the only exception being variable min that has a large bias. This might present a problem in terms of the parameter stability of the model. However, the most significant covariates (spatialCov, dd, slope) have small associated biases and the coefficient of the minimum temperature covariate is small enough that it does not have a large impact on the odds of infestation.

In order to get a better sense of accuracy of the MRF model one can examine the ROC curve, which appears is given below (Figure 11).

Figure 11 – ROC Curve for Reduced MRF Model



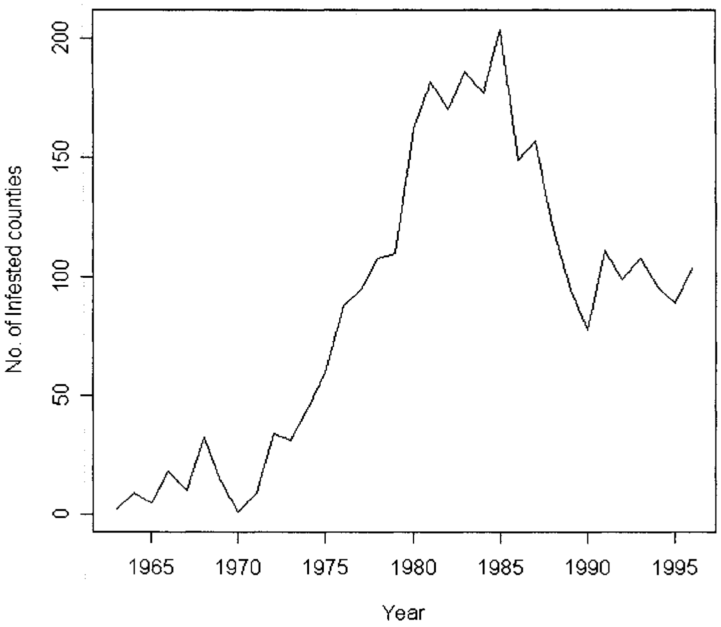
The AUC has improved significantly to a value of .954. This is an excellent value, which is 13.8% higher than the AUC of the logistic model. Although sensitivity and specificity for two different models should not be directly compared at different cutoff values, the logistic and the MRF models have very similar cutoffs at their highest pairing combination of sensitivity and specificity (.047 and .041). Compared to the logistic model, the MRF sensitivity has gone up by almost 7% to 87.4%, while the specificity has increased by 21% to 92.1%. That is, throughout the years, 87.4% of the true infestations are properly detected and 92.1% of the non-infested areas are correctly identified as not infested. The MRF model is thus much better at predicting the status of infestations.

c) The Spatial-Temporal MC Model

The MRF model is superior to the logistic model both in the dependencies it models as well as the quality of the predictions it makes. However, as in the case of the NC data, the BC data contains information gathered from all locations throughout the years and is therefore a spatial-temporal data set. One can expect that a certain degree of temporal dependence exists among the response values over time given that the infestation occurs in cycles. The plot of infestations over time (Figure 12), provided on the next page, depicts the positive correlation between the number of infestations of consecutive years. That is, periods of successive infestation increases, respectively decreases can be noticed (from 1970-1980, from 1986 to 1990).

The need for including time dependence terms is not just reflected by Figure 12, but makes sense from a biological standpoint. A model that takes into account past year neighboring infestation might be more useful in predicting whether infestation is currently present at a particular location due to the fact that the infestation takes time to spread to surrounding areas. The spatial MC model is thus a more intuitively appealing tool that better captures and uses the time lag between the original attacks and the final phase of at which infestation is detected. However, one has to be aware that the infestation status cannot always be properly assessed since it takes trees about a year time from the original attacks to turn red. This could in turn have an impact on the fit of the model, although less so for a binary model where any degree of infestation is coded with as infested.

Figure 12 - Number of Infested Locations throughout BC over Time



The spatial MC model used in conjunction with the BC data appears below:

$$\log \frac{P(Y_{i,t}=1\,|\,\underline{X}_i, Y_{i,t-1}, spatialCov_{i,t}, timeCov_{i,t})}{P(Y_{i,t}=0\,|\,\underline{X}_i, Y_{i,t-1}, spatialCov_{i,t}, timeCov_{i,t})} =$$

$$= \beta_0 + \beta_1 * DD_{t,i} + \ldots + \beta_8 * Aspect_i + \beta_9 * spatialCov_{i,t} + \beta_{10} * timeCov_{i,t}$$

Where:

$$spatialCov_{i,t} = \sum_{j=1}^{n} a_{j,i} * Y_{j,t-1}$$

$$timeCov_{i,t} = \begin{cases} 1, & if\ Y_{i,t-1} = 1 \\ 0, & if\ Y_{i,t-1} = 0 \end{cases} \qquad (22)$$

The spatial-temporal covariate (spatialCov) and the temporal covariate (timeCov) are almost identical to the once in the MC model from chapter 3, with the notable distinction that the $a_{i,j}$ values describe a regular lattice structure where each location has at most four neighbors (the first order neighbors). The model coefficients are given in (Table 9) and were obtained similarly to the coefficients of the MRF model in part b).

Table 9 – Jackknife Coefficients and St. Errors for the MC Model

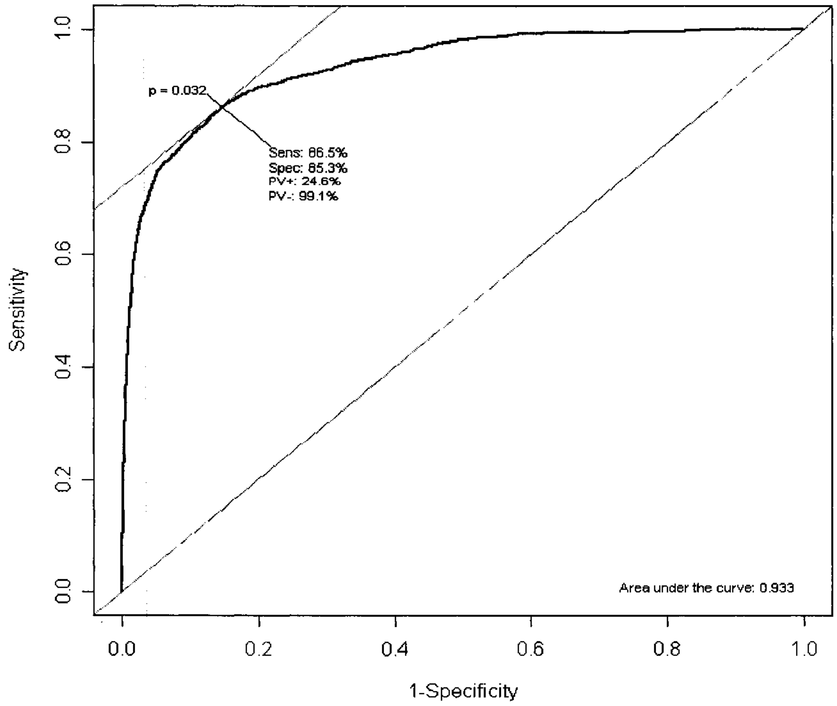
	Estimate	Std. Error	z value	Pr(>|z|)	Bias
(Intercept)	-6.716	3.563e-01	-18.846	3.162e-79	-8.628e-02
dd	0.00199	1.293e-04	15.367	2.734e-53	-2.856e-04
min	0.020	1.955e-03	10.236	1.369e-24	-1.551e-02
age	0.034	1.295e-02	2.649	8.076e-03	1.556e-03
pine	0.000372	3.261e-05	11.394	4.458e-30	5.597e-05
dem	0.000194	4.997e-05	3.876	1.059e-04	-3.986e-05
slope	0.0605	4.545e-03	13.309	2.040e-40	4.743e-03
aspect	-0.00272	6.298e-04	-4.321	1.554e-05	5.864e-04
spatialCov	0.968	1.507e-02	64.258	0.000e+00	-7.038e-02
timeCov	2.437	3.691e-02	66.035	0.000e+00	-9.042e-02

All covariates present in the model are significant after the consistent standard errors are computed; therefore, there is no need for backward selection to be performed. The temporal and the spatial-temporal covariates are once again the most significant covariates of the model. They both have positive coefficients and a large impact on the odds of infestation. Previously infested locations have exp{2.437}=11.439 higher odds of infestation than previously non infested locations with identical covariates; also, having an additional infested neighbor during the previous year increases its odds of infestation by a factor of exp{0.968}, or 2.633. Except for covariate aspect, which has a negative coefficient, all other variables positively impact on the probability of infestation.

