Modeling Historical Forest Landscape in the County of Halland, Sweden Using Data from the First National Forest Inventory

Cen Chen

Arbetsrapport 375 2012
Master thesis in environmental science 30hp D
Environmental monitoring and Assessment

Handledare:
Anna-Lena Axelsson

Sveriges lantbruksuniversitet
Institutionen för skoglig resurshushållning
901 83 UMEÅ
www.slu.se/srh
Tfn: 090-786 81 00

ISSN 1401–1204
ISRN SLU–SRG–AR–375–SE

Modeling Historical Forest Landscape in the County of Halland, Sweden Using Data from the First National Forest Inventory

Cen Chen

Master thesis in environmental science 30 hp
Environmental monitoring and Assessment
EX0627
Supervisor: Anna- Lena Axelsson, SLU, department of forest resource management
Examiner: Håkan Olsson, SLU, department of forest resource management

---

Sveriges lantbruksuniversitet
Institutionen för skoglig resurshushållning
Utgivningsort: Umeå
Utgivningsår: 2012

ISSN 1401–1204
ISRN SLU–SRG–AR–375–SE

Abstract

Spatial patterns is of core interest for landscape ecology, and tracking its temporal evolvement helps to attain a better understanding of the ecological effects of current ecosystems. The presence and recent digitization of the first Swedish National Forest Inventory data and the occurrence of a concurrent historical map offers a unique chance to take on this very challenging task. The objective was to describe and test a statistical model of the historical forest landscape for a study area in the county of Halland, Sweden during the 1920s by utilizing different spatial data sources in the model building towards a plausible methodological application of the model. Data from a detailed digital elevation model, thematic maps of soil type, and topographic maps were introduced into the modeling. Both deterministic and stochastic parts of response variables were extracted by combining Partial Least Squares and logistic regressions with ordinary Kriging. Conventional cross-validating was applied to judge the performance of the modeling, so were two compatible estimations, one made by applying k-Nearest Neighbor method, and the other origins from quartic kernel function. Both of them indicate the performance of the modeling is fairly good for the vegetation type part (with an accuracy no less than 0.75), and barely acceptable for the forest stand age and openness part (Root Mean Square Error is no bigger than 18.45 years for stand age, and 0.17 for stand openness). The modeling results agree poorly with information extracted from historical map. The nature of the data available is considered to be the determinant causal factor behind the results. Certainly, the modeling has the potential to be further improved methodologically.

Keywords: landscape, regression, Kriging, k-Nearest Neighbor, kernel, R

Contents

Introduction 3
Background 3
Methodological Glimpse 3
Purpose and Objectives 4
Study Area and Data 5
Study Area 5
Data Sources 5
Data Pre-processing 8
Allocation of NFI Sample Plots to the Gridded Study Area 9
A Glance at the Sampling Quality 9
Methods 12
Partial Least Squares Regression 13
Logistic Regression 14
Ordinary Kriging 15
k-Nearest Neighbor 16
Quartic Kernel 16
Results and Discussion 18
Separation of Forest and Non-forest 18
Dividing Non-forest into Mire and Heath 20
Distinguishing between Coniferous, Deciduous, and Mixed Forests 21
Estimating AGE and OPENNESS of Forest Stands 25
The k-NN Reference of Estimation 29
Final Outputs 31
Comparison of Predictions Based on Two Different Data Sources 32
Comparison on OPENNESS 33
Comparison on Vegetation Type 35
Conclusion and Perspective 37
Constraints of Data 37
Future Improvement of Modeling 38
Acknowledgement 40
References 41
Appendix: R Codes Used 43

Introduction

Background

To landscape ecology, which emphasizes broad spatial scales and the ecological effects of the spatial patterning of ecosystems (Turner, 1989), spatial patterns are of primary importance. As populations respond to changes in the structure of landscape with certain time lags, temporal perspectives are also important in landscape ecology (Hanski, 1998). Hence it is not sufficient to consider just the current area, quality and spatial connectivity of habitats for the purpose of assessing population viability, temporal changes in habitat structure should also be taken into account (Gu et al., 2002). This means historical landscape structure and spatial distribution of specific habitats are important for analyzing and understanding the current prevalence of species, and beneficial for predicting its future development. This theory, on which the practical meaning of this study lies, is relevant also when applied to a forest landscape context.

For the three landscape characteristics, namely structure, function, and change, this modeling specifically deals with the first one in the forest context. Since structure refers to spatial relationships between distinctive ecosystems (Turner, 1989), here it is the distribution of forest in relation to the various attributes (as response variables of this modeling) of its components (as grid cells in this modeling). So, amid all possible approaches, the attempt to modeling historical forest landscape here is to reconstruct forest structure at landscape scale built up on estimations of relevant forest attributes on each of its components.

Methodological glimpse

Various spatial regression methods ranging from Ordinary Least Squares (OLS) regression to Geographically Weighted Regression (GWR) and etc. have often been applied for modeling of forest attributes. Such methods are of especial relevance for modeling the relationship with environmental covariates (Hooten, 2001). Specific in the historical field, He et al. (2007) applied hierarchical Bayesian model in mapping pre-European settlement vegetation in Missouri, USA, and considered that this method outperforms simply logistic regression when sample size is small.

It is generally accepted that almost all natural processes are subject to some measure of spatial dependence. Environmental covariates may even account for some of this dependence (since spatial structure more or less is also contained in those covariates), in addition to the deterministic part of variation in forest attributes they contains, it mostly remains unexplained. Kriging and its variants, also non-parametric methods, e. g., Inverse Distance Weighting (IDW), are popular choices for this situation. There are numerous such applications, and Brown (1998) also used Kriging and co-Kriging to map historical forest types in Michigan, USA.

One problem is that those methods assume interpolated data are numerical and spatially continuous, yet many covariates, e. g., soil and landuse type, often cause spatial discontinuity. Therefore, these interpolation methods may ignore the ecological principles underlined by those environmental covariates (He et al., 2007). Also, because knowing the existence of spatial autocorrelation is nowhere near the revelation of it, performance of geostatistical methods like Kriging frequently stay under expectation. This may explain the boom of the application of k-Nearest Neighbor method (k-NN).

It should be noted that the ambition in modeling more historical forest attributes than just species composition, probably stand age, openness, and density too, implies this modeling's different methodological approach than those of the studies mentioned.

Purpose and objectives

The purpose of this study is to develop statistical models revealing historical forest landscape by combining different types of comprehensive data and spatial dependence with detailed historical information from the first Swedish National Forest Inventory (NFI).

Concrete objectives of this study are to:

1) describe and test a statistical model of the historical forest landscape for a study area in Halland, Sweden during the 1920s;

2) evaluate the effectiveness of utilizing different data in the model building towards a plausible methodological application.

Study area and data sources

Study area

The study area (Figure 1) of a size of about 10 × 10 km is situated in the county of Halland, which is located on the western coast of Sweden belonging to the nemoral zone. The vegetation change in Halland during the Holocene was drastic, and followed a general pattern of from deciduous forest to heathland, then to coniferous forest (dominated by spruce). This change responded to both climatic and human impacts, but majorly to the latter. In the mid-19th century, heathland reached its largest extent of 150,000 ha in Halland. After that, the reforestation movement started, and reduced heathland to an extent of 73,000 ha by 1913 – 1914. The continuous reforestation made coniferous forest the most prevailing vegetation type. (Blennow and Hammarlund, 1993)

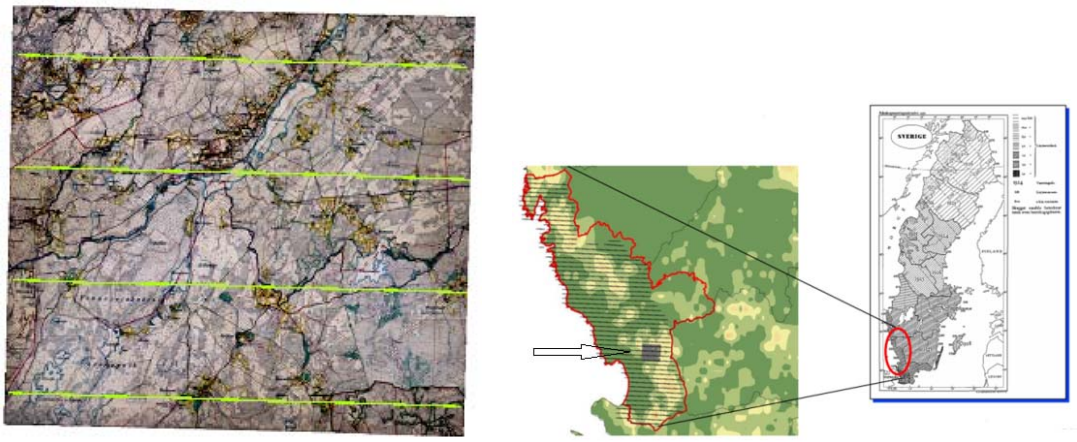

Figure 1. The study area (shown in the form of historical county economic map © Lantmäteriet, I2011/0032.) with transects of the 1st NFI imposed on it (left), its location in the county of Halland (middle) and its location in Sweden (right).

The selection of this study area was based on three reasons: 1) Its dramatic change of forest – the area of productive forest in the county of Halland increased by 70,000 ha or 35% during the past 50 years before early 1990s (Blennow and Hammarlund, 1993); 2) Its well documented forest history – Carl Malmström made the classic forest history study of "Forests of Halland during the last 300 years" (Lindbladh et al., 2011); 3) The historical economic map was surveyed during 1919 – 1925, which matches well with the first NFI data in terms of time.

Data sources

Data available for this study originated from five sources, namely the first Swedish NFI, the historical county economic map, the latest digital elevation model (DEM), thematic

maps of soil type from the Geological Survey of Sweden, and topographic maps from the Swedish mapping, cadastral and land registration authority (Lantmäteriet).

The first NFI

The first Swedish NFI started in 1923 (and was implemented in the study area in 1928). The NFI was performed on east-west directional parallel transects covering Sweden from border to border. The transects were 10 meters wide, and spaced variously from the south to the north of Sweden (at 2.5 kilometers in the study area) (Figure 1). Three different types of information was collected on the transects, namely stand characters, tree count and detailed measurements of individual sample trees (Thorell and Östlin, 1931). In this study stand characters: i.e. vegetation type, stand age, and stand openness were used.

The first NFI provides fundamental information on reconstructing historical forest landscape (from which the variables of interest of this study were extracted). The line transects sampling design provides information on landscape extent, patchiness, and connectivity, captures spatial dependencies better than inventories conducted on small circular plots. The main reason is that this design provides sufficient pairs of samples at consecutive distances to build spatial autocorrelation models as a function of distance on, and maintains the information regarding the dimensions of forest stands. This probably is not the case in current Swedish NFI, in which clustered circular plots are dispersed out.

The historical county economic map

The historical county economic map (surveyed in the study area during 1919 – 1925) that contains information on extents and borders of different land cover types, supplies another type of information than the first NFI data. The map provides full coverage of the study area that do not exist in the NFI sample data, despite its lacking of information on forest stand age, growing stock, and openness.

Besides marking land cover types, the historical county economic map also visualizes the distribution of coniferous and deciduous forests by using different symbols. A detailed example of the map used in this study is illustrated in Figure 2. The historical landcover borders and the tree symbols within the study area were digitized within another project. A wall-to-wall map created by merging the digitized historical and current digital layers were also provided (Axelsson et al 2012) and was used for further data processing and analysis. Information compatible with the modeling result was generated from the digitized tree symbols, and was used in a comparison with results from the modeling as performing the role of validation.

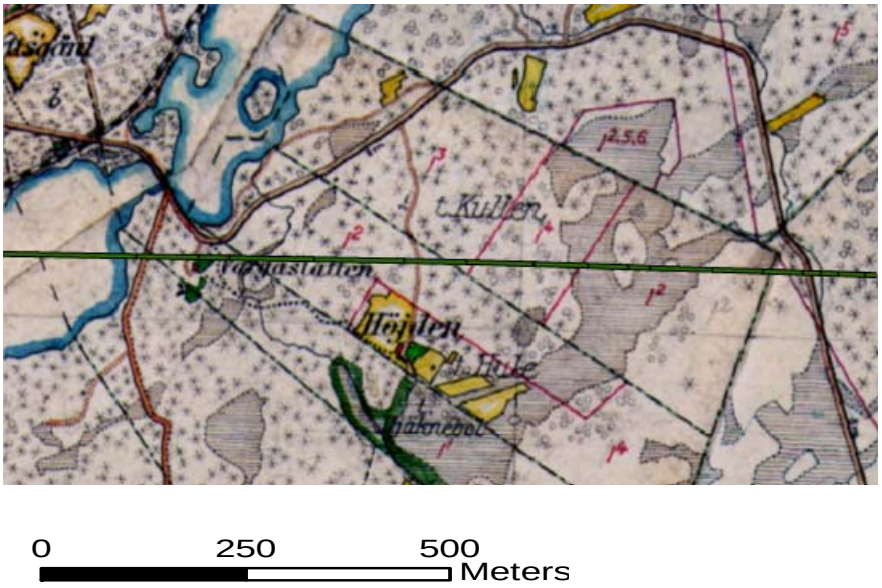

Figure 2. An example of a detailed part of the historical county economic map © Lantmäteriet, I2011/0032, in which coniferous forest is represented by the symbol "*", and deciduous forest is represented by the symbol "○". A part of the first NFI transect is also imposed on the map shown as the narrow green strip in the middle.

DEM

The DEM used in this modeling was produced by The Swedish mapping, cadastral and land registration authority through airborne laser scanning (LiDAR) with an aim to limit the standard error within 0.5 m. The density of laser points is 0.25-0.5 points/m^2 in the study area. In this way, there is at least one point on ground within one 2 × 2 m grid cell. Variables derived from this DEM are elevation, incoming solar radiation, aspect, plan curvature, profile curvature, slope, wetness index, X and Y coordinates.

Thematic maps of soil

Maps from the Geological Survey of Sweden present information about the distribution, structure, and properties of quaternary deposits at a scale of 1: 50 000 and 1:100 000 at a resolution of 5 meters. The maps are based mainly on interpretation of aerial photographs combined with field observations along roads. When the variable of soil type was generated, all those detailed soil types have been aggregated into three types, namely peat, sediment and rock, and moraine.

Topographic maps

The topographic maps at a scale of 1:50 000 with a standard error of approximately 10 m in plane are from the Swedish mapping, cadastral and land registration authority. They come in "shapefile" format with layers of administrative divisions, built-up areas and

buildings, hydrography, roads, etc. In this modeling, these maps have been used to generate variables of the distance (of study objects) to arable land and water.

Table 1. Overwiev of variables used in the modeling

Variable	Abbreviation	Type	Description
Response Variables (From the First NFI)			
Vegetation Type		Nominal	"C", "D", "MI"; "M", "J" for coniferous, deciduous, and mixed forest; mire, and heath respectively
Stand Age	AGE	Ordinal / Continuous	"0-20", "21-40", "41-60", "61-80", "81-100", and "101-120" years
Stand Openness	OPENNESS	Ordinal / Continuous	"0", "0.1-0.2", "0.3-0.4", "0.5-0.6", "0.7-0.9", "0.9-1.0", and "1.0+" ("1.0" for 100% canopy coverage)
Explanatory Variables (From DEM, Soil and Topographic Maps)			
Ground elevation	DEM	Continuous	Meter
Incoming Solar Radiation	SOLAR	Continuous	Watt / year · square meter
Aspect	ASPECT	Continuous	Degree
Plan Curvature	CURV_PL	Continuous	
Profile Curvature	CURV_PR	Continuous	
Slope	SLOPE	Continuous	Percentage
Wetness Index	WI	Continuous	Calculated by: $\ln\left(flow\ accumulation \cdot 20^2 / \frac{SLOPE}{100}\right)$
X Coordinate	X	Continuous	Meter
Y Coordinate	Y	Continuous	Meter
Soil Type	SOIL	Nominal	"1", "8", "93" for peat, sediment and rock, and moraine respectively
Distance to Arable Land	DIST_A	Continuous	Meter
Distance to Water	DIST_W	Continuous	Meter

* Variables will be represented by their abbreviations in this paper.

* Since data have been standardized before used for analyses in this study if necessary, their scale (unit) is not problematic.

Data pre-processing

All these original data are in "shapefile" format of ArcGIS, except the DEM coming in IMAGINE's "img" format with a cell size of 2 × 2 m. From these data sources, all of the variables (Table 1) were derived, and transformed to a uniform IMAGINE's "img" format with a cell size of 20 × 20 m in ArcGIS prior to being imported into R 2.12.2 (R Development Core Team, 2011) for analysis. Subject to different analyses performed in R,

these data were further manipulated and transformed into various suitable formats (All codes for these operations along with complete analyses can be found in Appendix: R Codes Used).

It should be noted that from the description in Table 1, AGE and OPENNESS are more of continuous variables with discrete values than being ordinal. Considering ordinal variables are often treated as continuous variables in research (Winship and Mare, 1984), it is reasonable to treat AGE and OPENNESS as continuous variables here. And it is indeed the practice in this study.

Allocation of NFI sample plots to the gridded study area

Conventionally, sample plots, regardless of their sizes, are reduced to points to be allocated to the study area during modeling. In the case of explanatory variables being presented in a gridded style, values of response variables of a sample plot will be assigned to a certain cell of the grid to be combined with respective values of explanatory variables. This admittedly is a convenient way to base the following analyses on, as long as the sample plots are not much larger than the grid cells on size. Otherwise, not only the dimensional information of the sample plots will get lost, but cells with erratic values of explanatory variables may more severely undermine the following analyses.

