

A Weighty Issue: Estimation of Fire Size with Geographically Weighted Logistic Regression

Julia Hjalmarsson

2016
Department of
Physical Geography and Ecosystem Science
Centre for Geographical Information Systems
Lund University
Sölvegatan 12
S-223 62 Lund
Sweden



Hjalmarsson, J., Lehsten, V., Lindström, J. (2016). A Weighty Issue: Estimation of Fire Size with Geographically Weighted Logistic Regression
Master degree thesis, 30 credits in Master in Geographical Information Science
Department of Physical Geography and Ecosystem Science, Lund University

A Weighty Issue: Estimation of Fire Size with Geographically Weighted Logistic Regression

Julia Hjalmarsson
Master Thesis, 30 credits, in Geographical Information
Science

Veiko Lehsten
Lund University

Johan Lindström
Lund University

Julia Hjalmarsson

A Weighty Issue: Estimation of Fire Size with Geographically Weighted Logistic Regression

Size estimations of fires that occurred centuries ago have been the subject of study for many decades. More accurate spatial fire histories from tree rings were possible by either drawing the sample location on detailed topographic maps or using GPS receivers. A popular method of delineating fire sizes is to draw an outline around the fire-scarred samples considering topographic and landscape features. This is a rather subjective method that cannot be replicated. Other more replicable methods have also been suggested to estimate fire size among them methods that use buffers (kernel ranges), grids, or Thiessen polygons. However, even those have a subjective component.

Geographically weighted logistic regression (GWLR), not previously used to estimate fire sizes, seemed promising since the method allows for the changing relationships between different topographic, landscape, or socioeconomic features to be considered in the analysis. Logistic regression is done with binomial data: fire/no fire. Geographically weighted regression (GWR) is a relatively new and more objective method that considers the geography of the data. Instead of using one regression coefficient over a whole study area, several regression coefficients are calculated for the different sample locations which might help explain the relationships better.

The GWLR analyses consistently found only one variable that explained fire location. This variable changed between the different analyses. The corrected Akaike Information Criterion (AICc) increased in every GWLR analysis when adding more variables (a lower AICc value means a higher quality model) while the R^2 value increased (more variables explain more of the variance). The optimal output of such analyses would be that the R^2 value increases and the AICc decreases which would mean that the added variables help explain more of the variance AND that the model has higher quality.

A probability analysis of whether close trees burn at the same time shows that trees closer to each other have a higher probability of burning compared to trees that are further apart. This is especially true between the years 1400 - 1700 (before human influences on the landscape). Between 1700 and 2000, this clear pattern partially vanished.

While GWR could be considered the most objective method of fire-size estimations (of the ones studied), it could not be used to estimate fire sizes. Fire size and location seems to depend more on the distance from fire-scarred trees than on different landscape features. Different methods of estimating fire sizes are more reliable before humans have added to the natural fire regimes. After human influence, the uncertainty of the fire sizes increases between the different methods of calculating fires sizes (up to 46% in this study).

Keywords: Physical Geography and Ecosystem analysis, GIS, geographically weighted logistic regression, fire reconstruction, fire size

Advisor: Veiko Lehsten, Johan Lindström

Master degree project 30 credits in Geographical Information Science, 2016
Department of Physical Geography and Ecosystem Science, Lund University.
Thesis nr 58

Uppskattning av Brandstorlek med Geografiskt Viktad Logistik Regression

Destruktiva skogsbränder är vanliga idag och genererar nyheter varje år. Förr i tiden var skogsbränder dock ett viktigt ursprung till förnyelse i många ekosystem. Många arter är anpassade och beroende av skogsbränder. Uppskattningen av skogsbrandstorleken är viktig för att återinsätta bränder i skogen med en naturlig rotation.

Storleksuppskattning av dåtida skogsbränder har undersökts i många årtionden. Mer exakta rumsliga brandhistorier från trädringar var möjligt genom att rita den geografiska platsen av alla prover på detaljerade topografiska kartor eller genom att använda GPS mottagare. En populär metod för att avgränsa brandstorlekar är att rita en kontur runt brandärrade prover genom att avväga topografiska och landskapselement. Detta är en ganska subjektiv metod som inte kan replikeras. Andra, mer reproducerbara, metoder har också föreslagits för att uppskatta brandstorlek bland annat metoder som använder buffer, rutnät eller Thiessen polygoner. Även de metoderna har dock en subjektiv komponent.

Geografiskt Viktad Logistik Regression har inte tidigare använts för att uppskatta brandstorlekar. Metoden verkade lovande eftersom den tillåter förändrade relationer mellan olika topografiska, landskaps-, och socioekonomiska element som ska betraktas i analysen. Logistik regression görs med binomiala data: brand/icke brand. Geografiskt Viktad Regression är en relativ ny och mer objektiv metod som avväger geografien i data. Istället för att använda en regressionskoefficient över hela studieområdet beräknas flera regressionskoefficienter för de olika provplatserna. Detta kan hjälpa förklara relationer bättre.

Olika metoder för att uppskatta brandstorlekar är mer tillförlitliga innan människor påverkade den naturliga brandrotationen. Efter den mänskliga påverkan har oöverensstämmelse mellan olika metoder som uppskattar brandstorleken ökat (med upp till 46 % i denna studie). Dessutom har det visats att även om Geografiskt Viktad Regression kan anses vara den mest objektiva metoden för att uppskatta brandstorlekar (bland dem som använts i avhandlingen) kan det faktiskt inte användas för att uppskatta brandstorlekar. Brandstorlek verkar bero mer på avstånd mellan brandärrade träd än på olika landskapselement.

Nyckelord: naturgeografi och ekosystemanalys, GIS, geografiskt viktade logistik regression, uppskattning av brandstorlek

Advisor: **Veiko Lehsten, Johan Lindström**

Master degree project 30 credits in Geographical Information Science, 2016
Department of Physical Geography and Ecosystem Science, Lund University.
Thesis nr 58

Table of Contents

1	Introduction	1
1.1	Research Objectives and Questions.....	2
2	Background	3
2.1	Dendrochronological Background Study	3
2.2	Method Comparison	5
2.2.1	Thiessen Polygons.....	5
2.2.2	Grid.....	5
2.2.3	Drawing.....	5
2.2.4	Geographically Weighted Logistic Regression.....	5
2.3	Geographically Weighted Regression	5
2.3.1	Why local spatial regression?.....	5
2.3.2	OLS vs. GWR	6
2.3.3	Geographically Weighted Logistic Regression.....	6
2.3.4	PCA and GWR.....	8
2.3.5	Advantages and Disadvantages	8
2.3.6	Methodological Considerations.....	11
3	Methodology	15
3.1	Study Area	15
3.2	Data.....	17
3.3	Data Preparation	19
3.4	Analyses.....	19
3.4.1	Thiessen Polygons.....	19
3.4.2	Grid Analysis.....	20
3.4.3	Drawing the Fire Perimeter	20
3.4.4	GWLR in R	21
3.4.5	INLA model	21
3.4.6	Probability Plot (Close Trees Burning at Same Time).....	22
4	Result.....	23
4.1	Multicollinearity between the Covariates.....	23
4.2	Principal Component Analysis (PCA).....	27
4.2.1	PCA 100m Grid.....	27
4.2.2	PCA 250 Grid.....	30
4.3	Stepwise GWLR.....	31
4.4	Method Comparison	33
4.5	Probability of Burning	35

5	Discussion	41
5.1	Stepwise GWLR	41
5.2	Method Comparison	42
5.3	Fire Size	42
5.3.1	Over- and Underestimate of Fire Size	42
5.3.2	Advantages and Disadvantages of the methods	43
5.4	Probability of Burning	44
6	Conclusion.....	47
7	References	49
8	Appendix A	53
9	Appendix B	55
10	Appendix C: Correlograms	57
10.1	Correlogram (100m grid – all variables).....	57
10.2	Correlogram (100m grid - 13 variables)	58
10.3	Correlogram (100m grid – 10 variables).....	59
10.4	Correlogram (100m grid PCA)	60
10.5	Correlogram (250m grid – all variables).....	61
10.6	Correlogram (250m grid - 12 variables)	62
10.7	Correlogram (250m grid PCA)	63
11	Appendix D: Maps of Fire Size Comparison	65
11.1	Year 1521	65
11.2	Year 1576.....	66
11.3	Year 1644.....	67
11.4	Year 1677.....	68
11.5	Year 1723.....	69
11.6	Year 1729.....	70
11.7	Year 1770.....	71
11.8	Year 1789.....	72
11.9	Year 1807.....	73
11.10	Year 1831.....	74
11.11	Year 1844.....	75
11.12	Year 1868.....	76

List of Figures

Figure 2.1: Study area's location.....	3
Figure 2.2: Pictures of a stump that was samples in 2009	4
Figure 2.3: Picture of a cross-section	4
Figure 2.4: Kernel shapes	7
Figure 2.5: A Web of Science search	10
Figure 3.1: Axe marks on a snag in Jämtgaveln Nature Reserve	15
Figure 3.2: Pictures of the center part of the Jämtgaveln Nature Reserve	16
Figure 3.3: Example correlogram and output of ArcGIS' exploratory regression.....	21
Figure 4.1: Legend (color scheme) for the correlograms	26
Figure 4.2: Correlogram for the 100m grid with 13 variables	26
Figure 4.3: Correlogram for the 100m grid with 10 variables	27
Figure 4.4: Results of the PCA with the 100m grid	29
Figure 4.5: Results of the PCA with the 250m grid	31
Figure 4.6: Method comparison (map 1868)	33
Figure 4.7: Fire size comparison	34
Figure 4.8: The probabilities of a pair of trees burning (time period 1400-2000)	35
Figure 4.9: Boxplots per distance band of the data in Figure 4.8	36
Figure 4.10: The probabilities of a pair of trees burning (time period 1400-1700)	37
Figure 4.11: Boxplots per distance band of the data in Figure 4.10	38
Figure 4.12: The probabilities of a pair of trees burning (time period 1700-2000)	39
Figure 4.13: Boxplots per distance band of the data in Figure 4.12	40
Figure 10.1: Correlogram (100m, all variables)	57
Figure 10.2: Correlogram (100m, 13 variables)	58
Figure 10.3: Correlogram (100m, 10 variables)	59
Figure 10.4: Correlogram (100m, PCA)	60
Figure 10.5: Correlogram (250m, all variables)	61
Figure 10.6: Correlogram (250m, 12 variables)	62
Figure 10.7: Correlogram (250m, PCA)	63
Figure 11.1: Method comparison between the different methods for the year 1521	65
Figure 11.2: Method comparison between the different methods for the year 1576	66
Figure 11.3: Method comparison between the different methods for the year 1644	67
Figure 11.4: Method comparison between the different methods for the year 1677	68
Figure 11.5: Method comparison between the different methods for the year 1723	69
Figure 11.6: Method comparison between the different methods for the year 1729	70
Figure 11.7: Method comparison between the different methods for the year 1770	71
Figure 11.8: Method comparison between the different methods for the year 1789	72
Figure 11.9: Method comparison between the different methods for the year 1807	73
Figure 11.10: Method comparison between the different methods for the year 1831	74
Figure 11.11: Method comparison between the different methods for the year 1844	75
Figure 11.12: Method comparison between the different methods for the year 1868	76

List of Tables

Table 2.1: Main differences between OLS and GWR	6
Table 3.1: The 12 fire years that had the most fire-scarred trees.	17
Table 3.2: Layer of the land cover (marktäcke) for Jämtgaveln Nature Reserve	18
Table 3.3: All and selected variables (and their names)	18
Table 4.1: Summary of multicollinearity for 250m and all variables	23
Table 4.2: Summary of multicollinearity for 250m grid with 12 variables	24
Table 4.3: Exploratory Regression output from ArcGIS 100m grid and all variables	24
Table 4.4: Exploratory Regression output from ArcGIS 100m grid and 13 variables	25
Table 4.5: Correlation coefficients (100m grid) for each variable and the first 7 PCs	28
Table 4.6: Correlation coefficients (250m grid) for each variable and the first 7 PCs	30
Table 4.7: Results of the stepwise regression with all of the 19 variables	32
Table 4.8: Results of the stepwise regression with selected variables with the 100m grid	32
Table 4.9: Results of the stepwise regression with 12 variables and the 250m grid	32
Table 4.10: Results of the stepwise regression with 7 PCs for the 100m and 250m grid.....	33
Table 4.11: Fire sizes for the three different methods including the sum and average.....	34

List of Abbreviations

AIC	Akaike Information Criterion
AICc	Corrected Akaike Information Criterion
BIC	Bayesian Information Criterion
Bw	Bandwidth
CV	Cross-validation
DEM	Digital Elevation Model
ESRI	Environmental Systems Research Institute
GIS	Geographic Information System
GLMM	Generalized Linear Mixed Model
GW PCA	Geographically Weighted Principal Component Analysis
GWLR	Geographically Weighted Logistic Regression
GWR	Geographically Weighted Regression
INLA	Integrated Nested Laplace Approximations
MDL	Minimum Description Length
OLS	Ordinary Least Squares
PC	Principal Component
PCA	Principal Component Analysis
VIF	Variance Inflation Factor

1 Introduction

Forest or wildland fire is often seen as destructive by the general public, which is certainly true in many countries such as some in southern Europe or in parts of the western United States. However, based on reconstructions of fire from tree rings and charcoal, we know that fire has been an integral part in, for example, forested landscapes and that some species depend on fire for their survival.

Reconstructing historical fires from tree rings is a method that has been used for many decades. These reconstructions can be very accurate in time as fire scars can be dated to their exact calendar year. When recording the geographic location of the samples, one would then have an exact temporal and spatial reconstruction of fire. However, until the late 20th century, spatial reconstructions were very rare. Instead, to delineate between larger and smaller fires, a larger or smaller number of fire scars were considered (Swetnam and Baisan 1996) or several watersheds were studied (Heyerdahl 1997). These methods lacked, of course, accuracy on smaller ecological scales as the spatial properties of fires were neglected.

As one of the first ecologists to do work on this problem, Per Linder (Linder 1988) reconstructed a spatial fire history of Jämtgaveln Nature Reserve (Figure 2.1). Because this report was published in Swedish, it did not generate as much attention as the later work of Niklasson and Granström (2000). In their often cited work, Niklasson and Granström (2000) reconstructed fires spatially over an area of about 19 x 32 km.

To delineate fires, Niklasson and Granström (2000, pg. 1489) drew outlines of the fire extent by relying on their expert knowledge and several rules/assumptions, e.g. “[f]or fires dated at only one point and not at the surrounding points, the border was drawn at half the distance to the nearest non-recording point in all directions, but never farther than 1 km from the fire-recording point.” This next quote from their paper (pg. 1489) shows the subjectivity in their method, which could, therefore, not be reproduced accurately: “If a *probable* fuel break (peatland, lake, or watercourse) occurred ...” (emphasis added). Mires can burn and rivers can be jumped in certain (very dry) years and knowing what happened in which year is impossible.

Storaunet et al. (2013) have tried to use a method that is more objective. Using a geographic information system (GIS), they estimated fire size and severity by using a kernel range application (i.e. buffers) for both trees that did not have a fire scar as non-burners and trees that had a fire scar in a particular year. Although this study is more objective, even here fire size is estimated by subjectively choosing a buffer size.

Another study, Hjalmarsson et al. (2015) used a grid over the study area to estimate fire size. They used a 10- and 25-hectare grid, and every cell that had a fire-scarred tree in it was recorded as having burned. Only cells that had a sample in them were included in the calculations for each fire year. Therefore, if 7 of 10 cells had a fire-scarred tree, Hjalmarsson et al. (2015) assumed that 70% of the study area had burned.

The fire extent in such spatial reconstructions is, therefore, often analyzed by drawing an outline around all samples that indicated a fire occurred during a certain year (Niklasson and Granström 2000), selecting a buffer distance and overlaying all buffered samples with and without fire indications (Storaunet et al. 2013), or using a grid Hjalmarsson et al. (2015). These methods give an indication of the size of a fire in the study area. They are, however, influenced by subjectively choosing a distance (for the outline, buffer, and grid) in which to include areas as burned or unburned, and they also do not allow for the assessment of the accuracy of the burned area.

In different analyses, geographically weighted regression (Fotheringham et al. 2002) has been used to delineate areas derived from point data based on varied sorts of phenomena and their attributes. Those phenomena have included housing prices in London, using additional information such as on, for example, apartment age or apartments having a garage or balcony (Fotheringham et al. 2002). Another study used graffiti occurrences in San Francisco as related to population age and other variables (Megler et al. 2014). Fire danger and fire risk in fire prone countries like Spain (Martínez-Fernández et al. 2011; Martínez-Fernández et al. 2013; Rodrigues et al. 2014) have also been studied using geographically weighted regression (GWR), where researchers tried to explain recent fire events based on topography, land cover, climate, or human activity among other explanatory variables (Koutsias et al. 2010; Oliveira et al. 2014).

1.1 Research Objectives and Questions

We reconstruct historical fire sizes in the Jämtgaveln Nature Reserve based on GWR and compare the results to other methods of estimating fire size. In addition, we analyze environmental factors with GWR to assess whether these could help predict fire locations. We, therefore, try to answer the following questions:

1. Are there variables that help predict fire location (e.g. elevation, land cover, soil type, soil depth, and distance to early dairy farms)?
2. Can fire size be estimated with a more objective approach (as opposed to, e.g., the rather subjective method of drawing a polygon around the fire-scarred samples)?

2 Background

2.1 Dendrochronological Background Study

The basis for our study is a dendrochronological fire history in Jämtgaveln Nature Reserve in Västernorrland, Sweden (Figure 2.1, Hjalmarsson et al. 2015). The samples (e.g. Figure 2.2) were from 1986 (Linder 1988) and from 2009/2010 (Hjalmarsson et al. 2015). Linder (1988) did not have a GPS receiver or a GIS available and the location of his samples were placed on a map by hand. This map was scanned and georeferenced to a GIS map in ArcGIS. For this data, we estimate the accuracy to approximately 25m. Sometimes, several samples were located on the map with a single dot. For later analyses, the location of the samples could not be on the exact same spot. So, each sample that overlapped spatially with another sample was moved by a few meters. The samples from 2009/10 were recoded with a GPS receiver.

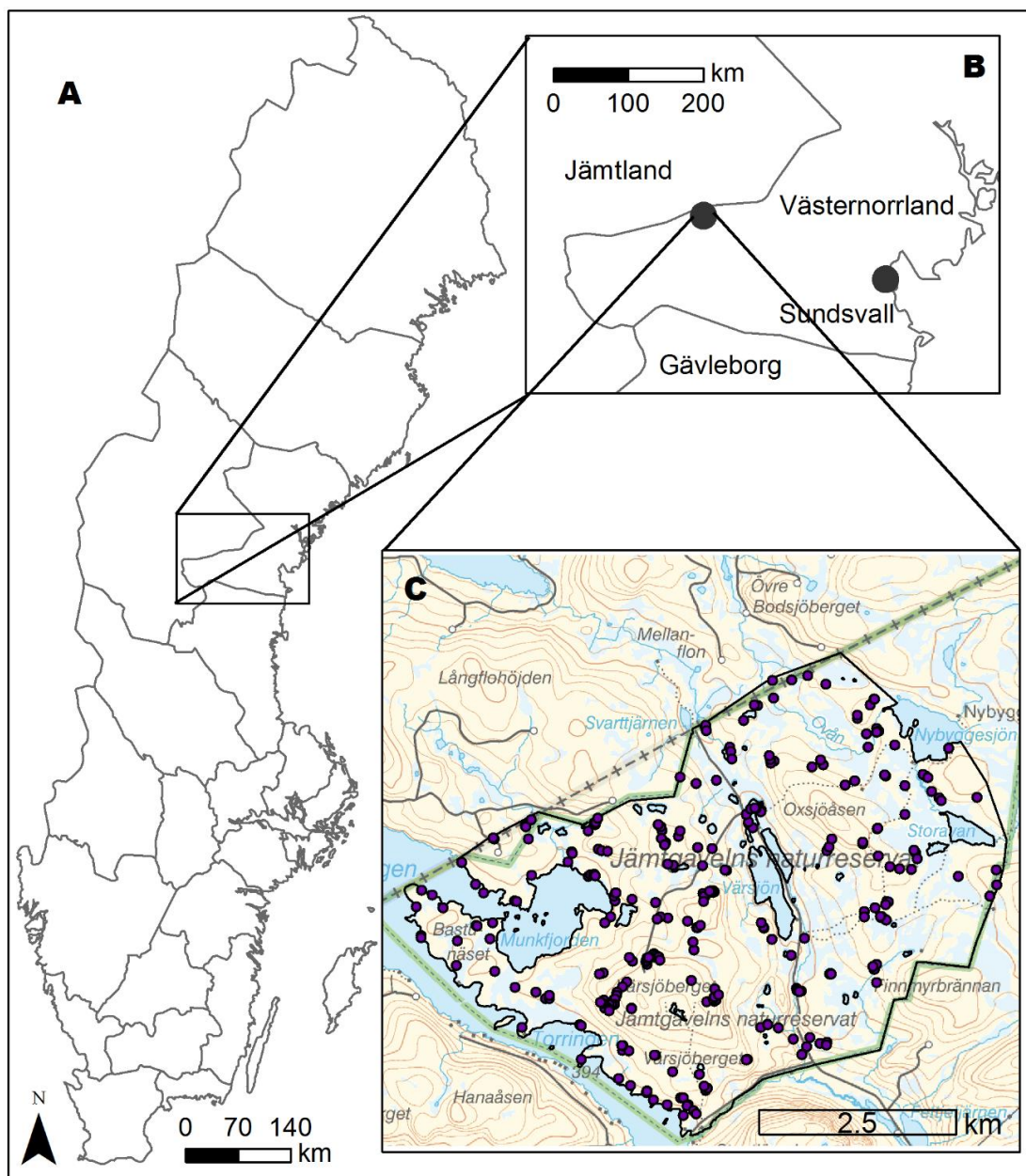


Figure 2.1: Study area's location in Sweden (A) and in the state of Västernorrland (B). C shows Jämtgaveln Nature Reserve (outlined in green), the study area (black line), and the location of all samples (purple dots).

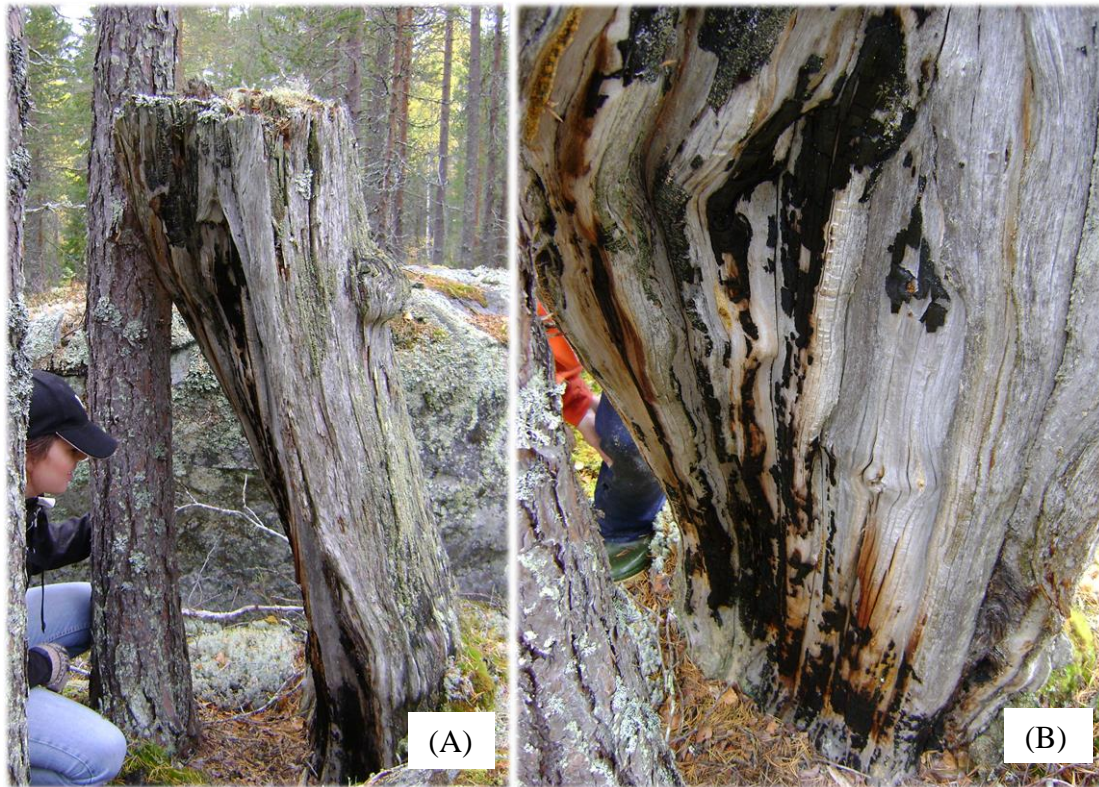


Figure 2.2: Pictures of a stump that was samples in 2009. (A) shows the size of the stump (selective cutting was done during winter and, therefore, stumps are often rather high) and (B) shows 4-5 fire scars on that particular stump.

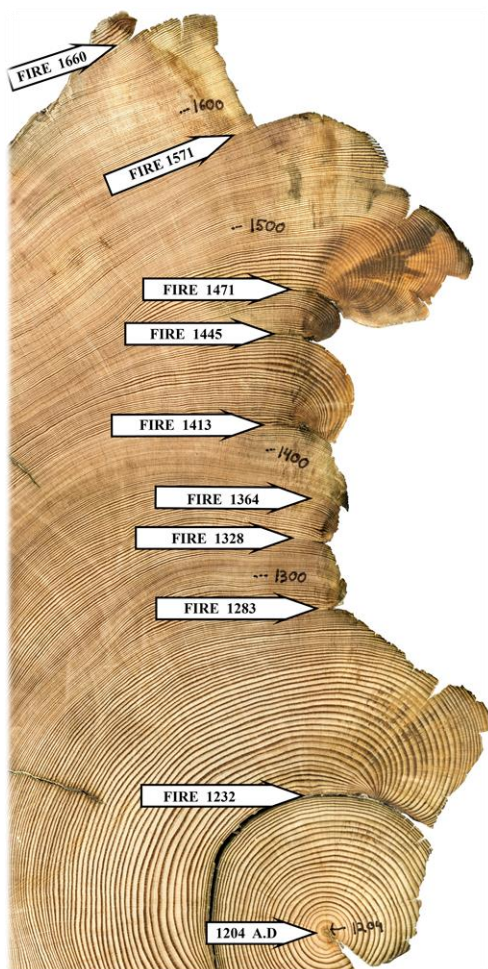


Figure 2.3: The cross-section of a sample taken just outside of the Jämtgaveln Nature Reserve in 2010 (Orelund 2015) with nine fire scars in approximately 450 years.

Because Linder (1988) did not use formal dendrochronological methods and that might have introduced some dating mistakes, most of his samples from 1986 were dated again with formal dendrochronological methods. The samples from 2009/10 and their fire scars were dated to the year with the appropriate dendrochronological methods. We dated 278 cross-sections (e.g. Figure 2.3) that had, altogether, 673 fire scars. The earliest fire scar was in 1311 and the oldest ring was recorded in 1274. In 1405, the chronology had 10 samples.

2.2 Method Comparison

In this study, several methods will be compared, i.e. Thiessen polygons, a grid analysis (Hjalmarsson et al. 2015), the drawing-outline method (Niklasson and Granström 2000), and geographically weighted logistic regression.

2.2.1 Thiessen Polygons

Thiessen polygons (a.k.a. Voronoi polygons) delineate areas of influence around the sample points so that each polygon only has one sample point in it. Every location within such a Thiessen polygon is closer to the point that it belongs to than to any other surrounding point.

2.2.2 Grid

Hjalmarsson et al. (2015) reconstructed a spatial fire history where the fire sizes are based on a grid that was overlaid on the study area. If a fire scar was located within a grid cell, that area that the grid cell covered was recorded as having experienced a fire that year. The area that had burned according to this method was summed and the percentage to the whole area with all samples that year calculated. Therefore, a percentage of the study area burned could be estimated.