As with the MRF model, the estimated bias of the model coefficients is small to moderate, ranging from 1% to 21%, with the exception of covariate min. Once again, this can be a sign of parameter instability and the removal of min could take care of this problem. However, the most highly significant covariates (spatialCov, timeCov) have small biases and their impact on the odds of infestation is much more pronounced.

Also the ROC can help us get a better sense of the adequacy of the model. As seen in Figure 13, the ROC has an AUC .933, which is characteristic to an excellent model. At the cutoff point of .032, the model has a sensitivity rate of 86.5% and a specificity rate of 85.3%, numbers that indicate a very good fit. Based on both the ROC and the dependencies modeled the spatial-temporal MC model seems to be a good candidate for predicting infestation.

Figure 13 – ROC for the spatial temporal MC model (for the BC data)



Although the overall results for the MC model do not differ considerably from the ones for the MRF model and the added complexity might seem inconsequential, one has to point out the fact that the main advantage of using the MC models is that one already has access to past information. Therefore, one can use these models to make predictions about the immediate future. This was not the case with the MRF models since the current state of infestation at neighboring locations was needed in order for predictions to be made. This is one reason why the MC model might constitute a better choice.

d) Variations on a theme - A brief look at Another Spatial-Temporal MC Model

The BC data set in its original form, as provided by Dr. de Camino-Beck and Dr. Lewis, divides the BC map into 1km by 1km unit areas. The model used in part c) was applied to a modified version of the BC data as described in Section I.b) of this chapter. However, a variety of models similar to model spatial-temporal model in part c) have been applied by Dr. de Camino-Beck to the original data, and the model having the best prediction accuracy has been selected and appears on the next page.

$$\log \frac{P(Y_{i,t}=1 \mid \underline{X}_{i,t}, \underline{X}_{i,t-1}, Y_{i,t-1}, spatialCov_{i,t})}{P(Y_{i,t}=0 \mid \underline{X}_{i,t}, \underline{X}_{i,t-1}, Y_{i,t-1}, spatialCov_{i,t})} =$$

$$= \beta_0 + \beta_1 * \min_{t,i} + \beta_2 * dd_{i,t} + \beta_3 * dd_{i,t-1} + \beta_{10} * spatialCov_{i,t}, \quad (23)$$

where spatialCov is defined as the number of infested neighbors within a radius of two unit areas of location i, from the previous year (t-1). The only covariates present in the model were the minimum temperature from the current year as well as the degree days from both the present and the previous year. The table of coefficients is given below:

Table 10 – Coefficients of the Model (23)

```
                  Coeff. Estimate
(Intercept)          -1.7209
    min               0.1905
    dd_t              0.0044
    dd_t-1           -0.0025
 spatialCov           0.7092
```

While different versions of the BC data were used with models (21) and (22), the coefficients provided in the above table have the same sign and have relatively similar values to the ones in Table 9. The spatial covariate (although somewhat different) has a similar coefficient to the one found in part c), and positively impacts the odds of infestation. Yet, given the differences in the neighboring information used as well as the number of locations considered, direct comparisons can be somewhat misleading. The model was found to have an ROC with an AUC of .85 and was deemed as good for making predictions.

III. Summary of Models and Results

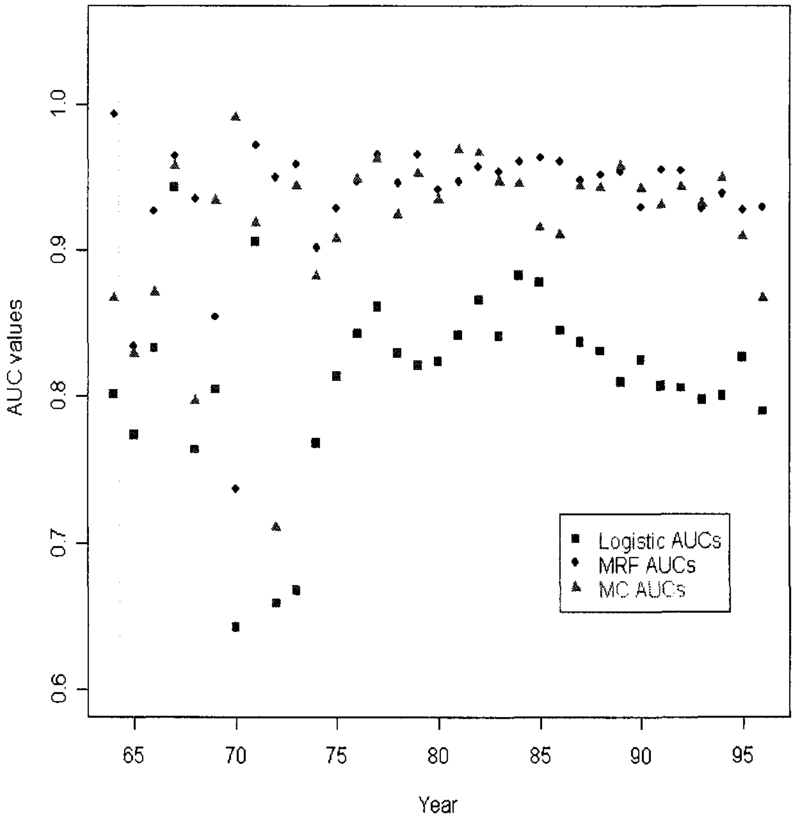
As in the case of the NC data a few models were used to predict the infestation (MPB infestation) status of various locations. The logistic model is easy to fit and interpret and, unlike its NC counterpart, it offers a much better fit to the data (AUC .816). All covariates considered were significant at 0.05 and together they accounted for a large amount of the variability in the infestation. However, the model fell short when it comes to accounting for various dependencies.

The MRF model used had an excellent fit. Spatial dependence among responses was modeled by including the neighboring infestation status at each location, thus greatly improving the accuracy of the predictions (AUC .954). However, its AUC is a conditional AUC in the sense that in order for future predictions to be made, the response values at the neighboring locations from the year of prediction have to be known. Once again, all covariates were significant at 0.05, with the spatial covariate having the largest on the odds of infestation.

Last but not least, the MC model allowed for temporal (as well as spatial) trends to be captured. This is often useful when data is gathered over time. Both the spatial-temporal and temporal covariates were found highly significant and they both positively impact the odds of infestation. Unlike the NC data, the AUC of the MC model (.933) was comparable to the AUC of the MRF model (.954). However, since the MC model relies only on previous information predictions into the near future are easy to make. Combined with the ability to model both temporal and spatial dependencies this is a better overall model than the MRF.

The plot given on the following page (Figure 14) displays the year by year AUC fit of the logistic, MRF and MC models. While, the logistic model has the worst AUCs of the three models for every single year, the MRF and MC models have similar AUC values for most years and outperform each other at times.

Figure 14 - AUC values per Year for the logistic (squares), MRF (diamonds) and MC (triangles) Models



Chapter 5: Conclusions

Good models for predicting the status of potentially MPB or SPB infested areas are of utter importance for the lumber and forestry sectors. The infestation is wide spread in many states and provinces throughout North America. Mountain pine beetle affected areas extend from BC and Alberta (Canada) to northern Mexico, while southern pine beetle infestation problems have been registered from Pennsylvania to Honduras. Due to the destructive nature of the infestation as well as the staggering economic costs associated with it, prevention measures (such as pheromone baiting, controlled burns, sanitation harvesting etc.) are of great consequence. However, to limit the cost of such interventions, accurate prediction and forecasts of infested areas play an important role.

Among the models examined in this paper one counts logistic, markov random field and markov chain models. All these models present advantages, but also have important limitations and one has to decide upon which models are more useful. An important role in obtaining good models is the choice of the covariates used. Weather related covariates, geo-terrain variables, forests characteristics and pine beetle reproductive cycle factors should all be considered as a way of improving the accuracy of the fit and predictions.

However, as it is apparent from the results provided throughout this paper the inclusion of spatial and/or temporal covariates can greatly improve the precision of simpler models. Taking into account neighboring information, as well as information from previous years can lead to much better predictions. This was evident from the superior fit of the Markov random field and Markov chain models, compared to the simple logistic models. However, while the fit of the MRF models was mostly excellent, their reliance on current neighboring response values was a clear disadvantage when making future predictions. The MC models did not have this problem and were more flexible in terms of modeling time dependencies.