In this study, a certain section of NFI sample strips will be allocated to all of the cells it contains in the gridded study area simultaneously, (specifically, when the centroid of a cell falls in an NFI sample plot, it is considered that the sample plot contains the cell). This equals to taking the dimension of the plots into account by giving larger plots more weights in the following analyses. The comparison of these two ideas is illustrated in Figure 3.

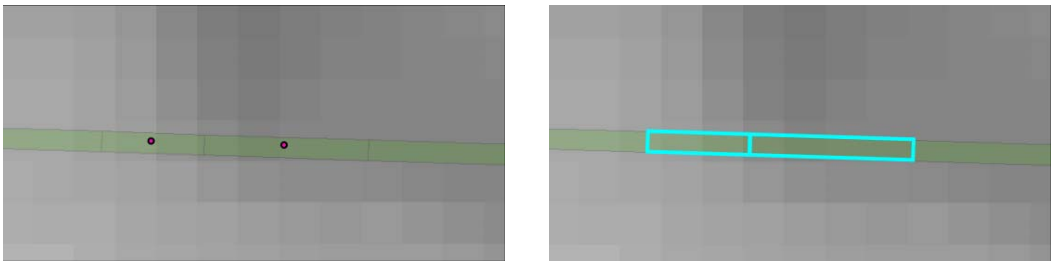

Figure 1. Illustration of two possible ways to allocate NFI sample plots to the grid. As individual points in a certain cell (left); or as polygons to all cells of the grid that they contain (right)

A glance at the sampling quality

As for anyone else, the performance of the modeling in this study depends not only on how well the sample dataset are fitted by the models, but how sufficient and representative the

sample dataset are. It is important to have a glance at this problem before the actual analyses are presented.
For the data sufficiency part, the major problem is the nonexistence of sample plots between NFI transects that were far apart, which will affect the prediction of Kriging severely. For the representativity part, except the obvious clustered spatial distribution of NFI sample plots, Figure 4 compares the distributions of each explanatory variable (except SOIL, X, and Y) in the sample and in the whole study area. Differences can be observed from all these pairs of distributions, even though the results from Kolmogorov-Smirnov test (Table 2) suggest only the distributions of CURV_PR and SLOPE are significantly different ($\alpha = 0.05$) in the sample and in the study area.

In this sense, models well fitted for the sample dataset do not guarantee their performance in the whole study area, and vice versa. However, quantifying the influence of the insufficiency of samples and the disparities between the two types of distributions on the following modeling does not seem to be possible. This type of study surely has more implications to the sampling design, which however is out of the scope of this thesis.

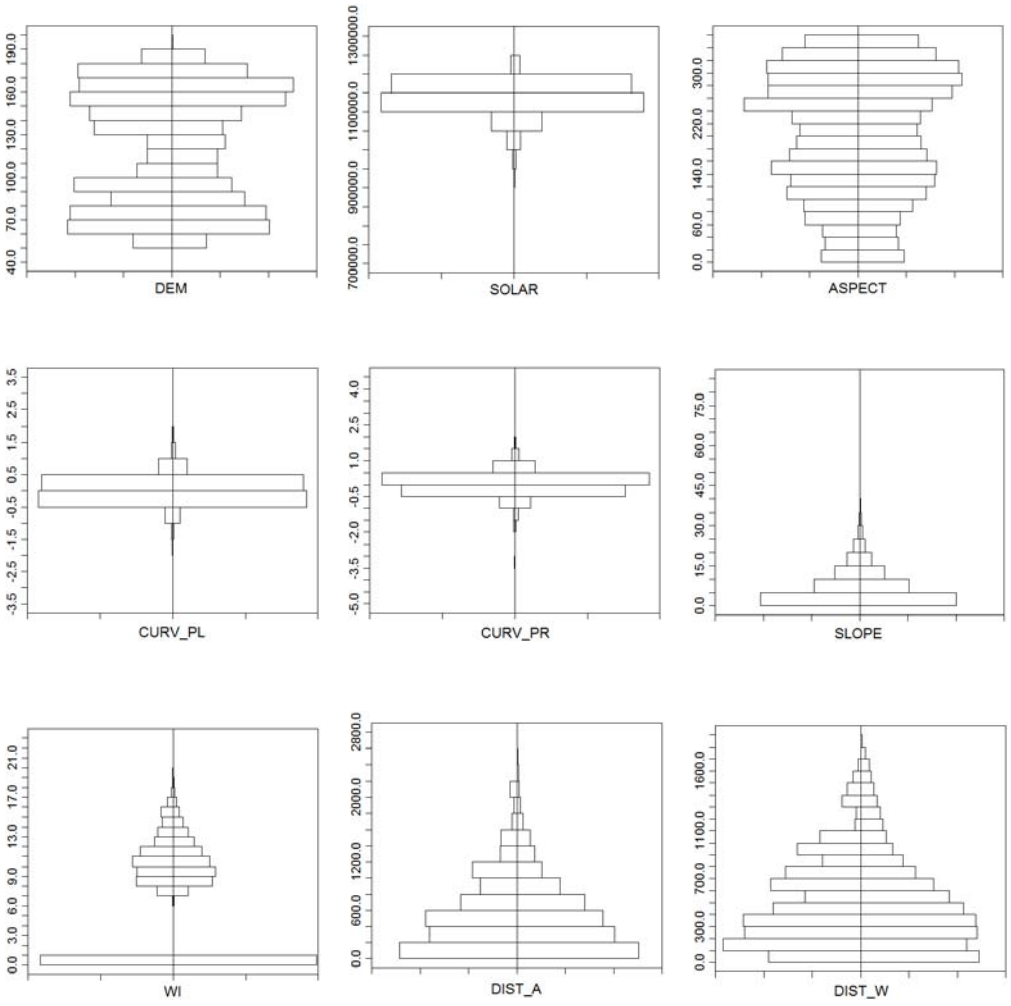

Figure 2. Back to back histograms of distributions of explanatory variables (except SOIL, X, and Y) in, left-hand side) the sample, right-hand side) the whole study area; horizontal scale of each histogram is of percentage.

Table 2. Results from Kolmogorov-Smirnov test on the distributions of each explanatory variable (except SOIL, X, and Y) in the sample and in the whole study area

	DEM	SOLAR	ASPECT	CURV_PL	CURV_PR	SLOPE	WI
p-value	0.6994	0.09956	0.964	0.1528	0.02816	0.01705	0.9897
	DIST_A	DIST_W					
p-value	0.9048	0.7937					

Methods

The possible lack of information with high explanatory power to the variables of interest may undermine the modeling effort enormously, that is the main reason why a rather delicate modeling design (Figure 5) has to be created here intended to compensate for this. Another driving force behind the design surely is the nature of the available data.

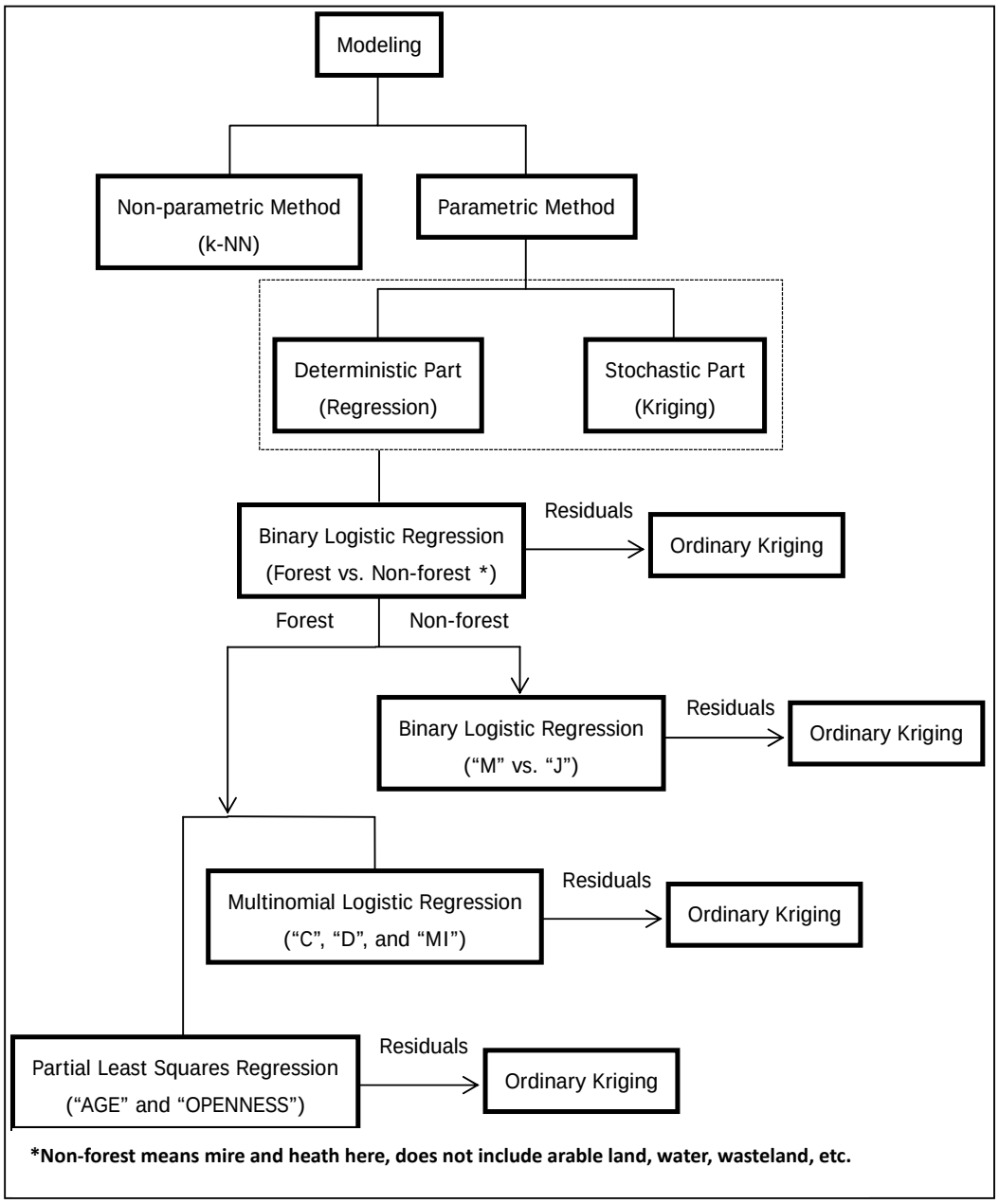

Figure 5. Diagram of the modeling design in this study

Partial least squares regression

One general idea of this modeling design is to combine regression and Kriging to explain both the deterministic and the stochastic parts of the response variables. The former part lies on the seemingly weak correlations between the explanatory and the response variables, which are illustrated in Figure 6. On the contrary, some correlations within the explanatory and the response variables seem not so weak, and this may adversely affect this modeling as well. That is the main reason why Partial Least Squares (PLS) regression, among other multivariate methods, was chosen to deal with this situation.

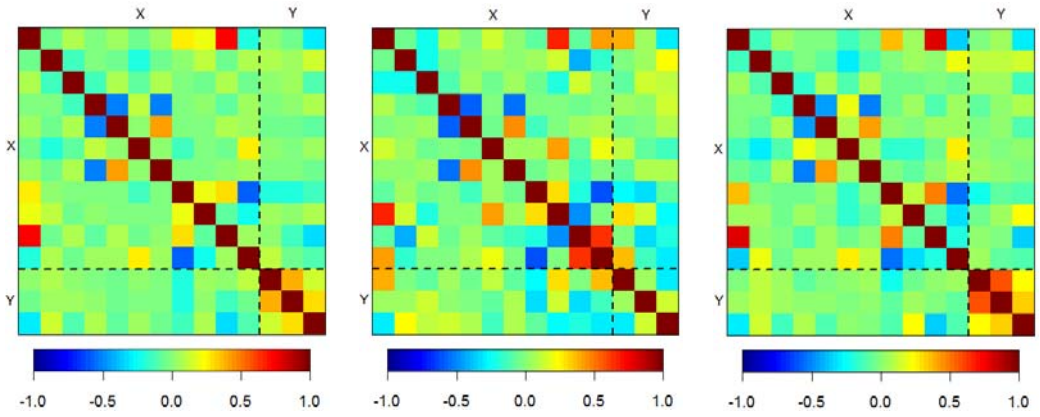

Figure 6. Correlation matrices of explanatory variables (*X*, from left to right and top to bottom following the order, in which they are listed in Table 1) and response variables (*Y*, from left to right and top to bottom, of AGE, DENSITY (discarded in this study because of poor data quality), and OPENNESS) in all SOIL types (left), SOIL type 8 (middle), and SOIL type 93 (right). SOIL type 1 is not individually included because of insufficient sample size on it

The main theory of PLS regression is to decompose (Singular Value Decomposition, SVD) the matrices of explanatory variables ($\boldsymbol{X}$) and response variables ($\boldsymbol{Y}$) simultaneously with the constraint that these decomposed components explain the covariance between $\boldsymbol{X}$ and $\boldsymbol{Y}$ as much as possible. It is the major difference between it and principal components regression, which is focused on explaining the variance in $\boldsymbol{X}$, even these two methods shares some common traits. Some also claim that PLS regression yields somewhat better results in terms of the predictive ability when compared to the other regression methods (Yeniay and Göktas, 2002).

$\boldsymbol{Y} = \boldsymbol{X}\mathbf{B} + \mathbf{E}$, where $\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}'$, $\widehat{\boldsymbol{Y}} = \boldsymbol{T}\boldsymbol{B}\boldsymbol{C}'$
$\boldsymbol{t} = \boldsymbol{X}\boldsymbol{w}$, $\boldsymbol{u} = \boldsymbol{Y}\boldsymbol{c}$, so $\boldsymbol{w}'\boldsymbol{w} = 1$ (or $\boldsymbol{W}'\boldsymbol{W} = \boldsymbol{I}$), $\boldsymbol{t}'\boldsymbol{t} = 1$ (or $\boldsymbol{T}'\boldsymbol{T} = \boldsymbol{I}$), and $\boldsymbol{t}'\boldsymbol{u}$ be maximized
$\widehat{\boldsymbol{B}} = \boldsymbol{W}(\boldsymbol{P}'\boldsymbol{W})^{-1}(\boldsymbol{T}'\boldsymbol{T})^{-1}\boldsymbol{T}'\boldsymbol{Y}$

$\boldsymbol{T}$ is the same scores of both $\boldsymbol{X}$ and $\boldsymbol{Y}$, and $\boldsymbol{P}$ is the loadings of $\boldsymbol{X}$ from SVD; $\boldsymbol{t}$ and $\boldsymbol{u}$ are column vectors of $\boldsymbol{X}$ and $\boldsymbol{Y}$ selected iteratively in a way to maximize the covariance

between $\boldsymbol{X}$ and $\boldsymbol{Y}$; $\boldsymbol{W}$ (and its component vectors $\boldsymbol{w}$) and $\boldsymbol{C}$ (not for the covariance matrix here; and its component vectors $\boldsymbol{c}$) is the weights; $\boldsymbol{B}$ is the regression coefficients.

It needs to be mentioned that one common feature of all regression models included in this study is that they are global, which means positions in the geographical space are treated in the same way. However some local regression methods, e. g., GWR, may benefit such kind of study considerably to response to the same stimulus differently by taking those positions into account. But due to the inadequate sample size and software's capability, it is not the case in this study yet.

Logistic regression

Since PLS regression is only suitable for continuous variables, a series of logistic regression models on a hierarchical style have been built for the categorical variable of vegetation type. When applying these logistic models, the problem of collinearity mentioned above may arise as well. To determine its influence on these models, the conditional number κ was calculated for each model (Table 3) following the method suggested by Belsley et al. (1980), according to whom, κ with a value between 0 and 6, around 15, and over 30 indicates no, medium, and harmful collinearity respectively. Obviously collinearity does not seem to be a problem here.

Table 3. Conditional number κ for each logistic model used in this study

	Binary Logistic Regression (Forest vs. Non-forest)	Binary Logistic Regression ("M" vs. "J")	Multinomial Logistic Regression ("C", "D", and "MI")
κ	4.41	5.97	4.21

The fundamental difference of logistic regression from least squares linear regression is that the response variable is constrained to a limited number of integer values (Peterson, 1998). Other than that, they share some common principles (Of course, the regression function for the parameters β is no longer linear.). Binary and multinomial logistic regressions are also quite similar, only response variable in the latter one is polytomous, thus $m-1$ other than one set of coefficients will be fitted.

$$\pi_{ij} = \frac{e^{g_{j(x_i;\beta)}}}{1+\sum_{j=2}^{m} e^{g_{j(x_i;\beta)}}},\ j = 2, \ldots, m$$

$$\pi_{i1} = 1 - \sum_{j=2}^{m} \pi_{ij}$$

π_{ij} is the probability that observation *i* is in response class *j* (m classes in total); g_j here is the logit, which is a linear function of β for class *j*.

Along the design of modeling continuous and categorical response variables separately,

another suspicion of whether this will violate the relations between these two types of variables emerges. It is not a problem for non-forest samples since they only have one attribute, which is the categorical variable of vegetation type. But for forest samples with attributes of both types of variables, there are potential violations. So ANOVA of different vegetation types on AGE and OPENNESS within forest samples have been carried out, and the results of which are shown in Table 4 and 5. The hypotheses that there is no difference among the means of different vegetation types either on AGE or on OPENNESS cannot be rejected at a significance level of 0.05. So it is justified to regress these two parts separately for forest samples.