2.2.3 Drawing

Niklasson and Granström (2000, pg. 1489) tried to be as objective as possible by following certain rules when they delineated fires, e.g. following “a probable fuel break (peatland, lake, or watercourse)”. The decision on what is a fuel break is rather subjective, as in relatively dry years peatland can also burn or in larger, intense fires watercourses could be jumped.

2.2.4 Geographically Weighted Logistic Regression

According to my literature research, applications of GWR to estimate fire size have not been published, yet. However, GWR have been used widely in other areas and have been explained in detail elsewhere (e.g. Charlton and Fotheringham 2009; Páez and Wheeler 2009). Logistic GWR has not been used to the same extent (but see, e.g., Atkinson et al. 2003; Wu and Zhang 2013; Zhang et al. 2014). Wu and Zhang (2013) have a good method description of logistic GWR (pg. 56-57).

Geographically weighted regression and its logistic counterpart, geographically weighted logistic regression, will be discussed in more detail in the next section.

2.3 Geographically Weighted Regression

2.3.1 Why local spatial regression?

Study areas like Spain or Southern Europe are far from homogeneous in topography, climate, land use/cover, population density or other phenomena (Koutsias et al. 2010). Yet, researchers apply global regression models to analyze larger study areas to explain occurrences of certain phenomena, e.g. Fernandes et al. (2014) who studied fuel and fire dynamics in the public forests of Portugal. Global regression models assume spatial stationarity in the relationship between dependent and independent variables, which is unreasonable in most ecological contexts, as nature is very heterogeneous. A concrete example of such spatial non-stationarity comes from modelling housing prices in London. In one part of the town, garages are valued very highly, whereas in other parts of the town, garages do not increase property value

(Fotheringham et al. 2002). Therefore, a global correlation value might express no or a low relationship between the two, even though in some locations there is a clear relationship. A more ecological example could be that the effect of temperature on species richness might change depending on precipitation (Jetz et al. 2005). GWR seems to be one solution to approaching these issues (Foody 2004).

Spatial autocorrelation between spatially dependent observations (two observations might have similar values because they are located closely to each other, e.g. elevation) violates classical statistical tests that are used in global statistical methods, as these assume independence between observations (Dale and Fortin 2002). GWR seems to be a possible solution here as well, because studies that compared GWR to global methods (i.e. ordinary least squares [OLS]) found a reduction in spatial autocorrelation in the residuals from those produced by the global methods as compared to the local GWR method (e.g. Zhang et al. 2005; Koutsias et al. 2010; Oliveira et al. 2014).

2.3.2 OLS vs. GWR

OLS (Ordinary Least Squares), a global linear regression model, is used to globally estimate which and to what extent independent variables influence the dependent variable (Table 2.1). The main goal is to find a set of variables that best describes the response to another variable, then predicts the response with the same variables in a similar population. Statistically, the goal is to minimize the difference between the predicted and the observed response.

GWR (Geographically Weighted Regression), using local linear regression models, is used to estimate how independent variables influence the response variable locally, which means that not one set of linear regression equations approximates the response, but rather, several locally applied regression models help estimate the response (locally). Statistically, the goal is also to minimize the difference between the predicted and observed response at each data point, and, thereby, to predict the relationship between variables. General inferences, however, from the study area to other study areas are difficult, because of this varying relationship with which hypothesis testing becomes impossible (Jetz et al. 2005).

Table 2.1: Main differences between OLS and GWR from an ESRI webinar on Regression Analysis (ESRI 2016).

OLS	GWR
Global regression model	Local regression model
One equation, which is calibrated using data from all features	One equation for each feature, which are calibrated using data from nearby features
Relationships are fixed across study area	Relationships can vary across study area

2.3.3 Geographically Weighted Logistic Regression

As GWR should only be used with data that have a wide spread in their values (ESRI 2016), Geographically Weighted Logistic Regression (GWLR) should be used for binary data such as fire data (fire/no fire). As with GWR, GWLR uses also local models, i.e. logistic regression models that are fitted at each sample point.

Within the geographically-weighted framework, a radius (a.k.a. bandwidth) around an observation is chosen and all observations within that bandwidth are then weighted. Most

often, the closer an observation is located to the observation in question, the more weight that observation is assigned. This is based on the premise that closer features are more related to each other than features further away from each other (Tobler 1970).

The weighting scheme is also known as a kernel (Figure 2.4). Commonly used kernels are the Gaussian and bi-square kernels, as only those two can be chosen in the program GWR4. The results have been found to be usually less sensitive to the choice of kernel than the selection of an accurate bandwidth (Brunsdon 2011; Charlton 2011).

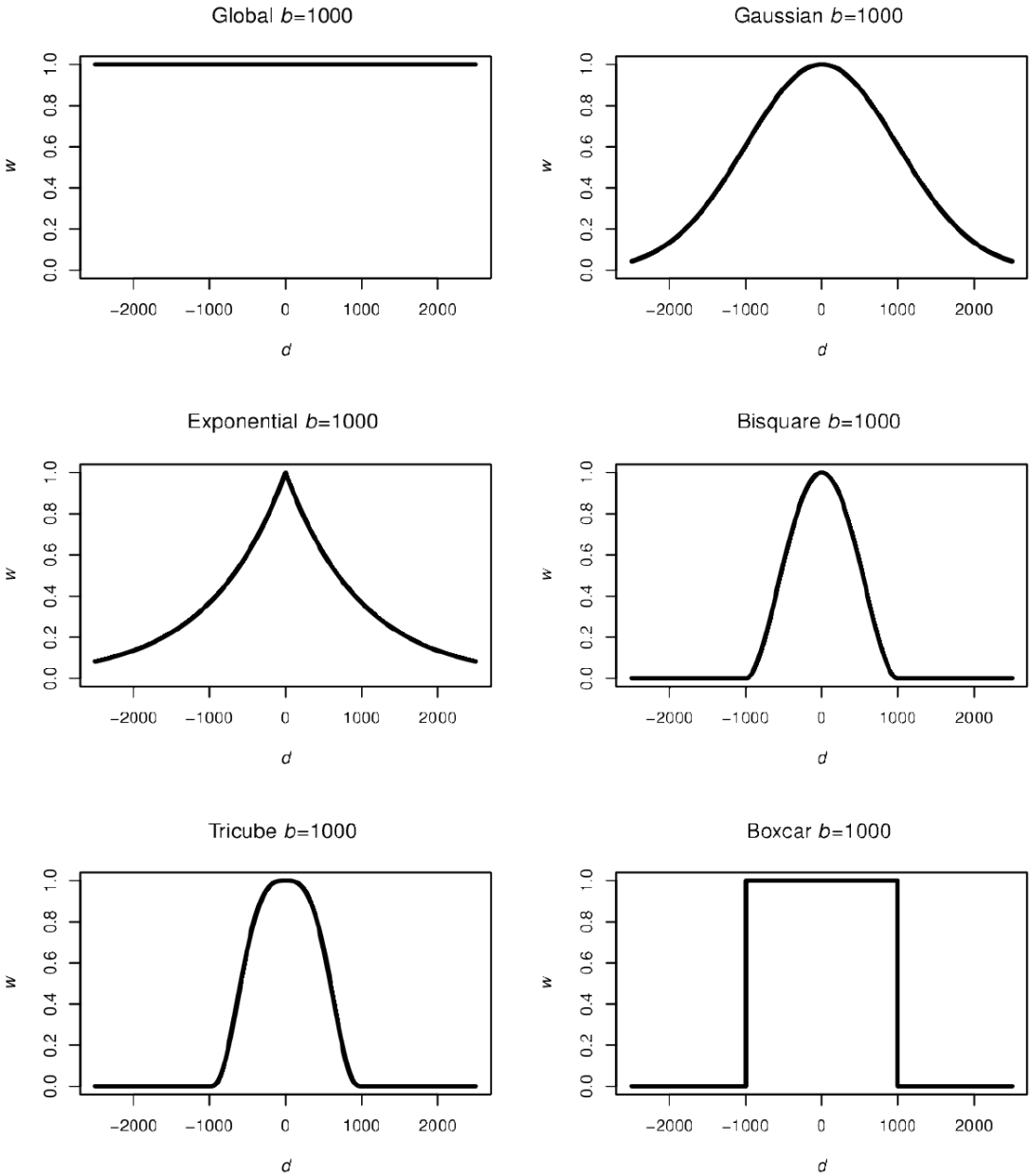


Figure 2.4: The different kernel shapes as can be selected in GWmodel (R package, Gollini et al. 2015).

The bandwidth can be fixed or adaptive. Fixed means that a certain radius, based on a Euclidian distance is used (e.g. 2000m). Adaptive means that a certain number of observations need to be included, so that the radius in meters might change based on the density of observations. Koutsias et al. (2010) decided in their study to work with the adaptive bandwidth, because it had slightly better coefficients of determination and because Cho et al.

(2009) found that extreme coefficients do not have as strong an influence on coefficient maps when using the adaptive approach.

The optimal bandwidth can be chosen based on different methods, of which three are most common. The researcher could (1) predefine the bandwidth, as e.g. based on expert knowledge, or could (2) base the bandwidth on a minimized (corrected) Akaike Information Criterion (AIC or AIC_C) (Akaike 1973), or could (3) base the bandwidth on a cross-validation method. Estimating the influence of observations further away can be rather difficult even with extensive expert knowledge. Therefore, more statistically grounded methods like CV or minimizing AIC might be a better solution.

2.3.4 PCA and GWR

Within the GWR framework, a principal component analysis (PCA) is also possible. As with OLS and GWR, a geographically weighted PCA (GW PCA), as compared to a standard PCA (where principal components are not dependent on location), might also be a better choice depending on the data used, because principal components (PCs) could also vary based on the location of the samples (see Gollini et al. 2015 for a succinct description with examples).

In this study, we did not run a GW PCA, but calculated the PCs from a standard PCA and used those in a GWR. A major problem in GWR is the multicollinearity between the independent variables. This multicollinearity is avoided when using principal components in GWR, as those PCs are, by definition, independent of each other.

2.3.5 Advantages and Disadvantages

GWR has been used in many projects and publications, and those in favor of using GWR have found significant improvements in their local model when it is compared to global models (like OLS). Those improvements, as compared to global models (such as OLS), include (1) no (or less) significant spatial autocorrelation in the residuals, which is indicated by normally distributed residuals, (2) the variance explained (of the dependent variable) increases, and (3) the size of the residuals in GWR is much less than that of the OLS residuals (Koutsias et al. 2010).

Atkinson et al. (2003) argued that GWR was developed for human geography uses, where socioeconomic variables are the basis for the statistical calculations. Such socioeconomic data and their relationship is likely to vary, e.g. between inner city and suburbs. Atkinson et al. (2003) continues that the assumption is different in physical geography, where a more positivistic view is taken, i.e. taking samples in one area, one should be able to infer the results to another area with the same characteristics because the relationship should not change. While the discussion about whether human and physical geography are distinctively different has a long history, this is not the place to engage in it. For more information on this topic see Goudie (1986) or Massey (1999).

Some papers that have found improved models when using GWR over OLS, have not been very critical of the method, although this method has been criticized widely. While most consider the multicollinearity of the covariates, multicollinearity among the coefficients, which is not unusual even when the covariates are not collinear (Wheeler and Tiefelsdorf

2005; Wu and Zhang 2013), is not always considered (e.g. Rodrigues et al. [2014] for a fire-related study). Wheeler and Tiefelsdorf (2005) criticized the use of independent variables that had high multicollinearity amongst each other, as this could possibly lead to a low correlation between the GWR coefficients, which, if undetected, could lead to misinterpretations of the spatial pattern. They also tested GWR for correlated coefficients that were based on uncorrelated independent input variables, as this is a violation of standard regression assumptions. As both of these multicollinearity problems are an issue in GWR, Wheeler and Tiefelsdorf (2005) suggest using well-established diagnostic tests, e.g. local VIF (variance inflation factor) maps applied to each variable, and to use caution when interpreting GWR results.

Jetz et al. (2005) criticized GWR and Foody's (2004) study in a sharp reply to Foody (2004) in which they pointed out several limitations of GWR that need to be addressed before being able to use it. Firstly, GWR does not account for spatial autocorrelation in the residuals directly. They agree that because the models are allowed to vary locally, the spatial patterns of the residuals are likely to be captured; however, much of that autocorrelation seems to remain, which in its turn undermines the validity of the regression method.

Even though GWR might capture all spatial patterns in the residuals (i.e. it accounts for nonspatiality), this method of modelling autocorrelation is “rather unparsimonious”, i.e. liberal, and other models, e.g. autoregressive ones, should be used (Jetz et al. 2005, pg. 97). Koutsias et al. (2010, pg. 234), in addressing this issue, counter these argument with the following: “Less structured residuals have been observed in other studies dealing with GWR, indicating that although the method does not directly address spatial autocorrelation issues, it provides a solution to the problem of spatially autocorrelated errors (Propastin and Kappas, 2008).”

Another limitation of GWR, according to (Jetz et al. 2005, pg. 98) is that with varying local parameters, “we can not make any general inferences from this analysis about how these variables affect [the dependent variable]” and that extrapolation to other regions is not possible (as only interpolations are possible in GWR). They argue that: “It is possible that the effects of these variables really do vary locally, [...]. It seems more likely, however, that the relationships are in fact global, but appear to vary locally due to missing variables or interactions terms” (pg. 98) and that strong correlations might just point to an excessively flexible GWR. To Jetz et al. (2005, pg. 98) “it seems more useful to fit global parameters” to be able to test biological hypotheses and make predictions for other geographical areas that then could be tested.

Jetz et al. (2005) also see a couple of benefits with GWR. However, this is mostly in terms of data exploration and in support of the global methods, e.g. using GWR to evaluate the changing relationship strength when the spatial resolution changes or identifying interaction terms or missing variables.

Páez et al. (2011, pg. 3007), comparing different aspects of the GWR method (bandwidth selection, sample size, collinearity, discrimination of stationarity/non-stationarity) using simulation methods, agree with Jetz et al. (2005) that the “findings of [their] experiments

weigh against” using GWR “as a tool for making spatial inferences”. Páez et al. (2011, pg. 3008) recommend: (1) using “judicious interpretation of the bandwidth size and/or the degree of spatial variability in the estimated coefficient surfaces [...] to discriminate between spatially stationary and non-stationary processes”, (2) using larger sample sizes, because smaller sample sizes could result in higher spurious correlations, or proceeding “with extreme caution” when interpreting results, and (3) if using larger sample sizes to visualize the estimated coefficients and other diagnostics to better assess the model and its potential problems.

Breyer (2013) summarized the problems that GWR has as being threefold: (1) Differentiating between non-stationary and stationary processes has not been consistently done using GWR, (2) multicollinearity in the calculated coefficients could result in biased interpretations, and (3) reliable diagnostics to identify model problems have not been identified. Roger Bivand (2012), who wrote a GWR script (spgwr) in R, went even so far as to call GWR “a notoriously unreliable technique” in a blog entry.

Considering the limitations of GWR, as presented above, and the fact that some researchers suggested other methods that more reliably describe the local relationships between independent and dependent variables (e.g. Griffith [2008], who suggested spatial filtering based on its better handling of the autocorrelation in the standard errors), GWR is more popular than ever (Figure 2.5). Many researchers seem to agree that even if GWR has its flaws, it is still “a useful exploratory analytical tool that can provide information on spatial non-stationarity in the relations between variables (Matthews and Yang 2012). Thus, most researchers concur that GWR can be reliably used as an exploratory technique to understand how covariates affect response variable of interest differently across geographical regions (Ogneva-Himmelberger et al., 2009)” (Wu and Zhang 2013, 53).

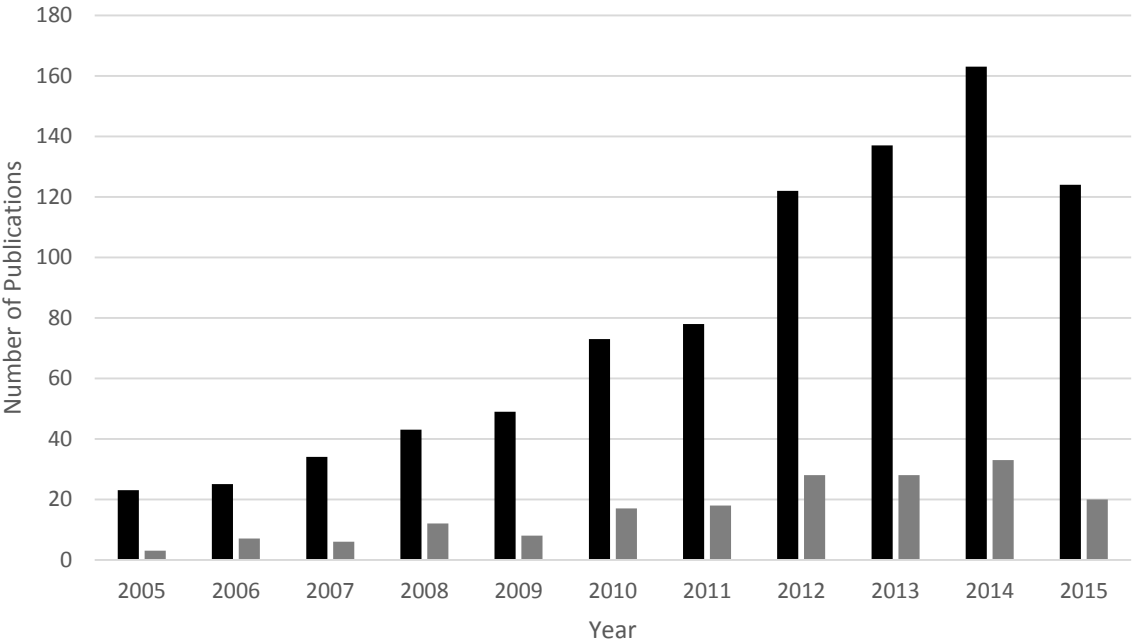


Figure 2.5: A Web of Science search of the term “geographically weighted regression” as it occurs in the topic (black bars) and title (grey bars) in the database (updated 2016-01-01).

2.3.6 Methodological Considerations

2.3.6.1 Six Checks

In a webinar from ESRI on regression analysis (ESRI 2016), the presenters talked about the GWR feature as it is implemented in ArcGIS 10.1. They introduced six checks when analyzing spatial patterns in spatial statistics, which I will base my results on:

1. Randomly distributed model residuals (over and under predictions should be random with no clustering)
2. Statistical significant coefficients (each/every explanatory variable should be important which is the case if coefficient is statistically significant)
3. Defensible variable relationships (each explanatory variable should have the relationship we would expect)
4. No multicollinearity (VIF needs to be less than 7.5)
5. Normally distributed residuals (model should perform as well for high as it does for low values and as well in one part of the study area as it does in the other parts; verify that residuals are normally distributed using Jarque-Bera test which should not be significant)
6. Strong adjusted R^2

2.3.6.2 Multicollinearity

Researchers that used GWR deal differently with multicollinearity in independent variables. Oliveira et al. (2014) calculated the Pearson product-moment correlation and removed variables based on a threshold of 0.7 in the correlation coefficients. In addition, they removed independent variables that had a “negligible relationship” with the dependent variable (Oliveira et al. 2014, pg. 147). The variance inflation factor (VIF, see 2.3.6.5) for the independent variables had to be close to one to be included in Oliveira's et al. (2014) study.

Zhang et al. (2014) removed multicollinearity between the covariates by running a principal component analysis (PCA) to get the principal components (PCs) of the data set. The PCs are independent from each other and do not contain any collinearity.

Rodrigues et al. (2014) calculated the Spearman's Rho correlation index to test for multicollinearity. Although they didn't have any collinear variables in their study, they discarded variables that were not significant in a Student's t-test and whose relationship was not consistent with the expectation as based on their (expert) experience and knowledge; just as was suggested by ESRI (2016). Rodrigues et al. (2014) did not take into consideration the problem of multiple hypothesis testing. Because in GWR, several linear regressions are run with (often) the same data, the significance values need to be adjusted (see section 2.3.6.3 below).

Sá et al. (2011) worked with VIF values, which in their study was under 1.5, indicated to them an absence of multicollinearity problems. They did not discard a coefficient that had a counter-intuitive sign in the global analysis. Instead, they saw that as an indication that the variable might better explain the response in a local analysis.

Multicollinearity is judged to be a rather influential aspect when models are created and interpreted. Surprisingly, reading through the fire GWR literature, many researchers are vague

in their research papers on what the acceptable level of multicollinearity is or how variables are judged to be included or not. While Oliveira et al. (2014) included a value (0.7) as a threshold in the correlation analysis, their VIF analysis needed to have values of “close to 1”. Martínez-Fernández et al. (2013) and Martínez et al. (2009, pg. 1244) used a lower threshold (0.5) for the correlation analysis and “[o]ther common statistical tests such as the tolerance coefficient, the variance inflation factor (VIF) and eigen-value analysis [...] to aid in variable selection”. No further explanations are given of the methods mentioned or values used. These vague descriptions do not help in reproducing the proposed methods and instead help critics of GWR to further criticize using this method.

2.3.6.3 Multiple Hypothesis Testing

Running many linear regressions with the same data could result in significant results just by chance as we increase the chances (likelihood) that rare event happen. This is a problem of multiple hypothesis testing. With multiple testing, we increase the likelihood of making a Type I error, i.e. falsely rejecting the null hypothesis.

Oliveira et al. (2014), as one of the few papers to consider this method, adjusted the p-values according to Byrne et al. (2009) and Sá et al. (2011) so that p became 0.000080 (for one study area) for an $\alpha = 0.05$.

2.3.6.4 AICc vs. CV

Akaike Information Criterion (AIC, Akaike 1974) measures the quality of a model for a data set relative to another statistical model of the same data set, rewarding better model fit and penalizing increasing model complexity. Using AIC, the researcher tries to find the model that approximates the data best, assuming that the model comes close to but is not the reality.

AIC is not an absolute quality measure, i.e. a null hypothesis cannot be tested. Several AIC values (from the same data set) are compared to each other and the statistical model with the lowest AIC value is the best model amongst all the ones tested. AIC is, therefore, a model selection measure and is calculated as a tradeoff between the complexity of models (a penalty is included when overfitting a model, where overfitting refers to improving the goodness of fit by increasing the number of independent variables) and the goodness of fit. AICc is then the AIC that is corrected for a low number of samples.

AIC is often seen as rather liberal, because overfitting (preferring a complex but wrong model over a simpler and true model) is a problem. To overcome this problem, cross-validation (CV) could be used as a model-selection method (Arlot and Celisse 2010). There are several types of CV methods, but what they have in common is a split into two subsets of the data for training and testing purposes. Splitting the data is usually done many times with results being averaged at the end.

CV has apparently been preferred in the GWR literature as it can more easily be decomposed into its components and tested (Farber and Páez 2007). However, AIC was used much more often in fire-related studies (e.g. Koutsias et al. 2010; Sá et al. 2011; Martínez-Fernández et al. 2013; Fernandes et al. 2014). In addition, Farber and Páez (2007) found that spatial outliers heavily influence the reliability of the cross-validation method.

BIC (Bayesian Information Criterion) and MDL (Minimum Description length) are other model-selection criteria. However, these two have not been commonly used in the GWR literature (although they can be chosen in the GWR4 program). The reason behind this might be due to the fact that BIC should be used to find a model when we know that a true model exists, whereas AIC and CV are used when trying to find a model that is the best approximation of the real world, i.e. no assumption is made that there exists a model that describes the data perfectly. Statistically, BIC might select a model that better fits theoretical data. However, real-world data might be better modelled after the AIC indicates that the ideal model rarely exists, and if it were to exist that model might not have the most predictive power.

2.3.6.5 Variance Inflation Factor (VIF)

The variance inflation factor (VIF) quantifies the influence multicollinearity has on an estimated regression coefficient. The higher the VIF, the more correlated the predictors. There are different guidelines that different researchers use in its application. Oliveira et al. (2014, pg. 147) used all covariates that had “VIF values close to 1”. ESRI (2016), in their webinar, suggested that all variables with a VIF value larger than 7.5 need to be excluded. Minitab (2015) published the following guidelines on their support pages: A VIF of one means that the predictors are not correlated; a VIF value between one and five means moderately-correlated predictors; and with a VIF value of five to ten, the predictors are highly correlated. Wu and Zhang (2013, pg. 58) indicate that a number less than 10 might be sufficient, as they wrote: VIFs “were all close to 1 (much less than 10)”.

An acceptable VIF value that is commonly or consistently used in different studies does not seem to exist. The reason behind that might be the fact that many researchers using VIF are vague about their application of it, or do not mention their threshold VIF value in their research papers (e.g. Martínez et al. 2009; Martínez-Fernández et al. 2013; Oliveira et al. 2014).

2.3.6.6 Programs

Several programs have the functionality to run GWR; among them ArcGIS 10.2.2 (ESRI 2014), GWR4 (Nakaya 2014), R (The R Foundation for Statistical Computing 2015), and SpaceStat (BioMedware 2016). As ArcGIS’ functionalities do not include logistic GWR, I decided to work with R, as several packages have been developed for running logistic GWR.

3 Methodology

3.1 Study Area

Jämtgaveln Nature Reserve (referred to as Jämtgaveln) is located in the middle of Sweden (62.67°N, 15.87°W). This Natura-2000 site (European Commission 2016) consist of approximately 3000 ha of forests, mires, and water bodies which comprise roughly 50%, 25%, and 25% of the Reserve, respectively. This study focuses on approximately 2680 ha of this nature reserve. The northeastern part was excluded because it is dominated by *Picea abies*. Such forests could potentially also have stumps of *Pinus sylvestris*; however, these stumps are often very decayed and soft compared to *Pinus sylvestris*-dominated forests.

Larger lakes dominate this part of the landscape and humans influenced this location relatively early (Länsstyrelsen Västernorrland 2007). For example, until the mid-17th century, Jämtgaveln bordered Norway and some battles were fought in this area. Also, a pilgrim's path went through this area from coast to coast, and a larger hut for overnight guests is assumed to have been located on Bastunäset (most western part of Jämtgaveln). Cultural artifacts are also still visible in some places like, for example, huts or axe marks (Figure 3.1). Widespread (partially selective) logging activities occurred in the late 19th century and in the 1930s (Linder 1988). In 1995, the area became a nature reserve. Controlled burning was introduced in the 1990s and is performed intermittently (Länsstyrelsen Västernorrland 2007).

Pinus sylvestris and *Picea abies* dominate the tree layer whereas *Vaccinium* spp. dominate the understory (Figure 3.2). Most of the forest in Jämtgaveln is about 100-140 years old, because in the middle of the 19th century, fires renewed the forest and selective cutting removed the huge and presumably old *Pinus sylvestris* (Länsstyrelsen Västernorrland 2007). Some old growth characteristics are still present in some parts of Jämtgaveln. However, material selected for this study was mostly from stumps.



Figure 3.1: Axe marks on a snag in Jämtgaveln Nature Reserve. Selective logging removed most of the larger and presumably older pines (*Pinus sylvestris*).



*Figure 3.2: Pictures of the center part of the Jämtgaveln Nature Reserve. These picture nicely illustrate the “blocky” (a.k.a. lots of boulders) nature of this reserve. Besides lots of blueberry (*Vaccinium* spp.), debris of former trees or standing snags are also common in the forested parts of the reserve.*

The bedrock consists mainly of three types; nutrient-poor granite (around the lake Torringen in the western part of Jämtgaveln), nutrient-rich diabase (mostly in the eastern part of Jämtgaveln), and in between these, granodiorite. Much of Jämtgaveln is very rocky, being covered by blocky moraine, whereas other parts are covered by peat. The elevation ranges from approximately 390m to 475m.