A number of changes could be made to these models in an attempt to improve their accuracy. One could consider new covariates, apply transformations to existing covariates and make use of interactions. Also, one could alter the way the spatial covariate is used by using a neighboring structure that gives different weights to various neighbors depending on the distance from a location to its neighbor. The spatial dependence could also be split into directions of dependence. This is particularly useful if one thinks of the impact wind direction has on the spread of infestation. Directional dependence (anisotropy) can looked at by dividing neighbors into classes of neighbors, based on location (north, south etc.), and examining their individual effect on predicting the infestation status.

As for the time covariate, one could take into account not just the information from the last year, but also from years before last year depending on the patterns that emerge from looking at the average length of the infestation cycle of the beetle population, as well as the periodicity of the infestations.

However, there exist other ways of modeling spatial and temporal dependence that can be explored. Hierarchical models are one such example. Random effects can be quite useful in capturing unexplained variability among observations by assuming that within the various levels of the hierarchy observations are more alike. Also variance covariance structures can be used to describe the spatial interdependence of responses at various locations within each year. In addition, the parameter variability caused by unaccounted factors could be modeled by assigning informative and non-informative priors to these parameters. One such model could have the following form:

$$\underline{Y}_j \mid \underline{\eta}_j \sim Bernoulli\left(\frac{\exp(\underline{\eta}_j)}{1+\exp(\underline{\eta}_j)}\right),\ \underline{\eta}_j \sim MVN\left(\underline{\mu}_j, \Sigma\right) \qquad (24)$$

$$\mu_{ij} = \beta_0 + \beta_1 * X_{i,j,1} + \ldots + \beta_p * X_{i,j,p}, \text{ where } \beta_0, \beta_1, \ldots, \beta_p \sim N(0, \sigma^2)$$

and $\underline{\Sigma}$ has a Wishart or a more structured form.

Improving the accuracy of predictions by using the abovementioned models together with their variants is somewhat of an art form. Each model offers benefits and has its own caveats. However, the guiding principles in choosing the right model are the quality of predictions and a better understanding of the underlying processes that impact infestation. It is based on these principles that proper measures can be taken to address the economic burdens caused by the pine beetle on society.

A multitude of actions can be taken to control and prevent the spread of infestation. Since the beetles spend a large part of their life under the bark of trees, control methods can focus on killing the developing beetles before they emerge as adults from under the bark. Currently infested trees can be cut and burned, can be buried under the soil until the beetles have been killed, or they can be debarked and chipped. Infested logs can also be transported to safe sites, located far away from any susceptible tree hosts. However, the implementation of such measures has associated economic costs and places a heavy burden on the limited resources that can be used to address the infestation problem. The models used throughout this paper allow for probabilities of infestation to be computed at all sites. These probability maps can then be used to address the spread of infestation, by removing all the trees located in areas adjacent to heavily infested sites, which have probabilities of infestation higher than a certain desired level of threshold. That is, the economic cost associated with the implemented control measures can be minimized by targeting areas of a specific size as determined based on the probability maps of infestation.

On the other hand, the probability maps themselves can be impacted on by the control measures taken. The removal of infested trees from a particular area can change the future infestation status of that area even if infestation has been registered within the area at the time of the tree removal. This is important for modeling purposes because, in some models, the infestation status is used as neighboring information and it directly impacts the model outcome. Also, the removal of red top trees does not guarantee that the infestation is not still present since it takes trees about one year from the original attacks to change color. This can once again affect the future infestation status, as infested locations may be perceived as not being infested. It becomes thus clear that the control measures used in handling infestation have an impact on the data used. Combining effective models applied to reliable data and appropriate control measures is the key to addressing the infestation problem.

Bibliography

1. Agesti, A. (2002) "Ch 4 : Introduction to Generalized Linear Models," pp.115-165 in *Categorical Data Analysis 2nd Ed.* Hoboken, New Jersey: Wiley & Sons

2. Amman, G.D., McGregor, M.D. and Dolph, R.E. (reprinted 1990) "Mountain Pine Beetle. Forest Insect and Disease Leaflet 2," US Department of Agriculture Forest Service. Retrieved July 25, 2008 from: http://www.fs.fed.us/r6/nr/fid/fidls/fidl2.htm

3. Besag, J. E. (1972) "Nearest-Neighbour Systems and the Auto-Logistic Model for Binary Data," *Journal of the Royal Statistical Society. Series B(Methodological)* 34: 75-83

4. Besag, J. E. (1974) (1972a) "Spatial Interaction and the Statistical Analysis of Lattice Systems," *Journal of the Royal Statistical Society. Series B(Methodological)* 36:192-236

5. Besag, J. E. (1975) "Statistical Analysis of Non-Lattice Data," *The Statistician* 24:179-195

6. British Columbia. Ministry of Forests and Range (update March 2008) "Pine Beetle attacks tops 700 million cubic meters." Retrieved July 25, 2008 from: http://www.for.gov.bc.ca/hfp/mountain_pine_beetle/Update_MPB_Volume_Estimate.pdf

7. British Columbia. Ministry of Forests and Range (2006) "Mountain Pine Beetle Action Plan 2006-2011." Retrieved July 25, 2008 from: http://www.for.gov.bc.ca/hfp/mountain_pine_beetle/actionplan/2006/Beetle_Action_Plan.pdf

8. Camino-Beck, T., Lele, S., Metes, D. and Lewis, M. (2007) "Markov process logistic regression to predict mountain pine beetle outbreaks," *Ecological Applications* (manuscript submitted).

9. Carroll, A., Regniere, J., Logan, J., Taylor, S., Bentz, B. and Powell, J. (2006) "Impacts of climate change in range expansion by the mountain pine beetle". Natural Resources Canada, Canadian Forest Service, Pacific Forestry Centre, Victoria, BC. Mountain Pine Beetle Initiative Working Paper. 27p. Retrieved on July 25, 2008 from: http://warehouse.pfc.forestry.ca/pfc/26601.pdf

10. Cressie, N. and Lele, S. (1992) "New Models for Markov Random Fields," *Journal of Applied Probability,* 29:877-884

11. Florida Department of Agriculture and Consumer Services & University of Florida. (revised 2008) "Featured Creatures: southern pine beetle." Based on work by Meeker, J.R., Dixon, W.N., Foltz, J.L. and Fasulo, T.R. Retrieved July 25, 2008: http://creatures.ifas.ufl.edu/trees/southern_pine_beetle.htm

12. Geizler, D.R., Gallucci, V.F. and Gara, R.I. (1980) "Modeling the dynamics of mountain pine beetle aggregation in a lodgepole pine stand," *Oecologia,* 46:244-253

13. Gibson, K. (2004) "Mountain Pine Beetle: Conditions and Issues in the Western United States," US Department of Agriculture Forest Service. Retrieved July 25, 2008 from: http://www.for.gov.bc.ca/hfd/library/MPB/gibson_2004_mount.pdf

14. Greig, D. M., Porteous, B. T. and Seheult, A. H. (1989) "Exact Maximum a Posteriori Estimation for Binary Data," *Journal of the Royal Statistical Society. Series B (Methodological),* 51:271-279

15. Greiner, M., Pfeiffer, D., Smith R.D. (2000) "Pricipals and application of the reciver operating characteristic analysis for diagnostic tests," *Preventive Veterinary Medicine* 45:23-41

16. Gumpertz, M.L., Wu, C.T., and Pye, J.M. (2000) "Logistic Regression for Southern Pine Beetle Outbreaks With Spatial and Temporal Autocorrelation." *Forest Science,* 46: 95-107

17. Lele, S. (1991) "Jackknifing linear estimating equations: Asymptotic theory and applications in stochastic processes," *Journal of the Royal Statistical Society, Ser. B* 53:253-268

18. Mitchell, R. and Preisler, H. (1991) "Analysis of spatial patterns of lodgepole pine attacked by outbreak populations of the mountain pine beetle," *Forest Science* 37: 1390-1408

19. Regniere, J. and Bentz, B.J. (2007) "Modeling cold tolerance in the mountain pine beetle," *Journal of Insect Physiology,* 53:559-572.