Table 4. Result of ANOVA of different vegetation types on AGE
H_0: The means of of different vegetation types on AGE are equal.

	df	**SS**	**p-value**
Vegetation Types	2	1754.8	0.106

Table 5. Result of ANOVA of different vegetation types on OPENNESS
H_0: The means of of different vegetation types on OPENNESS are equal.

	df	**SS**	**p-value**
Vegetation Types	2	0.07377	0.389

Ordinary Kriging

Concerned about the performance of those regression models, especially of the PLS model due to the absence of strong correlations between the explanatory and the response variables indicated by Figure 6, residuals from each of all those regression models were passed on to variogram modeling to track down spatial autocorrelations within them, on which Kriging can be based. Undoubtedly, better estimations can be obtained by this sort of combination of regression and Kriging than by either of these techniques alone.

The basic idea of Kriging is that the value of target variable at a new location ($Z(s_0)$) can be derived as a weighted average of values at neighboring locations ($Z(s_i)$), (for ordinary Kriging, an unknown constant mean μ of this variable is included). Hence, the critical part of Kriging is to obtain the weights. In order to do so, the covariance as a function of distance $\boldsymbol{h}$ has to be estimated first, and this can be achieved by estimating the (semi-)variogram $\gamma(\boldsymbol{h})$. The consideration behind choosing Kriging over other interpolation methods is that it is supposed to minimize the error variance of prediction (Isaaks and Srivastava, 1989).

$\hat{Z}(s_0) = \sum_{i=1}^{n} w_i(s_0) Z(s_i)$, where $E\big(\hat{Z}(s_0)\big) = \mu$

$\boldsymbol{w}(s_0) = \boldsymbol{C}^{-1}\boldsymbol{c}(s_0)$, where $\boldsymbol{c}(s_0) = (C(s_0, s_1), \dots, C(s_0, s_n))'$

$C(\boldsymbol{h}) = \sigma^2 - \gamma(\boldsymbol{h})$

$\gamma(\boldsymbol{h}) = \frac{1}{2} E\big(\big(z(s_i) - z(s_i + \boldsymbol{h})\big)^2\big)$

$\boldsymbol{w}$ is the weight vector; $\boldsymbol{C}$ is the covariance matrix; $\boldsymbol{c}(s_0)$ is the transposed covariance vector at point s_0; σ^2 is the variance.

k-Nearest Neighbour

The main reason to include k-NN method in this study is to set up a reference to the modeling. Because of its ability of multivariate imputation, more importantly its relaxation on those limitations, concerns, and delicate and arbitrary parameter estimation of parametric methods, the performance of applying k-NN method will help evaluate the worthiness of the modeling based on parametric methods.

In k-NN method, the values of target variables at a certain point are weighted by those values of its k nearest neighbors in the reference space defined by reference variables. The major difference among variants of k-NN method is how the nearness is constructed using different definitions of metrics. The specific k-NN variant used here is Mahalanobis, in which the weight is defined as the inversed covariance matrix $\boldsymbol{C}$ of reference variables, and k is set to be one (so that the imputation of vegetation type can be meaningful).

$$d(\boldsymbol{x},\boldsymbol{y}) = ((\boldsymbol{x}-\boldsymbol{y})'\boldsymbol{C}^{-1}(\boldsymbol{x}-\boldsymbol{y}))^{1/2}$$

$d(\boldsymbol{x},\boldsymbol{y})$ is the distance between points $\boldsymbol{x}$ and $\boldsymbol{y}$.

Quartic kernel

The predictions of OPENNESS through the 1920s historical economic maps and the first NFI data were compared. The former prediction was based on quartic kernel function since it is not a straightforward procedure. It started after the maps being digitized (Figure 7), which means symbols of coniferous and deciduous forests (Figure 1) had been transformed to corresponding points that each of them represents a certain amount of forest at a certain location. From this outset, point intensities, which reflect forest densities, can be computed.

$$\hat{\lambda}(x_0) = \frac{1}{h^2}\sum_{i=1}^{n} k\left(\frac{|x_0 - x_i|}{h}\right), \text{ where } k(u) = \begin{cases} \frac{3}{\pi}(1-|u|^2)^2 \; if \; u \in (-1,1) \\ 0 \qquad\qquad otherwise \end{cases}$$

λ is the point intensity of point x_0; x_i is a point within bandwidth h of point x_0.

Figure 7. Digitized historical county economic maps, in which points indicating forest densities are for coniferous forests (left); deciduous forests (right).

Results and Discussion

Separation of forest and non-forest

The first step towards modeling historical forest landscape is to separate forest and non-forest areas. This was achieved by building a binary logistic regression model on these two features using full predictors. After that, a subset of this model was selected among all of the possible subsets based on their scores on Akaike Information Criterion (AIC) in order to balance between goodness of fit and over-fitting. The selected one is shown in Table 6.

Table 6. Result of the selection of predictors of the regression model on forest and non-forest

	Predictors Discarded		**Predictors Retained**
	CURV_PR	CURV_PL	The Other Predictors
AIC	879.34	877.35	875.4

The overall performance of this selected model is displayed in Table 7. The results indicates that distinguishing forest and non-forest areas should be applicable, since the fitted values and the observed values are not significantly different according to the p-value of the χ^2 statistic of the residual deviance (D_M) (Menard (2001) considered the test of $D_0\ (null\ deviance) - D_M$ (referred to as G_M) be more appropriate, and did not recommend the statistic used here). An overall classification accuracy of 0.75 (coincide with a prediction error $\boldsymbol{\delta}$ of 0.265 of the adjusted leave-one-out cross-validation based on the cost function of $mean(|observed - predicted| > 0.5)$) with an un-weighted Kappa statistic of 0.48 on the sample dataset further confirms this possibility. However, the p-value may also arise concerns about the resemblance of the fitted values and observed values not being substantial enough, considering this model will set the scopes of the consequential modeling, hence its uncertainty will be added up to the uncertainties of the consequential modeling.

Table 7. Performance of the selected model on forest and non-forest
H_0: There is no difference between observed and fitted values.

$\boldsymbol{D_M}$	**df**	**p-value**	$\boldsymbol{\delta}$
851.4	840	0.3847	0.265
	Observed		
Predicted	Forest	Non-forest	
Forest	234	108	
Non-forest	105	405	

After extracting the deterministic part of observed values regarding forest and non-forest using the selected logistic model, the corresponding stochastic part was extracted from the residuals on probability through variogram modeling. The sample variogram and the fitted variogram model, which has the lowest value of the weighted Sum of Squared Errors (SSE) of 5.33e-06 among all candidate variogram models, (the selection of all variogram models

in this thesis was trying to follow this routine) are shown in Figure 8. It should be noted that the fitted variogram model has been isotropic, because the one directional line transect sampling design of the first NFI makes an anisotropic variogram model less than possible within a reasonable range. This is the common practice for all of the variogram models in this paper.

The most noticeable character of the fitted variogram model is its rather short range. This leads to many regions of the study area being uncovered by the prediction of Kriging. This can be observed also in the following output images of Kriging prediction. Even with repeated attempts of different ranges, the same situation occurs in parameter estimation using least squares method (also maximum likelihood method). This can be interpreted in two different ways, either little spatial autocorrelation remains after regression, or the spatial autocorrelation stays profoundly uncovered. But without truly understanding the mechanism behind it, which is the case here, it can never be sure, so as for the estimation of the other parameters. As reported in many other studies, this ambiguity in parameter estimation has been deteriorating Kriging's applicability.

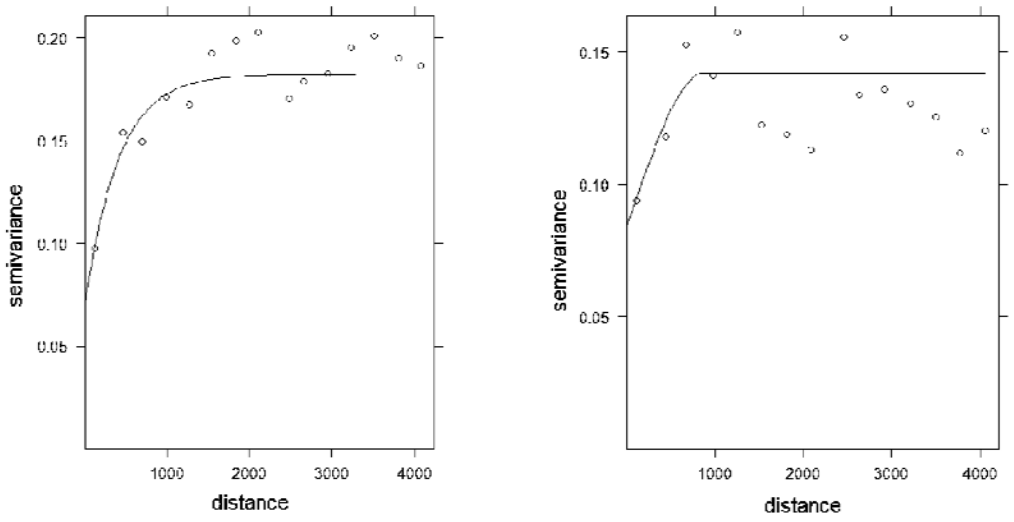

Figure 8. Sample variograms and fitted variogram models of the residuals from, left) the logistic model on forest and non-forest; right) the logistic model on mire and heath

When these two models were ready, they had been applied separately to predict the probabilities of regions in the study area being forest or non-forest. Then predicted probabilities were combined together to form the outputs, in one of which regions with probabilities of being forest less than 0.5 have been categorized as non-forest, otherwise as forest; in the other one of which only the probabilities of being forest have been illustrated, so forest regions can be separated using any suitable threshold value (Figure 8). The successive estimations of further forest and non-forest attributes, i.e. vegetation type, AGE, and OPENNESS will be limited in the corresponding scopes of forest or non-forest set up here. Surely, regions in the study area not belonging to any of those forest or non-forest

vegetation types, e.g. arable land, water, and wasteland etc. have not been clipped from these outputs, but will be clipped from the final outputs of this paper.

There are some patterns of the distribution of forest can be observed from Figure 9. One is that forest areas mainly are distributed along the main rivers where SOIL type is 8 (sediment and rock), and repels SOIL 1 (peat). This partially coincides with the result of the logistic model, in which three variables: Y (ΔAIC = 40.36), SOIL (ΔAIC = 39.43, here df equals 2), and SLOPE (ΔAIC = 18.55) count for most of the model's explanatory power. Also forest covers about 35% of the study area, which matches the proportion of forest in the sample.

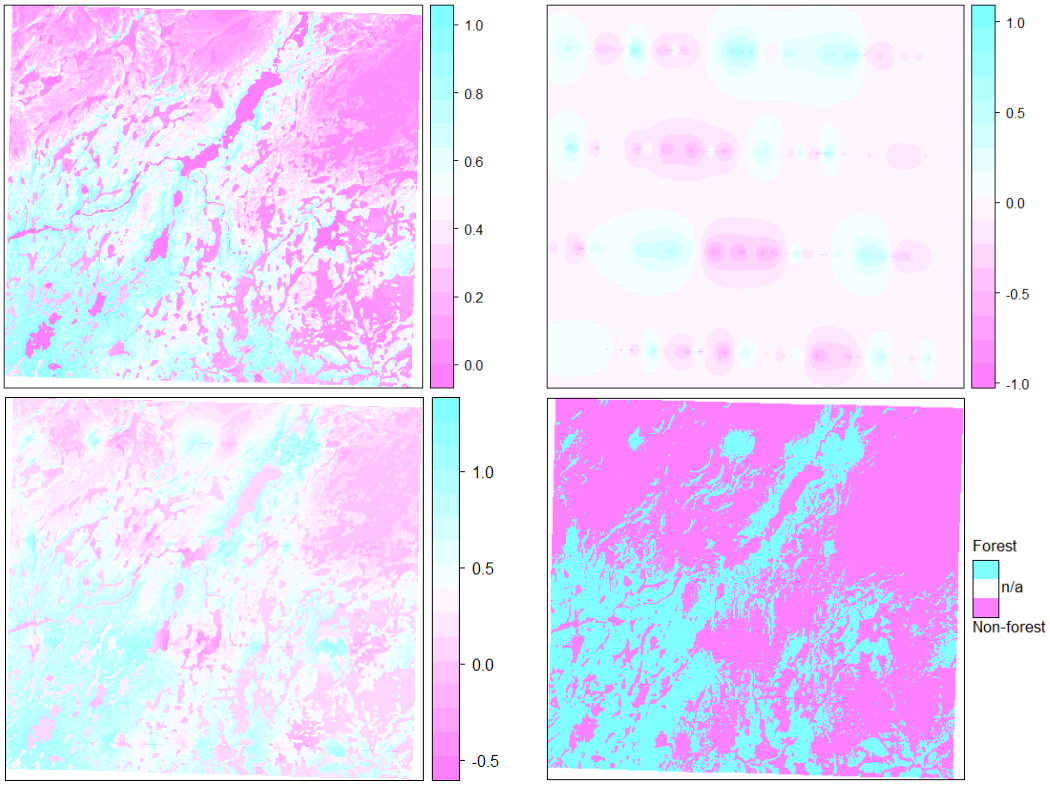

Figure 9. Outputs of the prediction on forest and non-forest, top left) using the logistic model; top right) using ordinary Kriging; bottom left) Combining two models (all these three outputs are scaled on the probability of success of forest); bottom right) Combining two models (mapped with the threshold probability of 0.5 of being forest)

Dividing non-forest into mire and heath

After forest and non-forest areas had been separated, the latter were further divided into areas of mire and heath following the same modeling approach as in the previous section. Table 8 shows the selected model, while Table 9 shows its overall performance, in which the p-value of the χ^2 statistic of the residual deviance suggests a high-level similarity

between the fitted and the observed values. This similarity results in an overall classification accuracy of 0.81 (coincide with a prediction error $\boldsymbol{\delta}$ of 0.199 of the adjusted leave-one-out cross-validation based on the cost function of $mean(|observed - predicted| > 0.5)$) with an un-weighted Kappa statistic of 0.61 on the sample dataset.

Table 8. Result of the selection of predictors of the regression model on mire and heath

	Predictors Discarded						**Predictors Retained**
	DIST_W	CURV_PL	X	DEM	CURV_PR	ASPECT	The Other Predictors
AIC	442.19	440.26	438.75	437.38	435.62	434.76	434.46

Table 9. Performance of the selected model on mire and heath

H_0: There is no difference between observed and fitted values.

$\boldsymbol{D_M}$	**df**	**p-value**	$\boldsymbol{\delta}$
418.46	505	0.998	0.199
	Observed		
Predicted	Heath	Mire	
Heath	201	53	
Mire	46	213	

The sample variogram and the fitted variogram model (SSE = 3.48e-06) for the residuals from the binary model on mire and heath are presented in Figure 8. Again, here predictions from both the binary model and Kriging were combined together to form the estimation for regions of mire and heath. The estimation results are shown in two ways (Figure 10), in one of which regions with probabilities of being mire less than 0.5 have been categorized as heath, otherwise as mire; in the other one of which only the probabilities of being mire have been illustrated, so regions of mire can be separated using any suitable threshold value. The same as in the first binary model, Y (ΔAIC = 12.78), SOIL (ΔAIC = 15.95, here df = 2), and SLOPE (ΔAIC = 4.39) are also the key determine factors here. The areas of mire highly accord with areas of SOIL 1 (peat) and concentrate at the southern part of the study area.

Distinguishing between coniferous, deciduous, and mixed forest

Forest areas separated from the first binary model were then classified into three vegetation types using a different multinomial logistic model, but the procedure of model selection is the same. Table 10 shows the selected model here, while its performance including an overall classification accuracy of 0.76 with an un-weighted Kappa statistic of 0.58 on the sample dataset can be found in Table 11.

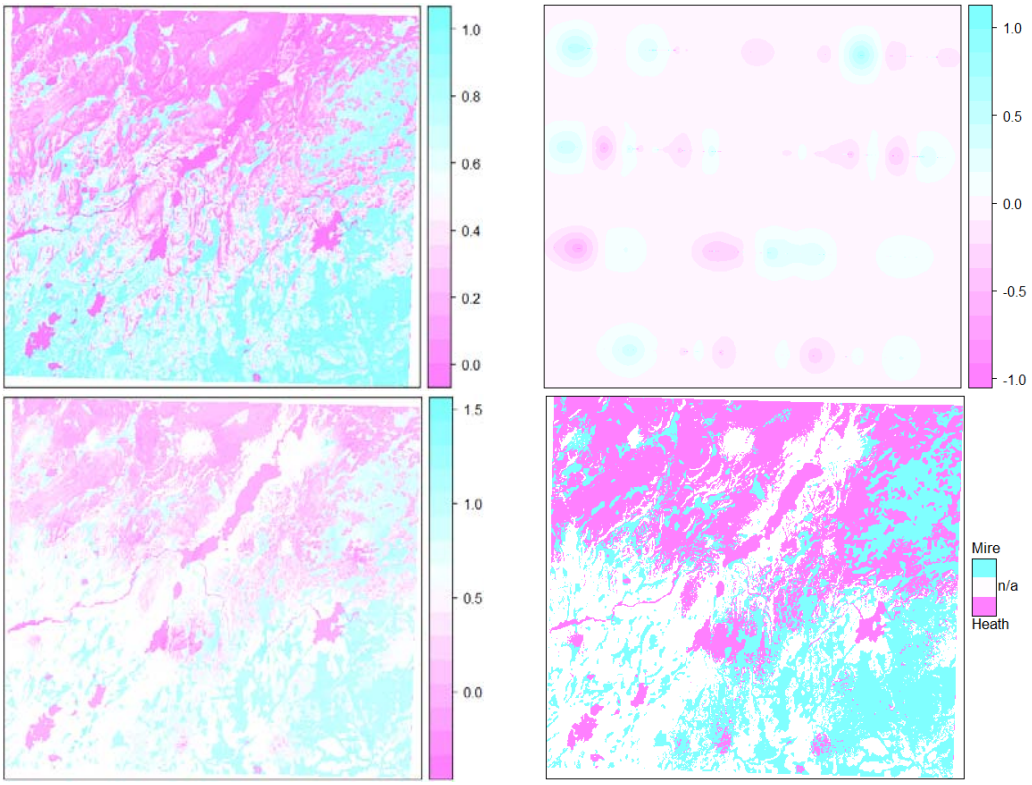

Figure 10. Outputs of the prediction on mire and heath, top left) using the logistic model; top right) using ordinary Kriging; bottom left) Combining two models (all these three outputs are scaled on the probability of success of mire); bottom right) Combining two models (mapped with the threshold probability of 0.5 of being mire, and clipped by the extent of non-forest).