The Köppen climate classification for this area is Dfc, which is a subarctic climate type with cool summers and no dry season (however, most precipitation comes during the summer) with about 603 mm precipitation on average per year (KNMI 2015). The highest average monthly temperature in July is 13.7°C whereas the coldest average monthly temperature in January is -9.6°C (KNMI 2015).

3.2 Data

The *dependent* variable is always a fire year. First, I worked with all the data combined, however, only 14 samples of the 278 included did not have a fire scar in any of the years. Second, I also worked with a subset of all fire years, which were the 12 fire years that had the most fire scars (Table 3.1).

The *independent* variables (downloaded from Lantmäteriet's data service) are the classes of a reclassified land cover layer that were represented in the study area (Table 3.2); 2m elevation data (DEM); soil depth; soil blocks (how many boulders) which were classified to poor (very little boulders), rich (many boulders), and normal (little to normal amount of boulders); and soil types which are rocks, moraine, peat, gravel moraine, water, river sediment sand, river sediment gravel, and top (surface) layer of peat (not continuous) (Table 3.3).

Table 3.1: The 12 fire years that had the most fire-scarred trees.

Fire year	Fire-scarred samples	Samples
1521	42	64
1576	53	69
1644	16	116
1677	79	126
1723	17	163
1729	51	167
1770	31	204
1789	28	206
1807	18	207
1831	38	211
1844	15	193
1868	41	176
All Fire Years	264	278

The analyses have been done with selected variables to reduce the multicollinearity (redundancy) among the independent variables (Table 3.3). The choice of variables was limited by the availability of data and by the multicollinearity. Aspect/slope, for example, were not considered as they are direct derivatives of the DEM. Unfortunately, a wetness index

was not available for every year. Instead of using a wetness index, I used a land cover classification that separated between forest, mires, and forest on mires. The location of two dairy farms were, however, available and this variable was included as a more socio-economic variable even though the dairy farms were likely not present during the whole study period (Johnson 2008, pgs. 77-78). Storaunet et al. (2013) found that dairy farms profoundly influenced the fire occurrences in their study area.

Table 3.2: Land cover (marktücke) for Jämtgaveln Nature Reserve and the surrounding area (from Lantmäteriet). Several layers in the original data were reclassified to reduce the number of layers.

Original Code	Name of Layer	New Name
40	Deciduous trees not on mire or exposed bedrock	
43	Coniferous forest on lichen	
44	Coniferous forest not on lichen (7-15m)	Forest
45	Coniferous forest not on lichen (>15m)	
47	Coniferous forest on exposed bedrock	
48	Mixed forest not on mire or exposed bedrock	
41	Deciduous trees on mire	Forest on mires
46	Coniferous trees on mire	
49	Mixed forest on mire	
54	Clearcutting	
55	Young forest	Rocks
59	Exposed bedrock	
70	Fresh-water wetlands	
71	Wet mires	mires
72	Remaining mires	
81	Lakes and dams (open area)	Lakes
82	Lakes and dams (overgrown area)	

Table 3.3: All and selected variables (and their names) as used in the analyses separated by grid size.

All variables	100m grid 13variables	100m grid 10variables	250m grid 12variables	Names in Analyses
Distance to dairy farms	Yes	Yes	Yes	Dairyfarm
2m elevation data (DEM)	Yes	No	Yes	dem
Land cover (forest)	Yes	No	Yes	Lc_forest
Land cover (forest on mires)	Yes	Yes	Yes	Lc_formir
Land cover (open water)	No	No	No	Lc_lakes
Land cover (mires)	Yes	No	Yes	Lc_mires
Land cover (exposed bedrock, clearcuts, young forest)	No	No	No	Lc_rocks
soil blocks (little to normal amount of boulders)	No	No	No	Sb_normal
Soil blocks (very little boulders)	No	No	No	Sb_poor
Soil blocks (many boulders)	Yes	Yes	Yes	Sb_rich
Soil depth	Yes	Yes	Yes	Soildepth
soil types (bedrock on surface)	No	No	No	St_berg
soil types (gravel moraine)	No	No	No	St_gravmo

soil types (moraine)	Yes	Yes	Yes	St_morain
soil types (peat)	Yes	Yes	Yes	St_peat
soil types (river sediment gravel)	Yes	Yes	Yes	St_sed_gr
soil types (river sediment sand)	Yes	Yes	Yes	St_sed_sa
soil types (water)	Yes	Yes	No	St_water
Surface layer of peat (non-continuous)	Yes	yes	yes	Sto_peat

3.3 Data Preparation

After downloading all layers, the projection had to be changed for some of the layers from RT90 to SWEREF99TM. I reclassified the land cover from 19 to 6 classes (Table 3.2). I created a fishnet with a cell size 250x250m (and 100mx100m) over the study area, used zonal statistics to tabulate the area, and joined the results to the fishnet. In that new file, I calculated the percentage land cover of each class, which I then converted to raster with each field having the percentage of land cover. Therefore, I created five raster (forest, forest on mires, mires, lakes, rocks). The other layers were similarly converted to rasters with a 250x250m grid (and a 100mx100m). Appendix A has detailed descriptions of this procedure using ArcGIS 10.2.2.

The layer with the dependent variable needed to be prepared in terms of attaching all values of the independent variables to the attribute table that included every fire year. This was done using the “Extract Multivalue to Point” tool in ArcGIS 10.2.2.

The fire history data was prepared in Excel so that every column represents one fire year and every row one sample. For each sample, a zero (0) was added for every year where no record for that sample was available. A one (1) was set for a fire year where this sample did not have a fire scar. Two (2) was then for a fire year where the sample had a fire scar. This was mainly done for ArcGIS and presentation purposes, but became applicable for almost all analyses as well.

3.4 Analyses

The analyses were mainly done in two programs: ArcGIS 10.2.2 and R 3.2.2. ArcGIS was used to prepare the data (projecting the data, georeferencing, reclassifying, creating raster, ...) and do some of the analyses, e.g. Thiessen polygon creation, the grid analysis, drawing an outline. The geographic weighted regression was done in R although ArcGIS also has GWR functionalities. However, these were not appropriate for logistic GWR. Several programs, e.g. GWR4, R (GWmodel, spgwr, gwrr), or Space Stats, were available for running GWLR. After weighing the pros and cons and trying to run a few GWLR analyses, I decided to use GWmodel among the R packages, as it had the best descriptive instructions in form of a publication (Gollini et al. 2015).

3.4.1 Thiessen Polygons

The Thiessen polygons were created by the “Create Thiessen Polygons” tool in ArcGIS. For this analysis, it is important to not have samples at the exact same location, because the program just chooses one point and disregards the others. Within the tool, only a rectangular

area could be applied as processing area (irregular polygons like the study area could not be applied). Therefore, every Thiessen surface needed to be clipped to the study area. For each of the 12 major fire years, the areas for all polygons were calculated and the area for all polygons that had a burned sample and an unburned sample, respectively, were summarized.

3.4.2 Grid Analysis

The grid analysis was done by Hjalmarsson et al. (2015). A grid was calculated for the study area using the Fishnet tool (25 ha, polygons). The grid was intersected with the study area (a polygon that represented the land area) to get the real dimensions of each grid cell, which could be at that point less than 25 ha. For every year, the spatial-join tool (Analysis → Overlay → Spatial Join) was used to get only the grid cells that had (a) all samples, and (b) all fire-scarred samples in it (by selecting either all samples or all fire-scarred samples and creating new shapefiles, respectively, and using those in the spatial join¹).

The resulting table (with # of samples, # of fire scars, # of sample grid cells, # of fire-scarred grid cells, area of all samples in m² based on the grid cells, and area of fire-scarred grid cells in m²) was then used to calculate several percentages:

1. Number of scars vs. samples
2. Number of scarred grid cells vs. sample grid cells
3. Hectare burned based on (2) and the total land area of 2336 ha
4. Area that had samples vs. area that had fire scars

The comparison in this study between the different methods of calculating the fire size is based on the percentage of grid burned and the total land area of 2336 ha (#3 above).

3.4.3 Drawing the Fire Perimeter

Hjalmarsson et al. (2015) drew the fire perimeter for the fire years with the most fire scars (Table 3.1) in ArcGIS and calculations of the fire size could be done easily. The perimeters were drawn before doing the grid analysis. This was done in this fashion so as to not be influenced by the results of the grid analysis when drawing the fire perimeter. A perimeter was drawn around fire-scarred samples. These polygons often also included samples that did not have a fire scar, but that were located within the area that did not show an apparent fire break. The results of a Finnish study (Piha et al. 2013) show that depending on size of tree and fire intensity, only a few trees actually scar during a fire.

If one area did have many fire scars, but a neighboring area had none (but many non-scarred samples), the fire perimeter was drawn between the two areas. In the same fashion, an area that had burned was delineated from a (larger) area that did not have any samples. Lakes were excluded from the area calculations and the fire perimeter was drawn along the shoreline of lakes. Mires were included in area possible to burn as the few samples located on the mire also showed repeated fire scars.

¹ The options used for spatial join were: (1) I unchecked “Keep all target features”, (2) target features = grid over Jämtgaveln, (3) join feature = sample points (for a) or fire-scarred samples (for b), (4) match options = contains, and (5) join operation = join_one_to_one.

3.4.4 GWLR in R

The R package I used to run GWLR is called GWmodel (Gollini et al. 2015). In preparation for running GWLR, I also calculated a principal component analysis using Minitab version 15.1.30.0 (Minitab 2007) to identify the most important PCs. The first 7 PCs were used in further analysis.

In order to test the dependency (i.e. collinearity) of the independent variables, two corroborating methods were used: a correlogram (Wright 2015) and the “exploratory regression” function in ArcGIS. A correlogram of the data will show correlation plots for all variables in a matrix (Figure 3.3 A). The output of an exploratory regression will show you which variables are correlated (Figure 3.3 B and C).

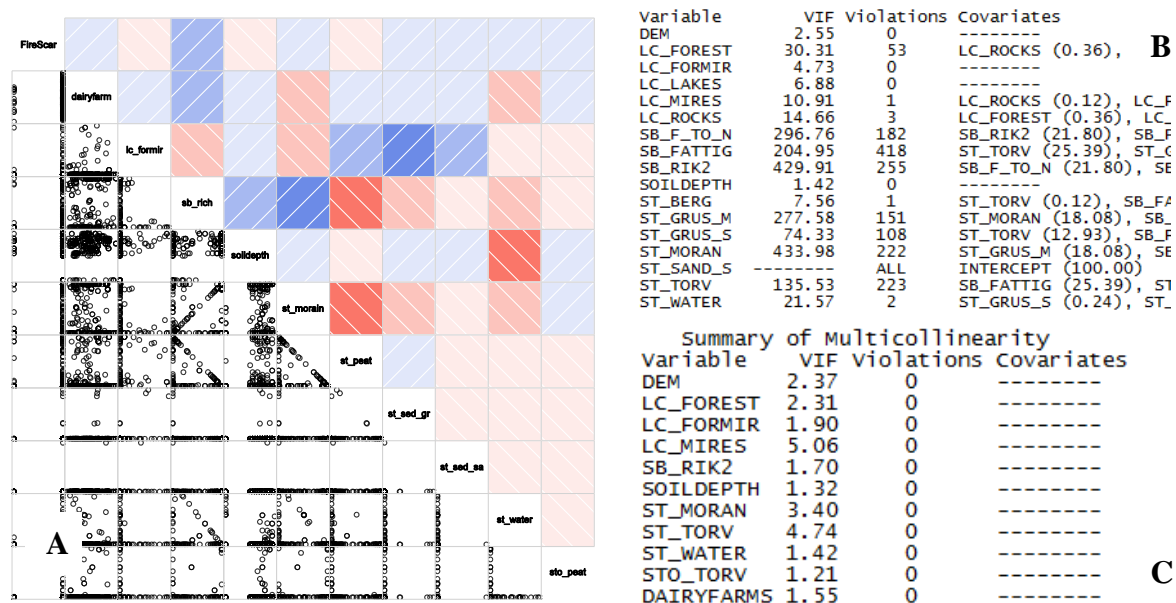


Figure 3.3: A correlogram from the analysis of my data: 100m grid and 10 independent variables (A), output of ArcGIS' exploratory regression with (B) and without (C) multicollinearity.

In GWmodel, a stepwise GWR is possible which would give spurious results when using with binary data. Therefore, a stepwise GWLR (with the command from GWmodel) was programmed (by Johan Lindström, Centre for Mathematical Sciences, Lund University). The stepwise GWLR was then run on two different scales (a 250m and 100m grid) and with all variables, a reduced number of variables, and the seven first principal components (Table 3.3).

3.4.5 INLA model

To test the project's feasibility at the very beginning of this work, Johan Lindström (Centre for Mathematical Sciences, Lund University) used Integrated Nested Laplace Approximations (INLA, Rue and Martino 2007; Rue et al. 2009), as implemented in the R-INLA package (Lindgren and Rue 2015) to estimate the likelihood of fire in each grid cell based on an model with binomial observations (fire/no fire) at the each observed tree. The model can be seen as a logistic regression with a spatial dependence structure (Lindgren et al. 2011) and regression coefficients fixed throughout space and time, this is equivalent to doing a GLMM

(Generalized linear mixed model) with spatially correlated random effects (Johan Lindström, personal [e-mail] communication, 2016-03-14 and 2016-03-21).

3.4.6 Probability Plot (Close Trees Burning at Same Time)

For this analysis, all samples that were located at the same location were moved by a few meters in ArcGIS (max. 15m). The distance between all points to all points were calculated in ArcGIS by using the Point Distance Tool, where I used the same data for input and near feature (shapefile with the data points for all samples = “fire data”). In Excel, I then calculated and visualized the probability of trees burning at the same time, as dependent upon their distance from each other (work-process description in Appendix B).

4 Result

4.1 Multicollinearity between the Covariates

In an exploratory regression in ArcGIS, the results show high multicollinearity between variables if all variables are considered (Table 4.1). This can be seen by the VIF values that are very high for most of the variables in the 250m analysis. In contrast, using only selected variables (250m grid), the VIF values drop significantly and are, according to ESRI (2016), acceptable (Table 4.2).

Table 4.1: Summary of multicollinearity for 250m and all variables (partial output of ArcGIS' exploratory regression analysis tool). The VIF is high for many variables and, therefore, multicollinearity is high between variables.

Variable	VIF	Violations	Covariates
DAIRYFARM	1.41	0	-----
SOILDEPTH	1.79	0	-----
STO_PEAT	1.1	0	-----
ST_SED_GR	5.38	0	-----
ST_SED_SA	1.23	0	-----
ST_WATER	16.79	835	LC_LAKES (99.88), SB_POOR (0.36), LC_FOREST (0.36), ST_MORAIN (0.24), SB_RICH (0.24), LC_MIRES (0.24), ST_PEAT (0.12), ST_GRAVMO (0.12), SB_NORMAL (0.12), LC_ROCKS (0.12)
ST_GRAVMO O	13.23	33	ST_MORAIN (3.95), SB_POOR (2.16), ST_WATER (0.12), LC_LAKES (0.12)
ST_PEAT	89.24	193	SB_POOR (16.77), ST_BERG (1.80), ST_WATER (0.12), LC_LAKES (0.12)
ST_MORAIN	23.65	83	ST_GRAVMO (3.95), SB_POOR (2.16), LC_FOREST (0.24), ST_WATER (0.24), LC_LAKES (0.24)
ST_BERG	19.58	15	ST_PEAT (1.80), SB_POOR (1.80)
SB_NORMAL	15.73	35	SB_RICH (4.19), SB_POOR (2.16), LC_FOREST (0.12), ST_WATER (0.12), LC_LAKES (0.12)
SB_RICH	24.96	95	SB_NORMAL (4.19), SB_POOR (2.16), LC_FOREST (0.24), ST_WATER (0.24), LC_LAKES (0.24)
SB_POOR	112.48	209	ST_PEAT (16.77), ST_MORAIN (2.16), ST_GRAVMO (2.16), SB_RICH (2.16), SB_NORMAL (2.16), ST_BERG (1.80), ST_WATER (0.36), LC_LAKES (0.36)
DEM	2.57	0	-----
LC_LAKES	18.39	837	ST_WATER (99.88), LC_FOREST (0.60), LC_MIRES (0.48), SB_POOR (0.36), ST_MORAIN (0.24), SB_RICH (0.24), LC_ROCKS (0.24), ST_PEAT (0.12), ST_GRAVMO (0.12), SB_NORMAL (0.12)
LC_MIRES	14.28	16	LC_FOREST (1.92), LC_ROCKS (0.48), LC_LAKES (0.48), ST_WATER (0.24)
LC_ROCKS	11	4	LC_MIRES (0.48), LC_FOREST (0.48), LC_LAKES (0.24), ST_WATER (0.12)
LC_FORMIR	3.06	0	-----

LC_FOREST	31.23	55	LC_MIRES (1.92), LC_LAKES (0.60), LC_ROCKS (0.48), ST_WATER (0.36), ST_MORAIN (0.24), SB_RICH (0.24), SB_NORMAL (0.12)
-----------	-------	----	--

Table 4.2: Summary of multicollinearity for 250m grid with 12 variables (partial output of ArcGIS' exploratory regression analysis). The VIF is low for many variables and, therefore, multicollinearity is low between variables (as suggested by ESRI [2016] who argue that a VIF value of below 7.5 is acceptable).

Variable	VIF	Violations	Covariates
DAIRYFARM	1.22	0	-----
SOILDEPTH	1.16	0	-----
STO_PEAT	1.06	0	-----
ST_SED_GR	1.18	0	-----
ST_SED_SA	1.05	0	-----
ST_PEAT	4.78	0	-----
ST_MORAIN	3.11	0	-----
SB_RICH	1.88	0	-----
DEM	2.51	0	-----
LC_MIRES	5.18	0	-----
LC_FORMIR	1.39	0	-----

For the 100m analysis, the results look similar in that most variables show high VIF values and multicollinearity when all variables were used, and no multicollinearity when only 13 variables were used (Table 4.3 and 4.4). Some differences can be seen between the 100m and 250m grid when using all variables. For example, in the 250m grid, the variables st_water and lc_lakes show collinearity issues, whereas in the 100m grid this is not the case. This could be due to the fact that many more cells are located next to water in the 250m grid, whereas in the 100m grid, the cells that have fire data are further away from water, because the cells are smaller.

The principal components from the PCA have per definition a VIF value of 1 and, therefore, no multicollinearity.

Table 4.3: Exploratory Regression output from ArcGIS for the 100m grid and all variables. Many variables have collinearity.

Variable	VIF	Violations	Covariates
DAIRYFARM	1.35	0	-----
DEM	2.26	0	-----
LC_FOREST	17.48	2676	LC_MIRES (1.74), SB_POOR (0.34), LC_ROCKS (0.21), ST_MORAIN (0.15), SB_RICH (0.15), SB_NORMAL (0.12), ST_GRAVMO (0.11), ST_PEAT (0.09)
LC_FORMIR	3.34	0	-----
LC_LAKES	3.8	0	-----
LC_ROCKS	7.79	46	LC_MIRES (0.21), LC_FOREST (0.21)
LC_MIRES	9.47	378	LC_FOREST (1.74), LC_ROCKS (0.21), SB_POOR (0.01), ST_MORAIN (0.01), SB_RICH (0.01), ST_PEAT (0.00), ST_GRAVMO (0.00), SB_NORMAL (0.00)

SB_NORMAL	19.79	6554	SB_RICH (30.09), SB_POOR (17.66), ST_PEAT (2.53), LC_FOREST (0.12), LC_MIRES (0.00)
SB_POOR	45.82	17547	ST_PEAT (41.11), ST_MORAIN (19.04), SB_RICH (18.89), SB_NORMAL (17.66), ST_GRAVMO (16.21), LC_FOREST (0.34), LC_MIRES (0.01)
SB_RICH	28.33	9552	SB_NORMAL (30.09), SB_POOR (18.89), ST_PEAT (3.06), LC_FOREST (0.15), LC_MIRES (0.01)
SOILDEPTH	2.14	0	-----
ST_BERG	5.24	0	-----
ST_GRAVMO	18.52	5387	ST_MORAIN (24.73), SB_POOR (16.21), ST_PEAT (1.99), LC_FOREST (0.11), LC_MIRES (0.00)
ST_MORAIN	27.54	9678	ST_GRAVMO (24.73), SB_POOR (19.04), ST_PEAT (3.08), LC_FOREST (0.15), LC_MIRES (0.01)
ST_PEAT	28.03	9110	SB_POOR (41.11), ST_MORAIN (3.08), SB_RICH (3.06), SB_NORMAL (2.53), ST_GRAVMO (1.99), LC_FOREST (0.09), LC_MIRES (0.00)
ST_SED_GR	3.43	0	-----
ST_SED_SA	1.45	0	-----
ST_WATER	5.23	0	-----
STO_PEAT	1.06	0	-----

Table 4.4: Exploratory Regression output from ArcGIS for the 100m grid and 13 variables. The VIF values are low (or at least below 7.5) and no collinearity is recorded.

Variable	VIF	Violations	Covariates
DAIRYFARM	1.22	0	-----
DEM	2.25	0	-----
LC_FOREST	2.4	0	-----
LC_FORMIR	1.63	0	-----
LC_MIRES	2.83	0	-----
SB_RICH	1.68	0	-----
SOILDEPTH	1.47	0	-----
ST_MORAIN	2.84	0	-----
ST_PEAT	2.5	0	-----
ST_SED_GR	1.34	0	-----
ST_SED_SA	1.08	0	-----
ST_WATER	1.65	0	-----
STO_PEAT	1.04	0	-----

The correlograms of the data (100m - all variables, 100m - 13 variables, 100m - 10 variables, 100m - PCA, 250m - all variables, 250m - 12 variables, and 250m - PCA) show a different story (Figures 4.1 and 4.3, Appendix C, the legend for the color code can be seen in Figure 4.1). Even though only variables with low VIF values (under 7.5) were included, the correlograms with only the selected 13 variables (for e.g. 100m) show high correlations between some of the remaining variables (Figure 4.2).

Correlation value (x 100)



Figure 4.1: Legend (color scheme) for the correlograms in Appendix D (Friendly 2002, 3).

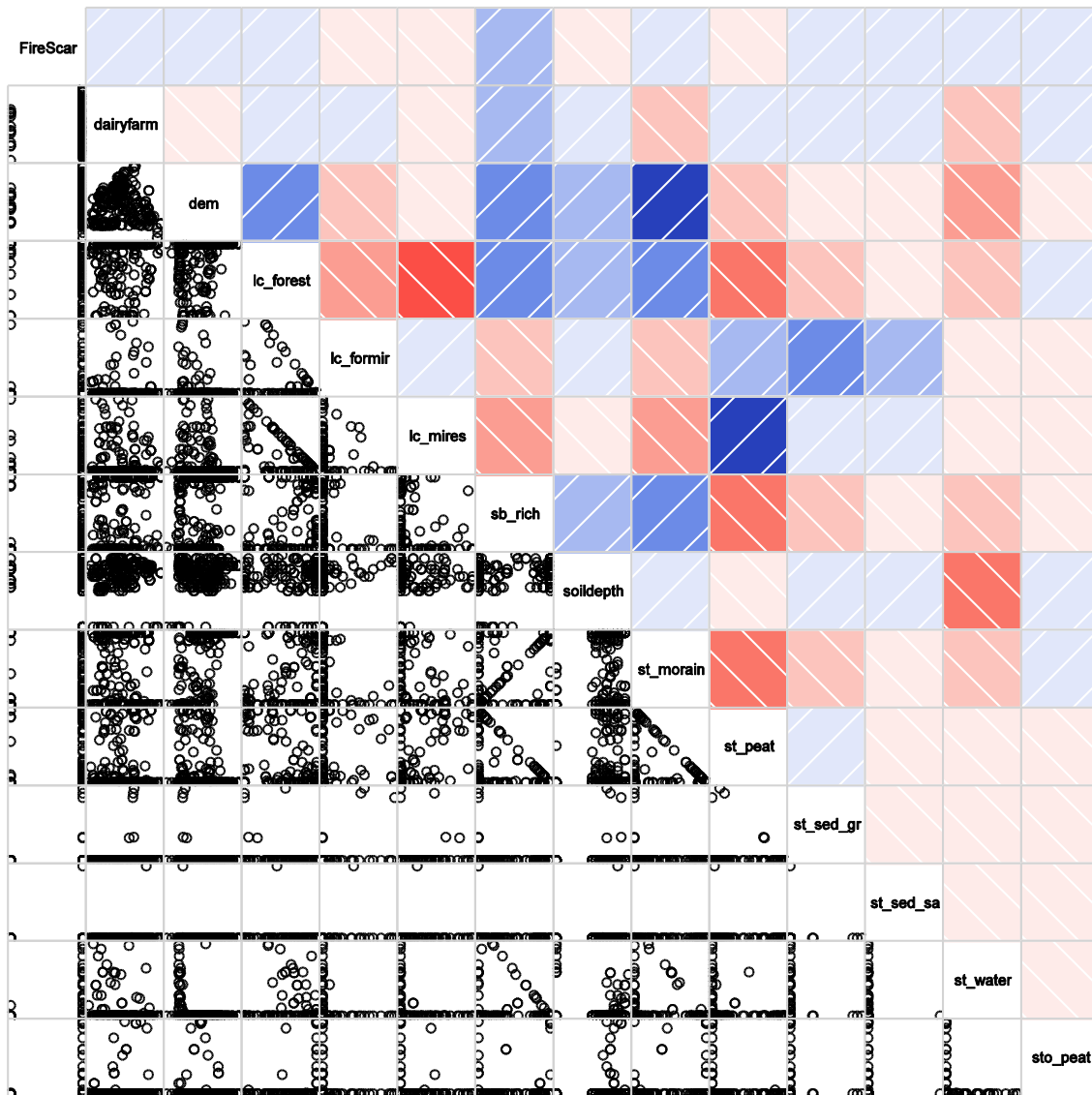


Figure 4.2: Correlogram for the 100m grid with 13 variables. High positive correlation can be seen between *dem* and *st_morain*, *lc_mires* and *st_peat*, and high negative correlation between *lc_forest* and *lc_mires*.

Therefore, disregarding three more variables from the 100m grid (Table 3.3), the resulting correlogram (Figure 4.3, Appendix C) shows that there is still some correlation between the variables, but not more than 0.7, which is the threshold suggested by Oliveira et al. (2015).