20. Thatcher, R.C. and Barry P.J. (1982) "Southern pine beetle. Forest and Disease Leaflet 49," US Department of Agriculture Forest Service. Retrieved July 25, 2008 from: http://www.na.fs.fed.us/spfo/pubs/fidls/so_pine_beetle/so_pine.htm

21. Thatcher, R.C., Searcy, J.L., Coster, J.E. and Hertel, G.D. (1980) " The Southern Pine Beetle," US Department of Agriculture Forest Service, Expanded Southern Pine Beetle Research and Application Program. Retrieved July 25, 2008 from: http://www.barkbeetles.org/spb/spbbook/Index.html

22. Youden, W.J. (1950) "Index for rating diagnostic tests," *Cancer,* 3:32-35

23. Zhu J., Huang, H.C. and Wu, J. (2005) "Modeling Spatial-Temporal Binary Data Using Markov Random Fields," *Journal of Agricultural, Biological and Environmental Statistics,* 10:212-225

Appendix

Section A: R code for the North Carolina data set

1. Computing the Neighborhood Matrix

<u>Neighbors (</u>*creation of a matrix whose elements describe if two counties are neighbours (=1) or not (=0))*

```
neighbors <- read.table(“C:\\TDATA\\NCN.txt”)
attach(neighbors)

matNeigh <- matrix(0, nrow=100, ncol=100)
for (j in 1:100) {
    for (i in 1:9)  {
        if (neighbors[j,i+2] >0) {
            neigh <- ( neighbors[j,i+2] + 1) / 2
            matNeigh[neigh,j] <- 1 }
    }
}
```

2. Creating the spatial covariate (information neighbors at present time) and adding it to the data frame that contains all other response and independent covariates.

```
library(MASS)
infestData <- read.table(“C:\\TDATA\\AllData1.txt”,header=TRUE)
infestData$ACRES = infestData$ACRES/100000
attach(infestData)
ninfest <- dimnames(infestData)[2]

noobs <- length(infestData[,1])
noyears <-37
nocounties <-100
prevyear <- c(rep(0,nocounties))
thisyear <- c(rep(0,nocounties))
spatialCov <- c(rep(0, noobs))
identVect <- c(rep(1, nocounties))

for (i in 1:(noyears-1)) {
     for (m in 1:nocounties) {
     thisyear[m] <- InfestStat[noyears*(m-1)+i+1] }
     for (j in 1:nocounties) {
     spatialCov[(j-1)*noyears+i+1] <- matNeigh[j,]%*%thisyear
     }
}
```

```
infestData <- as.data.frame(cbind(infestData,spatialCov))
noobsNew <- (noyears-1)*nocounties
infestDataNew <- infestData[1:noobsNew,]

infDataNew <- subset(infestData, Year != “Y60”)

infestData <- infDataNew
attach(infestData)
spatialCov <- infestData$spatialCov
timeCov <- infestData$timeCov
```

3. Fitting the logistic regression and using AIC selection to eliminate uninformative covariates

```
logitModel <- glm(InfestStat ~ SAW + XERIC + HYDRIC+ MAXTF+ PRCPF+
MAXTW+ PRCPW+ MAXTSU+ PRCPSU+ MAXTSP+ PRCPSP+ LNELEV+
NATFOR + ACRES, data = infestData,family="binomial")
summary(logitModel)

logitModelRed <- stepAIC(logitModel, direction=”backward”)
summary(logitModelRed)
```

4. Ploting the overall ROC and retaining the AUC values for each year

```
p = fits = fitted(logitModelRed)
ROC( test =p, stat=InfestStat, plot="ROC",MI=FALSE)
dat <- cbind(infestData,fits)
dat <- dat[order(dat[,2]),]
AUClog = rep(0,times=36)
infestperyear = table(infestData[,2],infestData[,3])[2:37,2]

for (i in 1:36) {
    if (infestperyear[i]!=0) {
        a = ROC(test=dat[(100*(i-1)+1):(100*i),20], stat=dat[(100*(i-
1)+1):(100*i),3],plot=”ROC”,MI=FALSE)$AUC
        AUClog[i] = round(a,3)
    }
}
plot(c(61:96),AUClog)
```

5. Fitting the MRF model

```
logitModelMRF <- glm(InfestStat ~ spatialCov + SAW + XERIC + HYDRIC+
MAXTF+ PRCPF+ MAXTW+ PRCPW+ MAXTSU+ PRCPSU+ MAXTSP+
PRCPSP+ LNELEV+ NATFOR +ACRES, family="binomial")
summary(logitModelMRF)
```

6. Obtaining Jackknife Estimates of the MRF model repeatedly while using backward selection to eliminate insignificant covariates after each round of estimation

```
logitModelMRF <- glm(InfestStat ~ spatialCov + SAW + MAXTW+ PRCPW+
PRCPSU+ LNELEV, family="binomial")

jackBetaJs <- function(removedJ) {
# the function removes county J from the data (analogous to removing component
j and
# its 36 obs from the log likelihood and refits the model to get new coefficients

data <- infestData[infestData[,1]!=removedJ,]
newlogit<- glm(data$InfestStat ~ data$spatialCov + data$SAW +
data$MAXTW+ data$PRCPW+ data$PRCPSU+ data$LNELEV,
family="binomial")
newCoeffs <- newlogit$coeff

return(newCoeffs)

}

jackEstimates <- function() {

origCoeffs <- logitModelMRF$coeff
nEsts = length(origCoeffs)
jackmatrix <- matrix(0, 100, nEsts)
jackEsts <- rep(0, times = nEsts)
Rn <- jackmatrix
Rnbar <- rep(0, times = nEsts)
jackVarEsts <- rep(0, times = nEsts)

for (j in 1:100) {
     jackmatrix[j,] = jackBetaJs(j)
     }

for (k in 1: nEsts) {
     Rn[,k] = jackmatrix[,k] – origCoeffs[k]
     Rnbar[k] = mean(Rn[,k])
     }

for (i in 1: nEsts) {
     jackEsts[i] = origCoeffs[i] - (100-1)/100 * sum(jackmatrix[,i]-origCoeffs[i])
     }

neigh = matNeigh +diag(100) #matrix of neighbours
for (k in 1: nEsts) {
    for (i in 1:100) {
        for (j in 1:100) {
            jackVarEsts[k] <- jackVarEsts[k] +(100-1)/100*neigh[i,j]*(Rn[i,k]-
Rnbar[k])* (Rn[j,k]-Rnbar[k])
            }
        }
    }


m=0
for (i in 1:100) {
    for (j in 1:100) {
        if (neigh[i,j]==1) m=m+1
        }
    }
print(m)

l = list(diffs =Rnbar, ests = jackEsts, stdev = sqrt(jackVarEsts))
return(l)
}

ests = jackEstimates()
ests
```

7. Creating the table of coefficients, p-values, biases after jackknifing

```
tablecoeff <- summary(logitModelMRF)$coeff
tablecoeff[,1] = ests[[2]]
tablecoeff[,2] = ests[[3]]
tablecoeff[,3] = ests[[2]]/ests[[3]]
tablecoeff[,4] = pnorm(abs(ests[[2]]/ests[[3]]),lower.tail=FALSE)*2
bias = (summary(logitModelMRF)$coeff)[,1]-ests[[2]]
tablec <- cbind(tablecoeff,bias)
dimnames(tablec)[[2]][5] = “Bias”
```

8. Obtaining the fitted values and ROC for the reduced MRF model and retaining the AUC values for each year

```
estims <- ests[[2]]
estims = c(-9.31706391,1.48809808,-.00844693,.11815228,.86162541,-
.44434204,-.21619010)
xcovvals <- cbind(c(rep(1,dim(infestData)[1])), infestData[,c(19,5,10,11,15,16)])
p = fits = exp(as.matrix(xcovvals) %*% estims)/(1+exp(as.matrix(xcovvals) %*%
estims))
```

```
ROC(test = p, stat=InfestStat, plot="ROC",MI=FALSE)

dat <- cbind(infestData,fits)
dat <- dat[order(dat[,2]),]
AUCmrf = rep(0,times=36)
infestperyear = table(infestData[,2],infestData[,3])[2:37,2]

for (i in 1:36) {
    if (infestperyear[i]!=0) {
        a = ROC(test=dat[(100*(i-1)+1):(100*i),20], stat=dat[(100*(i-
1)+1):(100*i),3],plot=”ROC”,MI=FALSE)$AUC
        AUCmrf[i] = round(a,3)
    }
}
plot(c(61:96),AUCmrf)
```