Table 10. Result of the selection of predictors of the regression model on vegetation types

	Predictors Discarded				**Predictors Retained**
	CURV_PL	SOLAR	CURV_PR	WI	The Other Predictors
AIC	461.66	459.32	457.51	456.63	455.38

Table 11. Performance of the selected model on vegetation types
H_0: There is no difference between observed and fitted values.

D_M	**df**	**p-value**	
415.38	997	>0.9999	
	Observed		
Predicted	Coniferous	Deciduous	Mixed
Coniferous	148	20	21
Deciduous	16	101	20
Mixed	1	4	8

Following the routine, the sample variograms and the fitted variogram models (individual models for vegetation types of coniferous, deciduous, and mixed forests with SSEs of 7.77e-06, 1.01e-05, and 1.21e-06 respectively) for the residuals from the multinomial

model here are plotted in Figure 11, and estimations of forest vegetation types again were obtained by combining predictions from both types of models. Figure 12 presents the outputs of those estimations, in which the vegetation type of a certain region is the one with the highest probability of success.

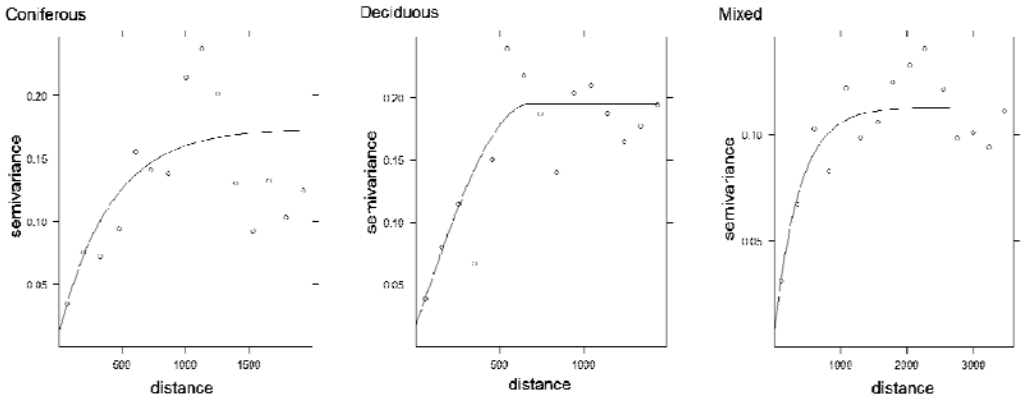

Figure 11. Sample variograms and fitted variogram models of the residuals from the multinomial logistic model, left) on coniferous forest; middle) on deciduous forest; right) on mixed forest

The single most important explanatory variable in this logistic model is DIST_A (ΔAIC = 28.86). Its coefficients of -2.21 and -2.23 imply the closer to arable land, the bigger chance to find deciduous or mixed forest compared to coniferous forest (the reference response class). SOIL (ΔAIC = 3.35) is of the second importance, which shows coniferous forests prefer sediment and rock, while deciduous forests prefer moraine. Mixed forest obviously is a minor element in the study area.

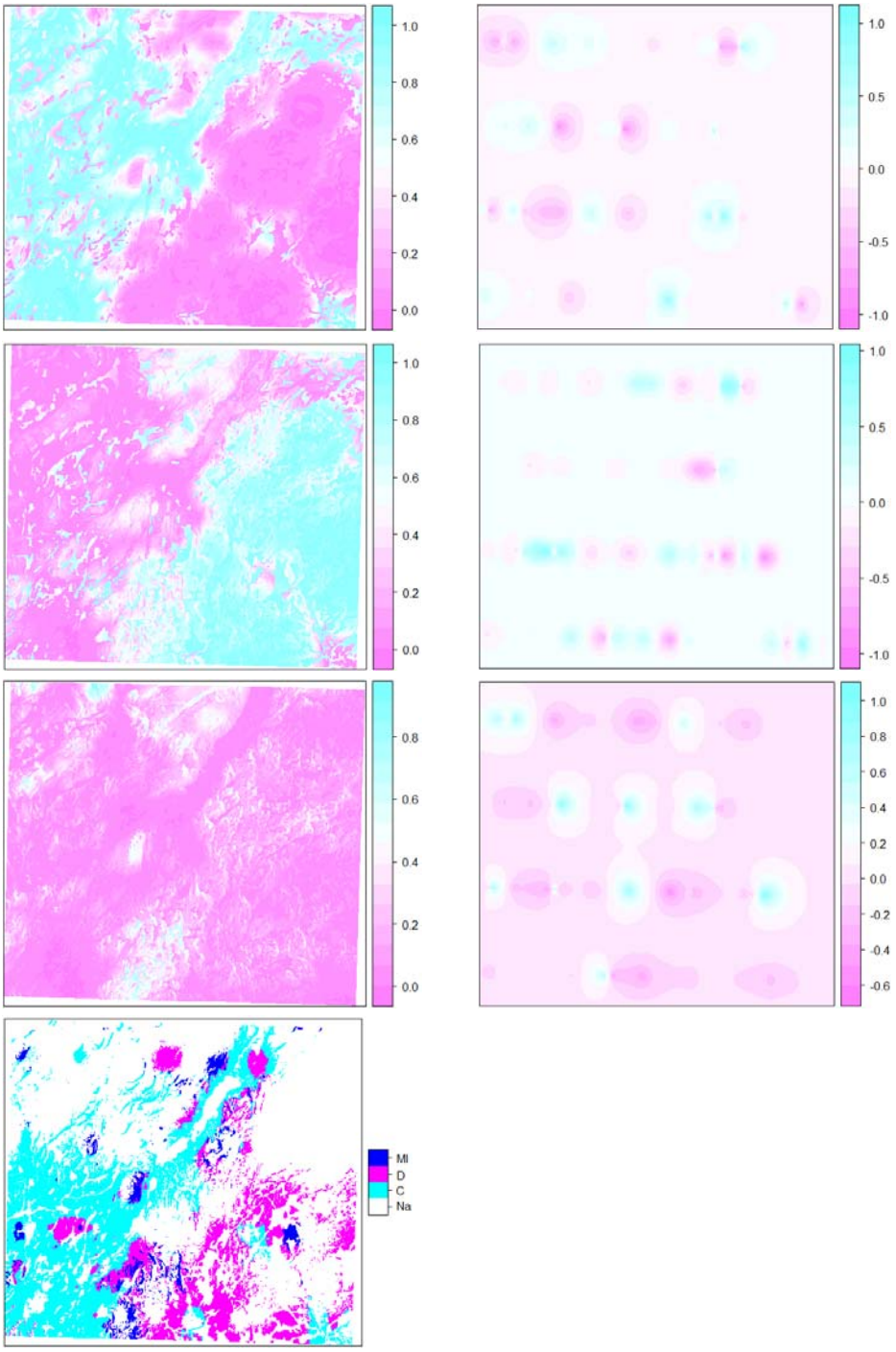

Figure 12. Outputs of the predictions on coniferous (1st row), deciduous (2nd row), and mixed forests (3rd row), left) using the logistic model; right) using ordinary Kriging (both scaled on the probability of success of a certain vegetation type); bottom) Combining two types of models (mapped with the vegetation type of the highest probability of success, and clipped by the extent of forest)

Estimating AGE and OPENNESS of forest stands

The last also the most challenging step in this modeling attempt is to model AGE and OPENNESS of forest stands, since all we can resort to are some seemingly weak correlations between the explanatory and the response variables plus the incomplete spatial coverage of Kriging prediction. In this step, except a general PLS model, individual PLS models for different SOIL types had also been applied, in which predictors had been standardized to avoid the influence of their scales on decomposition. The performances of these models by including different numbers of principal components are summarized by two means: 1) Figure 13 illustrates the changes of Root Mean Square Error of Prediction (RMSEP); 2) Table 12, 13, and 14 compare the percentages of variances of both the explanatory and the response variables being explained.

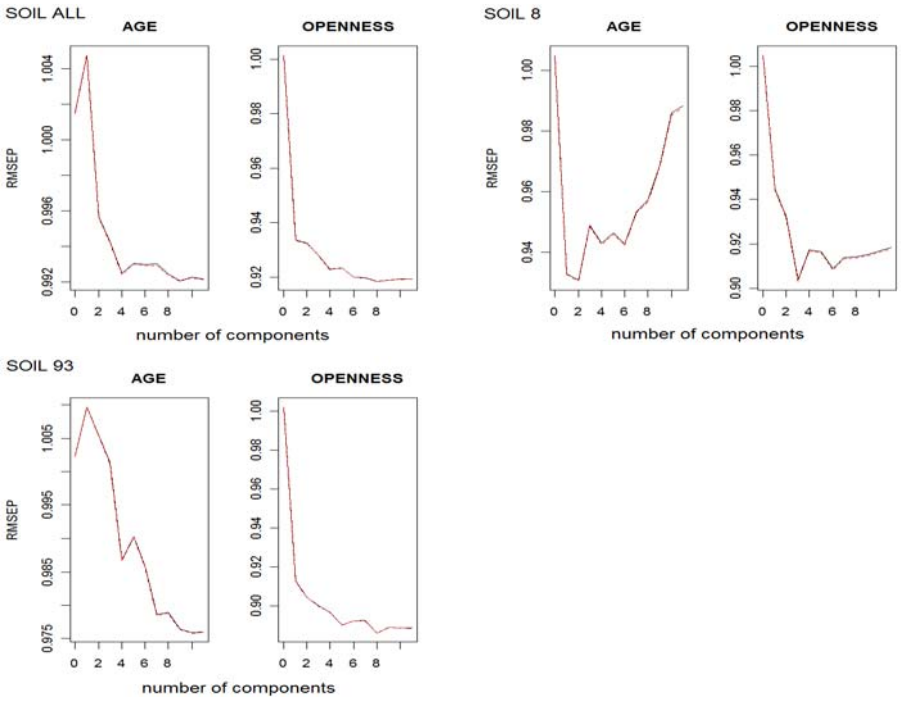

Figure 13. Standardized RMSEP (black solid lines for the ordinary leave-one-out cross-validation; red dotted lines (mostly overlap the black lines) for a bias-corrected one of that (Mevik and Cederkvist, 2004)) of the PLS models on, top left) all SOIL types; top right) SOIL type 8; bottom) SOIL type 93

Table 12. Percentages of variances explained by the PLS model on all SOIL types

	1 comps	2 comps	3 comps	4 comps	5 comps	6 comps	7 comps
Predictors	17.4589	30.376	40.701	46.914	61.018	68.681	75.050
AGE	0.2052	4.844	6.341	6.671	6.837	6.907	7.072
OPENNESS	16.0944	16.490	18.265	19.770	19.982	20.307	20.385
	8 comps	9 comps	10 comps	11 comps			
Predictors	81.920	88.957	94.530	100.000			
AGE	7.077	7.102	7.105	7.105			
OPENNESS	20.561	20.570	20.572	20.573			

Table 13. Percentages of variances explained by the PLS model on SOIL type 8

	1 comps	2 comps	3 comps	4 comps	5 comps	6 comps	7 comps
Predictors	18.25	27.96	39.45	52.30	60.39	70.40	84.44
AGE	20.62	23.87	24.67	26.14	26.37	26.54	26.58
OPENNESS	18.81	29.79	32.65	33.30	33.78	34.06	34.20
	8 comps	9 comps	10 comps	11 comps			
Predictors	90.34	94.81	99.51	100.00			
AGE	26.64	26.65	26.66	26.66			
OPENNESS	34.37	34.43	34.44	34.45			

Table 14. Percentages of variances explained by the PLS model on SOIL type 93

	1 comps	2 comps	3 comps	4 comps	5 comps	6 comps	7 comps
Predictors	17.6512	33.561	43.60	53.32	65.64	72.14	77.54
AGE	0.1639	4.134	8.91	11.17	11.18	11.95	12.96
OPENNESS	21.4819	23.238	25.73	26.47	27.64	28.00	28.03
	8 comps	9 comps	10 comps	11 comps			
Predictors	82.12	89.72	95.05	100.00			
AGE	13.13	13.24	13.24	13.25			
OPENNESS	28.24	28.27	28.28	28.28			

The summarization above indicates that candidate PLS models can only explain a maximum of 7 to 26% of the variation on AGE, 20 to 34% of that on OPENNESS, which is a quite poor performance. However there are two factors may still render the modeling useful: 1) The variables AGE and OPENNESS include more discrete values (see Table 1) than the main three classes, e. g., “low”, “medium”, and “high”, to which the final results will be aggregated. The aggregation will improve the estimation performance substantially, and the reason it will be performed afterwards is to keep its flexibility; 2) The predictions from Kriging will also be combined.

The selected PLS models on all SOIL types, SOIL type 8, and SOIL type 93 included 9, 3, and 10 principal components respectively, so maximum amounts of variances in both response variables of each model can be explained (the same as RMSEPs being minimized). The sample variograms and the fitted variogram models (SSE of 21 and

1.53e-07) combining residuals from all three PLS models on AGE and OPENNESS are presented in Figure 14. The sample variograms do not show clear patterns of spatial autocorrelation, which is preferred if the previous PLS models had explained most of the variances in response variables. Otherwise not much may be expected from including Kriging prediction.

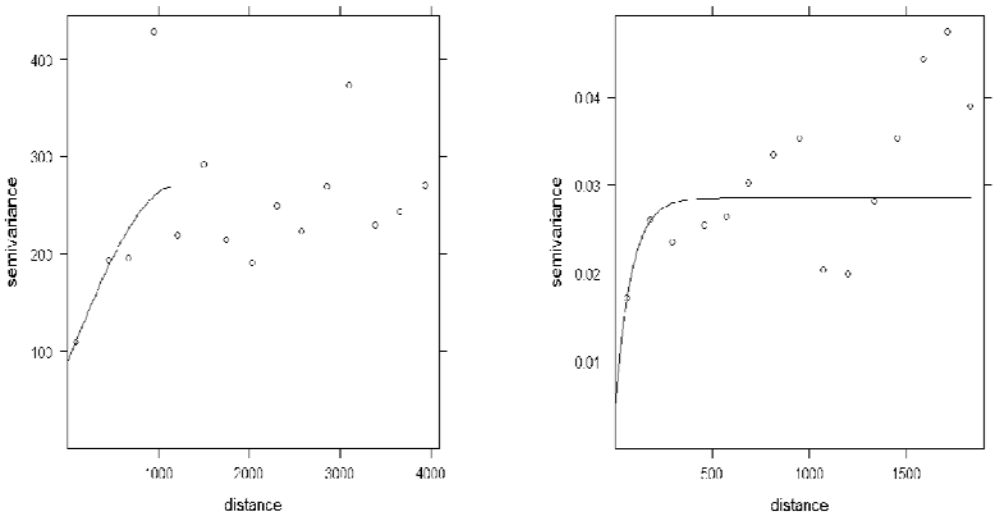

Figure 14. Sample variograms and fitted variogram models of the combined residuals from, left) the three PLS models on AGE; right) the three PLS models on OPENNESS

Prediction result on either AGE or OPENNESS from regression was obtained by applying three individual PLS models, while that from Kriging was obtained by applying a single model combining residuals from all three PLS models. In the outputs that merge predictions from those two types of models, predicted values had been grouped into classes in accordance with the classes of the original data shown in Table 1. Figure 15 and 16 present these three sorts of prediction results on AGE and OPENNESS, the former of which covers 5 of the 6 original AGE classes (except class "101 - 120") and suggests the classes of "21 - 40" and "41 - 60" are the dominant classes; whilst the latter of which covers all 7 OPENNESS classes, and indicates classes "0.5 – 0.6" and "0.7 – 0.8" have the largest extent. Either old forest or sparse forest is very uncommon in the study area. The most important predictors for each PLS model are summarized in Table 15.