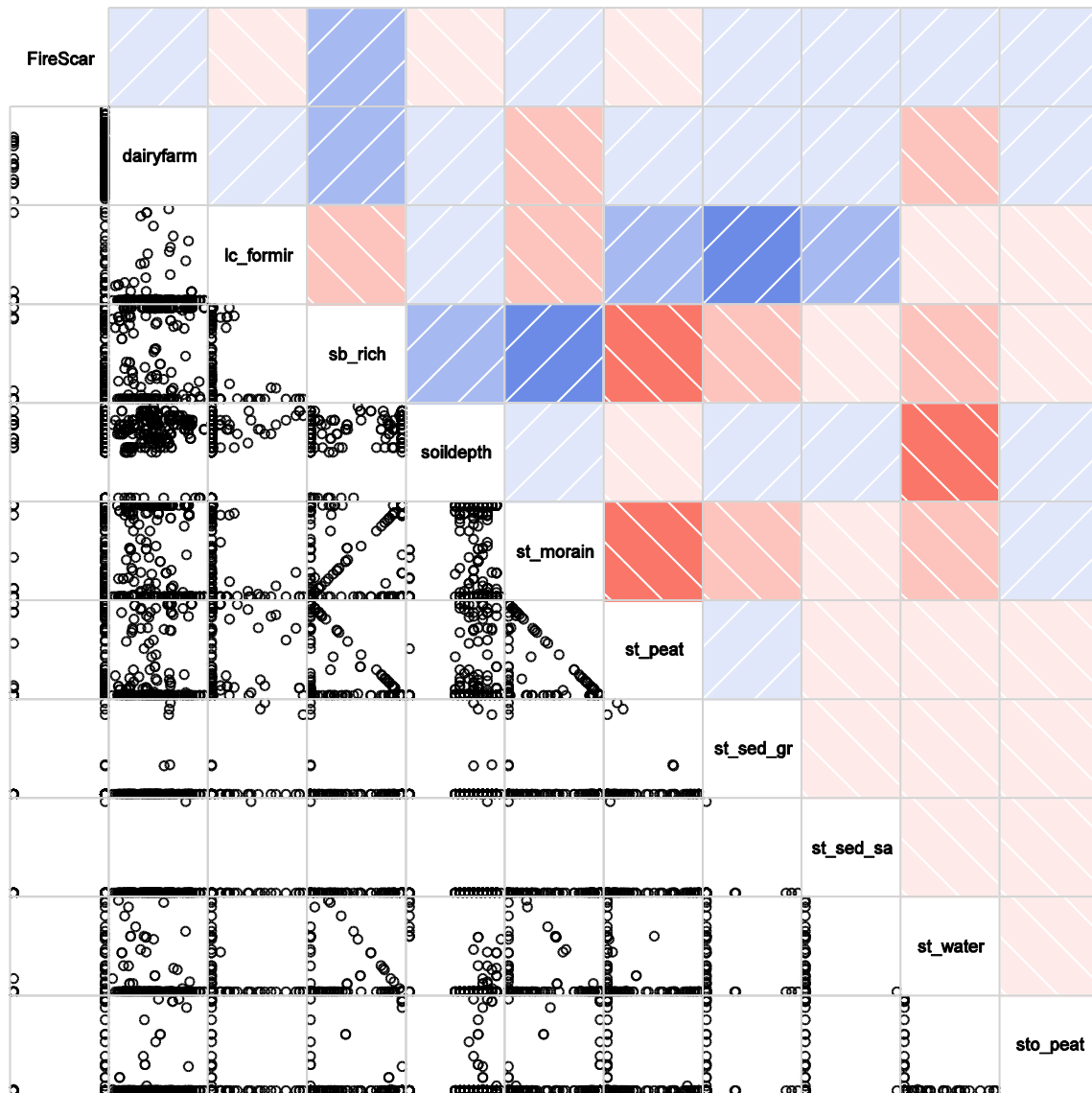


Figure 4.3: Correlogram for the 100m grid with 10 variables. After excluding 3 more variables, the correlation coefficients are below 0.7 which was deemed acceptable by Oliveira et al. (2015).

4.2 Principal Component Analysis (PCA)

4.2.1 PCA 100m Grid

For this analysis, the first seven PCs were extracted, although PC6 and PC7 do not add much to the explanatory power of the analysis (Figure 4.4, Scree Plot). The single PCs were interpreted with the help of the correlation coefficients of each variable with the PC (Table 4.5):

- PC 1. Mires (little forest, low on larger boulders [sb_poor], little moraine, lots of peat)
- PC 2. Low elevation and close to water/lakes
- PC 3. Far from dairy farms; rather shallow soil; not on gravel moraine (but positively with moraine?)
- PC 4. Close to dairy farms, rich in larger boulders and/or bedrock on surface (st_berg)

- PC 5. Bedrock and low soildepth; average amounts of boulders (sb_normal)
- PC 6. Forest on mires with gravel sediments; not on clearcuts, young forest, or exposed bedrock (lc_rocks)
- PC 7. Not forest on mires; not gravel sediments; not on clearcuts, young forest, or exposed bedrock (lc_rocks)

Table 4.5: The correlation coefficients (for the 100m grid) for each variable and the first 7 PCs (**values*** ≥ 0.35 , **values[§]** = values between 0.30 and 0.35; the loading thresholds of 0.30 and 0.35 was randomly set to ease the interpretation of the coefficients).

	Coeff PC1	Coeff PC2	Coeff PC3	Coeff PC4	Coeff PC5	Coeff PC6	Coeff PC7
dairyfarm	0.022	-0.125	-0.336[§]	0.380*	0.161	-0.038	-0.132
dem	-0.279	-0.321[§]	0.274	-0.061	-0.020	0.036	-0.043
lc_forest	-0.381*	0.064	-0.121	0.014	0.284	0.262	0.240
lc_formir	0.195	-0.093	-0.209	-0.196	-0.112	0.393*	-0.356*
lc_lakes	0.181	0.389*	0.210	0.133	-0.270	0.181	0.077
lc_rocks	0.064	-0.016	0.032	0.062	-0.202	-0.629*	-0.526*
lc_mires	0.313[§]	-0.260	0.148	-0.052	-0.026	-0.126	0.280
sb_normal	-0.002	0.212	0.047	-0.614*	0.309[§]	-0.112	-0.114
sb_poor	0.402*	-0.282	0.024	0.102	0.152	0.109	0.072
sb_rich	-0.328[§]	-0.108	-0.138	0.401*	-0.306[§]	0.007	0.049
soildepth	-0.145	-0.267	-0.342[§]	-0.255	-0.305[§]	0.052	0.180
st_berg	0.080	0.054	-0.023	0.351*	0.611*	0.025	-0.175
st_gravmo	0.041	0.292	-0.560*	-0.077	-0.069	-0.193	0.155
st_morain	-0.367*	-0.168	0.357*	-0.050	0.008	0.071	-0.173
st_peat	0.360*	-0.299	0.086	-0.001	-0.021	-0.054	0.299
st_sed_gr	0.150	-0.076	-0.113	-0.059	-0.138	0.444*	-0.418*
st_sed_sa	0.057	-0.053	-0.089	0.005	0.023	0.166	-0.152
st_water	0.112	0.476*	0.256	0.153	-0.185	0.164	0.076
sto_peat	-0.053	0.027	-0.109	-0.131	0.169	-0.072	0.076

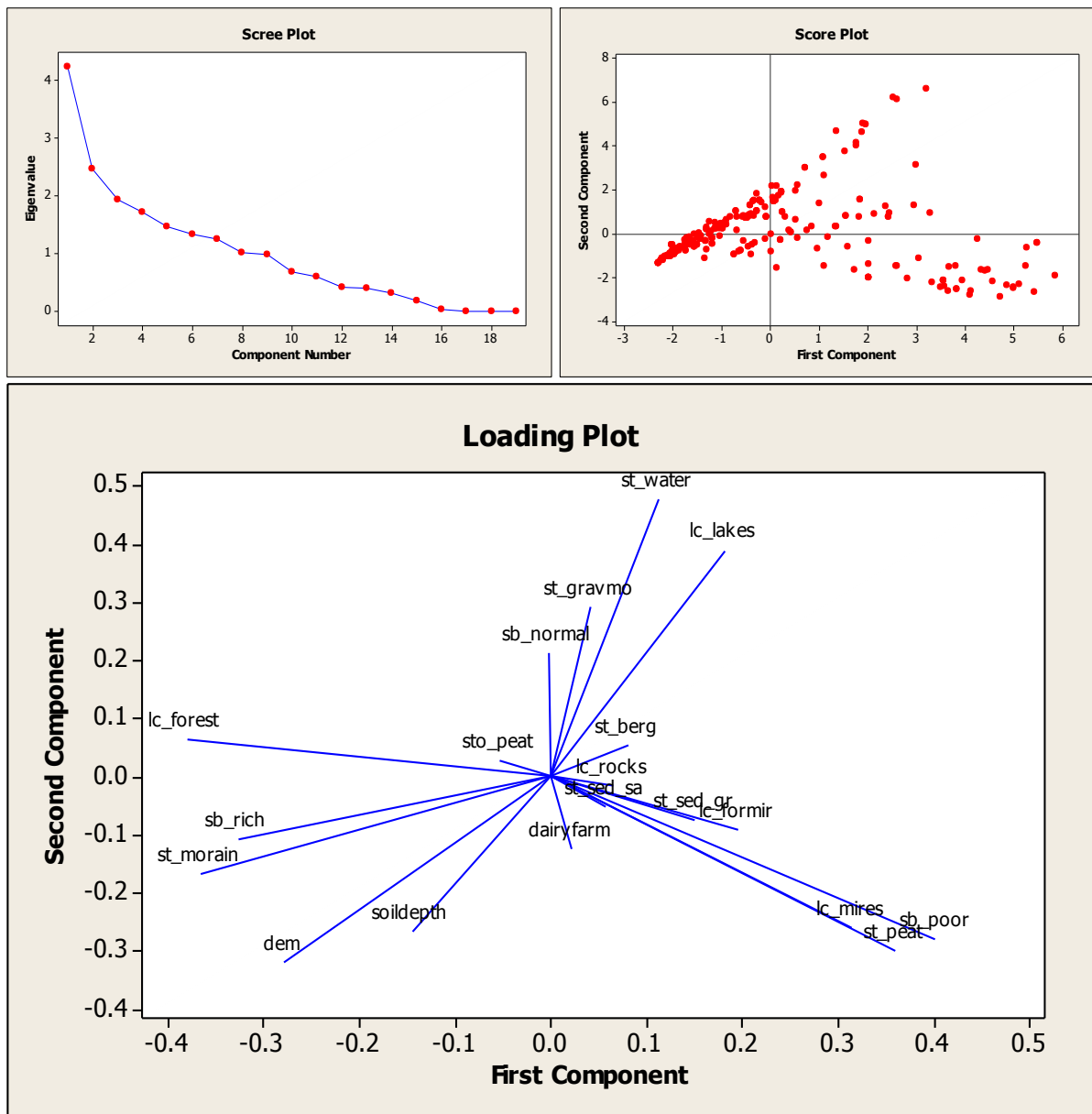


Figure 4.4: Results of the PCA with the 100m grid. All variables were included in the PCA.

4.2.2 PCA 250 Grid

The results for the 250m-grid PCA show that the PCs are more distinct (Figure 4.5, Loading Plot) as compared to the 100m-grid PCA (Figure 4.4, Loading Plot). The variables are much more dispersed in the 100m-grid PCA than with the 250m-grid PCA. Interpreting the PCs was a little less challenging in this analysis (Table 4.6):

- PC 1. Mires with lots of peat; low on larger boulders (sb_poor); not on morain
- PC 2. close to open water; not on mires/peat
- PC 3. far away from dairy farms; not on exposed bedrock or on gravel moraine; higher elevations
- PC 4. high amounts of larger boulders; no surface peat (sto_peat); on deeper soils
- PC 5. on exposed bedrock (st_berg); low soil depths
- PC 6. not on exposed bedrock, clearcuts and young forests (lc_rocks)
- PC 7. On gravel sediments; on forest on mire, but not on mires

As with the analysis of the 100m grid, PC 6 and PC 7 are not very important to the analysis either (Figure 4.5, Scree Plot).

Table 4.6: The correlation coefficients (for the 250m grid) for each variable and the first 7 PCs (**values*** ≥ 0.35 , **values[§]** = values between 0.30 and 0.35; the loading thresholds of 0.30 and 0.35 was randomly set to ease the interpretation of the coefficients).

	Coeff PC1	Coeff PC2	Coeff PC3	Coeff PC4	Coeff PC5	Coeff PC6	Coeff PC7
dairyfarm	-0.006	-0.115	-0.515*	0.155	0.021	-0.108	-0.029
soildepth	-0.125	0.166	0.085	0.290	-0.532*	-0.025	-0.141
sto_peat	-0.050	0.057	-0.083	-0.338[§]	-0.210	0.015	-0.244
st_sed_gr	0.179	-0.046	-0.027	0.101	-0.129	0.044	0.607*
st_sed_sa	0.066	-0.057	-0.075	0.107	-0.141	0.079	0.025
st_water	0.118	0.517*	0.145	0.156	0.187	0.141	-0.089
st_gravmo	0.078	0.229	-0.464*	-0.180	-0.311[§]	-0.213	-0.118
st_peat	0.341[§]	-0.298	0.156	0.119	-0.162	0.070	-0.199
st_morain	-0.370*	-0.186	0.288	0.022	0.114	-0.001	0.146
st_berg	0.077	-0.093	-0.382*	-0.153	0.515*	0.268	0.035
sb_normal	0.045	0.055	0.190	-0.661*	-0.141	-0.069	0.118
sb_rich	-0.337[§]	-0.059	-0.202	0.407*	-0.003	-0.100	-0.035
sb_poor	0.369*	-0.311[§]	-0.030	0.073	0.026	0.185	-0.035
dem	-0.312[§]	-0.251	0.245	0.006	-0.005	0.018	0.050
lc_lakes	0.170	0.491*	0.163	0.161	0.155	0.146	0.031
lc_mires	0.333[§]	-0.291	0.137	0.066	-0.074	-0.008	-0.318[§]
lc_rocks	0.057	-0.040	0.042	0.003	0.247	-0.795*	0.070
lc_formir	0.185	-0.008	-0.088	0.048	-0.272	0.031	0.583*

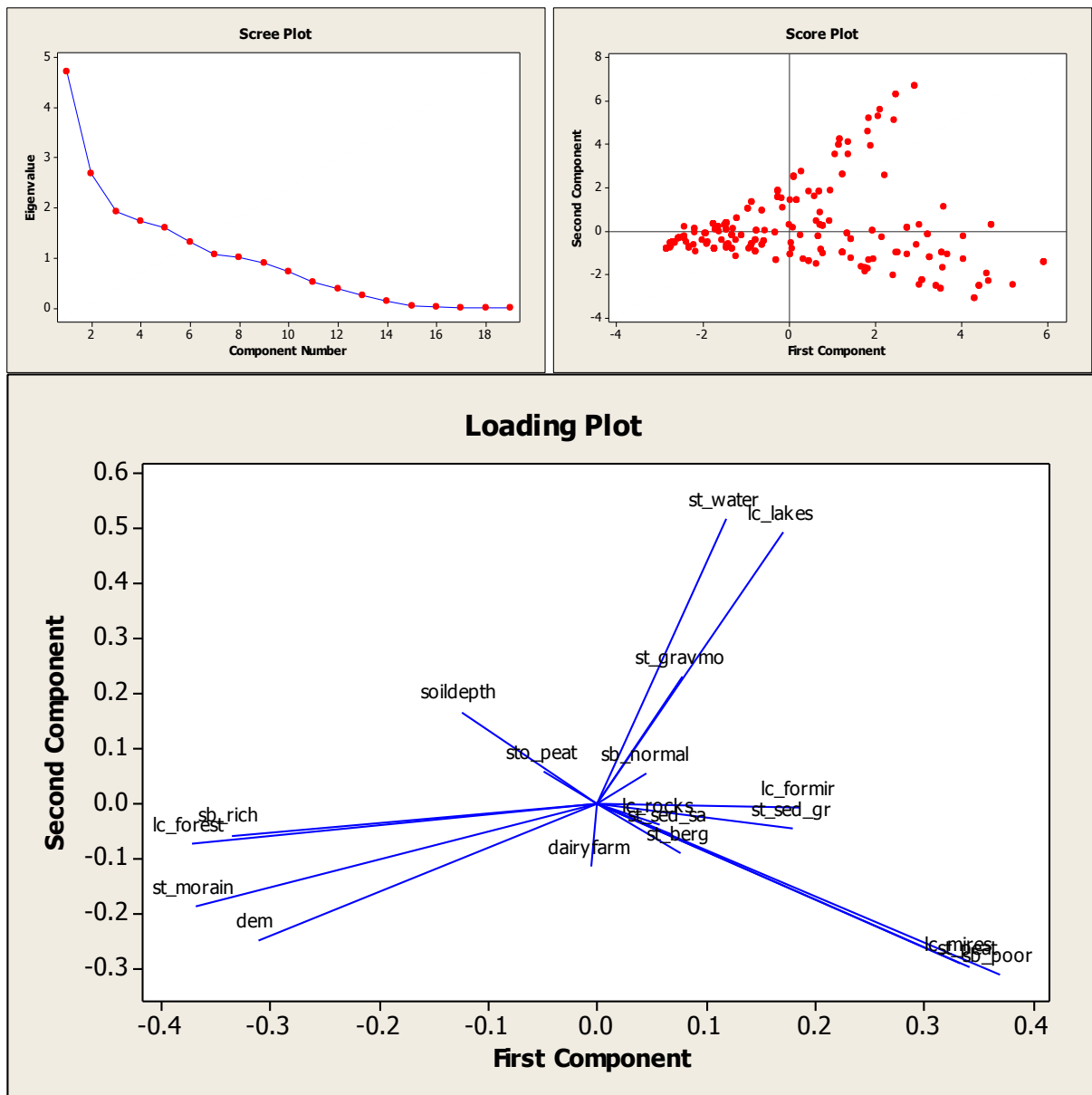


Figure 4.5: Results of the PCA with the 250m grid. All variables were included in the PCA.

4.3 Stepwise GWLR

To find the set of variables that can be used to predict fire location and size, 7 stepwise GWLR model sequences were used. Firstly, two grids were used (100m and 250m). Secondly, several sets of variables were used: (1) All 19 variables for both grids, (2) 13 variables for the 100m grid, (3) 10 variables for the 100m grid, 4) 12 variables for the 250m grid, and (5) all principal components. For all analyses, a bandwidth of 275 was used, as this was the bandwidth calculated as optimal with the bandwidth calculation command that minimized the AICc values in GWmodel (Gollini et al. 2015).

The results show that the best “set of variables” is consistently just one variable, as the AICc value increases with each added variable (Table 4.7 – 4.10). These variables change between different stepwise regressions.

Table 4.7: Results of the stepwise regression with all of the 19 variables (see Table 3.3 for a short description of the variables).

100m, all (19) variables, bw=275			250m, all (19) variables, bw=275		
Vars	AICc	Adj. R ²	Vars	AICc	Adj. R ²
Sb_normal	33.29	0.078	Sb_normal	32.11	0.098
Dairyfarm	36.13	0.118	Dairyfarm	34.48	0.138
Sto_peat	38.93	0.191	Soildepth	37.14	0.146
Soildepth	42.74	0.194	Sto_peat	39.47	0.181
Lc_formir	45.54	0.208	St_gravmo	42.59	0.180

Table 4.8: Results of the stepwise regression with selected variables with the 100m grid (see Table 3.3 for which variables were included and a short description of each variable).

13 Variables, 100m (bw=275)			10 Variables, 100m (bw=275)		
Vars	AICc	Adj. R ²	Vars	AICc	Adj. R ²
Dairyfarm	34.76	-0.001	Dairyfarm	34.76	-0.001
Sb_rich	37.36	0.066	Sb_rich	37.36	0.066
Lc_forest	40.38	0.071	St_peat	41.12	0.066
Sto_peat	42.99	0.159	Sto_peat	43.84	0.169
St_gravmo	45.48	0.257	Soildepth	47.12	0.229

Table 4.9: Results of the stepwise regression with 12 variables and the 250m grid (see Table 3.3 for which variables were included and a short description of each variable)

12 Variables, 250m (bw=275)		
Vars	AICc	Adj. R ²
Soildepth	34.49	-0.003
Sto_peat	37.72	0.015
Sb_rich	40.41	0.123
Lc_forest	43.26	0.144
St_morain	45.94	0.170

Table 4.10: Results of the stepwise regression with the seven principal components of the 19 variables for the 100m and 250m grid.

7 PCs (100m, bw=275)			7 PCs (250m, bw=275)		
Vars	AICc	Adj. R ²	Vars	AICc	Adj. R ²
PC4	33.65	0.063	PC4	33.77	0.041
PC6	37.05	0.066	PC5	35.45	0.068
PC3	40.39	0.092	PC7	37.93	0.066
PC5	43.81	0.095	PC3	41.58	0.072
PC1	47.18	0.096	PC1	44.90	0.088

4.4 Method Comparison

The fire size was compared between the different methods used: Drawing an outline, based on a grid, and Thiessen polygons (Figure 4.6, Appendix D for all 12 maps). In the absence of usable results from the GWLR, an INLA model was run in R (script provided by Johan Lindström, Centre for Mathematical Sciences, Lund University).

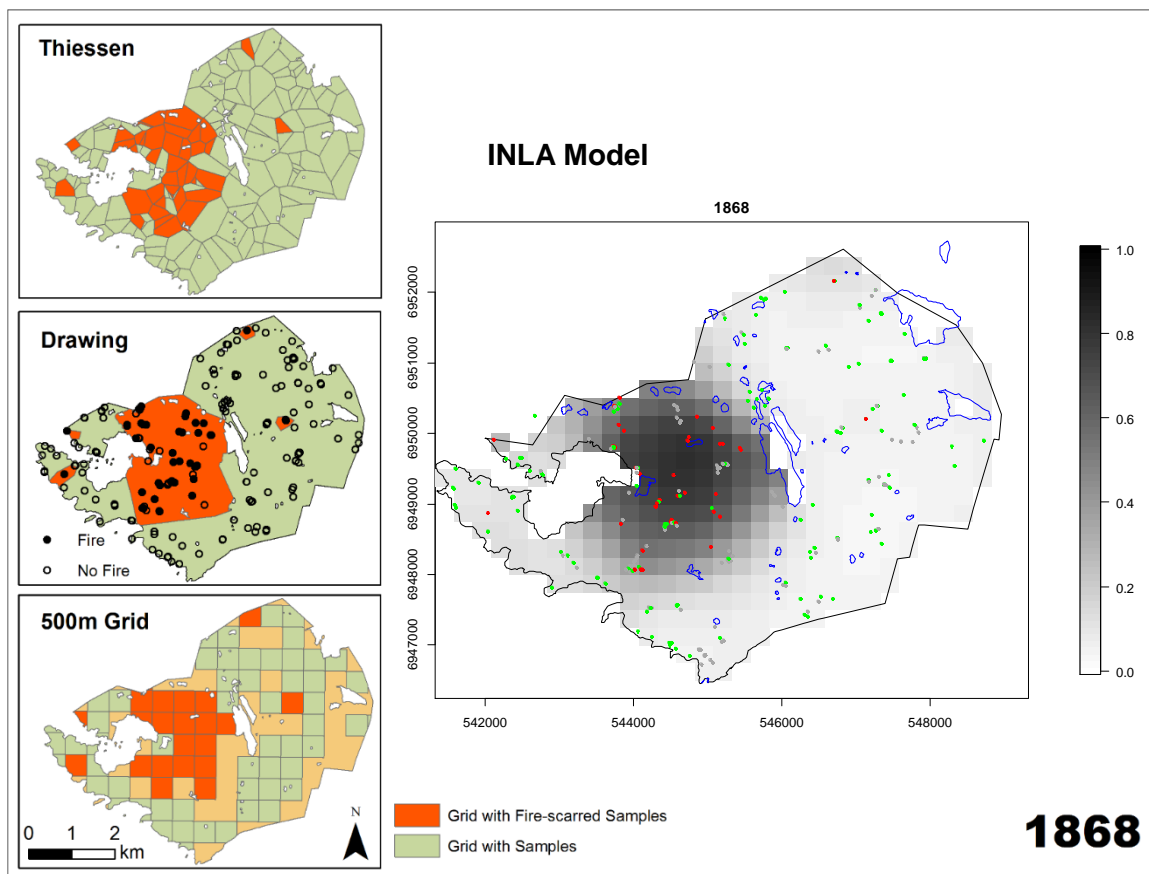


Figure 4.6: Method comparison between the different methods: Thiessen polygons, drawing an outline, grid method, and an INLA model.

The fire size was largest with the drawing-outline method in 7 out of 12 fire years (Figure 4.7). Overall, most area was shown as burned using the outline method; however, this number

is influenced by the large fires. In 5 out of 12 fire years, more area was burned according to the grid method; however, this was mainly during the smaller fire years in the latter part of the chronology (Table 4.11).



Figure 4.7: Fire size comparison as based on the grid method, drawing-outline method, and Thiessen polygons.

Table 4.11: Fire sizes for the three different methods including the sum and average.

years	Sample Depth	Grid	Drawing Outlines	Thiessen	Average (range)
1521	64	1673	1995	1410	1693 (+ 18%, - 17%)
1576	69	1562	1352	1505	1473 (+ 6%, - 8%)
1644	116	355	437	307	367 (+ 19%, - 16%)
1677	126	1482	2018	1126	1542 (+ 31%, - 27%)
1723	163	162	241	119	174 (+ 38%, - 31%)
1729	167	903	1000	625	843 (+ 19%, - 26%)
1770	204	451	374	297	374 (+ 21%, - 21%)
1789	206	446	242	227	305 (+ 46%, - 26%)
1807	207	340	412	216	323 (+ 28%, - 33%)
1831	210	809	947	504	753 (+ 26%, - 33%)
1844	192	330	168	173	224 (+ 47%, - 27%)
1868	175	685	615	443	581 (+ 18%, - 24%)
sum		9198	9802	6954	8651 (+ 13%, - 20%)
average		767	817	579	721 (+ 13%, - 20%)

4.5 Probability of Burning

The probability that a pair of trees burned at the same time as based on their distance from each other seems to follow Tobler's law (Tobler 1970) (Figure 4.8 and 4.9). Trees that are close or far from each other have, in general, higher or lower probabilities for burning at the same time, respectively. This means that the distance of trees to each other does matter when it comes to computing the probability. It also shows that high average probabilities are only reached at rather low distances, while low probabilities of co-occurrence are reached over the whole range.

A probability of one means in this context that the tree had an overlapping period, and that all fires within that period occurred in the same year. A probability of zero means then that in an overlapping period there were fires that effected either one or both of the compared trees that did not occur in the same years. The situations where a division error would occur (no overlapping period or no fires in an overlapping period) were not considered in these graphs (Figures 4.8 to 4.13).

Box plots summarize the probability data for different time periods: 1400-2000 (Figure 4.9), 1400-1700 (Figure 4.11), and 1700-2000 (Figure 4.13). The sample size of the probability of burning in the 1400-1700 period is much smaller ($N=7994$) than in the 1700-2000 period ($N=29563$). However, when taking out all zero probabilities, the sample size is about the same ($N_{1400-1700}=4282$, $N_{1700-2000}=4078$).

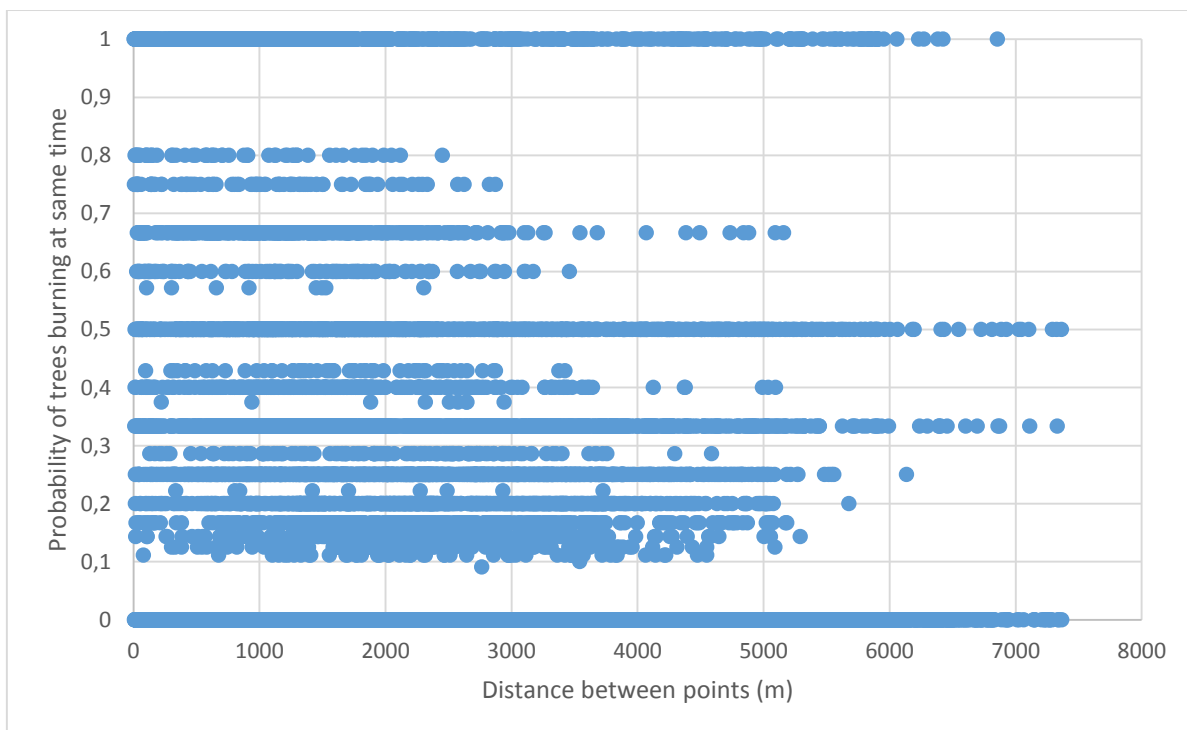


Figure 4.8: The probabilities of a pair of trees burning at the same time as based on the distance to each other (for the whole time period 1400-2000).