9. Creating the spatial covariate (inf of neighbors at previous time), and temporal covariate – inf at the existing location at previous time)) – for the simple Markov Chain and fitting the model

```
library(MASS)
infestData <- read.table(“C:\\TDATA\\AllData1.txt”,header=TRUE)

infestData$ACRES = infestData$ACRES/100000
attach(infestData)
ninfest <- dimnames(infestData)[2]

noobs <- length(infestData[,1])
noyears <-37
nocounties <-100
prevyear <- c(rep(0,nocounties))
thisyear <- c(rep(0,nocounties))

spatialCov <- c(rep(0, noobs))
identVect <- c(rep(1, nocounties))

for (i in 1:(noyears-1)) {
     for (m in 1:nocounties) {
     prevyear[m] <- InfestStat[noyears*(m-1)+i]
     thisyear[m] <- InfestStat[noyears*(m-1)+i+1] }
     for (j in 1:nocounties) {
     spatialCov[(j-1)*noyears+i+1] <- matNeigh[j,]%*%prevyear
     }
}

timeCov <- c(rep(0, noobs))
```

```
for (i in 1:(noyears-1)) {
     for (m in 1:nocounties) {
          prevyear[m] <- InfestStat[noyears*(m-1)+i]
          thisyear[m] <- InfestStat[noyears*(m-1)+i+1] }
     for (j in 1:nocounties) {
           if (prevyear[j]==1)  timeCov[(j-1)*noyears+i+1] = 1
     }
}
infestData <- as.data.frame(cbind(infestData,spatialCov,timeCov))
noobsNew <- (noyears-1)*nocounties

infestDataNew <- infestData[1:noobsNew,]
infDataNew <- subset(infestData, Year != "Y60")

infestData <- infDataNew
attach(infestData)
spatialCov <- infestData$spatialCov
timeCov <- infestData$timeCov
```

10. Fitting the MC model

```
logitModelMC <- glm(InfestStat ~  spatialCov + timeCov + SAW + XERIC +
HYDRIC+ MAXTF+ PRCPF+ MAXTW+ PRCPW+ MAXTSU+ PRCPSU+
MAXTSP+ PRCPSP+ LNELEV+ NATFOR + ACRES, family="binomial")
summary(logitModelMC)
```

11. Obtaining Jackknife Estimates of the MRF model repeatedly while using backward selection to eliminate insignificant covariates after each round of estimation

```
logitModelMC <- glm(InfestStat ~  spatialCov + timeCov + SAW  + PRCPW+
PRCPSU+ MAXTSP+ LNELEV, family="binomial")

jackBetaJs <- function(removedJ) {
# the function removes county J from the data (analogous to removing component
j and
# its 36 obs from the log likelihood and refits the model to get new coefficients

data <- infestData[infestData[,1]!=removedJ,]
newlogit<- glm(data$InfestStat ~ data$spatialCov + data$timeCov + data$SAW +
data$PRCPW+ data$PRCPSU+ data$MAXTSP+ data$LNELEV ,
family="binomial")
newCoeffs <- newlogit$coeff

return(newCoeffs)
```

```
}

jackEstimates <- function() {

origCoeffs <- logitModelMC$coeff
nEsts = length(origCoeffs)
jackmatrix <- matrix(0, 100, nEsts)
jackEsts <- rep(0, times = nEsts)
Rn <- jackmatrix
Rnbar <- rep(0, times = nEsts)
jackVarEsts <- rep(0, times = nEsts)

for (j in 1:100) {
     jackmatrix[j,] = jackBetaJs(j)
     }

for (k in 1: nEsts) {
     Rn[,k] = jackmatrix[,k] – origCoeffs[k]
     Rnbar[k] = mean(Rn[,k])
     }

for (i in 1: nEsts) {
     jackEsts[i] = origCoeffs[i] - (100-1)/100 * sum(jackmatrix[,i]-origCoeffs[i])
     }

neigh = matNeigh +diag(100) #matrix of neighbours

for (k in 1: nEsts) {
    for (i in 1:100) {
         for (j in 1:100) {
              jackVarEsts[k] <- jackVarEsts[k] +(100-1)/100*neigh[i,j]*(Rn[i,k]-
Rnbar[k])* (Rn[j,k]-Rnbar[k])
              }
         }
    }


m=0
for (i in 1:100) {
    for (j in 1:100) {
        if (neigh[i,j]==1) m=m+1
        }
     }
print(m)

l = list(diffs =Rnbar, ests = jackEsts, stdev = sqrt(jackVarEsts))
```

```
return(l)
}

ests = jackEstimates()
ests
```

12. Creating the table of coefficients, p-values, biases of the MC model after jackknifing

```
tablecoeff <- summary(logitModelMC)$coeff
tablecoeff[,1] = ests[[2]]
tablecoeff[,2] = ests[[3]]
tablecoeff[,3] = ests[[2]]/ests[[3]]
tablecoeff[,4] = pnorm(abs(ests[[2]]/ests[[3]]),lower.tail=FALSE)*2
bias = (summary(logitModelMC)$coeff)[,1]-ests[[2]]
tablec <- cbind(tablecoeff,bias)
dimnames(tablec)[[2]][5] = “Bias”
```

13. Obtaining the fitted values and ROC for the reduced MC model and retaining the AUC values for each year

```
estims <- ests[[2]]
estims = c(-6.514676379,.360959885,1.471268377,-.003997193,.519654290,-
.458554532,.073069097,-.111390885)
xcovvals <- cbind(c(rep(1,dim(infestData)[1])),
infestData[,c(19,20,5,11,15,12,16)])
p = fits = exp(as.matrix(xcovvals) %*% estims)/(1+exp(as.matrix(xcovvals) %*%
estims))
ROC(test = p, stat=InfestStat, plot="ROC",MI=FALSE)

dat <- cbind(infestData,fits)
dat <- dat[order(dat[,2]),]
AUCmc = rep(0,times=36)
infestperyear = table(infestData[,2],infestData[,3])[2:37,2]

for (i in 1:36) {
    if (infestperyear[i]!=0) {
        a = ROC(test=dat[(100*(i-1)+1):(100*i),21], stat=dat[(100*(i-
1)+1):(100*i),3],plot=”ROC”,MI=FALSE)$AUC
        AUCmc[i] = round(a,3)
    }
}
plot(c(61:96),AUCmc)
```

14. Zhu model - redefining neighboring structure based on neighbors within 30 miles

```
#Find neighbours within 30 miles;
#read latitudes: lat
#red longitudes: long

coords <- read.csv(“m:\\new thesis\\latlongforNCdata.csv”)
lat = coords[,1]
long = coords[,2]

n = 100
dist = matrix(0,n,n)
neighdist = matrix(0,n,n)

for (i in 1:n) {
    for (j in 1:n) {
        R =6371
        torad = pi/180
        dlat = lat[i]-lat[j]
        dlong = long[i]-long[j]
        a = sin(dlat/2*torad)^2 + cos(lat[i]*torad)*cos(lat[j]*torad)*
(sin(dlong/2*torad)^2)
        c = 2*atan2(sqrt(a),sqrt(1-a))
        d = R *c
    dist[i,j] = d
    if (dist[i,j] <= 30*1.6 & dist[i,j]!=0) {neighdist[i,j]=1
    }
    }
}
matNeigh = neighdist
```

15. Computing the spatial and temporal covariates for the Zhu model and fitting the model

```
infestData <- read.table(“m:\\new thesis\\AllData1.txt”,header=TRUE)
infestData$ACRES = infestData$ACRES/100000
attach(infestData)
ninfest <- dimnames(infestData)[2]

noobs <- length(infestData[,1])
noyears <-37
nocounties <-100
prevyear <- c(rep(0,nocounties))
nextyear <- c(rep(0,nocounties))
thisyear <- c(rep(0,nocounties))
```

```
spatialCov <- c(rep(0, noobs))
identVect <- c(rep(1, nocounties))

for (i in 1:(noyears-1)) {
    for (m in 1:nocounties) {
    thisyear[m] <- InfestStat[noyears*(m-1)+i+1] }
    for (j in 1:nocounties) {
    #if(thisyear[j]==1) spatialCov[(j-1)*noyears+i+1] <-
matNeigh[j,]%*%(2*thisyear-1)
      spatialCov[(j-1)*noyears+i+1] <-matNeigh[j,]%*%(2*thisyear-1)
    }
}

timeCov <- c(rep(0, noobs))
for (i in 1:(noyears-1)) {
    for (m in 1:nocounties) {
        prevyear[m] <- InfestStat[noyears*(m-1)+i]
        thisyear[m] <- InfestStat[noyears*(m-1)+i+1]
        if (i!=36) nextyear[m] <- InfestStat[noyears*(m-1)+i+2] }
    for (j in 1:nocounties) {
        # timeCov[(j-1)*noyears+i+1] = thisyear[j]*(2*prevyear[j] +
2*nextyear[j] -2 )
        timeCov[(j-1)*noyears+i+1] = 2*prevyear[j] + 2*nextyear[j] -2
    }
}

infestData <- as.data.frame(cbind(infestData,spatialCov,timeCov))
noobsNew <- (noyears-1)*nocounties
infestDataNew <- infestData[1:noobsNew,]

infDataNew <- subset(infestData, Year != “Y60”)

infestData <- infDataNew
attach(infestData)
spatialCov <- infestData$spatialCov
timeCov <- infestData$timeCov

sqSAW =sqrt(SAW)
sqXERIC =sqrt(XERIC)
sqHYDRIC =sqrt(HYDRIC)
sqNATFOR =sqrt(NATFOR)

logitModelZhu <- glm(InfestStat ~  sqSAW + sqHYDRIC + sqXERIC
+MAXTF+ PRCPF+ MAXTSU +sqSAW*MAXTSU+ spatialCov+ timeCov,
family="binomial")
summary(logitModelMC)
```