Table 15. Standardized predictors with the highest weights in each PLS models (coefficients in parentheses)

Model		**Predictor**		
All SOIL Types	**AGE**	DIST_A (-0.295)	X (0.103)	SOLAR (0.090)
	OPENNESS	Y (-0.332)	DEM (-0.297)	DIST_A (-0.237)
SOIL Type 8	**AGE**	DEM (0.195)	DIST_W (0.181)	Y (0.173)
	OPENNESS	CURV_PL (0.281)	DIST_W (-0.258)	SOLAR (0.209)
SOIL Type 93	**AGE**	DIST_A (-0.478)	X (0.367)	Y (-0.293)
	OPENNESS	X (-0.359)	Y (-0.309)	SOLAR (0.262)

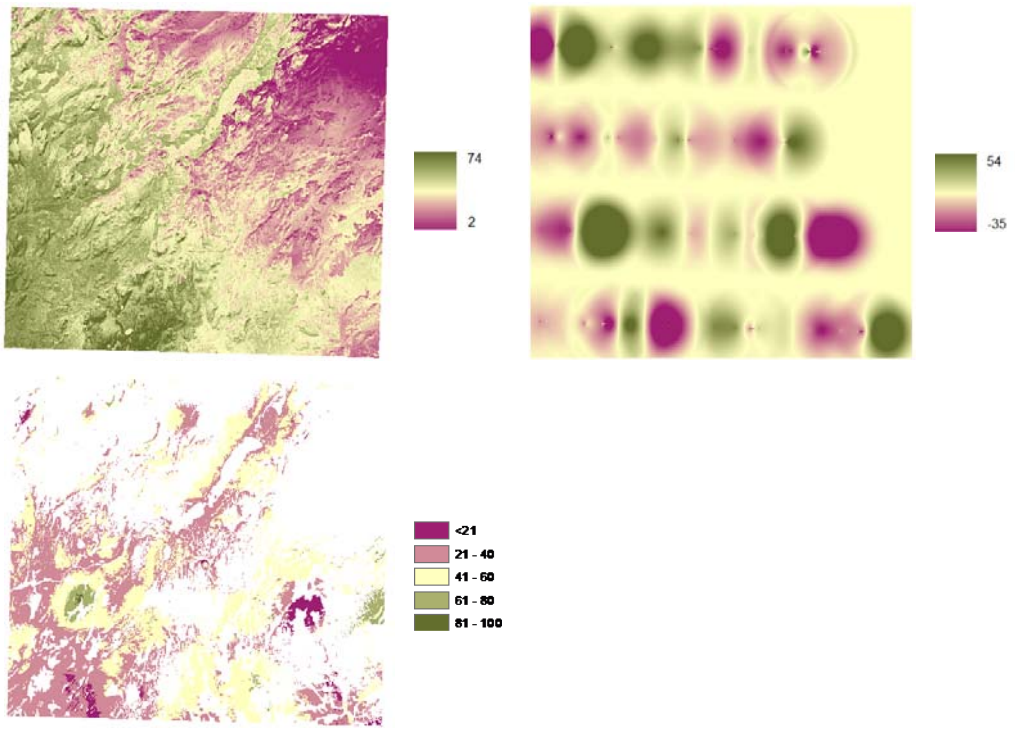

Figure 15. Outputs of the prediction on AGE, top left) using the three PLS models; top right) using ordinary Kriging; bottom) Combining two types of models

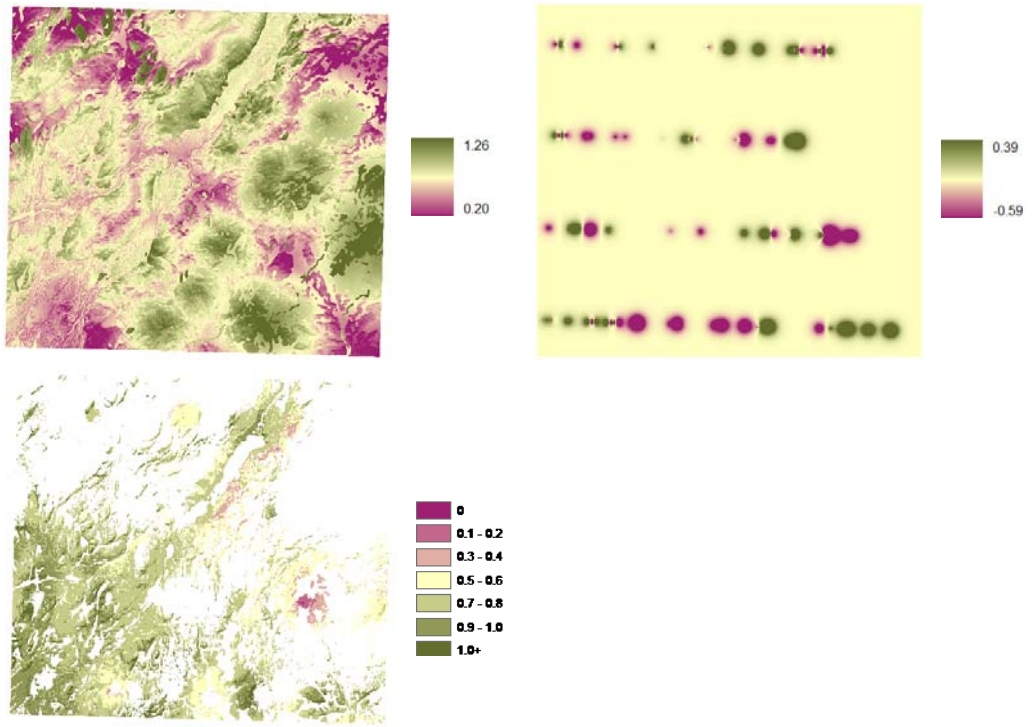

Figure 16. Outputs of the prediction on OPENNESS, top left) using the three PLS models; top right) using ordinary Kriging; bottom) combining two types of models

The k-NN reference of estimation

As planned, estimations had also been made by implementing k-NN method in order to set up a reference to the modeling attempt, and the results of which are manifested in Figure 17. It is noticed that they share some common features, albeit more dissimilarities with the modeling outputs. Despite these, the outputs here look patchier and less smooth, which probably not agree with the real landscape structure. This is a drawback of k-NN method since the exact known values of target variables will be imputed, which ends up in discrete and enveloped estimation results. Such a problem is often dealt with by enlarging the support, which is not the case here since it is just used as a reference. Also because categorical variables cannot be included in the weights of distances that is the inversed covariance matrix used in the k-NN method here, from which the predictor of SOIL is excluded, and this is a loss of important information.

A direct comparison between the performances of k-NN method and the modeling effort on AGE and OPENNESS can be made by using corresponding RMSEs (Crookston and Finley (2008) consider it is more appropriate to name it Root Mean Square Difference (RMSD) in the k-NN case). Such a comparison reveals the modeling effort considerably outperforms k-NN method (Table 16). The performance of k-NN method on classification of vegetation type is summarized in Table 17, which leads to an overall classification accuracy of 0.72 with an un-weighted Kappa statistic of 0.62 on the sample dataset. This performance is so

close to that of the modeling that hardly a judgment can be made upon them.

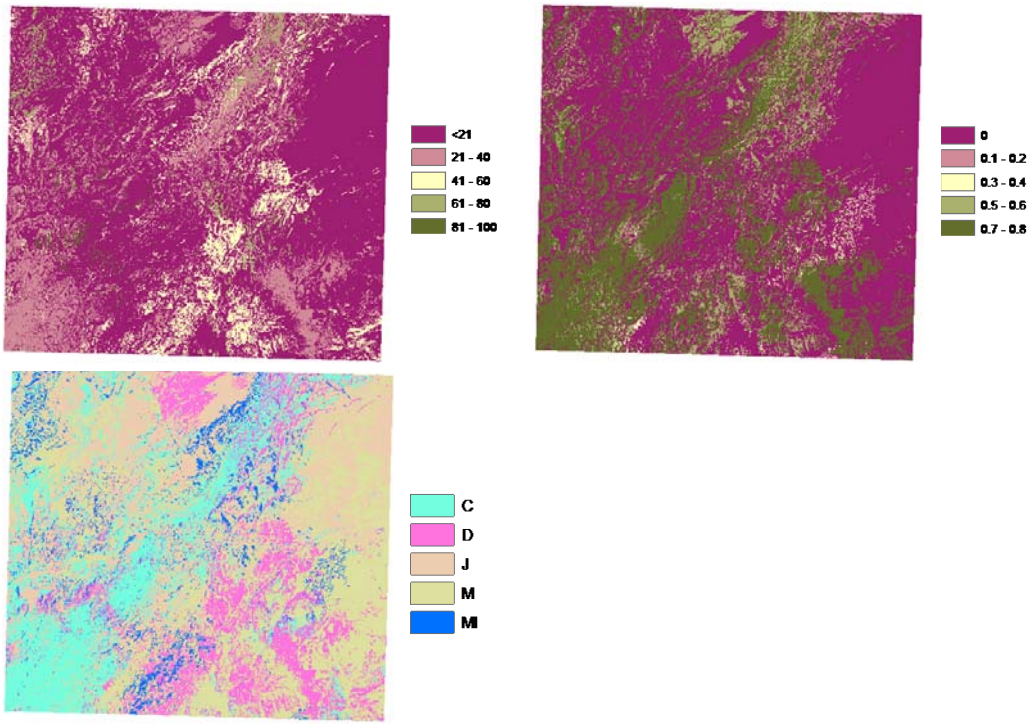

Figure 17. Outputs of predictions from implementing k-NN method on, top left) AGE; top right) OPENNESS; bottom) vegetation type.

Table 16. RMSEs on AGE and OPENNESS of k-NN method and the modeling

	AGE (year)		**OPENNESS**	
	The Modeling	k-NN	The Modeling	k-NN
RMSE	18.45*	23.64	0.17*	0.36

* It should be noted that the RMSEs of the modeling is that of PLS models. After Kriging was implemented to the residuals of those models, the Kriging variance (hence RMSE) should be considerably lower. Although to which extent the RMSEs will be lowered is hard to be reliably summarized (which partially is a drawback of the way that NFI sample plots had been allocated (Figure 3), since the attribute value of each of these plots (over 100 m long on average) was assigned to all the grid cells it contains, which causes a certain grid cell is surrounded by its neighbors with the same attribute value and high weighs, finally leads to underestimated and unrealistic low Kriging variance).

Table 17. Classification performance of k-NN method on vegetation type

	Observed				
Predicted	**C**	**D**	**MI**	**M**	**J**
C	115	4	15	17	3
D	10	81	11	14	10
MI	22	11	187	22	5
M	17	21	29	205	7
J	1	8	5	8	24

Final outputs

Combining all elements, final outputs of this modeling are manifested in Figure 18 and 19.

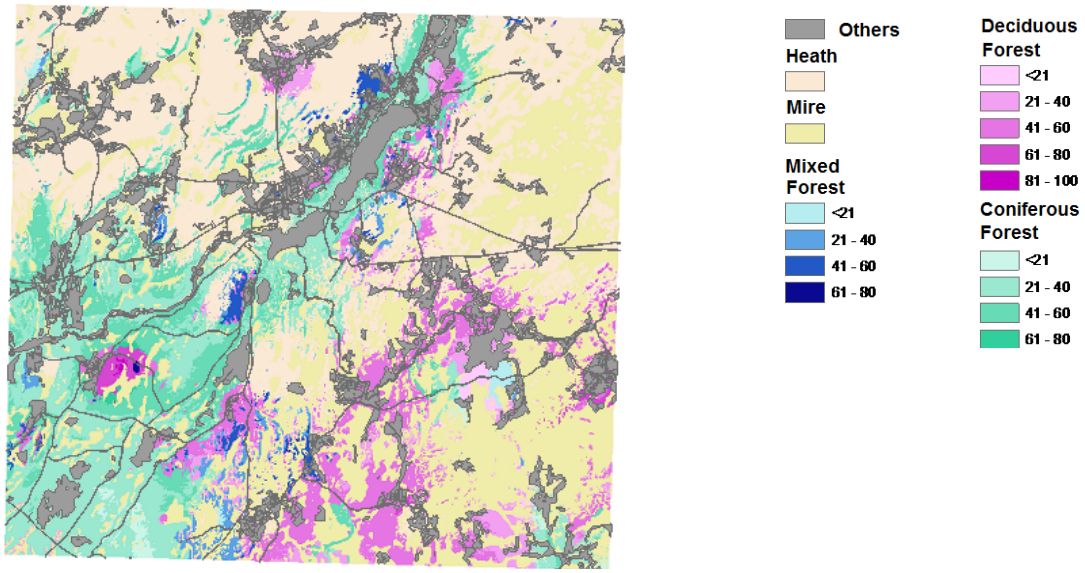

Figure 18. Final output of AGE on different vegetation types; areas in grey are anything but out of the scope of this study, e. g., arable land, water, roads, and etc. © Lantmäteriet, I2011/0032.

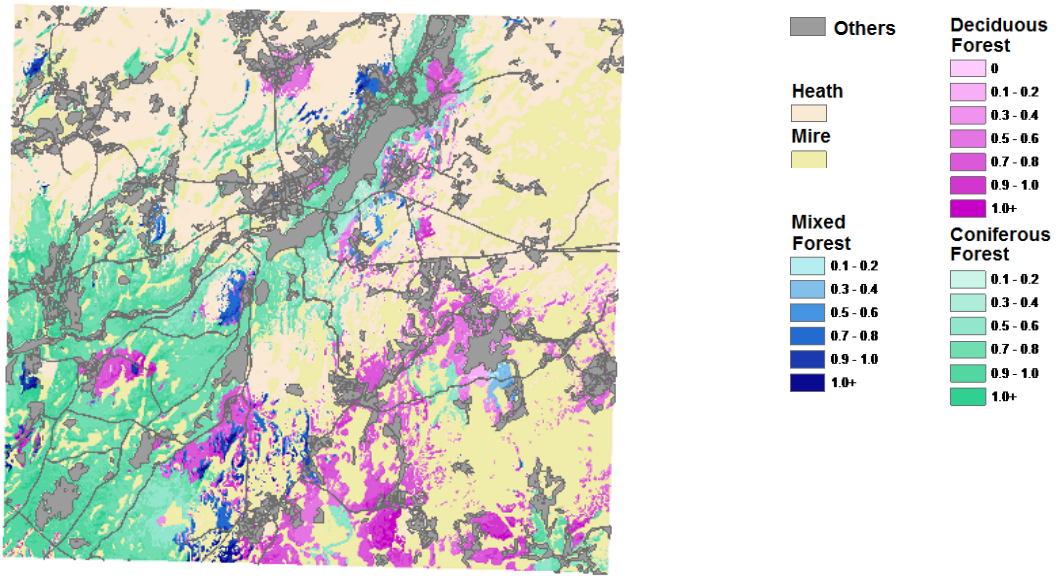

Figure 3. Final output of OPENNESS on different vegetation types; areas in grey are anything but out of the scope of this study, e. g., arable land, water, roads, and etc. © Lantmäteriet, I2011/0032.

Comparison of predictions based on two different data sources

The predictions made by the modeling are based on the first NFI data. These same predictions can also be obtained by using information from the historical economic maps. A comparison of predictions based on these two different data sources will serve as a sort of validation. Cross-validations had already been implemented for individual models before.

The main reason why not following the conventional way to divide samples into two groups as cases and controls to perform the validation is the inadequate sample size. Among a total of 1019 sample cells, 164 of them fall in water, arable land, and wasteland etc., which is not a part of this study; within the 855 sample cells left, only 342 are of forest, which are further separated by different types of soil. Also, the relatively large number of predictors means fewer samples will fall on a certain combination of predictors. This is a limitation of using parametric approaches where large amount of data is required. Instead, nonparametric approaches could be considered.

Besides, even a few samples on each soil type had been selected to form a control group, validation based on which probably would not be reliable because the fitness is more likely a result of chance (Kravchenko, 2003). The same author suggests permuting the samples of cases and controls a number of times, and performing validation on the simulated control groups to obtain more generalized and reliable information on the performance of the modeling. However, this is way beyond the computational capacity given the number of models included in this study.

Comparison on OPENNESS

It has been considered that the key choice of the computation of point intensity is not the specific kernel function but the bandwidth used (Bivand et al., 2008). When applying the quartic kernel function, we selected the bandwidths by following the proposal of Berman and Diggle (1989) on minimizing the Mean Square Error (MSE) of the kernel estimator. Figure 20 shows such a selection of bandwidths and Figure 21 shows the estimations of point intensities based on the selected optimal bandwidths.

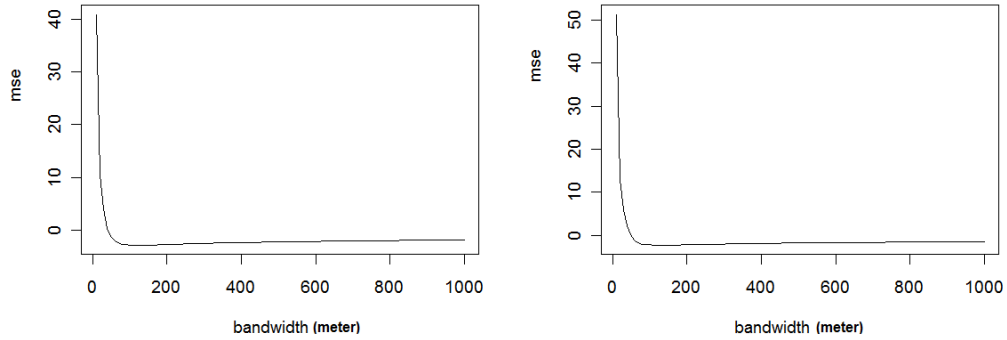

Figure 20. MSEs (rescaled (Anonymous, 2011)) of point intensity at different bandwidths using quartic kernel, left) for coniferous forests, minimized at 120 m; right) for deciduous forests, minimized at 150 m.