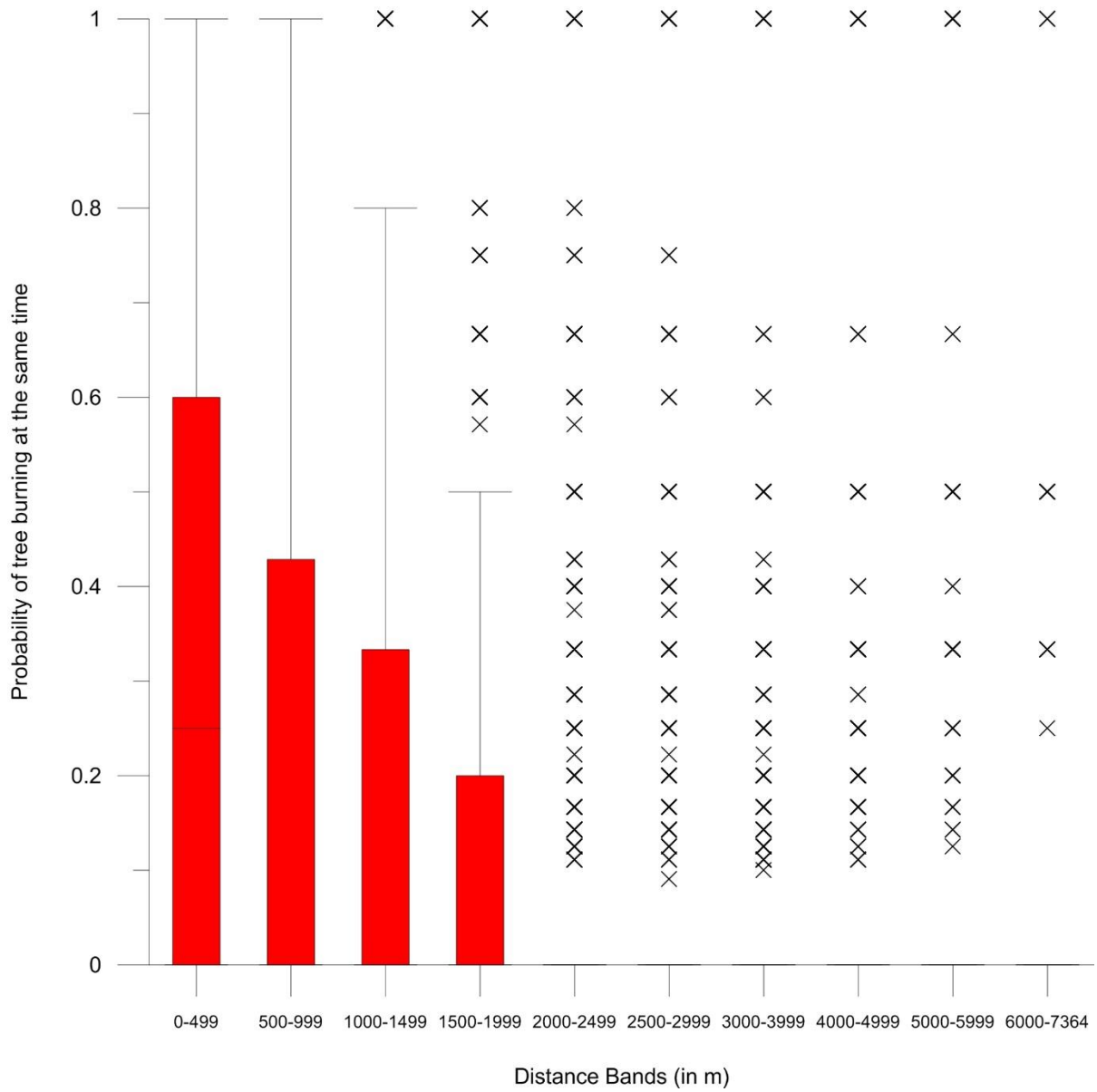


Figure 4.9: Boxplots per distance band of the data in Figure 4.8 (probabilities of a pair of trees burning at the same time for the whole time period 1400-2000).

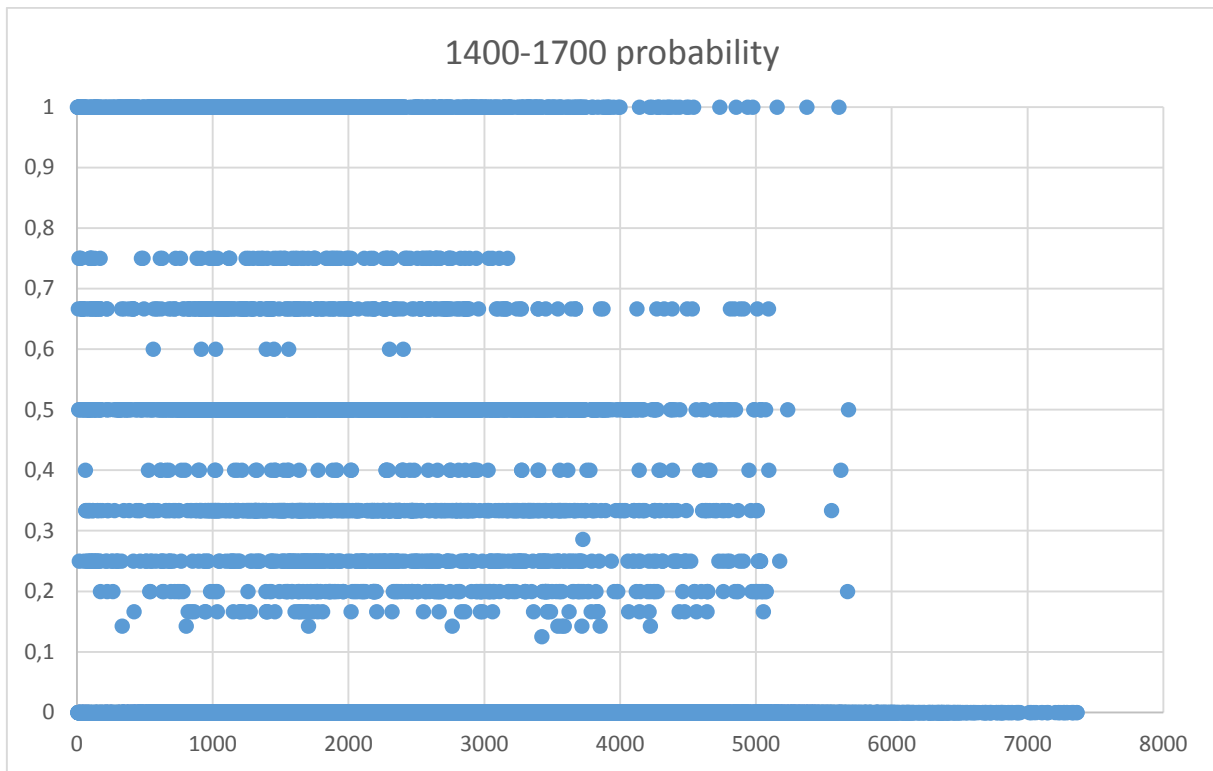


Figure 4.10: The probabilities of a pair of trees burning at the same time as based on the distance to each other (for the time period 1400-1700).

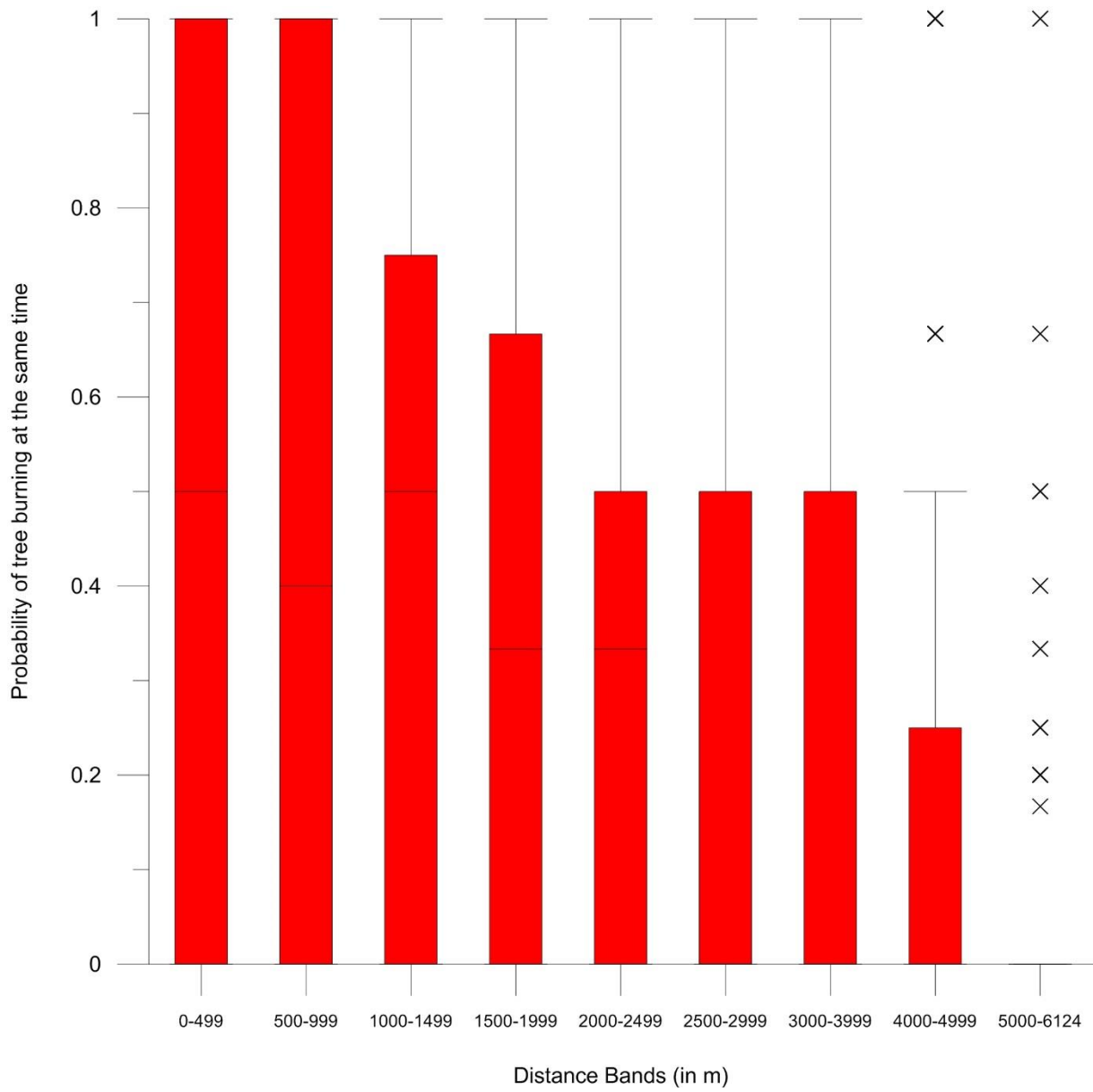


Figure 4.11: Boxplots per distance band of the data in Figure 4.10 (probabilities of a pair of trees burning at the same time for the time period 1400-1700).

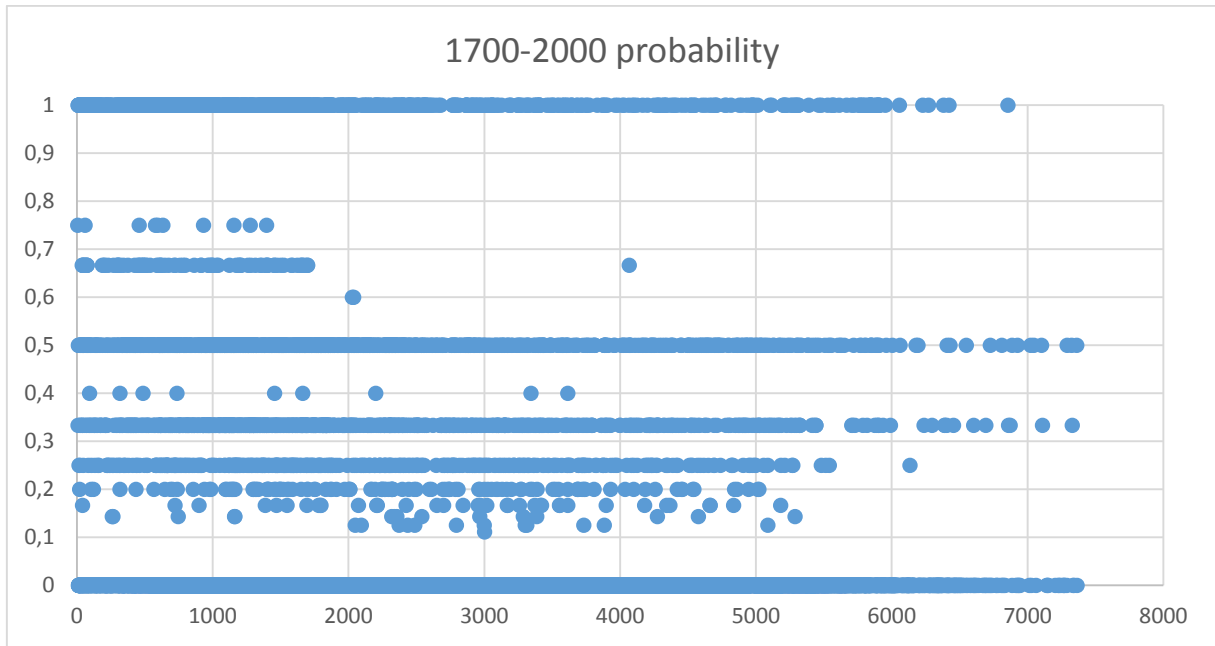


Figure 4.12: The probabilities of a pair of trees burning at the same time as based on the distance to each other (for the time period 1700-2000).

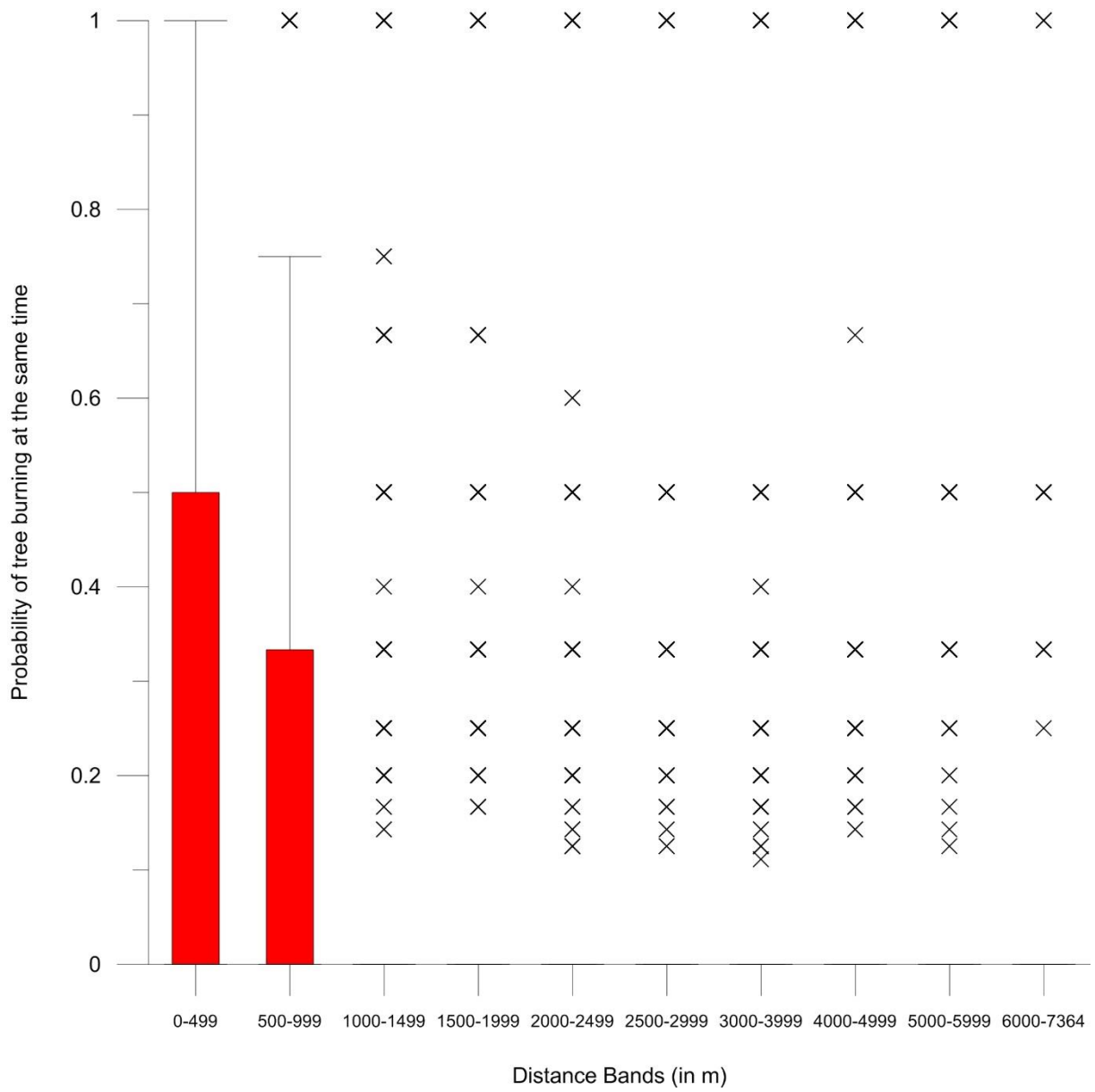


Figure 4.13: Boxplots per distance band of the data in Figure 4.12 (probabilities of a pair of trees burning at the same time for the whole time period 1700-2000).

5 Discussion

5.1 Stepwise GWLR

As Jetz et al. (2005) pointed out, variable selection is critical for both local and global methods. The initial selection of variables to include in this study was difficult. Some variables were not included, because they were to 100% derivatives of another variable (e.g. aspect and slope as derived from a DEM), others (e.g. wetness indices) were not available and replaced by approximations (land cover classes separated between forest, mires, and forest on mires), and some were included even though there were issues. For example, variables of human influence were not available except for two locations of early dairy farms. While this layer is more relevant in some centuries (approximately 1600s to 1800s) than in others, GWLR is supposed to extract the variables' influence as based on the fire data.

Variables that then passed the multicollinearity analysis (the VIF test) were input to the stepwise GWLR analysis. However, in the stepwise analysis, some of them could not be used as (almost) no sample was situated within a grid cell with a value (> than 0%) for that variable, e.g. `st_sed_gr`, `st_sed_sa`, or `lc_rocks`. In such cases, where variables were added to the model, no R^2 value could be calculated and the variable was dropped.

All AICc values for the first variables for every stepwise regression are so close to each other (min = 32.11 and max = 34.76), i.e. less than 3 (Fotheringham et al. 2002), that the models are to be seen as equal in quality and could be exchanged with each other, i.e. each variable is equally useful for predicting fire location/size. In addition, when adding new variables to the model, the AICc values increase, which also indicates that the model does not improve sufficiently with an increasing number of variables to offset the penalty associated with potential overfitting.

The adjusted R^2 values are all below 0.26. The best model describes, therefore, only 26% of its variance with five independent variables (Table 4.9). This larger R^2 value is, however, from an analysis that still has some multicollinearity between the variables. Additionally, a model that only explains 26% of its variance is according to ESRI (2016) not a passing model for what they call a “strong adjusted R^2 value” in their six checks (see section 2.3.6.1). This indicates that fire size cannot be estimated based on the variables used in this analysis.

While the AICc values increased with more and more variables added (a lower AICc value means a higher quality model), the R^2 values increased as well (which means more variables explain more of the variance). We only focused on the first 5 variables that were added in a stepwise GWLR. While models become more difficult to explain with more and more variables, adding more variables to the GWLR model might help improve the variance explained (although the AICc value would surely increase as well).

Even the stepwise GWLR of the PCA does not show results that allow for further interpretation. PC4 is the first component to be chosen in the stepwise GWLR in both grids. However, I have interpreted PC4 in the 100m grid as “Close to dairy farms, rich in larger boulders and/or bedrock on surface (`st_berg`)” and in the 250m grid as “high amounts of larger boulders; no surface peat (`sto_peat`); on deeper soils”, because they had different

loadings. So, as with the stepwise GWLR with individual variables, the stepwise GWLR of the PCA seem to show problems with multiple hypothesis testing (see section 2.3.6.3) as the analyses do not produce in consistent results.

5.2 Method Comparison

The three methods compared show rather different fire sizes in one year. The uncertainty in one year can be as high as 46% (year 1789). It seems that the three fires in the earlier period (before 1650) have a smaller range of uncertainty than the nine fires after 1650. However, it is difficult to say something with certainty with this small sample size.

5.3 Fire Size

5.3.1 Over- and Underestimate of Fire Size

The different methods depend on sample size to a large degree. The main problem with such a fire size analysis is the lower sample size in earlier parts of the chronology. In 1521, 64 samples were used in the analysis, which could, however, be considered an adequate number of samples. The average Thiessen-polygon size in 1521 was 36.5 ha. In 1868, 176 samples were used and the average Thiessen-polygon size was 13.3 ha.

In the fire-reconstruction community, several aspects have been discussed when reconstructing fire. One subject that has gotten much attention is the fact that most researchers do target scarred trees in the landscape which could result in a misrepresentation of the fire history. This hypothesis could, however, be rejected based on a comparative study by Van Horne and Fulé (2006). In their study, Van Horne and Fulé (2006) compared different sampling techniques (targeted, grid-based, and random sampling) using a census data set. Their results indicate that the sampling approaches were “essentially indistinguishable from census data” when ≥ 50 samples were used (Van Horne and Fulé 2006, pg. 866). Also, as researchers are interested in reconstructing a fire history, a systematic search for samples is quite inefficient and unreasonable when the efficient expenditure of resources is considered (Swetnam and Baisan 1996).

In addition, if we have a fire scarred tree in an area, that area has clearly burned even if we find other trees in the same area that have no scar. A tree that has no scar is not evidence that it has not experienced fire, because not all trees get scarred during a fire (Piha et al. 2013). Hence, while there is very little chance of a commission error (all fire scars are caused by fires) there is a quite high omission error (not all fires leave a scar). One could reason from the fact that not every tree scars during a fire (Piha et al. 2013), that targeting both scarred and unscarred trees to the same degree might overestimate the area that has not burned. In fact only fires with a medium intensity leave detectable fire scars. Hot fires might burn the tree and lead to the absence of a sampling point, while too cold fires might not leave a scar.

The changing severity of a single fire on the landscape and with changing weather could, on the other hand, lead to overestimating fire sizes because of its patchiness, as discussed with respect to sustainable management and harvests in Bergeron et al. (2002). This patchiness leads in many methods to an overestimation of the fire size, as not every square meter burns within the fire perimeter (Bergeron et al. 2002).

5.3.2 Advantages and Disadvantages of the methods

Since the true fire size for each year is not available, an optimal method cannot be chosen based on ground truth. However, the advantages and disadvantages of the different methods will be discussed.

5.3.2.1 Grid Method

Although the results produced by the grid method depend somewhat on the size of the grid, Hjalmarsson et al. (2015) found in preliminary analyses that the difference between the two different grid sizes used might only play a minor role. The final fire size is a percentage of study area burned as based on the ratio between cells with fire scars and all cells with a sample (in a particular year). Therefore, even if only half of the study area is covered by cells, the ratio is extended to all of the study area. Because the grid size does not seem to matter, this method could be considered rather objective.

5.3.2.2 Drawing outlines

Drawing outlines has been applied to almost every spatial fire history study. In earlier studies, this method had been applied based on expert knowledge and arbitrarily set rules (Linder 1988; Niklasson and Granström 2000). The advantages of this method are that the fire does not end at arbitrary lines in the landscape. The fire will follow the contours and features of the study area. However, how much the fire expands beyond or stops before such an arbitrary line is based on rather subjectively set rules which are difficult to consistently apply in each and every case. In addition, a fire break, e.g. a river, might be a fire break in one year, but not in another (very dry) year.

This method is today used after applying more objective methods first, i.e. buffer zones around the fire scarred samples, as in Storaunet et al. (2013), and a replication of the methods are better possible. Given that my GWLR analysis found only a low correlation of fire extent to the environmental conditions of the area, I would not suggest using this method on its own.

5.3.2.3 GWLR

GWLR was another method that I thought could possibly be used to calculate fire sizes. With a set of variables used to explain fire occurrences in a landscape, you would be able to see where all these variables converge and the fire occurred. However, in our analyses, the variables included in the different calculations (Table 3.3) do not show a pattern that could help estimate the fire size. Therefore, based on my GWLR analyses, the scientific basis for the outline method is not in evidence.

5.3.2.4 Thiessen Polygons

The advantages of calculating Thiessen polygons to get the fire size are its objectivity and replicability, its ease of use, and relative inexpensiveness of computation in GIS programs like ArcGIS. While all methods are to some degree affected by the sample size, this method seems to have the most problems related to low sample sizes. The average Thiessen-polygon size can differ considerably from one fire year to the next, and the area that is covered by the polygons is directly affected by the density of points. In an extreme example, the fire year 1311, which has one sample (which happened to record a fire), is represented by only one polygon, and the whole study area burned.

5.4 Probability of Burning

Trees closer to each other have a higher probability of burning compared to trees that are further apart. This is especially true in the first half of the time period. In the second half of the time period, this clear pattern is partially vanishing. Therefore, the box plots that summarize the probability data for each time periods (Figures 4.9, 4.11, 4.13) look strikingly different, which could be explained by a couple of reasons: 1) Different sample sizes between these two time periods, or 2) a fire-regime shift between the two time periods as Hjalmarsson et al. (2015) suggest are potential explanations.

At first, when including all probability values, the sample sizes between the two time periods differ significantly, which could explain the differences between the data. However, after taking out all zero probability values, the samples size in the earlier and later time period are nearly the same, which indicates to me that, in fact, a shift in fire-regime occurred between these two time periods. In the earlier period, larger and less frequent fires occurred, whereas in the latter period, smaller and more frequent fires occurred. Therefore, many more zero probabilities of trees burning at the same time exist in the latter period.

This shift from larger, less frequent fires to smaller, more frequent fires can also be seen in the fire history, as partially seen in Figure 4.7 (Hjalmarsson et al. 2015). In other studies, such a shift has been attributed to the increased density of humans in the area (Niklasson and Granström 2000; Storaunet et al. 2013). In the Jämtgaveln and surrounding area, humans have been present for a long time (Johnson 2008). First anthropogenic findings date to 880 AD (Johnson 2008). Orelund (2015) and Johnson (2008) thoroughly described the human impact around the Jämtgaveln area, detailing a “main” road passing through (as early as 1000 A.D.), armed conflicts, and slash and burn practices (in the middle of the 17th century). The human population had rather low densities in Sweden around 1350 when the first wave of Black Death raged through Sweden (Orelund 2015): Population size in the early 1400s was about 347,000 inhabitants, just 1/3 of the population present at the beginning of 1300s. The population recovered slowly, and once again reached its pre-Black Death levels around 1650 (Orelund 2015).

The population density in the Jämtgaveln Nature Reserve is not known for any time period. However, Orelund (2015, pg. 13) argued that the time between 1400-1600 “is as close as we can get to natural conditions when it comes to fire regimes with a minimum of anthropogenic impacts”. Therefore, the 1600s are characterized by increasing densities of settlements and huts that used fire in this area to clear the forest of its understory or for slash-and-burn practices (Orelund 2015).