16. Ploting the ROC of the ZHU model and retaining the AUCs per year

```
library(Epi)
p <- exp(fitted(logitModelZhu))/(1+exp(fitted(logitModelZhu)))
ROC( test=p, stat=infestStat, plot="ROC")

dat <- cbind(infestData,p)
dat <- dat[order(dat[,2]),]

AUCzhu = rep(0,times=36)
infestperyear = table(infestData[,2],infestData[,3])[2:37,2]

for (i in 1:36) {
    if (infestperyear[i]!=0) {
        a = ROC(test=dat[(100*(i-1)+1):(100*i),21], stat=dat[(100*(i-
1)+1):(100*i),3],plot=”ROC”,MI=FALSE)$AUC
        AUCzhu[i] = round(a,3)
    }
}
plot(c(61:96),AUCzhu)
```

17. Plotting the AUCs by year of the logistic, MRF, MC and Zhu models (years with no infestations excluded)

```
AUClog1 = AUClog[AUClog!=0]
AUCmrf1 = AUCmrf[AUCmrf!=0]
AUCmc1 = AUCmc[AUCmc!=0]
AUCzhu1 = AUCzhu[AUCzhu!=0]

windows()
plot(c(61:77,79,80,86:96),rep(-1,30),main="AUCs per Year for the logistic (red),
MRF(blue), MC(green) models", xlab="Year",ylab="AUC values",
ylim=c(0.2,1.05)) #tck=1,

points(c(61:77,79,80,86:96),AUClog1,pch=15,col=2)
points(c(61:77,79,80,86:96),AUCmrf1,pch=16,col=4)
points(c(61:77,79,80,86:96),AUCmc1,pch=17,col=3)
points(c(61:77,79,80,86:96),AUCzhu1,pch=20,col=1)
legend(86,.4,c(“Logistic AUCs”,”MRF AUCs”,”MC AUCs”,”Zhu AUCs”),
pch=c(15,16,17,20), col = c(2,4,3,1), text.col = c(2,4,3,1)) #,bg=8
```

Section B: R code for the British Columbia data set

1. Reading in the data

```
BCinfestdata <- read.table(“C:\\BCdata.txt”,header=TRUE)
attach(BCinfestdata)
```

2. Creating the spatial covariate (information neighbors at present time) and adding it to the data frame that contains all other response and independent covariates.

```
BCdata <- BCinfestdata[order(BCinfestdata$year),]
attach(BCdata)
years <- 34
n <- dim(BCdata)[1]
nsquares <- n/years
ncols <- 74
nrows <- 55

spatialCov <- rep(0,times=n)

for (i in (nsquares+1):n) {
    if (dd[i] != 0) {
       if (blck1[i] != 1  &&  blck1[i] != nrows && blck2[i] != 1 && blck2[i] !=
ncols) {
           if (infest[i -1]==1) spatialCov[i]=spatialCov[i]+1
           if (infest[i+1] ==1) spatialCov[i]=spatialCov[i]+1
           if (infest[i-ncols] ==1) spatialCov[i]=spatialCov[i]+1
           if (infest[i+ncols] ==1) spatialCov[i]=spatialCov[i]+1
       }
   else if (blck1[i] == 1  &&  blck1[i] != nrows && blck2[i] != 1 && blck2[i] !=
ncols) {
           if (infest[i -1] ==1) spatialCov[i]=spatialCov[i]+1
           if (infest[i+1] ==1) spatialCov[i]=spatialCov[i]+1
           if (infest[i+ncols] ==1) spatialCov[i]=spatialCov[i]+1
       }
   else if (blck1[i] != 1  &&  blck1[i] == nrows && blck2[i] != 1 && blck2[i] !=
ncols) {
           if (infest[i -1] ==1) spatialCov[i]=spatialCov[i]+1
           if (infest[i+1] ==1) spatialCov[i]=spatialCov[i]+1
           if (infest[i-ncols] ==1) spatialCov[i]=spatialCov[i]+1
       }
   else if (blck1[i] != 1  &&  blck1[i] != nrows && blck2[i] != 1 && blck2[i] ==
ncols) {
           if (infest[i -1] ==1) spatialCov[i]=spatialCov[i]+1
```

```
            if (infest[i-ncols] ==1) spatialCov[i]=spatialCov[i]+1
            if (infest[i+ncols] ==1) spatialCov[i]=spatialCov[i]+1
        }

    }
}

BCdata <- BCinfestdata[order(BCinfestdata$year),]
BCdata <- cbind(BCdata, spatialCov)
BCdata <- BCdata[4071:n,]
BCdata <- subset(BCdata,dd!=0)
attach(BCdata)
spatialCov <- BCdata$spatialCov
```

3. Fitting the logistic regression

```
logitBC <- glm(infest ~ dd + min + age + pine + dem + slope + aspect, family =
"binomial")
summary(logitBC)
```

4. Ploting the overall ROC and retaining the AUC values for each year

```
p = fits = fitted(logitBC)
ROC( test =p, stat=infest, plot="ROC",MI=FALSE)

dat <- cbind(BCdata,fits)
dat <- dat[order(dat[,12]),]

AUClog = rep(0,times=33)
infestperyear = table(BCdata[,12],BCdata[,4])[1:33,2]

for (i in 1:33) {
    if (infestperyear[i]!=0) {
        a = ROC(test=dat[(1706*(i-1)+1):(1706*i),14], stat=dat[(1706*(i-
1)+1):(1706*i),4],plot=”ROC”,MI=FALSE)$AUC
        AUClog[i] = round(a,3)
    }
}
plot(c(64:96),AUClog)
```

5. Fitting the MRF model

```
logitBCMRF <- glm(infest ~ dd + min + age + pine + dem + slope + aspect +
spatialCov, family = "binomial")
summary(logitBCMRF)
```

6. Obtaining Jackknife Estimates of the MRF model repeatedly while using backward selection to eliminate insignificant covariates after each round of estimation

```
matNeight <- matrix(0,1706,1706)
colBlock = unique(BCdata[,1])
for (i in 1:1706) {
    print(colBlock[i])
    neighsI = c(colBlock[i]-1, colBlock[i]+1, colBlock[i]-74,colBlock[i]+74)
    print(neighsI)
    #print(matNeight[i,is.element(colBlock, neighsI)])
    #Print(colBlock[1:100])
    #print(is.element(colBlock, neighsI)[1:100])
    matNeight[i,is.element(colBlock, neighsI)] = 1
}
mn <- as.data.frame(matNeight)
dimnames(mn)[2] = colBlock

jackBetaJs <- function(removedJ) {
# the function removes county J from the data (analogous to removing component
j and
# its 33 obs from the log likelihood) and refits the model to get new coefficients

data <- BCdata[BCdata [,1]!=removedJ,]
newlogit<- glm(data$infest ~ data$dd + data$min + data$age+ data$pine+
data$dem+ data$slope+ data$aspect + data$spatialCov, family="binomial")
newCoeffs <- newlogit$coeff

return(newCoeffs)

}

jackEstimates <- function() {

origCoeffs <- logitBCMRF$coeff
nEsts = length(origCoeffs)
jackmatrix <- matrix(0, 1706, nEsts)
jackEsts <- rep(0, times = nEsts)
Rn <- jackmatrix
Rnbar <- rep(0, times = nEsts)
jackVarEsts <- rep(0, times = nEsts)

for (j in 1:1706) {
     jackmatrix[j,] = jackBetaJs(j)
     print(j)
     }
for (k in 1: nEsts) {
    Rn[,k] = jackmatrix[,k] – origCoeffs[k]
    Rnbar[k] = mean(Rn[,k])
   }

for (i in 1: nEsts) {
    jackEsts[i] = origCoeffs[i] - (1706-1)/1706 * sum(jackmatrix[,i]-
origCoeffs[i])
   }

neigh = matNeight +diag(1706) #matrix of neighbours

for (k in 1: nEsts) {
print(k)
   for (i in 1:1706) {
        for (j in 1:1706) {
           if (neigh[i,j]==1) {
            jackVarEsts[k] <- jackVarEsts[k] +(1706-
1)/1706*neigh[i,j]*(Rn[i,k]-Rnbar[k])* (Rn[j,k]-Rnbar[k])
             }
           }
        }
   }

l = list(matjack = jackmatrix, diffs =Rnbar, ests = jackEsts, stdev =
sqrt(jackVarEsts))
return(l)
}

ests = jackEstimates()
ests
```