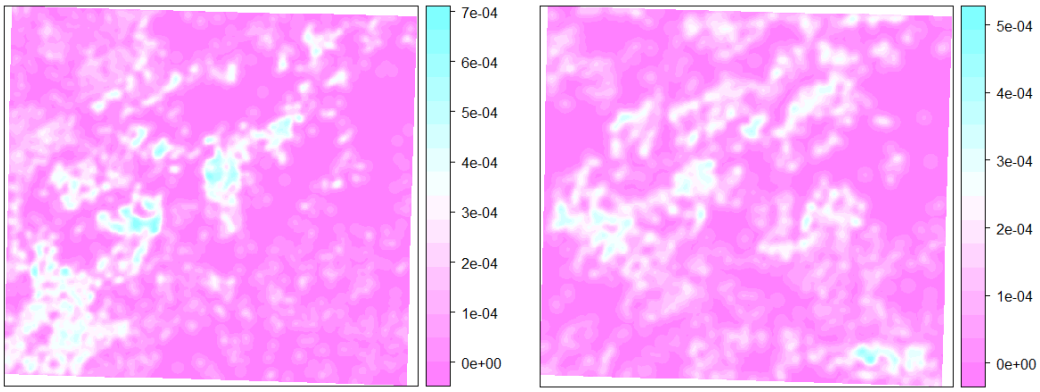

Figure 21. Estimated point intensities of the digitized historical county economic maps based on quartic kernel function with selected optimal bandwidths, left) for coniferous forests; right) for deciduous forests.

Once obtained point intensities, its connection with OPENNESS has to be built in order for it to be used for the comparison. The bridge of this connection is the NFI samples used in this study. One problem is that point intensities here represent forest densities, and the same density does not necessarily lead to the same OPENNESS, which is also heavily decided by AGE. The adjusted R^2 of 0.3004 of the connection between OPENNESS and intensity, of 0.7563 between OPENNESS and intensity plus AGE confirmed this argument.

Unfortunately, AGE is one of the variables to be estimated, not already known in this study. In this case, estimated values of AGE had been used to transfer point intensity to OPENNESS. Although not preferred, this operation nevertheless still helps improve the precision of the transfer to a certain extent.

Finally, the prediction of OPENNESS through the digitized 1920s economic map had been generated (clipped by the extent of forest in these maps). It should be noted that heath is a part of forest in these maps, and assigned with the attributes of OPENNESS and AGE. However this was not the case in the first NFI data, on which the estimation was based. This disparity certainly contributed much to the differences between OPENNESS estimated from these two different data sources (Figure 22). Yet this inconsistency is an unchangeable fact, and historical county economic maps will set the par for the comparisons here and afterwards, hence the modeling outputs will be retailored.

Except the visual comparison on Figure 22, the agreement on the estimations of OPENNESS from these two different data sources is summarized in Table 18, while that between using k-NN method and historical county economic maps is in Table 19. The former presents an overall agreement of 0.52 with a weighted Kappa statistic of 0.11, while those of the latter are 0.27 and 0.09.

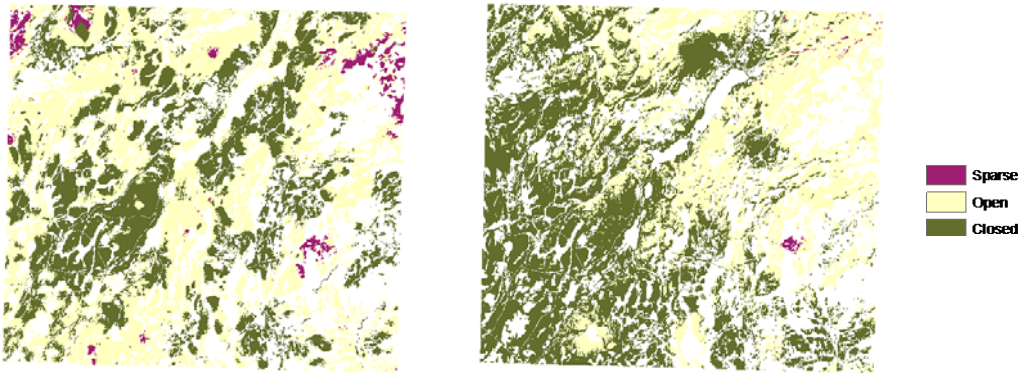

Table 18. Summary of the agreement on the estimations of OPENNESS between using 1920s economic maps and the 1928 NFI data

	Historical County Economic Maps			
The Modeling (First NFI)	Sparse	Open	Closed	Total
Sparse	0.2%	0.1%	0.0%	0.3%
Open	2.0%	23.7%	14.8%	40.5%
Closed	0.8%	30.3%	28.0%	59.1%
Total	3.0%	54.1%	42.8%	100.0%

Table 19. Summary of the agreement on the estimations of OPENNESS between using 1920s economic maps and k-NN method

	Historical County Economic Maps			
k-NN	Sparse	Open	Closed	Total
Sparse	2.4%	33.6%	20.2%	56.2%
Open	0.3%	5.7%	3.9%	9.9%
Closed	0.3%	14.9%	18.8%	34.0%
Total	3.0%	54.2%	42.9%	100.0%

Comparison on vegetation type

The agreement on the estimations of vegetation type from the two different data sources is summarized in Table 20 (and can be visually compared from Figure 23), while Table 21 is that between using k-NN method and historical county economic maps. The overall agreement for the former is 0.34. The misinterpretation mostly rests on mixed forest, which is the major vegetation type in historical county economic maps. However, there are merely 49 out of 1019 NFI samples are of mixed forest, and when the modeling is based on this, the result here is really not a surprise. And again, the modeling outperforms the k-NN method, which has an overall accuracy of 0.19 here.

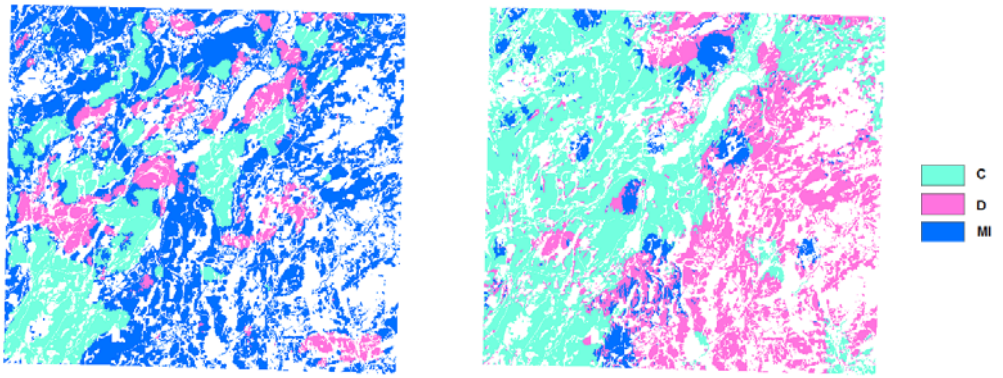

Figure 23. Estimated vegetation classes ("C" for coniferous forest; "D" for deciduous forest; "MI" for mixed forest), left) from historical county economic maps; right) from the modeling result of this study based on the first NFI data.

Table 20. Summary of the agreement on the estimations of vegetation type between using historical county economic maps and the first NFI data

	Historical County Economic Maps			
The Modeling (First NFI)	Coniferous	Deciduous	Mixed	Total
Coniferous	21.3%	7.1%	26.4%	54.8%
Deciduous	2.9%	5.5%	27.9%	36.3%
Mixed	1.3%	0.9%	6.9%	9.1%
Total	25.5%	13.5%	61.2%	100.0%

Table 21. Summary of the agreement on the estimations of vegetation types between using historical county economic maps and k-NN method (*Others* here means non-forest vegetation types, since they had not been imputed separately with forest vegetation types)

	Historical County Economic Maps			
k-NN	Coniferous	Deciduous	Mixed	Total
Coniferous	11.7%	3.0%	8.9%	23.6%
Deciduous	1.9%	3.0%	10.5%	15.4%
Mixed	1.3%	1.0%	4.0%	6.3%
Others	*10.4%*	*6.4%*	*37.8%*	*54.6%*
Total	*25.3%*	*13.4%*	*61.2%*	*100.0%*

Conclusion and Perspective

In summary, the performance of the modeling works fairly well for the vegetation type while it is barely acceptable for the forest stand age and openness part. Quite common as for similar studies, this situation is mainly caused by the nature of the available data, to which there hardly is a foreseeable solution. Nevertheless, the way the existing data soruces are combined and utilized is also a causal factor behind the results, and this may be improved.

Constraints of data

In the modeling, a fundamental assumption is that the distributions of forest attributes of interest are varying with environmental covariates, which could be topographical, pedological, climatic, spectral etc. However, the lack of relevant data containing such covariates, or from which such covariates could be derived, is the most troublesome constraint being encountered (Figure 6).

To make things worse, such limited data regarding those covariates are seldom available from the same era of this study. Thus, contemporary surrogates or proxy data have to be considered, and this may be problematic. For instance, spectral information from remote sensing supplies valuable direct measurement subject to this study, but there hardly is a means to utilize it rationally in the historical context. To put the rationality of using specific contemporary surrogates or proxies in this study under scrutiny, the changes of terrain (DEM) and soil types are admittedly extremely slow processes, so as for the changes of water areas especially in the rural part of Sweden. He et al. (2007) confirmed this point of view. But it may not be reasonable to neglect the changes in areas of arable land, since it decreased by about 1.2% (or about 5,000 km^2) of the total land area of Sweden just from 1967 to 2008 (Trading Economics, 2011). At a regional level changes in area of arable land were even larger.

Therefore, even the utilization of some of those covariates (their surrogates) can be justified; they probably have already lost much of their explanatory power to factors concerning human involvement in forest. For instance, the long history of forest management in Sweden makes the distribution of forest more as consequences of the supply and demand of forestry products, human preference, social development, and etc., less affected by environmental factors even in the 1920s.

Another major constraint is for the two data sources containing information of forest attributes in this study, i.e. the first NFI and the 1920s economic map. Since the interest of the former lay on the living stock and increment of forest, while forest is just one of many components of the latter, thereby it was not always the case that these two data sources can supply sufficient information to generate those attributes from, not to mention being of the same caliber.

This not only makes the modeling less productive and reliable, but makes verifying the predictions from these two data sources of less value. For instance, with its full coverage of the study area, the historical county economic maps are supposed to supply the ground truth of the study objects, against which a validation of the modeling result can be made. However, in order to achieve an estimation of forest stand openness — one of the variables of interest, forest density had to be estimated from symbols in these maps first, this estimation then was used to facilitate a linear function to obtain the estimation intended. Along the process of estimation on estimation, uncertainties accumulated intolerably; and the only hypothesis can be rejected by comparing estimations from these two data sources is that they both fit the sample well.

Another example regarding this constraint is about heath in the study area, which is treated separately from forest in the first NFI data, but not in the historical county economic maps. Attempting to include heath in forest in the first NFI data to make these two data sources compatible did not seem possible, because although NFI sample plots of heath have the attribute of stand openness, but not stand age (for most of them), PLS regression that estimates multiple response variables simultaneously thereby treats all plots of heath as missing values (absent stand age values of heath plots could be imputed firstly, but the extra uncertainty arisen surely devalues the effort).

After all, these constrains are quite common for studies of historical landscapes. To deal with them, some compromises and premises have to be made before any progress can be achieved. But this whole procedure must be guided by proper historical methods, e. g. explication of the interconnectedness of variables, contextualization and causation of events, tolerance of ambiguity, and etc. (Lewenson and Herrmann, 2008), which is the exact practice all through this modeling.

Future improvement of modeling

Despite all the constraints with the used approach, the modeling methods certainly could be further improved. So far two sorts of potential improvement approaches have been perceived (but not tried). Besides these there are some other premature and rather wild notions.

1) For data preparation

Grid cell size has heavy influence on the modeling result that larger size generally leads to less noisy hence statistically better performance (at the cost of the compromise of some information), also the computational convenience, but may also make the sample size less than sufficient to build all the models on. Also, different grid cell sizes offer different opportunities to reveal spatial autocorrelations within samples. In this study, concerns about grid cell size were placed on trying to enlarge it so two soil types (the minimum area of any certain soil type is 400 m^2), so as for two NFI sample plots (the minimum distance between any two plots is 30 m) will not fall into the same cell. But this enlargement should

optimize, not undermine the usability of this modeling. Although, whether this is the case can only be found out after repeated trials of different cell sizes (meaning the whole time consuming modeling process), which have yet been done.

2) For statistical techniques

The motivation to apply Kriging after regression in this modeling is that we believe there is autocorrelation left in the residuals of regression, which needs to be caught in order to reach a better estimation. As a matter of fact, this modeling do benefit from this design. But this however contradicts with the assumption of parameter estimation of the regression models used in this study that residuals of each of those models are independent (having constant error variance). Therefore, the formerly estimated regression model parameters are biased. To deal with this problem and further enhance the estimation, covariance matrix of errors estimated during applying Kriging can be used to re-estimate error variance, hence parameters of regression models. This process then should be iterated until convergence. The "NeweyWest" function from package "sandwich" (Zeileis, 2004) implementing heteroskedasticity and autocorrelation consistent (HAC) estimator, and the "gls" function from package "nlme" (Pinheiro et al., 2011) performing Generalized Least Squares (GLS) regression of R make this theory generally applicable, but maybe not in the specific PLS regression context.

Another statistical technique that might be applied is Bayesian inference. As the study area of Halland has an interesting and dramatic forest history, there are many studies (both new and older) that could supply additional valuable information for enhanced modeling (constructing priors). The broad use of Bayesian inference also boosted the development of relevant programs and software. Specifically in R, there are plenty of functions suitable for many different statistical models (all the models used in this study).

Acknowledgement

I thank my supervisor, Dr. Anna-lena Axelsson for her tremendous support, patience, and kindness. I also thank the examinor of this thesis, Professor Håkan Olsson, and Professor Jun Yu, who acted as assistant supervisor, Mr. Sören Holm and Mr. Mikael Egberth for providing valuable advices. I am also grateful for Dr. Torgny Lind's generous help all through my study at SLU.

Mr. Nelson Sherman digitized tree symbols and land-cover information from the 1920 economic map. Historical NFI data from 1920s was supplied in database format by RINFI (Research Infrastructure NFI) www.slu.se/histtax. Digital maps were provided by the Swedish mapping, cadastral and land registration authority. © Lantmäteriet, I2011/0032. Soil data was provided by the Geological Survey of Sweden under the contract 721-1636/2011.

References

Anonymous, 2011. http://www.rni.helsinki.fi/~jmh/ss03/points2.pdf

Axelsson, A-L, Egberth, M and Olsson , H. 2012. Modelling historical forest landscapes by combining sample based forest inventory data and maps. Submitted manuscript.

Baayen, R. H., 2011. languageR: Data sets and functions with "Analyzing Linguistic Data: A practical introduction to statistics". R package version 1.1. http://CRAN.R-project.org/package=languageR

Belsley, D. A., Kuh, E., Welsch, R. E., 1980. Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. Wiley Series in Probability and Mathematical Statistics. Wiley, New York, USA.

Berman, M., Diggle, P. J., 1989. Estimating Weighted Integrals of the Second-order Intensity of a Spatial Point Process. Journal of the Royal Statistical Society 51: 81–92.

Bivand, R. S., Pebesma, E. J., Gómez-Rubio, V., 2008. Applied Spatial Data Analysis with R. Springer, New York, USA.

Blennow, K., Hammarlund, K., 1993. From Heath to Forest: Land-Use Transformation in Halland, Sweden. Ambio 22(8): 561-567.

Brown, D. G., 1998. Mapping Historical Forest Types in Baraga County Michigan, USA as Fuzzy Sets. Plant Ecology 134: 97–111.

Canty, A., Ripley, B., 2011. boot: Bootstrap R (S-Plus) Functions. R package version 1.3-2.

Crookston, N. L., Finley, A. O., 2007. yaImpute: An R Package for k-NN Imputation. Journal of Statistical Software 23(10): 1–16.

González, I., Déjean, S., 2009. CCA: Canonical correlation analysis. R package version 1.2. http://CRAN.R-project.org/package=CCA

Gu, W., Heikkilä, R., Hanski, I., 2002. Estimating the Consequences of Habitat Fragmentation on Extinction Risk in Dynamic Landscapes. Landscape Ecology 17: 699–710.

Hanski, I., 1998. Metapopulation Dynamics. Nature 396: 41–49.

Harrell, F. E. Jr., 2010. Hmisc: Harrell Miscellaneous. R package version 3.8-3. http://CRAN.R-project.org/package=Hmisc

He, H. S., Dey, D. C., Fan, X., Hooten, M. B., Kabrick, J. M., Wikle, C. K., Fan, Z., 2007. Mapping Pre-European Settlement Vegetation at Fine Resolution Using a Hierarchical Bayesian Model and GIS. Plant Ecology 191: 85–94.

Hooten, M. B., 2001. Modeling the Distribution of Ground Flora on Large Spatial Domains in the Missouri Ozarks. Master's Thesis, University of Missouri-Columbia.

Isaaks, E. H., Srivastava, R. M., 1989. An Introduction to Applied Geostatistics. Oxford University Press, New York, USA.

Keitt, T. H., Bivand, R., Pebesma, E., Rowlingson, B., 2011. rgdal: Bindings for the Geospatial Data Abstraction Library. R package version 0.7-1. http://CRAN.R-project.org/package=rgdal

Kravchenko, A. N., 2003. Influence of Spatial Structure on Accuracy of Interpolation Methods. Soil Science Society of America Journal 67: 1564–1571.