In the first part of the chronology, it was more likely to have higher probabilities of trees burning at the same time, even at larger distances between trees, because larger fires have burned at the same time. In the latter part of the chronology, humans likely influenced (or set fires) so that more frequent, but smaller, fires occurred. This could result in either overestimation of the high probability in distant trees, since the trees would not have burned at the same time under natural conditions; or in underestimation of the high probability in close trees, as fires were contained.

Depending on the size of the tree and thickness of the bark, a tree might or might not scar. No scar is left when the area has not burned, but also when the tree had too thick of a bark that the fire did not reach the cambium (or the fire burned too hot and killed the tree). Therefore, even if a tree without a scar is next to a tree with a scar, the probability that the tree without a scar has burned could be high. Therefore, dependent on fire severity, we might underestimate the probability of burning in nearby trees. In other words, the pattern in the early period could be much clearer if all trees that experienced fires would have left evidence of burning.

6 Conclusion

Historical fire sizes for the 12 largest fires have been reconstructed for Jämtgaveln Nature Reserve. Fire sizes could be estimated using three methods; Thiessen polygons, a grid method, and drawing the outlines. Geographically Weighted Logistic Regression could not be used to estimate fire sizes, since a set of variables to model fire occurrences as based on landscape variables was not found in this study. The assumption that landscape features influence fire behavior has been made explicitly when drawing an outline around fire-scarred samples. However, in this analysis in the Jämtgaveln nature reserve, these assumptions seems incorrect, and the probability of fire depends more on distance from a fire-scarred tree (at least in the earlier record).

Even though the adjusted R^2 values increased with added variables, the AICc values also increased, indicating a decrease in the quality of the models. Future work needs to include logistic regression/OLS analyses to see whether all variables together can better estimate fire size. In this analysis, only five variables were considered to calculate the R^2 and AICc values and more variables might be necessary.

The variable selection process was challenging as some variables were not available and others could not be included (due to multicollinearity issues). Future work could include a geographically weighted PCA to see if variables that do not have multicollinearity issues might produce better results.

In this study, only locally changing variables that did not change temporally in our analyses were considered. In future work, even global variables might be included as, for example, a climate reconstructions that is derived from another chronology than the one that includes the fire scars. With this climate reconstruction, wet/dry years could be distinguished and included as a global variable that changes throughout time.

Because we do not have the actual fire size, the correctness of the fire-size estimation methods cannot be established. However, comparing those three methods, it seems that a pattern emerges. Before 1650, fire sizes had less spread between the methods compared to after 1650. The period around 1650 is, incidentally, a time where the density of the human population increased in both Sweden and the Jämtgaveln area. Future work needs to check whether this pattern holds even when more fires are considered.

7 References

- Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. In *Proceedings of the Second International symposium on information theory*, ed. B. Petrov and F. Csaki, 267–281. Tsahkadsor, Armenian SSR: Budapest: Akademiai Kiado.
- Akaike, H. 1974. A new look at the statistical model identification. *IEEE Transactionson Automatic Control* 19: 716–723. doi:10.1109/TAC.1974.1100705.
- Arlot, S., and A. Celisse. 2010. A survey of cross-validation procedures for model selection. *Statistics Surveys* 4: 40–79. doi:10.1214/09-SS054.
- Atkinson, P. M., S. E. German, D. A. Sear, and M. J. Clark. 2003. Exploring the relations between riverbank erosion and geomorphological controls using geographically weighted logistic regression. *Geographical Analysis* 35: 58–82.
- Bergeron, Y., A. Leduc, B. D. Harvey, and S. Gauthier. 2002. Natural fire regime: A guide for sustainable management of the Canadian boreal forest. *Silva Fennica* 36: 81–95. doi:10.14214/sf.553.
- BioMedware. 2016. SpaceStat.
- Bivand, R. 2012. notoriously unreliable technique. <http://r-sig-geo.2731867.n2.nabble.com/A-question-about-gwr-morantest-pvalue-td7292670.html>.
- Breyer, B. 2013. Exploring local variability in statistical relationships with geographically weighted regression.
- Brunsdon, C. 2011. Geographically Weighted Analysis - Review and Prospect.
- Byrne, G., M. Charlton, and S. Fotheringham. 2009. Multiple Dependent Hypothesis Tests in Geographically Weighted Regression: 5.
- Charlton, M. 2011. Geographically weighted regression: Modeling spatial heterogeneity (Lecture at Hamilton Institute).
- Charlton, M., and A. S. Fotheringham. 2009. *Geographically Weighted Regression. White Paper*. Maynooth, Co Kildare, IRELAND: National University of Ireland Maynooth.
- Cho, S., D. M. Lambert, S. G. Kim, and S. Jung. 2009. Extreme Coefficients in Geographically Weighted Regression and Their Effects on Mapping. *GIScience & Remote Sensing* 46: 273–288.
- Dale, M. R. T., and M.-J. Fortin. 2002. Spatial autocorrelation and statistical tests in ecology. *Écoscience* 9: 162–167.
- ESRI. 2014. ArcGIS.
- ESRI. 2016. Beyond Where: Using Regression Analysis to Explore Why. <http://training.esri.com/gateway/index.cfm?fa=catalog.webCourseDetail&courseid=2586> . Accessed February 29.
- European, C. 2016. Natura 2000.
- Farber, S., and A. Páez. 2007. A systematic investigation of cross-validation in GWR model estimation: empirical analysis and Monte Carlo simulations. *Journal of Geographical Systems* 9: 371–396. doi:10.1007/s10109-007-0051-3.

- Fernandes, P. M., C. Loureiro, N. Guiomar, G. B. Pezzatti, F. T. Manso, and L. Lopes. 2014. The dynamics and drivers of fuel and fire in the Portuguese public forest. *Journal of Environmental Management* 146: 373–382.
- Foody, G. M. 2004. Spatial nonstationarity and scaledependency in the relationship between species richness and environmental determinants for the sub-Saharan endemic avifauna. *Global Ecology and Biogeography* 13: 315–320.
- Fotheringham, A. S., B. Chris, and M. Charlton. 2002. *Geographically Weighted Regression - The analysis of spatially varying relationships*. West Sussex, England: Wiley.
- Gollini, I., B. Lu, M. Charlton, C. Brunsdon, and P. Harris. 2015. GWmodel: An R Package for Exploring Spatial Heterogeneity Using Geographically Weighted Models. *Journal of Statistical Software* 63: 1–50.
- Goudie, A. S. 1986. The Integration of Human and Physical Geography. *Transactions of the Institute of British Geographers* 11, 545–458.
- Griffith, D. A. 2008. Spatial-filtering-based contributions to a critique of geographically weighted regression (GWR). *Environment and Planning A* 40: 2751–2769.
- Heyerdahl, E. K. 1997. Spatial and temporal variation in historical fire regimes of the Blue Mountains, Oregon and Washington: the influence of climate. University of Washington.
- Hjalmarsson, J., B. G. Jonsson, and P. Linder. 2015. A 3000 ha spatial explicit fire history illustrating the mosaic structure of boreal landscapes in Northern Europe. In *27th International Congress of Conservation Biology/4th European Congress of Conservation Biology*. Montpellier, France.
- Hjalmarsson, J., B. G. Jonsson, and P. Linder. A 3000 ha spatially explicit fire history illustrating the mosaic structure of boreal landscapes in Northern Europe.
- Van Horne, M. L., and P. Z. Fulé. 2006. Comparing methods of reconstructing fire history using fire scars in a southwestern United States ponderosa pine forest. *Canadian Journal of Forest Research* 36: 855–867. doi:10.1139/x05-289.
- Jetz, W., C. Rahbek, and J. W. Lichstein. 2005. Local and global approaches to spatial data analysis in ecology. *Global Ecology and Biogeography* 14: 97–98.
- Johnson, I. 2008. Torringen och Jämtgaveln: En Historisk Studie.
- KNMI, (Koninklijk Nederlands Meteorologisch Instituut). 2015. KNMI Climate Explorer.
- Koutsias, N., J. Martínez-Fernández, and B. Allgöwer. 2010. Do Factors Causing Wildfires Vary in Space? Evidence from Geographically Weighted Regression. *GIScience & Remote Sensing* 47: 221–240. doi:10.2747/1548-1603.47.2.221.
- Linder, P. 1988. *Jämtgaveln : en studie av brandhistorik, kulturpåverkan och urskogsvärden i ett mellannorrländskt skogsområde*. Härnösand: Länsstyr. i Västernorrlands län.
- Lindgren, F., and H. Rue. 2015. Bayesian Spatial Modelling with R-INLA. *Journal of Statistical Software* 63: 1–25.
- Lindgren, F., H. Rue, and J. Lindström. 2011. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73: 423–498.

- Länsstyrelsen, V. 2007. Jämtgaveln – brandpåverkad naturskog.
- Martínez, J., C. Vega-García, and E. Chuvieco. 2009. Human-caused wildfire risk rating for prevention planning in Spain. *Journal of Environmental Management* 90: 1241–1252. doi:<http://dx.doi.org/10.1016/j.jenvman.2008.07.005>.
- Martínez-Fernández, J., N. Koutsias, and E. Union. 2011. *Modelling Fire Occurrence Factors In Spain. National Trends And Local Variations*. Edited by Jesus San-Miguel Ayanz, Ioannis Gitas, Andrea Camia, and Sandra Oliveira. *Advances in Remote Sensing and GIS Applications in Forest Fire Management - From local to global assessments*. Luxembourg: European Commission, Joint Research Centre, Institute for Environment and Sustainability. doi:10.2788/67097.
- Martínez-Fernández, J., E. Chuvieco, and N. Koutsias. 2013. Modelling long-term fire occurrence factors in Spain by accounting for local variations with geographically weighted regression. *Natural Hazards and Earth System Sciences* 13: 311–327. doi:10.5194/nhess-13-311-2013.
- Massey, D. 1999. Space-Time, “Science” and the Relationship between Physical Geography and Human Geography. *Transactions of the Institute of British Geographers* 24: 261–276. <http://doi.org/10.1111/j.0020-2754.1999.00261.x>
- Matthews, S. A., and T. Yang. 2012. Mapping the results of local statistics: using geographically weighted regression. *Demographic Research* 26: 151–166.
- Megler, V., D. Banis, and H. Chang. 2014. Spatial analysis of graffiti in San Francisco. *Applied Geography* 54: 63–73.
- Minitab. 2007. No Title. Minitab Inc.
- Minitab. 2015. What is a variance inflation factor (VIF)?
- Nakaya, T. 2014. GWR.
- Niklasson, M., and A. Granström. 2000. Numbers and sizes of fires: long-term spatially explicit fire history in a Swedish boreal landscape. *Ecology* 81: 1484–1499.
- Ogneva-Himmelberger, Y., H. Pearsall, and R. Rashit. 2009. Concrete evidence & geographically weighted regression: A regional analysis of wealth and the land cover in Massachusetts. *Applied Geography* 29: 478–487. doi:10.1016/j.apgeog.2009.03.001.
- Oliveira, S., J. M. C. Pereira, J. San-Miguel-Ayanz, and L. Lourenço. 2014. Exploring the spatial patterns of fire density in Southern Europe using Geographically Weighted Regression. *Applied Geography* 51: 143–157.
- Orelund, J. 2015. Fire regimes and anthropogenic impacts in a boreal forest landscape. Mid Sweden University.
- Páez, A., and D. C. Wheeler. 2009. Geographically Weighted Regression. In *International Encyclopedia of Human Geography*, ed. R. Kitchin and N. Thrift, 1:407–414. Oxford: Elsevier.
- Páez, A., S. Farber, and D. C. Wheeler. 2011. A simulation-based study of geographically weighted regression as a method for investigating spatially varying relationships. *Environment and Planning A* 43: 2992–3010. doi:10.1068/a44111.

- Piha, A., T. Kuuluvainen, H. Lindberg, and I. Vanha-Majamaa. 2013. Can scar-based fire history reconstructions be biased? An experimental study in boreal Scots pine. *Canadian Journal of Forest Research* 43: 669–675. doi:10.1139/cjfr-2012-0471.
- Propastin, P. A., and M. Kappas. 2008. Reducing Uncertainty in Modeling the NDVI-Precipitation Relationship: A Comparative Study Using Global and Local Regression Techniques. *GIScience & Remote Sensing* 45: 47–67. doi:10.2747/1548-1603.45.1.47.
- Rodrigues, M., J. de la Riva, and S. Fotheringham. 2014. Modeling the spatial variation of the explanatory factors of human-caused wildfires in Spain using geographically weighted logistic regression. *Applied Geography* 48: 52–63.
- Rue, H., S. Martino, and N. Chopin. 2009. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 71: 319–392.
- Rue, Hå., and S. Martino. 2007. Approximate Bayesian inference for hierarchical Gaussian Markov random field models. *Journal of Statistical Planning and Inference* 137: 3177–3192. doi:10.1016/j.jspi.2006.07.016.
- Sá, A. C. L., J. M. C. Pereira, M. E. Charlton, B. Mota, P. M. Barbosa, and A. S. Fotheringham. 2011. The pyrogeography of sub-Saharan Africa: a study of the spatial non-stationarity of fire–environment relationships using GWR. *J Geogr Syst* 13: 227–248. doi:10.1007/s10109-010-0123-7.
- Storaunet, K. O., J. Rolstad, M. Toeneiet, and Y. Blanck. 2013. Strong anthropogenic signals in historic forest fire regime: a detailed spatiotemporal case study from south-central Norway. *Canadian Journal of Forest Research* 43: 836–845. doi:10.1139/cjfr-2012-0462.
- Swetnam, T. W., and C. H. Baisan. 1996. *Historical fire regime patterns in the southwestern United States since AD 1700*. Edited by Craig D Allen. *Fire Effects in Southwestern Forest : Proceedings of the 2nd La Mesa Fire Symposium*. USDA Forest Service, Rocky Mountain Research Station.
- The R Foundation for Statistical Computing. 2015. Wooden-Christmas-Tree.
- Tobler, W. 1970. A computer movie simulating urban growth in the Detroit region. *Economic Geography* 46: 234–240.
- Wheeler, D. C., and M. Tiefelsdorf. 2005. Multicollinearity and correlation among local regression coefficients in geographically weighted regression. *Journal of Geographical Systems* 7: 161–186. doi:10.1007/s10109-005-0155-6.
- Wright, K. 2015. corrram: Plot a correlogram. R package version 1.7.
- Wu, W., and L. Zhang. 2013. Comparison of spatial and non-spatial logistic regression models for modeling the occurrence of cloud cover in north-eastern Puerto Rico. *Applied Geography* 37: 52–62.
- Zhang, H., P. Qi, and G. Guo. 2014. Improvement of fire danger modelling with geographically weighted logistic model. *International Journal of Wildland Fire* 23: 1130–1156.
- Zhang, L., J. H. Gove, and L. S. Heath. 2005. Spatial residual analysis of six modeling techniques. *Ecological Modelling* 186: 154–177. doi:10.1016/j.ecolmodel.2005.01.007.

8 Appendix A

Instructions for Data Rasterization (example with 100m grid cells):

Landcover

- Create a fishnet (Data management – Feature Class – Create Fishnet)
 - o 100m, template extent = study area
- Tabulate Area (Spatial Analyst – Zonal – tabulate Area)
 - o Input zone data = 100m fishnet
 - o Zone field = FID
 - o Input raster or feature class data = e.g. landcover
 - o Class field = (if landcover) value
 - o Output table = enter a name
 - o Processing cell size 1 x 1m
 - Not the cell size of output table or so, but how big cells that gets averaged
- Join resulting table to 100m fishnet
- Export 100m fishnet (with tabulated-area table attached)
- Add fields to new shapefile (to attribute table)
 - o Double
 - o E.g. lc_forest, lc_formir, lc_mires, ...
- Calculate the percentage of each landcover type in the new fields
- Polygon to Raster (Conversion – to Raster)
 - o Cell assignment type = maximum_area; cellsize = 100
 - o Each field with % of landcover creates one raster (in my case 5 raster created: 100_lc_forest, 100_lc_formir, 100_lc_rocks, 100_lc_mires, 100_lc_lakes)

DEM

- Data management – Raster – Raster processing – Resample
 - o Output cell size 100m
 - o Resampling technique: bilinear

Soil Types

- o Soil layers to Raster (Conversion – to Raster – Polygon to Raster)
 - 1m cell size
 - Max_area
 - Processing extent (under Environment) = study area
- o Tabulate area (some above, but with class field, e.g. BL_TX)
- o Join with fishnet
 - Keep matching data
- o Export joined fishnet to shapefile
- o Add fields
- o Calculate % of each soil type
- o Polygon to raster

Soil depth

- Zonal Statistics (calculates the average value of raster in each polygon)
 - o Environmental Settings: cell size = 100m
 - o Input raster/zone feature = 100m fishnet
 - o Zone field = FID
 - o Input value raster = jorddjup...
 - o Output raster = insert new name
 - o Statistic type = mean
 - o Check "Ignore NoData in calculations"

Dairy Farms

- Euclidean Distance (Spatial Analyst – Distance)
 - o Environmental Settings
 - Snap to Raster = 100m_raster (it doesn't matter which one)
 - Cell size = 100m

9 Appendix B

Step-by-step instruction for calculating the probability plot (close trees burn at same time):

In the Point-Distance tool (Analysis → Proximity → Point Distance), the input feature was also the near feature) and I chose to only export the FID (feature ID). These FID numbers were then used to join the “fire data” to the distance table which had about 77000 possible combinations. These combinations included the distance between two points twice as it calculated the distance between 1 – 2 and 2 – 1. However, I used this fact to calculate the overlapping fires:

After importing the table to Excel, I sorted the distance column (lowest to highest) and added a column to filter the data so that only one distance between two points were included. (This was done by adding a blank, 1, blank, 2, and so on in the first, second, third, and fourth row, respectively, and filling the cells automatically with this pattern.) After copying this table to a new sheet, getting rid of the fire data in the new table, and filtering (Data → Filter) the data so that only the blank lines were visible (so only one distance per sample pair), I could then calculate (by adding each cell of the visible row to the next invisible one) a number that represented one of five option:

- (0)no overlap for of the recording sample,
- (1)one sample is recording,
- (2)both samples are recording or one sample has a fire scar and the other sample is not recording,
- (3)One sample with a fire scar with the other sample recording, and
- (4)both samples with a fire scar

In this case, I used 3 and 4 to calculate (with countif) how many fires there were that did (#4) and did not (#3) overlap. The probability was then calculated by the overlapping fires (#4) and total number of fires in overlapping period (#3 + #4). I replaced all division errors with a zero.

10 Appendix C: Correlograms

10.1 Correlogram (100m grid – all variables)

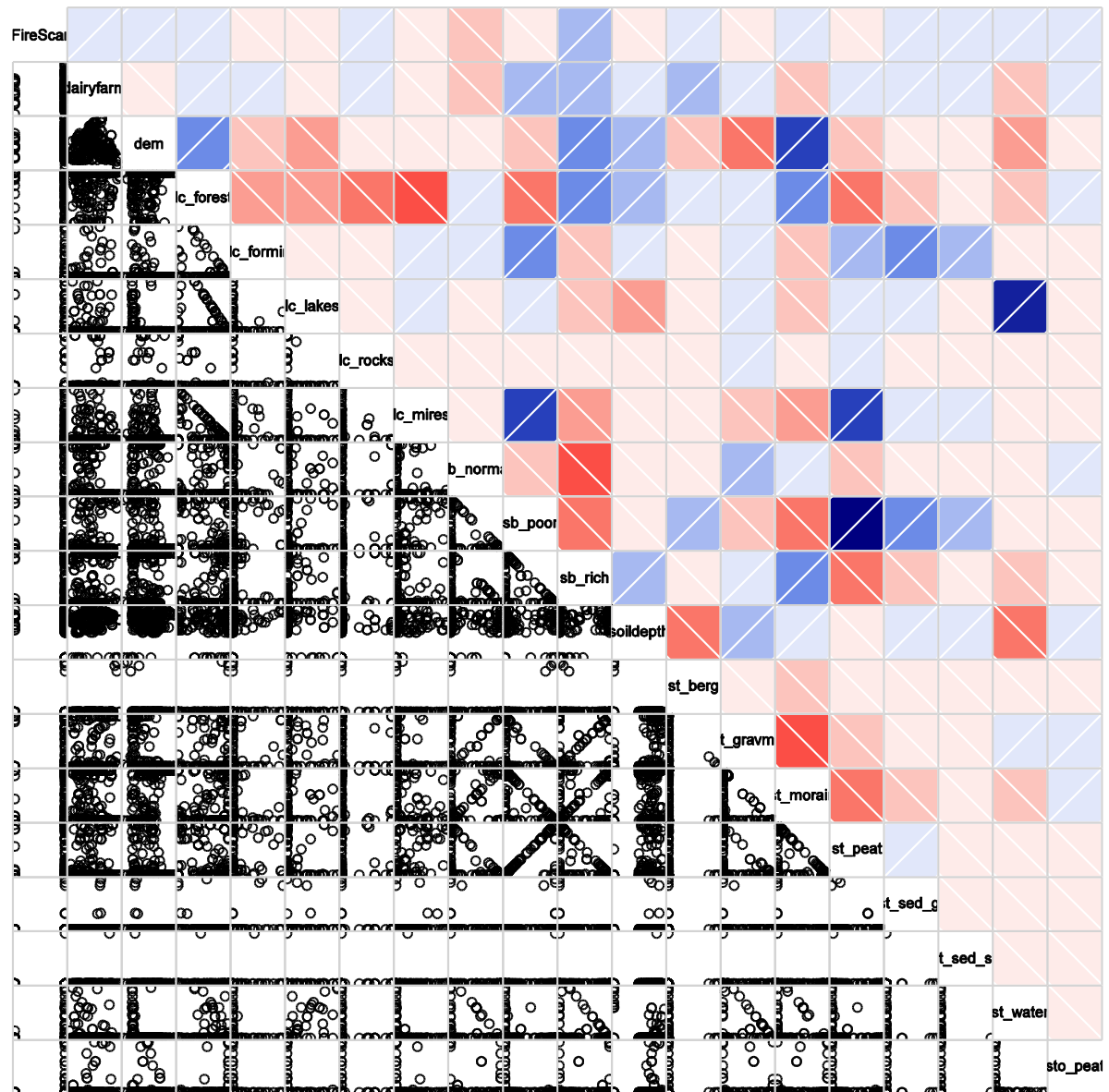


Figure 10.1: Correlogram (100m, all variables) showing the correlation values as based on the color scheme of Figure 4.1 (upper panel) and the point distribution between variables (lower panel).

```
#Set working directory
setwd("C:/temp/gwr_100grid/corrgram")

#load datafile in R
jamt.corr <- read.table("jamt100av.txt", header = T)

#Convert to data frame
jamt.corr <- as.data.frame(jamt.corr)

# corrgram with scatter plots
corrgram(jamt.corr, lower.panel = panel.pts)
```

10.2 Correlogram (100m grid - 13 variables)

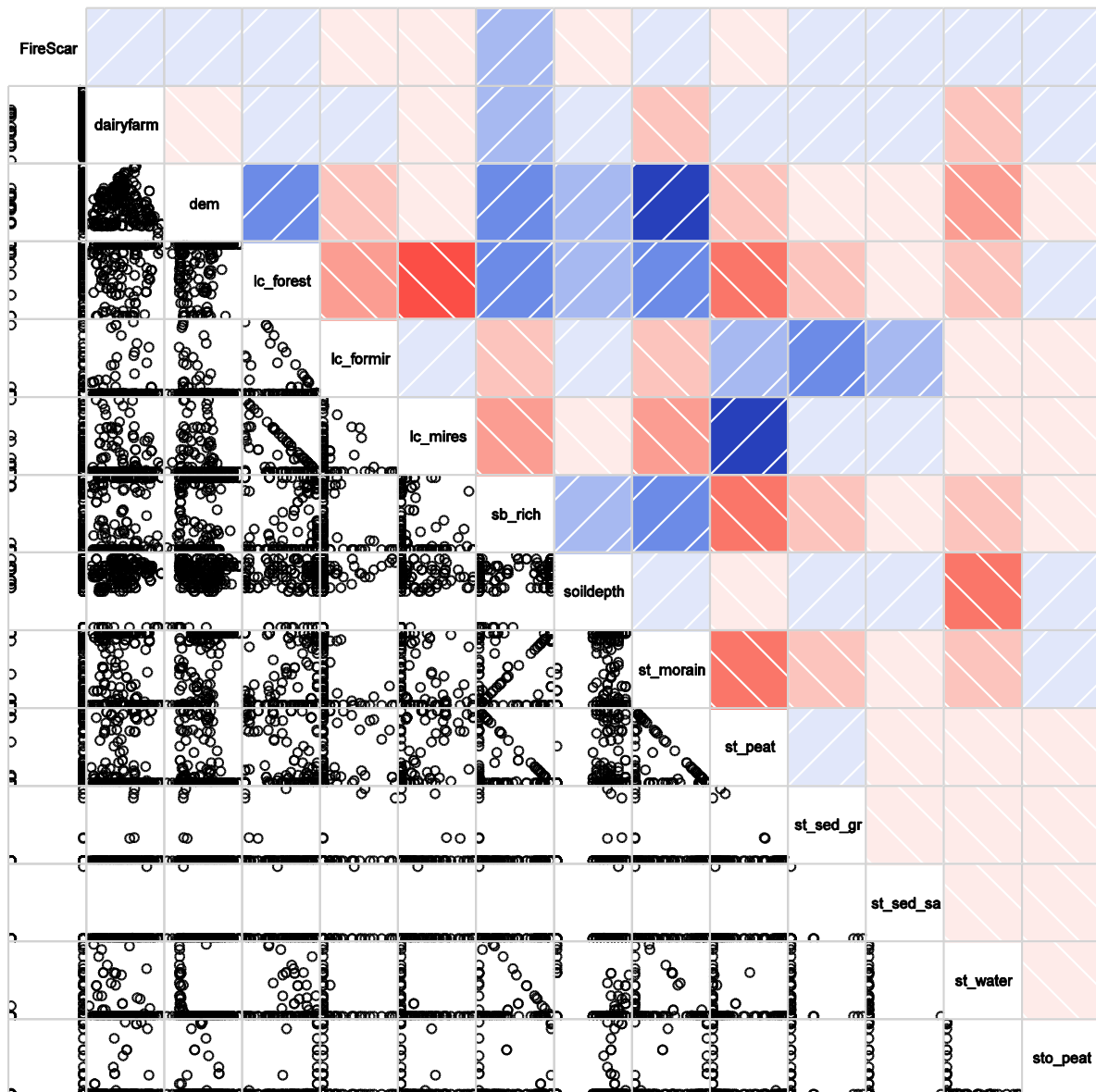


Figure 10.2: Correlogram (100m, 13 variables) showing the correlation values as based on the color scheme of Figure 4.1 (upper panel) and the point distribution between variables (lower panel).

```
#Set working directory
setwd("C:/temp/gwr_100grid_sel_var/corrgram")

#load datafile in R
jamt.corr <- read.table("jamt100_13v.txt", header = T)

#Convert to data frame
jamt.corr <- as.data.frame(jamt.corr)

# corrgram with scatter plots
corrgram(jamt.corr, lower.panel = panel.pts)
```