7. Creating the table of coefficients, p-values, biases after jackknifing

```
tablecoeff <- summary(logitBCMRF)$coeff
tablecoeff[,1] = ests[[3]]
tablecoeff[,2] = ests[[4]]
tablecoeff[,3] = ests[[3]]/ests[[4]]
tablecoeff[,4] = pnorm(abs(ests[[3]]/ests[[4]]),lower.tail=FALSE)*2
bias = (summary(logitBCMRF)$coeff)[,1]-ests[[3]]
tablec <- cbind(tablecoeff,bias)
dimnames(tablec)[[2]][5] = “Bias”
```

8. Obtaining the fitted values and ROC for the reduced MRF model and retaining the AUC values for each year

```
tab = tablec[,1]
tab = c(-7.3076961507, 0.0020610115, 0.0059791199, 0.0746890109,
0.0003050399, -0.0003394390, 0.0963487875, -0.0024714363, 2.0381429153)
datBC <-cbind(rep(1,times=dim(BCdata)[1]),BCdata[,c(5,6,7,8,9,10,11,13)])
fittedvals = rep(0, times =dim(datBC)[1])
fittedvals = as.matrix(datBC)%*%tab
p = exp(fittedvals)/(1+exp(fittedvals))
ROC(test=p, stat = BCdata$infest , plot="ROC",MI=FALSE)

dat <- cbind(BCdata,p)
dat <- dat[order(dat[,12]),]
AUCmrf = rep(0,times=33)
infestperyear = table(BCdata[,12],BCdata[,4])[1:33,2]

for (i in 1:33) {
    if (infestperyear[i]!=0) {
        a = ROC(test=dat[(1706*(i-1)+1):(1706*i),14], stat=dat[(1706*(i-
1)+1):(1706*i),4],plot="ROC",MI=FALSE)$AUC
        AUCmrf[i] = round(a,3)
    }
}
plot(c(64:96),AUCmrf)
```

9. Creating the spatial covariate (inf of neighbors at previous time), and temporal covariate – inf at the existing location at previous time)) – for the simple Markov Chain and fitting the model

```
BCinfestdata <- read.table("C:\\BCdata.txt",header=TRUE)
attach(BCinfestdata)

BCdata <- BCinfestdata[order(year),]
attach(BCdata)
years <- 34
n <- dim(BCdata)[1]
nsquares <- n/years
ncols <- 74
nrows <- 55

spatialCov <- rep(0,times=n)
timeCov <- rep(0,times=n)

for (i in (nsquares+1):n) {
     if (dd[i] != 0) {
        if (blck1[i] != 1  &&  blck1[i] != nrows && blck2[i] != 1 && blck2[i] !=
ncols) {
            if (infest[i -1-nsquares]==1) spatialCov[i]=spatialCov[i]+1
            if (infest[i+1-nsquares]==1) spatialCov[i]=spatialCov[i]+1
            if (infest[i-ncols-nsquares]==1) spatialCov[i]=spatialCov[i]+1
            if (infest[i+ncols-nsquares]==1) spatialCov[i]=spatialCov[i]+1
        }
      else if (blck1[i] == 1  &&  blck1[i] != nrows && blck2[i] != 1 && blck2[i]
!= ncols) {
            if (infest[i -1-nsquares]==1) spatialCov[i]=spatialCov[i]+1
            if (infest[i+1-nsquares]==1) spatialCov[i]=spatialCov[i]+1
            if (infest[i+ncols-nsquares]==1) spatialCov[i]=spatialCov[i]+1
        }
   else if (blck1[i] != 1  &&  blck1[i] == nrows && blck2[i] != 1 && blck2[i] !=
ncols) {
            if (infest[i -1-nsquares]==1) spatialCov[i]=spatialCov[i]+1
            if (infest[i+1-nsquares]==1) spatialCov[i]=spatialCov[i]+1
            if (infest[i-ncols-nsquares]==1) spatialCov[i]=spatialCov[i]+1
        }
   else if (blck1[i] != 1  &&  blck1[i] != nrows && blck2[i] != 1 && blck2[i] ==
ncols) {
            if (infest[i -1-nsquares]==1) spatialCov[i]=spatialCov[i]+1
            if (infest[i-ncols-nsquares]==1) spatialCov[i]=spatialCov[i]+1
            if (infest[i+ncols-nsquares]==1) spatialCov[i]=spatialCov[i]+1
        }
      }
}

for (i in (nsquares+1):n) {
    if (dd[i] != 0) {
        if (infest[i -nsquares]==1) timeCov[i]=timeCov[i]+1
    }
}

BCdata <- cbind(BCdata, spatialCov, timeCov)
BCdata <- BCdata[4071:n,]
BCdata <- subset(BCdata,dd!=0)
attach(BCdata)
spatialCov <- BCdata$spatialCov
timeCov <- BCdata$timeCov
```

10. Fitting the MC model

```
logitModelMC <- glm(infest ~ dd + min + pine + slope + aspect +
spatialCov+timeCov, family = "binomial")
summary(logitModelMC)
```

11. Obtaining Jackknife Estimates of the MRF model repeatedly while using backward selection to eliminate insignificant covariates after each round of estimation

```
logitBCMC<- glm(infest ~ dd + min + age + pine + dem + slope+ aspect +
spatialCov +timeCov, family="binomial")

jackBetaJs <- function(removedJ) {
# the function removes county J from the data (analogous to removing component
j and
# its 33 obs from the log likelihood) and refits the model to get new coefficients

data <- BCdata[BCdata [,1]!=removedJ,]
newlogit<- glm(data$infest ~ data$dd + data$min + data$age + data$pine +
data$dem + data$slope+ data$aspect + data$spatialCov +data$timeCov,
family="binomial")
newCoeffs <- newlogit$coeff

return(newCoeffs)

}

jackEstimates <- function() {

origCoeffs <- logitBCMC$coeff
nEsts = length(origCoeffs)
jackmatrix <- matrix(0, 1706, nEsts)
jackEsts <- rep(0, times = nEsts)
Rn <- jackmatrix
Rnbar <- rep(0, times = nEsts)
jackVarEsts <- rep(0, times = nEsts)

for (j in 1:1706) {
    jackmatrix[j,] = jackBetaJs(j)
    print(j)
    }

for (k in 1: nEsts) {
    Rn[,k] = jackmatrix[,k] – origCoeffs[k]
    Rnbar[k] = mean(Rn[,k])
    }

for (i in 1: nEsts) {
    jackEsts[i] = origCoeffs[i] - (1706-1)/1706 * sum(jackmatrix[,i]-
origCoeffs[i])
```

```
    }

neigh = matNeight +diag(1706) #matrix of neighbours

for (k in 1: nEsts) {
print(k)
    for (i in 1:1706) {
        for (j in 1:1706) {
           if (neigh[i,j]==1) {
            jackVarEsts[k] <- jackVarEsts[k] +(1706-
1)/1706*neigh[i,j]*(Rn[i,k]-Rnbar[k])* (Rn[j,k]-Rnbar[k])
              }
           }
        }
    }

l = list(matjack = jackmatrix, diffs =Rnbar, ests = jackEsts, stdev =
sqrt(jackVarEsts))
return(l)
}

ests = jackEstimates()
ests
```