Leisch, F., Hornik, K., Ripley, B. D.,. 2011. mda: Mixture and flexible discriminant analysis. R package version 0.4-2. http://CRAN.R-project.org/package=mda

Lewenson, S. B., Herrmann, E. K., editors, 2008. Capturing Nursing History: A Guide to Historical

Methods in Research. Springer, New York, USA.
Lindbladha, M., Hultberga, T., Widerberga, M. K., Felton, A., 2011. Halland's forests during the last 300 years: a review of Malmström (1939). Scandinavian Journal of Forest Research 26(S10): 81-90.
Menard, S., 2001. Applied Logistic Regression Analysis. Second Edition. Sage University Papers Series on Quantitative Applications in the Social Sciences 07-106. Sage, Thousand Oaks, CA, USA.
Mevik, B. H., Cederkvist, H. R., 2004. Mean Squared Error of Prediction (MSEP) Estimates for Principal Component Regression (PCR) and Partial Least Squares Regression (PLSR). Journal of Chemometrics, 18(9): 422–429.
Mevik, B. H., Wehrens, R., 2011. pls: Partial Least Squares Regression (PLSR) and Principal Component Regression (PCR). R package version 2.2-0. http://CRAN.R-project.org/package=pls
Meyer, D., Zeileis, A., Hornik, K., 2011. vcd: Visualizing Categorical Data. R package version 1.2-12. http://CRAN.R-project.org/package=vcd
Pebesma, E. J., 2004. Multivariable geostatistics in S: the gstat package. Computers & Geosciences 30: 683–691.
Peterson, J, K., 1998. Logistic Regression Applications and Cluster Analysis. Master's Thesis, Texas Tech University.
Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., R Development Core Team, 2011. nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-98.
R Development Core Team, 2011. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
Ribeiro, P. J. Jr., Diggle, P. J., 2001. geoR: a package for geostatistical analysis. R-NEWS, 1(2): 15-18.
Rowlingson, B., Diggle, P., Bivand, R., 2010. splancs: Spatial and Space-Time Point Pattern Analysis. R package version 2.01-27. http://CRAN.R-project.org/package=splancs
Thorell, K. E., Ostlin, E. O., 1931. The National Forest Survey of Sweden. Journal of Forestry 29: 585–591.
Trading Economics, 2011. Sweden – National Statistical Data. http://www.tradingeconomics.com/sweden/indicators
Turner, M. G., 1989. Landscape Ecology: The Effect of Pattern on Process. Annual Review of Ecology and Systematics 20: 171–197.
Venables, W. N., Ripley, B. D., 2002. Modern Applied Statistics with S. Fourth Edition. Springer, New York, USA.
Winship, C., Mare, R. D., 1984. Regression Models with Ordinal Variables. American Sociological Review 49: 512–525.
Yeniay, Ö., Göktas, A., 2002. A Comparison of Partial Least Squares Regression with Other Regresssion Methods. Hacettepe Journal of Mathematics and Statistics 31: 99–111.
Zeileis, A., 2004. Econometric Computing with HC and HAC Covariance Matrix Estimators. Journal of Statistical Software 11(10): 1-17.