10.3 Correlogram (100m grid – 10 variables)

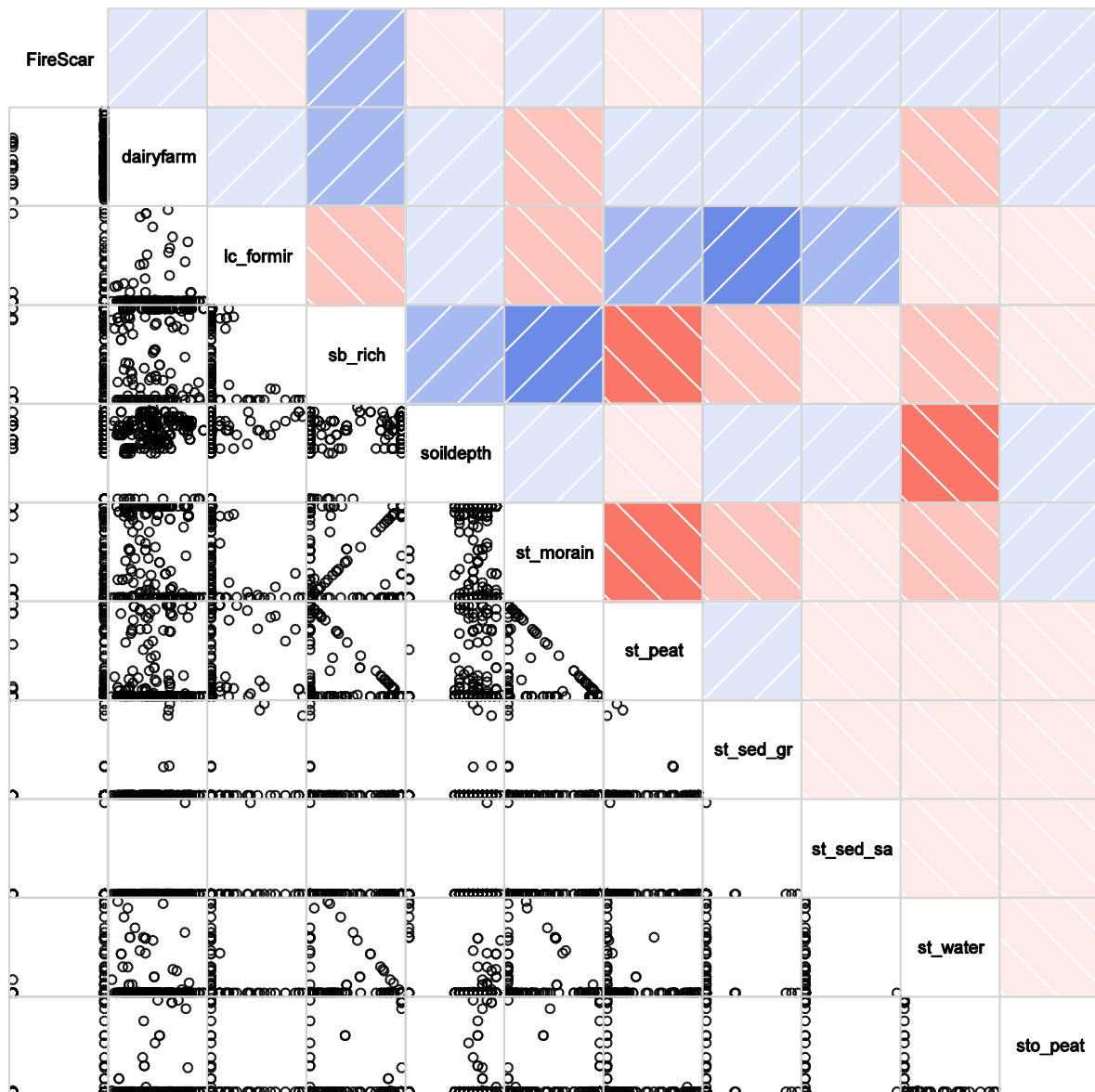


Figure 10.3: Correlogram (100m, 10 variables) showing the correlation values as based on the color scheme of Figure 4.1 (upper panel) and the point distribution between variables (lower panel).

```
# Correlogram (100m grid - nine variables)
#Set working directory
setwd("C:/temp/gwr_100grid_9var/corrgram")

#load datafile in R
jamt.corr <- read.table("jamt100_10v.txt", header = T)

#Convert to data frame
jamt.corr <- as.data.frame(jamt.corr)

# corrgram with scatter plots
corrgram(jamt.corr, lower.panel = panel.pts)
```

10.4 Correlogram (100m grid PCA)

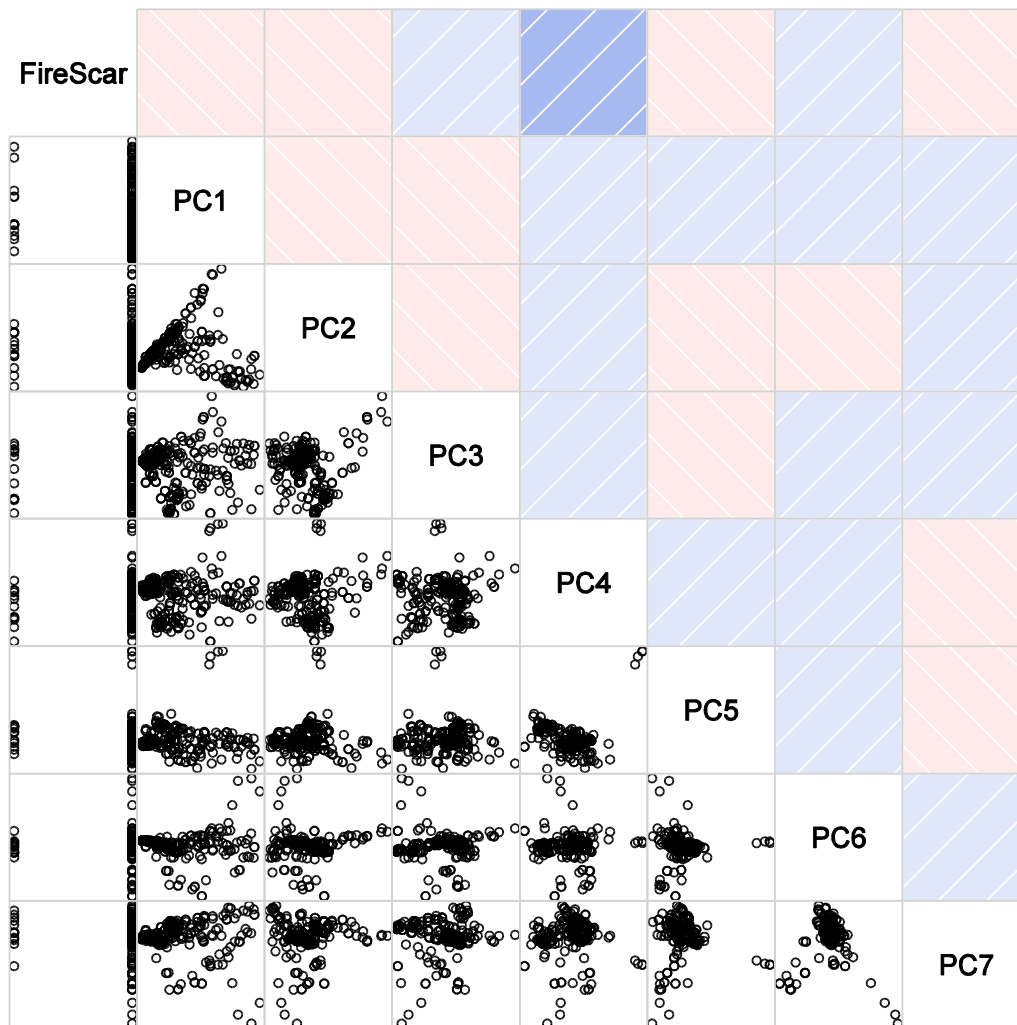


Figure 10.4: Correlogram (100m, PCA) showing the correlation values as based on the color scheme of Figure 4.1 (upper panel) and the point distribution between principal components (lower panel).

```
#Set working directory
setwd("C:/temp/gwr_100grid_PCs/corrgram")

#load datafile in R
jamt4.corr <- read.table("jamt100PC.txt", header = T)

#Convert to data frame
jamt4.corr <- as.data.frame(jamt4.corr)

# corrgram with scatter plots
corrgram(jamt4.corr, lower.panel = panel.pts)
```


10.5 Correlogram (250m grid – all variables)

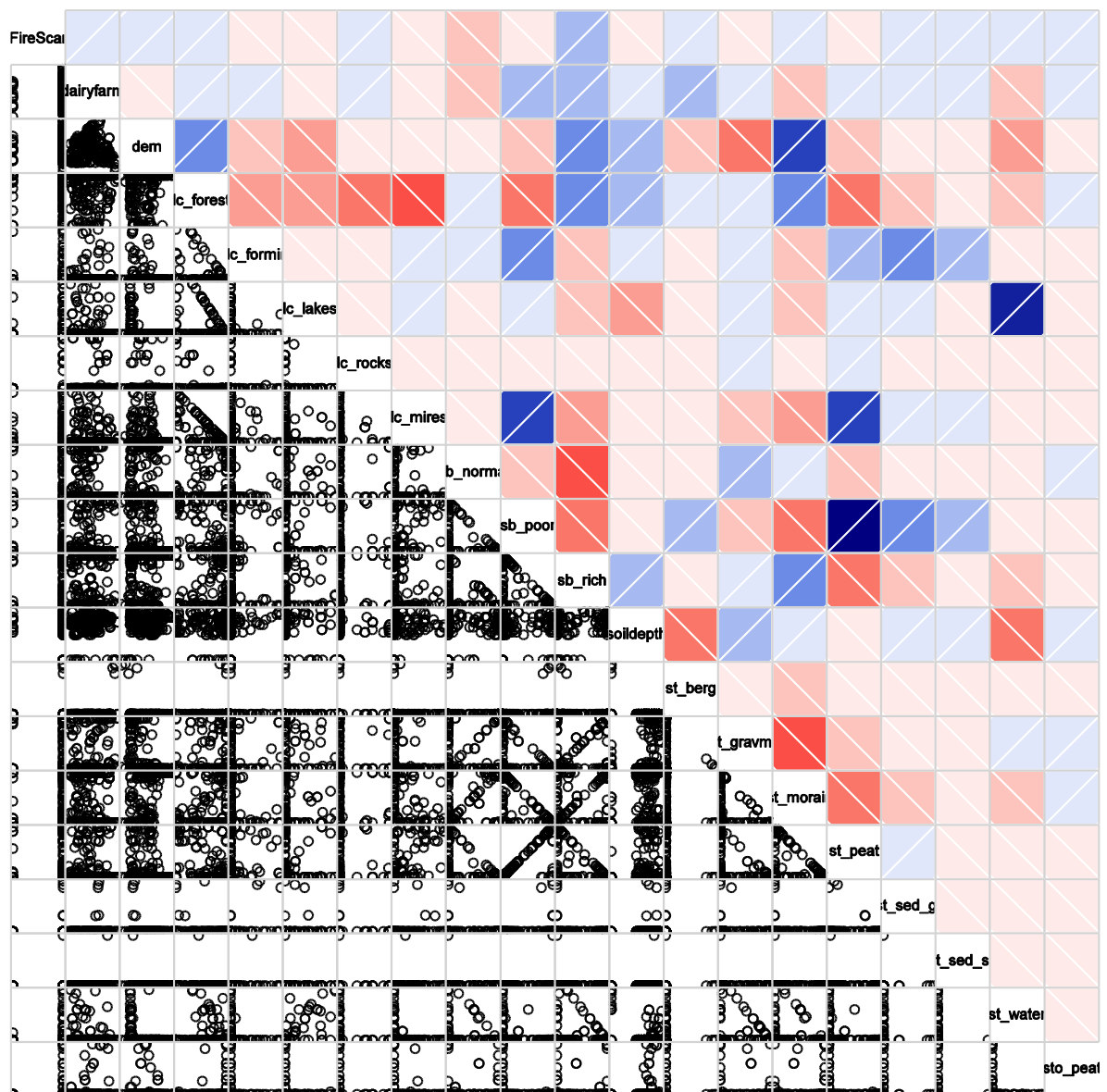


Figure 10.5: Correlogram (250m, all variables) showing the correlation values as based on the color scheme of Figure 4.1 (upper panel) and the point distribution between variables (lower panel).

```
#Set working directory
setwd("C:/temp/gwr_250grid/corrgram")

#load datafile in R
jamt2.corr <- read.table("jamt250av.txt", header = T)

#Convert to data frame
jamt2.corr <- as.data.frame(jamt2.corr)

# corrgram with scatter plots
corrgram(jamt2.corr, lower.panel = panel.pts)
```

10.6 Correlogram (250m grid - 12 variables)

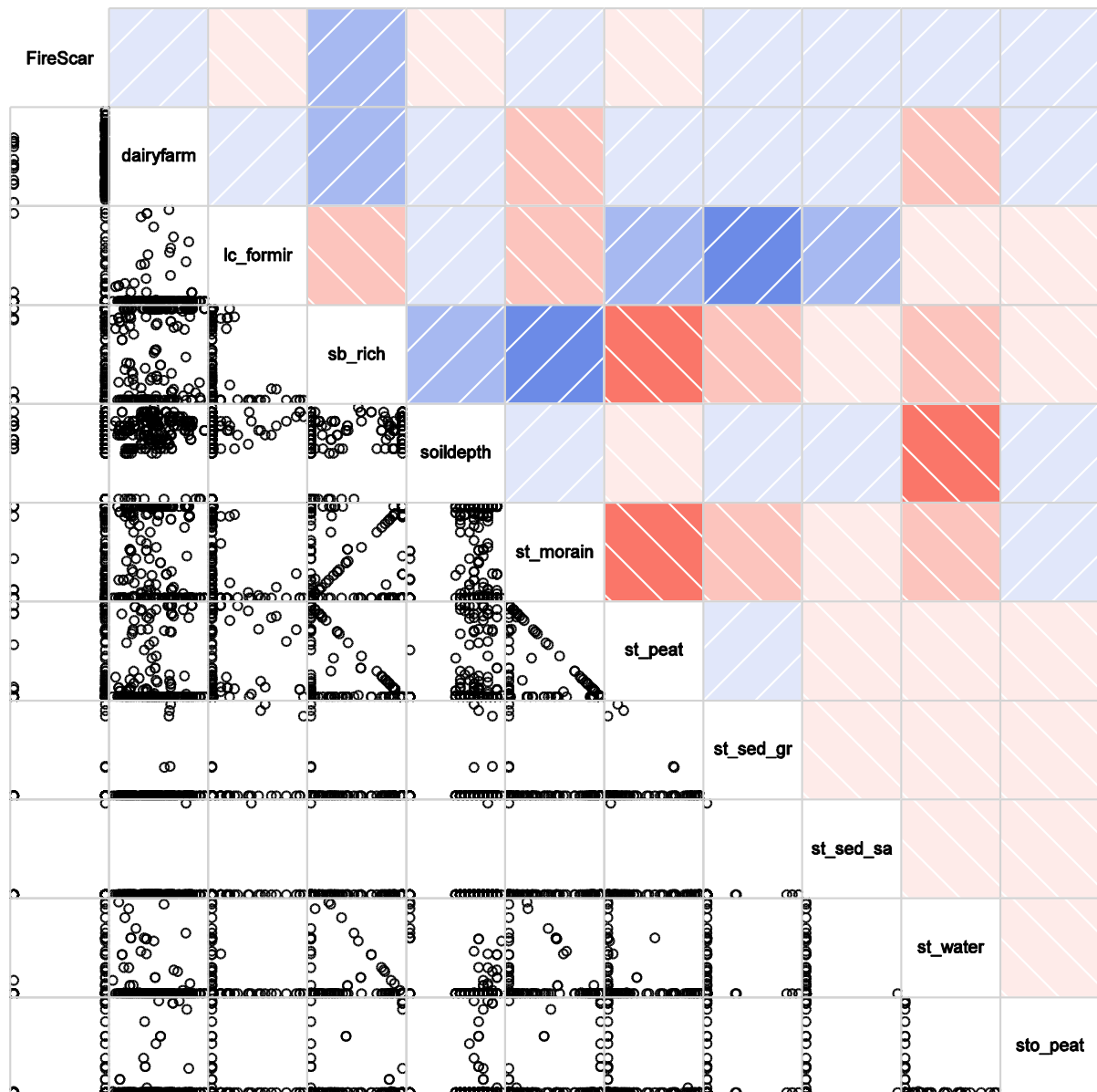


Figure 10.6: Correlogram (250m, 12 variables) showing the correlation values as based on the color scheme of Figure 4.1 (upper panel) and the point distribution between variables (lower panel).

```
#Set working directory
setwd("C:/temp/gwr_250grid_sel_var/corrgram")

#load datafile in R
jamt2.corr <- read.table("jamt250_12v.txt", header = T)

#Convert to data frame
jamt2.corr <- as.data.frame(jamt2.corr)

# corrgram with scatter plots
corrgram(jamt2.corr, lower.panel = panel.pts)
```

10.7 Correlogram (250m grid PCA)

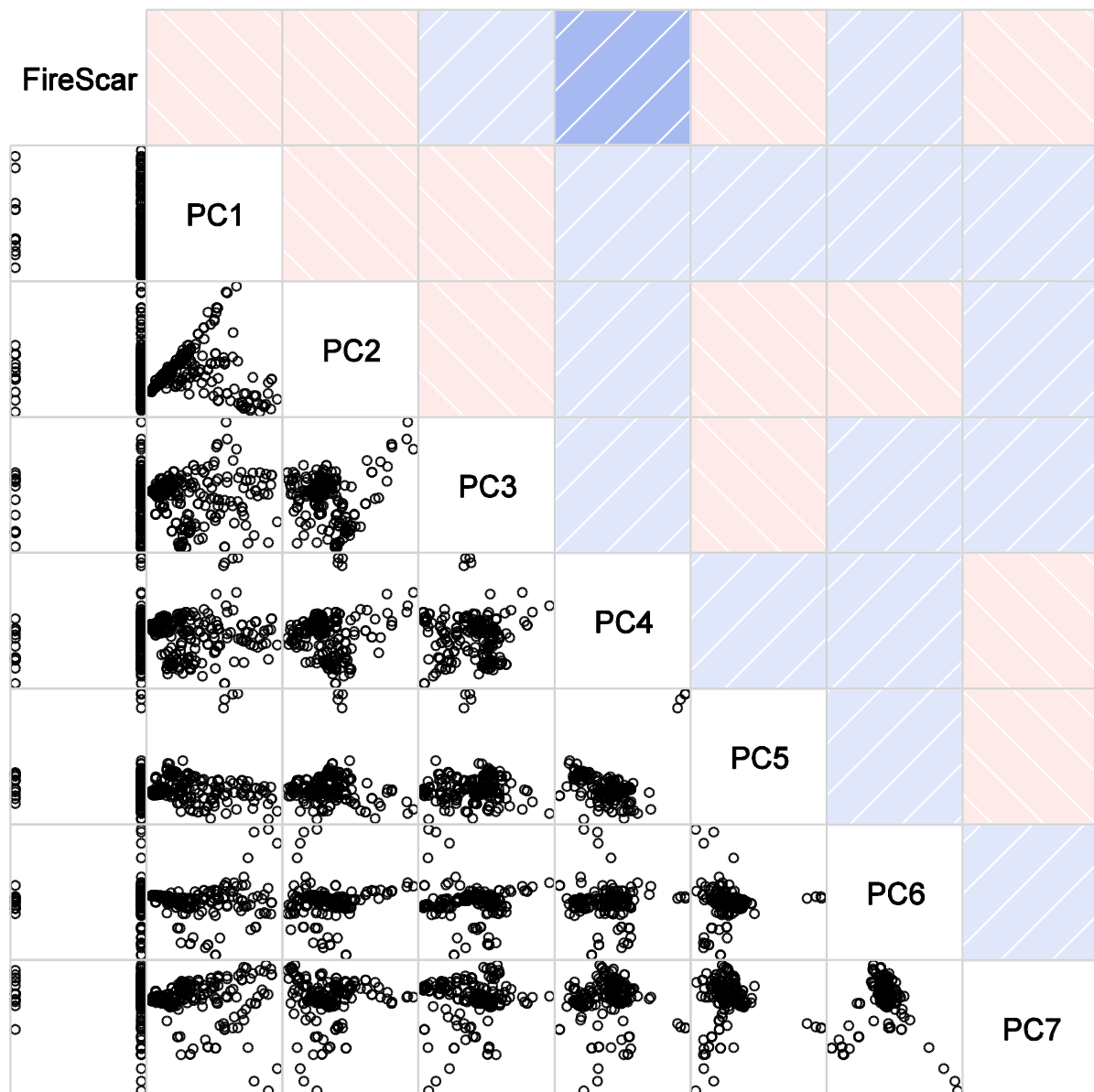


Figure 10.7: Correlogram (250m, PCA) showing the correlation values as based on the color scheme of Figure 4.1 (upper panel) and the point distribution between principal components (lower panel).

```
#Set working directory
setwd("C:/temp/gwr_250grid_PCs/corrgram")

#load datafile in R
jamt3.corr <- read.table("jamt250PC.txt", header = T)

#Convert to data frame
jamt3.corr <- as.data.frame(jamt3.corr)

# corrgram with scatter plots
corrgram(jamt3.corr, lower.panel = panel.pts)
```


11 Appendix D: Maps of Fire Size Comparison

11.1 Year 1521

65

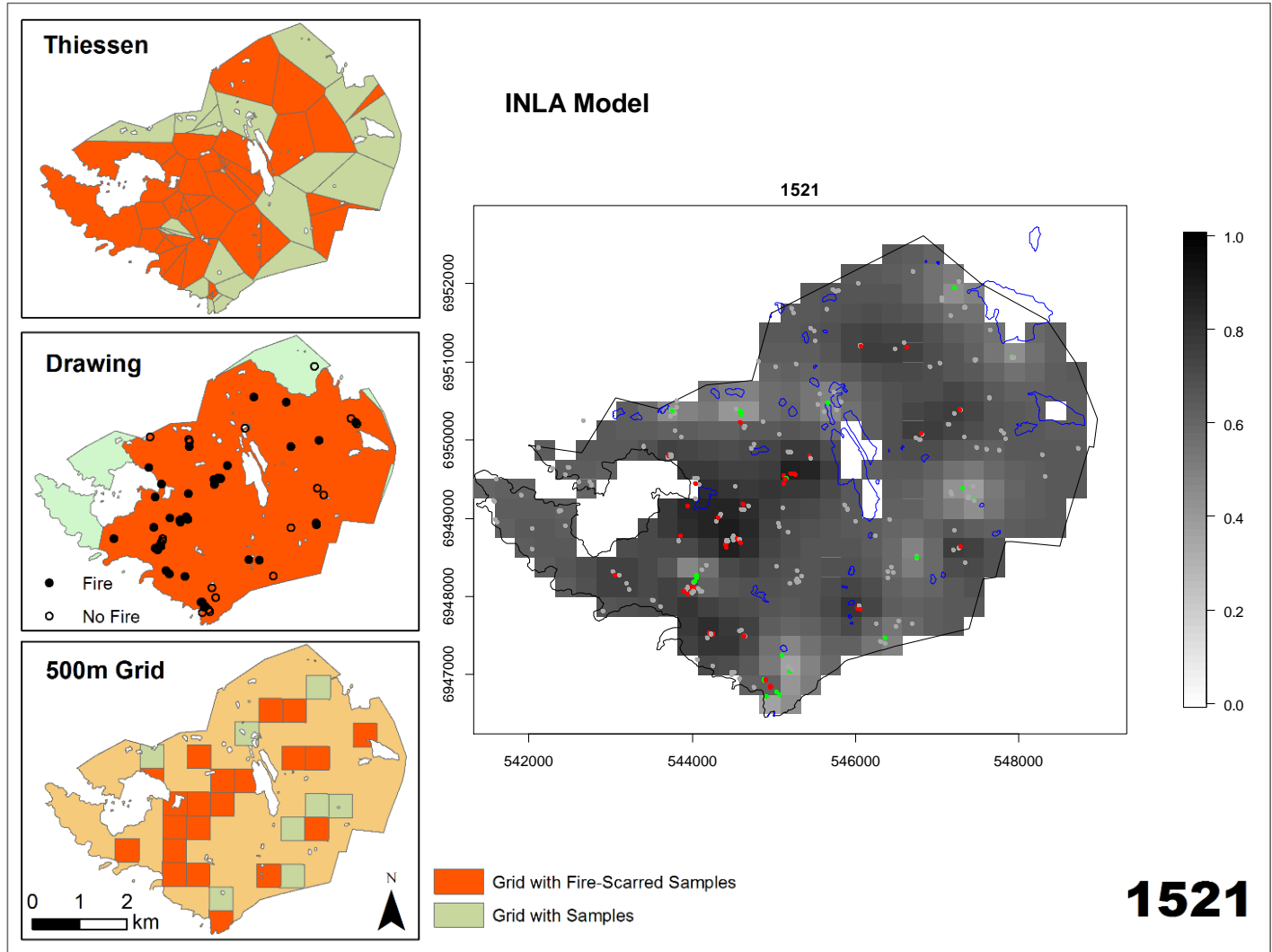


Figure 11.1: Method comparison between the different methods for the year 1521: Thiessen polygons, drawing an outline, grid method, and an INLA model.

11.2 Year 1576

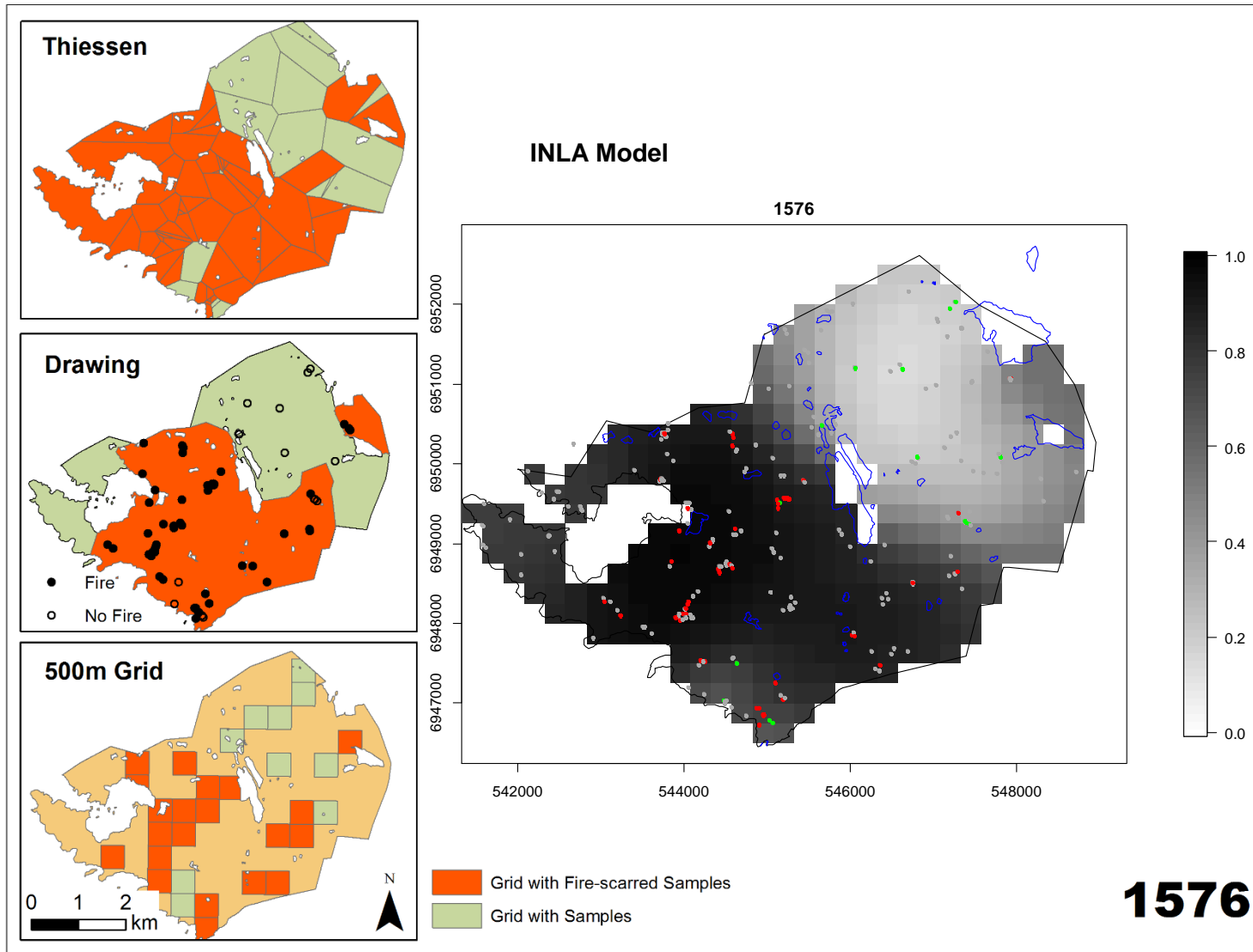


Figure 11.2: Method comparison between the different methods for the year 1576: Thiessen polygons, drawing an outline, grid method, and an INLA model.