12. Creating the table of coefficients, p-values, biases of the MC model after jackknifing

```
tablecoeff <- summary(logitBCMC)$coeff
tablecoeff[,1] = ests[[3]]
tablecoeff[,2] = ests[[4]]
tablecoeff[,3] = ests[[3]]/ests[[4]]
tablecoeff[,4] = pnorm(abs(ests[[3]]/ests[[4]]),lower.tail=FALSE)*2
bias = (summary(logitBCMC)$coeff)[,1]-ests[[3]]
tablec <- cbind(tablecoeff,bias)
dimnames(tablec)[[2]][5] = "Bias"
```

13. Obtaining the fitted values and ROC for the reduced MC model and retaining the AUC values for each year

```
tab = tablec[,1]
tab = c(-6.7157704658, 0.0019862001, 0.0200122613, 0.0342946031,
0.0003715340, 0.0001936922, 0.0604973049, -0.0027210991, 0.9681973743,
2.4370812889)
datBC <-cbind(rep(1,times=dim(BCdata)[1]),BCdata[,c(5,6,7,8,9,10,11,13,14)])
fittedvals = rep(0, times =dim(datBC)[1])
fittedvals = as.matrix(datBC)%*%tab
```

```
p = exp(fittedvals)/(1+exp(fittedvals))
ROC(test=p, stat = BCdata$infest , plot="ROC",MI=FALSE)

dat <- cbind(BCdata,p)
dat <- dat[order(dat[,12]),]
AUCmc = rep(0,times=33)
infestperyear = table(BCdata[,12],BCdata[,4])[1:33,2]

for (i in 1:33) {
    if (infestperyear[i]!=0) {
        a = ROC(test=dat[(1706*(i-1)+1):(1706*i),15], stat=dat[(1706*(i-
1)+1):(1706*i),4],plot=”ROC”,MI=FALSE)$AUC
        AUCmc[i] = round(a,3)
    }
}
plot(c(64:96),AUCmc)
```

14. Plotting the AUCs by year of the logistic, MRF, and MC models (years with no infestations excluded)

```
AUClog1 = AUClog[AUClog!=0]
AUCmrf1 = AUCmrf[AUCmrf!=0]
AUCmc1 = AUCmc[AUCmc!=0]

windows()
plot(c(64:96),rep(-1,33),main="AUCs per Year for the logistic (red), MRF(blue),
MC(green) models", xlab="Year",ylab="AUC values",ylim=c(0.6,1.05)) #,tck=1

points(c(64:96),AUClog1,pch=15,col=2)
points(c(64:96),AUCmrf1,pch=16,col=4)
points(c(64:96),AUCmc1,pch=17,col=3)
legend(86,.72,c(“Logistic AUCs”,”MRF AUCs”,”MC AUCs”), pch=c(15,16,17),
col = c(2,4,3), text.col = c(2,4,3)) #,bg=8
```

Section C: Table of Symbols Used Throughout the Thesis

Symbol	Description
i, j, s_i or s_j	- locations on the map at which covariate information and response values are gathered, or where predictions are made (s_i ,i are used interchangeably, with i used to simplify notation)
t	- the time at which covariate information and response values are gathered, or where predictions are made
$\eta_i, \eta_{i,t}$	- the logits of infestation at location i and/or time t
$p_i, p_{i,t}$	- the probability of infestation at location i and/or time t
$Y_i, Y_{i,t}, Y(s_i), Y(s_j)$	- the infestation status at location $i/s_i/\ldots$, and/or time t
$X_i, X_{i,t}, X_{i,t,k}$	- covariate information for the kth covariate at location i and/or time t
S	- the set of all locations on the map
N_i	- neighbourhood of location s_i (or i)
Y(S)	- the infestation status of the whole map
Pr(Y(S))	- the probability of infestation of the entire map
β_0	- model intercept
β_k	- the coefficient of covariate X_k, k=1,…,p
$\beta_{i,j}$	- the spatial dependence auto-regressive coefficient between locations i and j
β	- the common spatial dependence auto-regressive coefficient for all pairs of locations i and j
α	- coefficient of the temporal auto-regressive lag 1 covariate
θ_k	- the coefficient of covariate X_k, k=1,…,p
θ_{p+1}	- the coefficient of the spatial covariate
θ_{p+2}	- the coefficient of the temporal covariate
$i \sim j$	- denotes two neighbouring locations i and j
δ_n	- the original pseudo likelihood estimates
$\delta_{n,-j}$	- the estimates obtained from removing the jth component of the pseudo-likelihood
B_{nj}	- the bias between δ_n and $\delta_{n,-j}$
$JK\delta_n$	- the modified jackknife model parameters
$a_{i,j}$	- describes the relationship between two locations; 1 if i,j are neighbours of each other, 0 otherwise
Σ	- the variance covariance matrix of the errors
σ^2	- the standard error of the parameter estimates

Section D: Summary of Models Used and Their Features

The Logistic Model

$$\eta_{i,t} = \log \frac{P(Y_{i,t} = 1 \mid \underline{X}_{i,t}, spatialCov_{i,t})}{P(Y_{i,t} = 0 \mid \underline{X}_{i,t}, spatialCov_{i,t})} =$$

$$= \beta_0 + \beta_1 * X_{i,t,1} + \ldots + \beta_p * X_{i,t,p}$$

Logistic Model Features

- allows for inclusion of covariates (discrete and continuous) of various types (climate, terrain properties, forest characteristics)
- estimation is done by maximum likelihood
- parameter interpretation in terms of log odds
- allows for future predictions to be made

The MRF model (Autologistic Regression)

$$\eta_{i,t} = \log \frac{P(Y_{i,t} = 1 \mid \underline{X}_{i,t}, spatialCov_{i,t})}{P(Y_{i,t} = 0 \mid \underline{X}_{i,t}, spatialCov_{i,t})} =$$

$$= \beta_0 + \beta_1 * X_{i,t,1} + \ldots + \beta_p * X_{i,t,p} + \beta_{p+1} * spatialCov_{i,t}$$

Where:

$$\text{->}\ spatialCov_{i,t} = \sum_{j \in N_i} a_{j,i} * Y_{j,t} \qquad \text{->}\ a_{j,i} = \begin{cases} 1, & if\ j \in N_i \\ 0, & if\ j \notin N_i \end{cases}$$

MRF Model Features

- spatial dependence among responses accounted for via neighbouring response information
- flexibility in defining the neighbouring structure (regular, irregular lattice; first and second order neighbours, neighbours within a certain radius of current location)
- spatial dependence can be split into directions of dependence (e.g east neighbours separated from west neighbours etc.)
- covariate information can be included (as with the logistic regression)
- parameter estimates obtained by pseudo-likelihood; proper standard errors can be obtained by jackknifing the pseudo-likelihood
- parameter interpretation - log odds
- future predictions cannot be made due to the fact that the model relies on neighbouring response information from the very same year we are trying to predict

The Spatial MC Model

$$\eta_{i,t} = \log \frac{P(Y_{i,t} = 1 \mid \underline{X}_{i,t}, spatialCov_{i,t}, timeCov_{i,t})}{P(Y_{i,t} = 0 \mid \underline{X}_{i,t}, spatialCov_{i,t}, timeCov_{i,t})} =$$

$$= \beta_0 + \beta_1 * X_{i,t,1} + \ldots + \beta_p * X_{i,t,p} + \beta_{p+1} * spatialCov_{i,t} + \beta_{p+2} * timeCov_{i,t}$$

Where:

$$\text{->}\ spatialCov_{i,t} = \sum_{j=1}^{n} a_{j,i} * Y_{j,t-1} \qquad \text{->}\ timeCov_{i,t} = \begin{cases} 1, & if\ Y_{i,t-1} = 1 \\ 0, & if\ Y_{i,t-1} = 0 \end{cases}$$

$$\text{->}\ a_{i,j} = \begin{cases} 1, & if\ j \in N_i \\ 0, & if\ j \notin N_i \end{cases}$$

Spatial MC Model Features

- temporal and spatial dependence can be accounted for
- the neighbouring structure can be defined as with the MRF model
- the model uses neighbouring response information from the previous time state(s)
- information from the current location but from the previous year is used
- the periodicity of infestations can offer clues regarding the temporal dependencies that could be used
- covariates can be incorporated as with logistic regression and previous time state values for these covariates can be used
- pseudo-likelihood and jackknifing of the pseudo-likelihood can be used to obtain parameter estimates and standard errors
- model parameters are interpreted in terms of log odds
- future predictions can be made since the model uses only the known past information