Appendix: R Codes Used

* Except the packages come with R 2.12.2, the following packages have been used in this study: boot (Canty and Ripley, 2011), CCA (González and Déjean, 2009), geoR (Ribeiro and Diggle, 2001), gstat (Pebesma, 2004), Hmisc (Harrell, 2010), languageR (Baayen, 2011), mda (Leisch et al., 2011), nnet (Venables and Ripley, 2002), pls (Mevik and Wehrens, 2011), rgdal (Keitt et al., 2011), splancs (Rowlingson et al., 2010), vcd (Meyer et al., 2011), yaImpute (Crookston and Finley, 2007).
* R codes used for repeated trials, image production, and quoting results are not included here.

```
### Import all the data into R, and perform necessary pre-processing
################################################################################
> predictors = readGDAL("D:/Documents during in
Sweden/Thesis/Thesis_draft/Scratch/dem_clip_20.img")
> predictors$DEM = readGDAL("D:/Documents during in
Sweden/Thesis/Thesis_draft/Scratch/dem_clip_20.img")$band1
### (Also import data of SOIL, SOLAR, ASPECT, CURV_PL, CURV_PR, SLOPE, WI, DIST_A,
DIST_W, X, Y)
> predictors@data[[1]] <- 1:dim(predictors@data)[1]
> indices_1 <- is.na(predictors$WI)
> predictors$WI = replace(predictors$WI, indices_1, 0)
> indices_2 <- which(predictors$SOIL == "888")
> indices_3 <- which(predictors$SOIL == "200" | predictors$SOIL == "66" |
predictors$SOIL == "91")
> indices_4 <- is.na(predictors$SOIL)
> predictors$SOIL = replace(predictors$SOIL, indices_2, "8")
> predictors$SOIL = replace(predictors$SOIL, indices_3, "NA")
> predictors$SOIL = replace(predictors$SOIL, indices_4, "NA")
> predictors$SOIL = as.factor(predictors$SOIL)
> co <- coordinates(predictors)
> predictors$X = co[1:248040, 1]
> predictors$Y = co[1:248040, 2]
> polygons = readOGR(dsn="D:/Documents during in
Sweden/Thesis/Thesis_draft/Scratch/30", layer="NFI_polygon_simp")
> proj4string(predictors) = CRS("+init=epsg:3854")
> proj4string(polygons) = CRS("+init=epsg:3854")

### Combine data of predictors and response variables into one dataset
################################################################################
> over1 <- over(predictors, polygons)
```

```
> over1$ID = predictors$band1
> over1$DEM = predictors$DEM
### (Also add in the data of SOIL, SOLAR, ASPECT, CURV_PL, CURV_PR, SLOPE, WI, DIST_A,
DIST_W, X, Y)
> variables = over1[!is.na(over1$OBJECTID_1),]

### Check the representativity of the sample
##############################################################################
> options(digits=1)
> DEM.histbb = histbackback(variables$DEM, predictors$DEM, prob=TRUE)
### (Also perform this on SOIL, SOLAR, ASPECT, CURV_PL, CURV_PR, SLOPE, WI, DIST_A,
DIST_W, X, Y)
> options(digits=7)
> ks.test(DEM.histbb$left, DEM.histbb$right)
### (Also perform this on SOIL, SOLAR, ASPECT, CURV_PL, CURV_PR, SLOPE, WI, DIST_A,
DIST_W, X, Y)

### Check the correlations between variables, and separate predictors from response
variables for the PLS regression
##############################################################################
> v_f = variables[variables$Vegetation == "C" | variables$Vegetation == "D" |
variables$Vegetation == "MI",]
> v_f_8 = v_f[v_f$SOIL == "8",]
> v_f_1 = v_f[v_f$SOIL == "1",]
> v_f_93 = v_f[v_f$SOIL == "93",]
> X <- as.matrix(v_f[, c(9, 11:20)])
> X8 <- as.matrix(v_f_8[, c(9, 11:20)])
> X1 <- as.matrix(v_f_1[, c(9, 11:20)])
> X93 <- as.matrix(v_f_93[, c(9, 11:20)])
> Y <- as.matrix(v_f[, c(5:7)])
> Y8 <- as.matrix(v_f_8[, c(5:7)])
> Y1 <- as.matrix(v_f_1[, c(5:7)])
> Y93 <- as.matrix(v_f_93[, c(5:7)])
> cor <- matcor(X, Y)
> cor8 <- matcor(X8, Y8)
> cor1 <- matcor(X1, Y1)
> cor93 <- matcor(X93, Y93)

### Build the PLS regression models
##############################################################################
> YAGEs = scale(Y)$AGE
> YDENSITYs = scale(Y)$DENSITY
> YOPENNESSs = scale(Y)$OPENNESS
```

```
> Y8AGEs = scale(Y8)$AGE
> Y8DENSITYs = scale(Y8)$DENSITY
> Y8OPENNESSs = scale(Y8)$OPENNESS
> Y1AGEs = scale(Y1)$AGE
> Y1DENSITYs = scale(Y1)$DENSITY
> Y1OPENNESSs = scale(Y1)$OPENNESS
> Y93AGEs = scale(Y93)$AGE
> Y93DENSITYs = scale(Y93)$DENSITY
> Y93OPENNESSs = scale(Y93)$OPENNESS
> plsss <- plsr(cbind(YAGEs, YDENSITYs, YOPENNESSs) ~ X, scale=TRUE, validation="LOO")
> pls8ss <- plsr(cbind(Y8AGEs, Y8DENSITYs, Y8OPENNESSs) ~ X8, scale=TRUE,
validation="LOO")
> pls1ss <- plsr(cbind(Y1AGEs, Y1DENSITYs, Y1OPENNESSs) ~ X1, scale=TRUE,
validation="LOO")
> pls93ss <- plsr(cbind(Y93AGEs, Y93DENSITYs, Y93OPENNESSs) ~ X93, scale=TRUE,
validation="LOO")
> plss <- plsr(cbind(YAGEs, YOPENNESSs) ~ X, scale=TRUE, validation="LOO")
> pls8s <- plsr(cbind(Y8AGEs, Y8OPENNESSs) ~ X8, scale=TRUE, validation="LOO")
> pls93s <- plsr(cbind(Y93AGEs, Y93OPENNESSs) ~ X93, scale=TRUE, validation="LOO")

### Perform predictions using the PLS models built
###############################################################################
> pre_8 <- predict(pls8s, ncomp=3, as.matrix(over1[which(over1$SOIL == "8"), c(9,
11:20)]))
> pre_all <- predict(plss, ncomp=9, as.matrix(over1[which(over1$SOIL == "1" |
over1$SOIL == "NA"), c(9, 11:20)]))
> pre_93 <- predict(pls93s, ncomp=10, as.matrix(over1[which(over1$SOIL == "93"), c(9,
11:20)]))
> pre_8_df <- as.data.frame(pre_8)
> pre_93_df <- as.data.frame(pre_93)
> pre_all_df <- as.data.frame(pre_all)
> colnames(pre_8_df) <- c("AGEs", "OPENNESSs")
> colnames(pre_93_df) <- c("AGEs", "OPENNESSs")
> colnames(pre_all_df) <- c("AGEs", "OPENNESSs")
> pre_8_df$ID = over1[which(over1$SOIL == "8"), "ID"]
> pre_93_df$ID = over1[which(over1$SOIL == "93"), "ID"]
> pre_all_df$ID = over1[which(over1$SOIL == "1" | over1$SOIL == "NA"), "ID"]
> pre_8_df$AGEre = pre_8_df$AGEs * sd(as.data.frame(Y8)$AGE) +
mean(as.data.frame(Y8)$AGE)
> pre_all_df$AGEre = pre_all_df$AGEs * sd(as.data.frame(Y)$AGE) +
mean(as.data.frame(Y)$AGE)
> pre_93_df$AGEre = pre_93_df$AGEs * sd(as.data.frame(Y93)$AGE) +
mean(as.data.frame(Y93)$AGE)
```

```
> pre_8_df$OPENNESSre = pre_8_df$OPENNESSs * sd(as.data.frame(Y8)$OPENNESS) +
mean(as.data.frame(Y8)$OPENNESS)
> pre_all_df$OPENNESSre = pre_all_df$OPENNESSs * sd(as.data.frame(Y)$OPENNESS) +
mean(as.data.frame(Y)$OPENNESS)
> pre_93_df$OPENNESSre = pre_93_df$OPENNESSs * sd(as.data.frame(Y93)$OPENNESS) +
mean(as.data.frame(Y93)$OPENNESS)
> pre_df <- rbind(pre_8_df, pre_93_df, pre_all_df)
> pre_df = pre_df[with(pre_df, order(ID)),]
> predictors$AGEre = pre_df$AGEre
> predictors$OPENNESSre = pre_df$OPENNESSre

### Perform ordinary Kriging on the residuals of the PLS models
##############################################################################
> pls8s_df <- as.data.frame(pls8s$residuals[1:100, 1:2, 3])
> colnames(pls8s_df) <- c("A_resi", "O_resi")
> pls8s_df$X = v_f_8$X
> pls8s_df$Y = v_f_8$Y
> pls8s_df$ID = v_f_8$ID
> pls8s_df$A_resi = pls8s_df$A_resi * sd(as.data.frame(Y8)$AGE)
> pls8s_df$O_resi = pls8s_df$O_resi * sd(as.data.frame(Y8)$OPENNESS)
> pls93s_df <- as.data.frame(pls93s$residuals[1:223, 1:2, 10])
> colnames(pls93s_df) <- c("A_resi", "O_resi")
> pls93s_df$X = v_f_93$X
> pls93s_df$Y = v_f_93$Y
> pls93s_df$ID = v_f_93$ID
> pls93s_df$A_resi = pls93s_df$A_resi * sd(as.data.frame(Y93)$AGE)
> pls93s_df$O_resi = pls93s_df$O_resi * sd(as.data.frame(Y93)$OPENNESS)
> plss_df <- as.data.frame(plss$residuals[1:342, 1:2, 9])
> colnames(plss_df) <- c("A_resi", "O_resi")
> plss_df$X = v_f$X
> plss_df$Y = v_f$Y
> plss_df$ID = v_f$ID
> plss_df$SOIL = v_f$SOIL
> plss_df$A_resi = plss_df$A_resi * sd(as.data.frame(Y)$AGE)
> plss_df$O_resi = plss_df$O_resi * sd(as.data.frame(Y)$OPENNESS)
> plss_sub_df <- subset(plss_df, plss_df$SOIL == "1" | plss_df$SOIL == "NA",
select=c(1:5))
> plss_all_df = rbind(pls8s_df, pls93s_df, plss_sub_df)
> coordinates(plss_all_df) = ~ X + Y
> proj4string(plss_all_df) = CRS(proj4string(predictors))
> out1 = plss_all_df[c(1:53, 56:59, 66:176, 214:328, 332:342),]
> A_vgm_o <- variogram(A_resi ~ 1, out1)
> A_vgm_o_fit <- fit.variogram(A_vgm_o, model=vgm(150, "Sph", 2500, 100))
```

```
> krige_A_o_resi <- krige(A_resi ~ 1, out1, predictors, model=A_vgm_o_fit)
> O_vgm_o <- variogram(O_resi ~ 1, out1, cutoff=1900)
> O_vgm_o_fit <- fit.variogram(O_vgm_o, model=vgm(0.03, "Exp", 1900, 0.01))
> krige_O_o_resi <- krige(O_resi ~ 1, out1, predictors, model=O_vgm_o_fit)
> predictors$A_K = krige_A_o_resi@data$var1.pred
> predictors$O_K = krige_O_o_resi@data$var1.pred

### Build the logistic models
################################################################################
> variables_temp <- subset(variables, !(variables$Vegetation == "NA")
& !(variables$SOIL == "NA"))
> collin.fnc(scale(variables_temp[, c(9, 11:20)]))$cnumber
> variables_sub <- data.frame(scale(variables_temp[, c(9, 11:20)]), variables_temp[,
c(2, 5:8, 10)])
> variables_sub$FOREST = variables_sub$ID
> indices_5 <- which(variables_sub$Vegetation == "C" | variables_sub$Vegetation ==
"D" | variables_sub$Vegetation == "MI")
> variables_sub$FOREST = replace(variables_sub$FOREST, indices_5, "1")
> indices_6 <- which(variables_sub$Vegetation == "J" | variables_sub$Vegetation ==
"M")
> variables_sub$FOREST = replace(variables_sub$FOREST, indices_6, "0")
> variables_sub$FOREST = as.factor(variables_sub$FOREST)
> m_FOREST <- glm(FOREST ~ . -Vegetation -OPENNESS -AGE -DENSITY -ID, family=binomial,
variables_sub)
> m_FOREST_step <- step(m_FOREST)
> m_FOREST_all <- glm(FOREST ~ . -Vegetation -OPENNESS -AGE -DENSITY -ID -SOIL,
family=binomial, variables_sub)
> m_FOREST_all_step <- step(m_FOREST_all)
> v_vege <- subset(variables_sub, !(variables_sub$Vegetation == "J")
& !(variables_sub$Vegetation == "M"))
> collin.fnc(v_vege[, c(1:11)])$cnumber
> v_mj <- subset(variables_sub, variables_sub$Vegetation == "J" |
variables_sub$Vegetation == "M")
> collin.fnc(v_mj[, c(1:11)])$cnumber
> m_mj <- glm(Vegetation ~ . -FOREST -OPENNESS -AGE -DENSITY -ID, family=binomial,
v_mj)
> m_mj_step <- step(m_mj)
> m_mj_all <- glm(Vegetation ~ . -FOREST -OPENNESS -AGE -DENSITY -ID -SOIL,
family=binomial, v_mj)
> m_mj_all_step <- step(m_mj_all)
> v_vege$Vegetation = as.numeric(v_vege$Vegetation)
> v_vege$Vegetation = as.factor(v_vege$Vegetation)
> m_vege <- multinom(Vegetation ~ . -FOREST -OPENNESS -AGE -DENSITY -ID, v_vege)
```

```
> m_vege_step <- step(m_vege)
> m_vege_all <- multinom(Vegetation ~ . -FOREST -OPENNESS -AGE -DENSITY -ID -SOIL,
v_vege)
> m_vege_all_step <- step(m_vege_all)

### Perform predictions using the logistic models built
################################################################################
> over1_temp <- subset(over1, !(over1$SOIL == "NA"))
> over1_sub <- data.frame(scale(over1_temp[, c(9, 11:20)]), over1_temp[, c(2, 5:8,
10)])
> pre_FOREST <- predict(m_FOREST_step, over1_sub, type="response")
> over1_sub$FOREST = pre_FOREST
> pre_FOREST_all <- predict(m_FOREST_all_step, over1, type="response")
> over1$FOREST = pre_FOREST_all
> pre_mj <- predict(m_mj_step, over1_sub, type="response")
> over1_sub$mj = pre_mj
> pre_mj_all <- predict(m_mj_all_step, over1, type="response")
> over1$mj = pre_mj_all
> pre_vege <- predict(m_vege_step, over1_sub, type="probs")
> over1_sub <- data.frame(over1_sub, pre_vege)
> pre_vege_all <- predict(m_vege_all_step, over1, type="probs")
> over1 <- data.frame(over1, pre_vege_all)
> FOREST_sub <- subset(over1, over1$SOIL == "NA", c(2, 5:25))
> over1_FOREST <- rbind(over1_sub, FOREST_sub)
> over1_FOREST = over1_FOREST[with(over1_FOREST, order(ID)),]
> predictors$FOREST = over1_FOREST$FOREST
> predictors$mj = over1_FOREST$mj
> predictors$C = over1_FOREST$C
> predictors$D = over1_FOREST$D
> predictors$MI = over1_FOREST$MI

### Perform ordinary Kriging on the residuals of the logistic models
################################################################################
> FOREST_resi <- as.data.frame(residuals(m_FOREST_step, type="response"))
> FOREST_resi$X = variables_temp$X
> FOREST_resi$Y = variables_temp$Y
> coordinates(FOREST_resi) = ~ X + Y
> proj4string(FOREST_resi) = CRS(proj4string(predictors))
> FOREST_vgm <- variogram(residuals(m_FOREST_step, type="response") ~ 1, FOREST_resi)
> FOREST_vgm_fit <- fit.variogram(FOREST_vgm, model=vgm(0.1, "Exp", 2000, 0.1))
> krige_FOREST <- krige(residuals(m_FOREST_step, type="response") ~ 1, FOREST_resi,
predictors, model=FOREST_vgm_fit)
> predictors$FOREST_K = krige_FOREST@data$var1.pred
```

```
> mj_resi <- as.data.frame(residuals(m_mj_step, type="response"))
> mj_resi$X = subset(variables_temp, variables_temp$Vegetation == "J" |
variables_temp$Vegetation == "M")$X
> mj_resi$Y = subset(variables_temp, variables_temp$Vegetation == "J" |
variables_temp$Vegetation == "M")$Y
> coordinates(mj_resi) = ~ X + Y
> proj4string(mj_resi) = CRS(proj4string(predictors))
> mj_vgm <- variogram(residuals(m_mj_step, type="response") ~ 1, mj_resi)
> mj_vgm_fit <- fit.variogram(mj_vgm, model=vgm(0.05, "Sph", 1100, 0.1))
> krige_mj <- krige(residuals(m_mj_step, type="response") ~ 1, mj_resi, predictors,
model=mj_vgm_fit)
> predictors$mj_K = krige_mj@data$var1.pred
> vege_resi <- data.frame(subset(variables_temp, !(variables_temp$Vegetation == "J")
& !(variables_temp$Vegetation == "M"))$X,
subset(variables_temp, !(variables_temp$Vegetation == "J")
& !(variables_temp$Vegetation == "M"))$Y, m_vege_step$residuals)
> colnames(vege_resi) <- c("X", "Y", "C", "D", "MI")
> coordinates(vege_resi) = ~ X + Y
> proj4string(vege_resi) = CRS(proj4string(predictors))
> vege_vgm_C <- variogram(C ~ 1, vege_resi, cutoff=2000)
> vege_vgm_C_fit <- fit.variogram(vege_vgm_C, model=vgm(0.15, "Exp", 800, 0))
> krige_C <- krige(C ~ 1, vege_resi, predictors, model=vege_vgm_C_fit)
> vege_vgm_D <- variogram(D ~ 1, vege_resi, cutoff=1500)
> vege_vgm_D_fit <- fit.variogram(vege_vgm_D, model=vgm(0.2, "Sph", 1000, 0))
> krige_D <- krige(D ~ 1, vege_resi, predictors, model=vege_vgm_D_fit)
> vege_vgm_MI <- variogram(MI ~ 1, vege_resi, cutoff=3600)
> vege_vgm_MI_fit <- fit.variogram(vege_vgm_MI, model=vgm(0.1, "Exp", 1000, 0))
> krige_MI <- krige(MI ~ 1, vege_resi, predictors, model=vege_vgm_MI_fit)
> predictors$C_K = krige_C@data$var1.pred
> predictors$D_K = krige_D@data$var1.pred
> predictors$MI_K = krige_MI@data$var1.pred

### Check the performance of the logistic models
################################################################################
> cost_FOREST <- function(FOREST, pi=0) mean(abs(FOREST - pi) > 0.5)
> cv.m_FOREST_step <- cv.glm(variables_sub, m_FOREST_step, cost_FOREST,
K=nrow(variables_sub))
> cost_mj <- function(Vegetation, pi = 0) mean(abs(Vegetation - pi) > 0.5)
> cv.m_mj_step <- cv.glm(v_mj, m_mj_step, cost_mj, K=nrow(v_mj))
> accu_FOREST <- data.frame(m_FOREST_step$fitted.values, variables_sub$FOREST)
> colnames(accu_FOREST) <- c("FITTED", "OBSERVED")
> coordinates(variables_temp) = ~ X + Y
> proj4string(variables_temp) = CRS(proj4string(predictors))
```

```
> krige_FOREST_accu <- krige(residuals(m_FOREST_step, type="response") ~ 1,
FOREST_resi, variables_temp, model=FOREST_vgm_fit)
> accu_FOREST$FITTED = accu_FOREST$FITTED + krige_FOREST_accu$var1.pred
> indices_7 <- which(accu_FOREST$FITTED >= 0.5)
> accu_FOREST$FITTED = replace(accu_FOREST$FITTED, indices_7, "1")
> indices_8 <- which(accu_FOREST$FITTED < 0.5)
> accu_FOREST$FITTED = replace(accu_FOREST$FITTED, indices_8, "0")
> accu_FOREST$FITTED = as.factor(accu_FOREST$FITTED)
> confusion(accu_FOREST$FITTED, accu_FOREST$OBSERVED)
> Kappa(confusion(accu_FOREST$FITTED, accu_FOREST$OBSERVED))
> accu_mj <- data.frame(m_mj_step$fitted.values,
as.factor(as.numeric(v_mj$Vegetation)))
> colnames(accu_mj) <- c("FITTED", "OBSERVED")
> coordinates(v_mj) = ~ X + Y
> proj4string(v_mj) = CRS(proj4string(predictors))
> krige_mj_accu <- krige(residuals(m_mj_step, type="response") ~ 1, mj_resi, v_mj,
model=mj_vgm_fit)
> accu_mj$FITTED = accu_mj$FITTED + krige_mj_accu$var1.pred
> indices_9 <- which(accu_mj$FITTED >= 0.5)
> indices_10 <- which(accu_mj$FITTED < 0.5)
> accu_mj$FITTED = replace(accu_mj$FITTED, indices_9, "4")
> accu_mj$FITTED = replace(accu_mj$FITTED, indices_10, "3")
> accu_mj$FITTED = as.factor(accu_mj$FITTED)
> confusion(accu_mj$FITTED, accu_mj$OBSERVED)
> Kappa(confusion(accu_mj$FITTED, accu_mj$OBSERVED))
> accu_vege <- data.frame(m_vege_step$fitted.values,
as.factor(as.numeric(v_vege$Vegetation)), v_vege$ID)
> accu_vege$max <- whatsMax(m_vege_step$fitted.values)$fitted.values.maxCol
> confusion(accu_vege$max, accu_vege[,4])
> Kappa(confusion(accu_vege$max, accu_vege[,4]))

### Generate modeling outputs
################################################################################
> predictors$FOREST = predictors$FOREST + predictors$FOREST_K
> indices_25 <- which(predictors$FOREST >= 0.5)
> indices_26 <- which(predictors$FOREST < 0.5)
> predictors$FOREST = replace(predictors$FOREST, indices_25, "Forest")
> predictors$FOREST = replace(predictors$FOREST, indices_26, "Non-forest")
> predictors$FOREST = as.factor(predictors$FOREST)
> predictors$mj = predictors$mj + predictors$mj_K
> indices_27 <- which(predictors$mj >= 0.5)
> indices_28 <- which(predictors$mj < 0.5)
> predictors$mj = replace(predictors$mj, indices_27, "Mire")
```

```
> predictors$mj = replace(predictors$mj, indices_28, "Heath")
> predictors$mj = as.factor(predictors$mj)
> df_1 <- data.frame(predictors$FOREST, predictors$mj)
> indices_29 <- which(!(df_1[,1] == "Non-forest"))
> df_1[,2] = replace(df_1[,2], indices_29, "NA")
> predictors$mj_c = df_1[,2]
> predictors$C = predictors$C + predictors$C_K
> predictors$D = predictors$D + predictors$D_K
> predictors$MI = predictors$MI + predictors$MI_K
> predictors$max = whatsMax(predictors@data[, c("C", "D",
"MI")])$fitted.values.maxCol
> predictors$max = as.factor(predictors$max)
> df_2 <- data.frame(predictors$FOREST, predictors$max)
> indices_33 <- which(df_2[,1] == "Non-forest")
> df_2[,2] = replace(df_2[,2], indices_33, "NA")
> predictors$max_c = df_2[,2]
> df_3 <- data.frame(predictors$FOREST, predictors$OPENNESS)
> indices_34 <- which(df_3[,1] == "Non-forest")
> df_3[,2] = replace(df_3[,2], indices_34, "NA")
> predictors$OPENNESS = df_3[,2]
> predictors$OPENNESS = as.numeric(predictors$OPENNESS)
> df_4 <- data.frame(predictors$FOREST, predictors$AGE)
> indices_35 <- which(df_4[,1] == "Non-forest")
> df_4[,2] = replace(df_4[,2], indices_35, "NA")
> predictors$AGE = df_4[,2]
> predictors$AGE = as.numeric(predictors$AGE)

### Apply k-NN method
################################################################################
> x = variables_sub[,c(1:11)]
> y = variables_sub[,c(12, 13, 15)]
> mah <- yai(x=x, y=y, method="mahalanobis")
> xfiles <- list(DEM="D:/Documents during in
Sweden/Thesis/Thesis_draft/Scratch/DEM.asc", SOLAR="D:/Documents during in
Sweden/Thesis/Thesis_draft/Scratch/SOLAR.asc", ASPECT="D:/Documents during in
Sweden/Thesis/Thesis_draft/Scratch/ASPECT.asc", CURV_PL="D:/Documents during in
Sweden/Thesis/Thesis_draft/Scratch/CURV_PL.asc", CURV_PR="D:/Documents during in
Sweden/Thesis/Thesis_draft/Scratch/CURV_PR.asc", SLOPE="D:/Documents during in
Sweden/Thesis/Thesis_draft/Scratch/SLOPE.asc", WI="D:/Documents during in
Sweden/Thesis/Thesis_draft/Scratch/WI.asc", DIST_A="D:/Documents during in
Sweden/Thesis/Thesis_draft/Scratch/DIST_A.asc", DIST_W="D:/Documents during in
Sweden/Thesis/Thesis_draft/Scratch/DIST_W.asc", X="D:/Documents during in
Sweden/Thesis/Thesis_draft/Scratch/X.asc", Y="D:/Documents during in
```

```
Sweden/Thesis/Thesis_draft/Scratch/Y.asc")
> outfiles <- list(Vegetation="D:/Documents during in
Sweden/Thesis/Thesis_draft/Scratch/Vegetation.asc", AGE="D:/Documents during in
Sweden/Thesis/Thesis_draft/Scratch/AGE.asc", OPENNESS="D:/Documents during in
Sweden/Thesis/Thesis_draft/Scratch/OPENNESS.asc")
> AsciiGridImpute(mah, xfiles, outfiles)

### Estimate point intensity from the historical county economic maps
################################################################################
> points_Barr = readOGR(dsn="D:/Documents during in Sweden/Thesis/From the
Supervisor/Master thesis/Historical map 1920", layer="Barr_punkt")
> points_Lov = readOGR(dsn="D:/Documents during in Sweden/Thesis/From the
Supervisor/Master thesis/Historical map 1920", layer="Lov_punkt")
> Extent = readOGR(dsn="D:/Documents during in Sweden/Thesis/Thesis_draft/Scratch",
layer="Extent")
> boundary <- slot(slot(slot(Extent, "polygons")[[1]], "Polygons")[[1]], "coords")
> bandwidth_Barr <- mse2d(as.points(coordinates(points_Barr)), boundary, 100, 1000)
> bandwidth_Lov <- mse2d(as.points(coordinates(points_Lov)), boundary, 100, 1000)
> bandwidth_Barr$h[which.min(bandwidth_Barr$mse)]
> bandwidth_Lov$h[which.min(bandwidth_Lov$mse)]
> grid <- slot(predictors, "grid")
> int_Barr <- spkernel2d(points_Barr, boundary, h0=120, grid)
> int_Lov <- spkernel2d(points_Lov, boundary, h0=150, grid)
> kernel_Barr <- SpatialGridDataFrame(grid, data=data.frame(int_Barr))
> kernel_Lov <- SpatialGridDataFrame(grid, data=data.frame(int_Lov))

### Comparisons on OPENNESS and vegetation types
################################################################################
> kernel = kernel_Barr$int_Barr + kernel_Lov$int_Lov
> validation <- data.frame(kernel, over1$AGE, over1$OPENNESS, over1$Vegetation)
> validation_sub <- validation[!is.na(validation[,4]) & !(validation[,4] == "NA"),]
> colnames(validation_sub) = c("kernel", "AGE", "OPENNESS", "Vegetation")
> kernel2OPENNESS <- lm(OPENNESS ~ kernel + AGE, validation_sub)
> validation2 <- data.frame(kernel)
> validation2$AGE = predictors$AGEre + predictors$A_K
> validation2$k2O = predict(kernel2OPENNESS, validation2)
> validation2$O_group = predictors$OPENNESSre + predictors$O_K
> indices_38 <- which(validation2$O_group <= 0.4)
> indices_37 <- which(validation2$O_group > 0.4 & validation2$O_group <= 0.8)
> indices_36 <- which(validation2$O_group > 0.8)
> indices_41 <- which(validation2$k2O <= 0.4)
> indices_40 <- which(validation2$k2O > 0.4 & validation2$k2O_group <= 0.8)
> indices_39 <- which(validation2$k2O > 0.8)
```

```
> validation2$O_group <- replace(validation2$O_group, indices_36, 30)
> validation2$O_group <- replace(validation2$O_group, indices_37, 20)
> validation2$O_group <- replace(validation2$O_group, indices_38, 10)
> validation2$k2O <- replace(validation2$k2O, indices_39, 30)
> validation2$k2O <- replace(validation2$k2O, indices_40, 20)
> validation2$k2O <- replace(validation2$k2O, indices_41, 10)
> predictors$O_group = validation2$O_group
> predictors$k2O = validation2$k2O
> writeGDAL(predictors["O_group"], "D:/Documents during in
Sweden/Thesis/Thesis_draft/Scratch/O_group.img", drivername="HFA", type="Float32",
mvFlag=999)
### (Also export “k2O”)
> va = readGDAL("D:/Documents during in
Sweden/Thesis/Thesis_draft/Scratch/k2O_group_Clip.img")
> va$k2O = readGDAL("D:/Documents during in
Sweden/Thesis/Thesis_draft/Scratch/k2O_group_Clip.img")$band1
> va$O = readGDAL("D:/Documents during in
Sweden/Thesis/Thesis_draft/Scratch/O_group_Clip.img")$band1
> va$kNN2O = readGDAL("D:/Documents during in
Sweden/Thesis/Thesis_draft/Scratch/O_asc_C2.img")$band1
> indices_45 <- which(va$kNN2O <= 0.4)
> indices_44 <- which(va$kNN2O > 0.4 & va$kNN2O <= 0.8)
> indices_43 <- which(va$kNN2O > 0.8)
> va$kNN2O <- replace(va$kNN2O, indices_43, 30)
> va$kNN2O <- replace(va$kNN2O, indices_44, 20)
> va$kNN2O <- replace(va$kNN2O, indices_45, 10)
> va$k2O = as.factor(va$k2O)
> va$O = as.factor(va$O)
> va$kNN2O = as.factor(va$kNN2O)
> confusion(va$O, va$k2O)
> Kappa(confusion(va$O, va$k2O))
> confusion(va$kNN2O, va$k2O)
> Kappa(confusion(va$kNN2O, va$k2O))
> predictors$max = as.numeric(predictors$max)
> writeGDAL(predictors["max"], "D:/Documents during in
Sweden/Thesis/Thesis_draft/Scratch/max.img", drivername="HFA", type="Float32",
mvFlag=999)
> va$max = readGDAL("D:/Documents during in
Sweden/Thesis/Thesis_draft/Scratch/max_Clip.img")$band1
> va$Tradslag = readGDAL("D:/Documents during in
Sweden/Thesis/Thesis_draft/Scratch/Tradslag.img")$band1
> va$Vege = readGDAL("D:/Documents during in
Sweden/Thesis/Thesis_draft/Scratch/V_asc_C.img")$band1
```

```
> indices_46 <- which(va$Vege == 3 | va$Vege == 4 | va$Vege == 7)
> va$Vege <- replace(va$Vege, indices_46, "NA")
> va$Vege <- as.numeric(va$Vege)
> indices_50 <- which(va$Tradslag == 0)
> va$Tradslag <- replace(va$Tradslag, indices_50, 5)
> va$max = as.factor(va$max)
> va$Tradslag = as.factor(va$Tradslag)
> va$Vege = as.factor(va$Vege)
> confusion(va$max, va$Tradslag)
> Kappa(confusion(va$max, va$Tradslag))
> confusion(va$Vege, va$Tradslag)
```