11.3 Year 1644

67

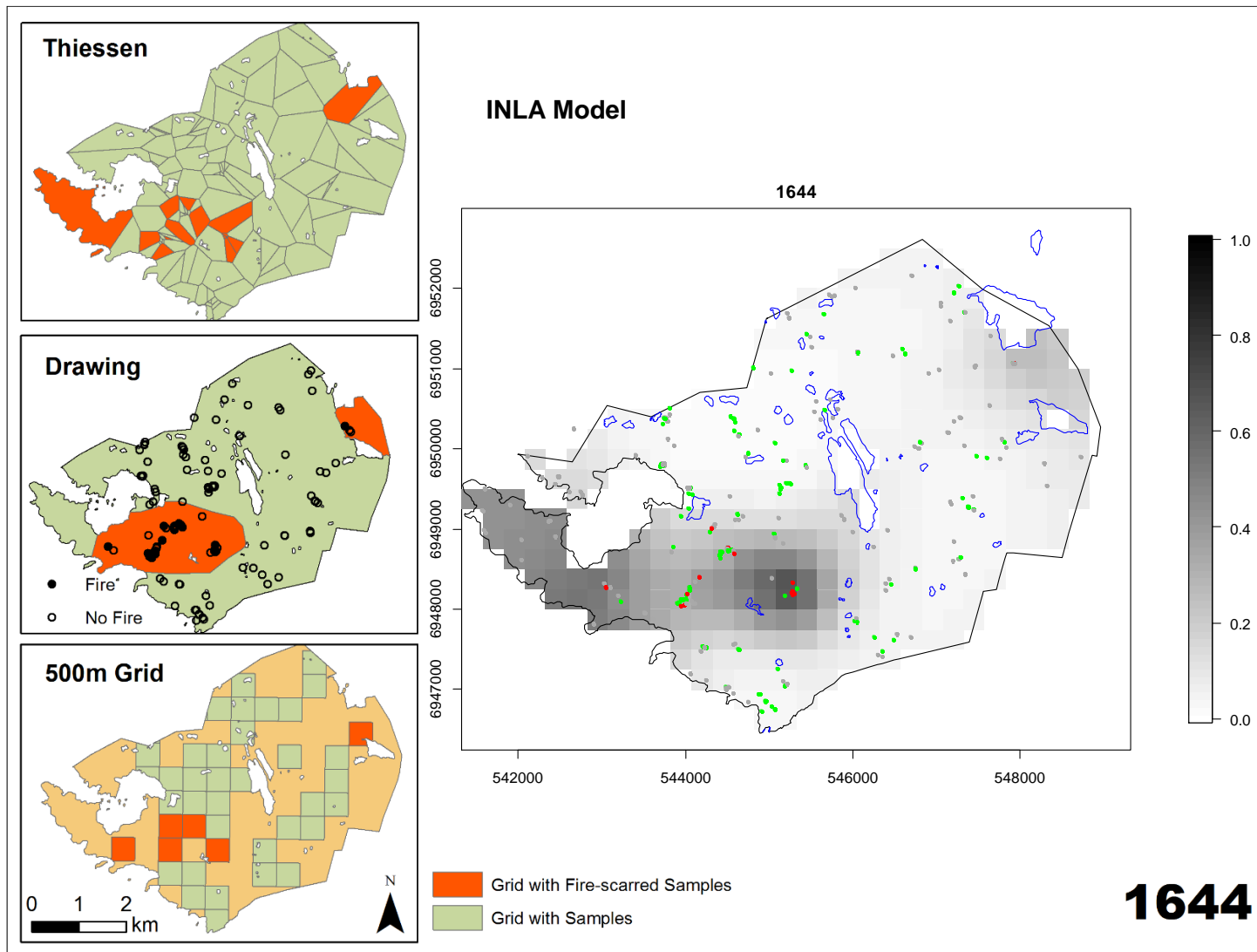


Figure 11.3: Method comparison between the different methods for the year 1644: Thiessen polygons, drawing an outline, grid method, and an INLA model.

11.4 Year 1677

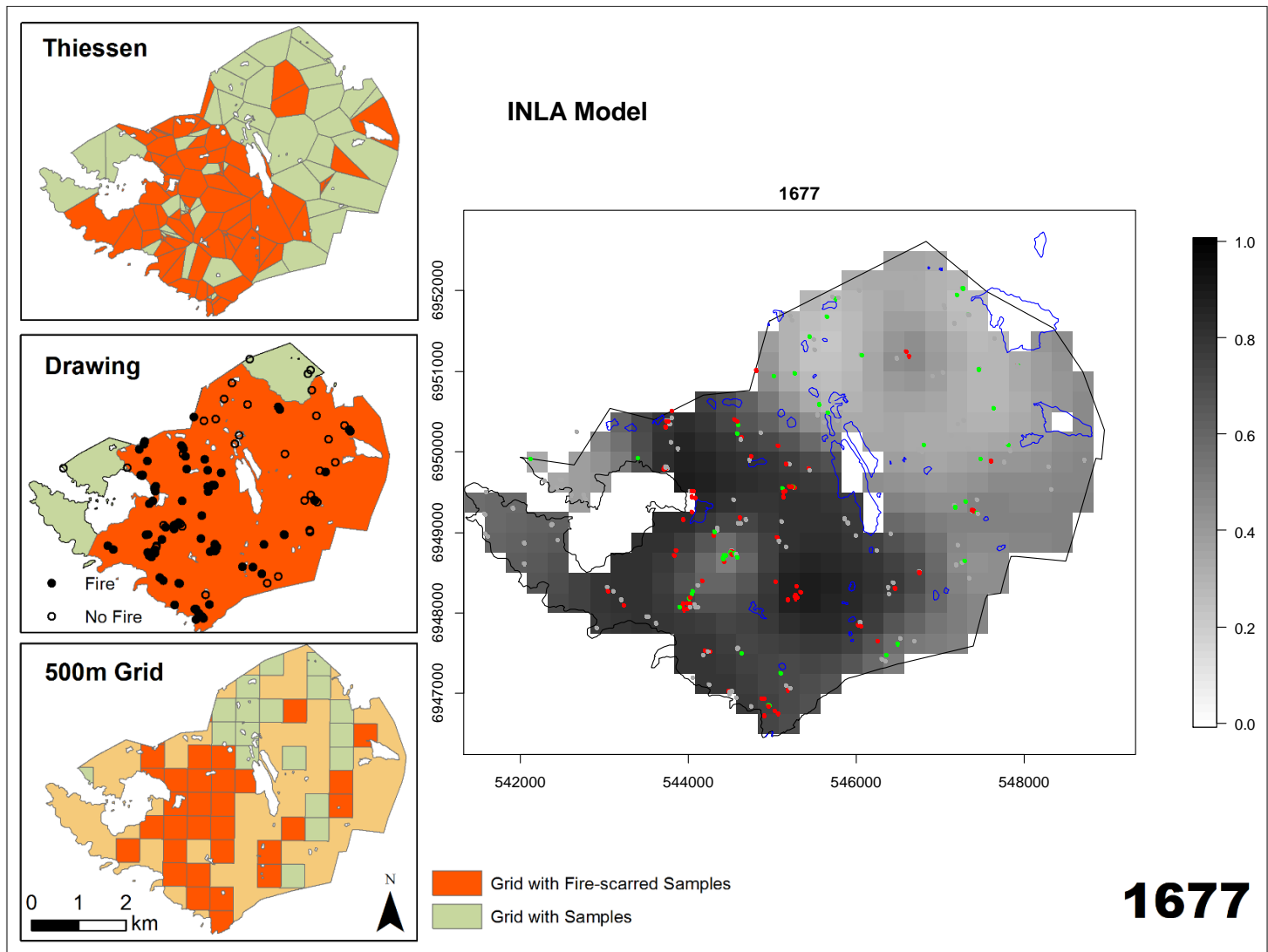


Figure 11.4: Method comparison between the different methods for the year 1677: Thiessen polygons, drawing an outline, grid method, and an INLA model.

11.5 Year 1723

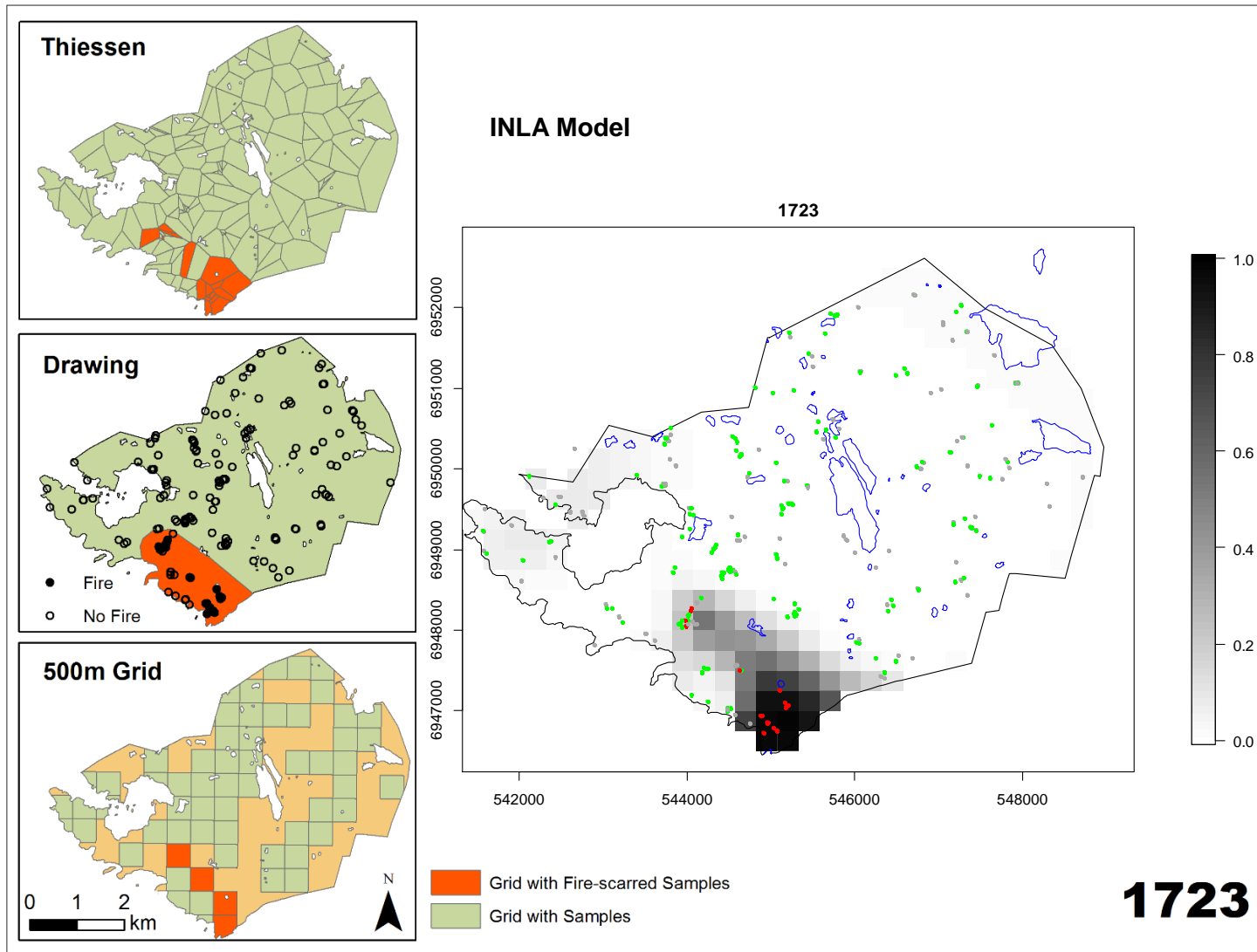


Figure 11.5: Method comparison between the different methods for the year 1723: Thiessen polygons, drawing an outline, grid method, and an INLA model.

11.6 Year 1729

70

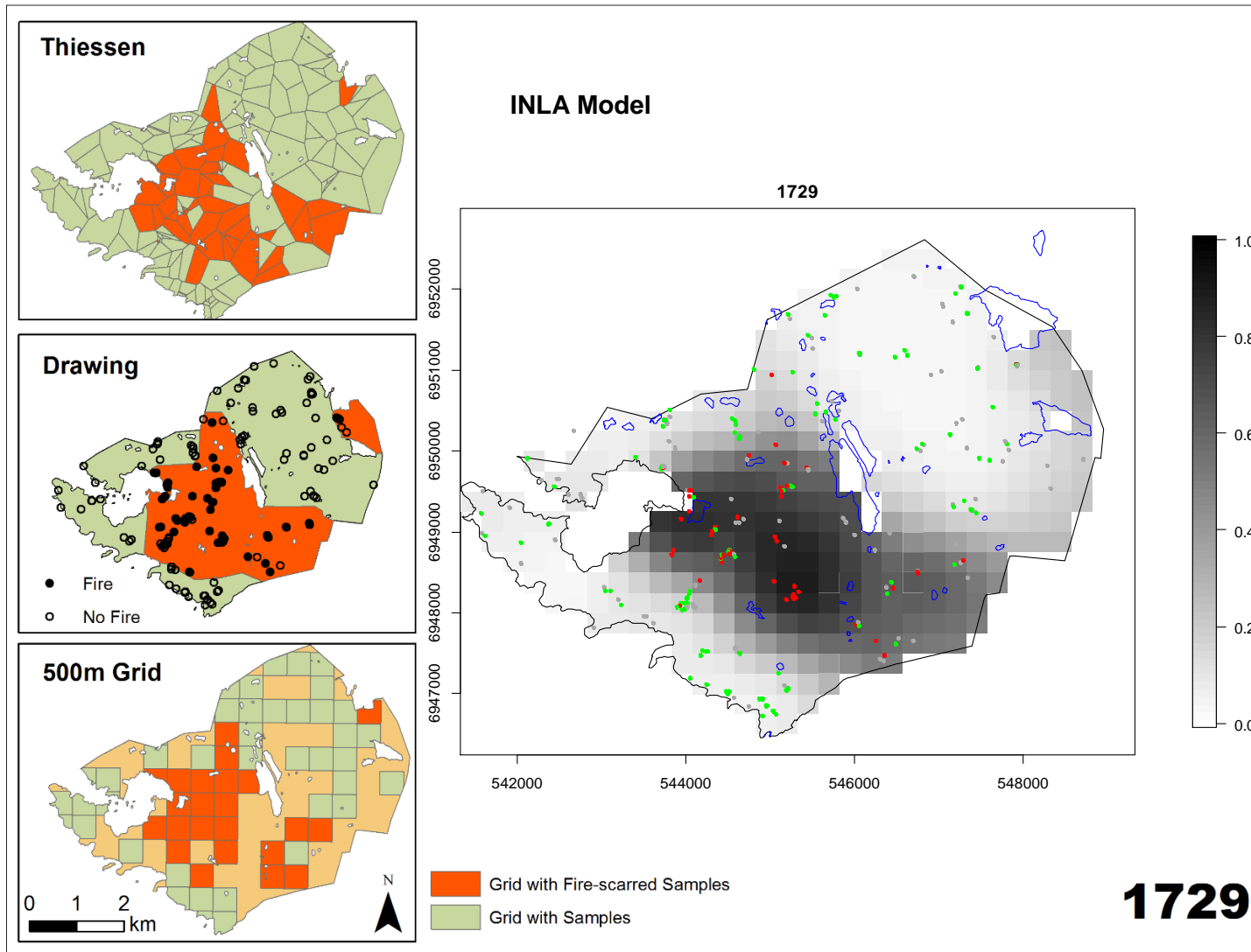


Figure 11.6: Method comparison between the different methods for the year 1729: Thiessen polygons, drawing an outline, grid method, and an INLA model.

11.7 Year 1770

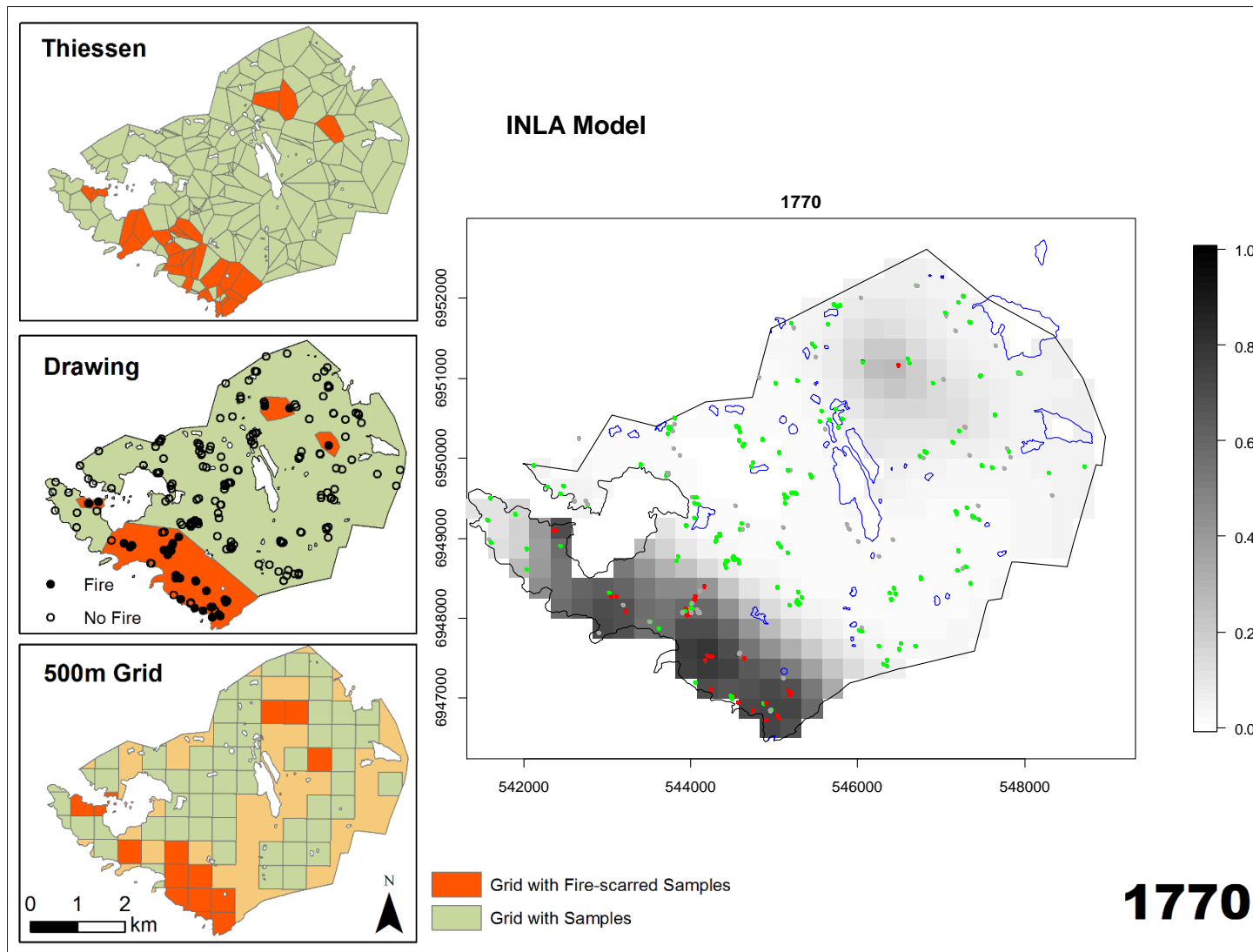


Figure 11.7: Method comparison between the different methods for the year 1770: Thiessen polygons, drawing an outline, grid method, and an INLA model.

11.8 Year 1789

72

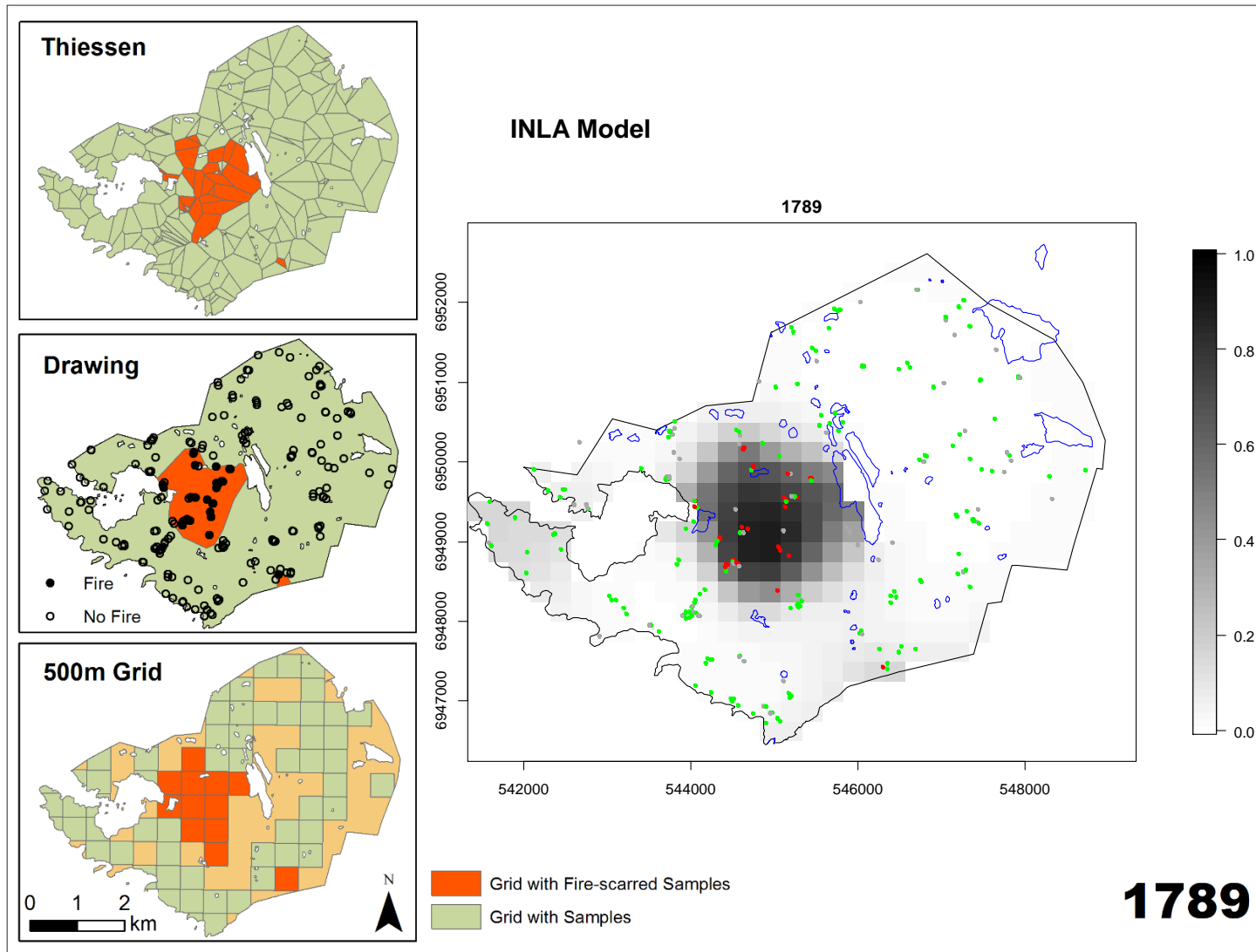


Figure 11.8: Method comparison between the different methods for the year 1789: Thiessen polygons, drawing an outline, grid method, and an INLA model.

11.9 Year 1807

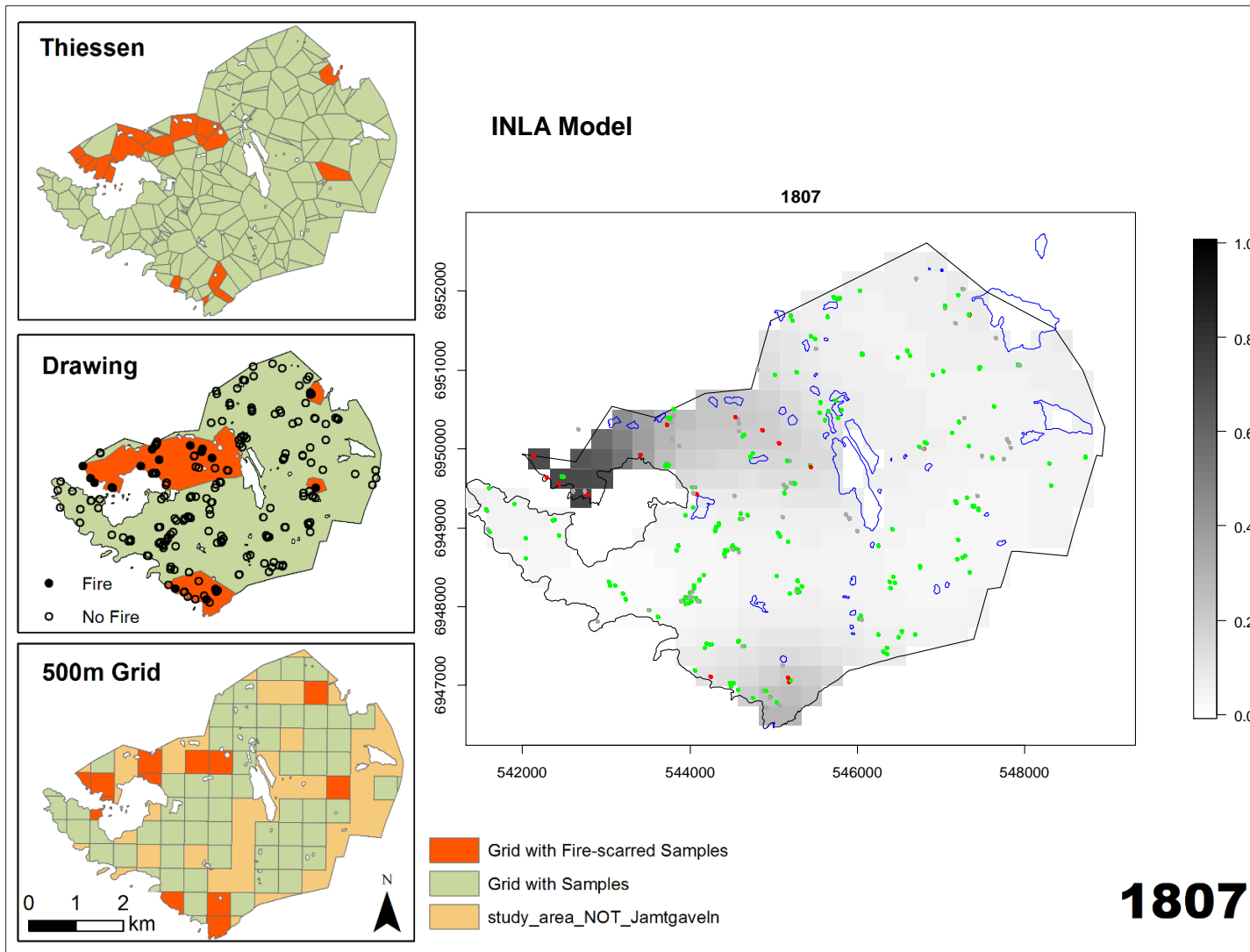


Figure 11.9: Method comparison between the different methods for the year 1807: Thiessen polygons, drawing an outline, grid method, and an INLA model.

11.10 Year 1831

74

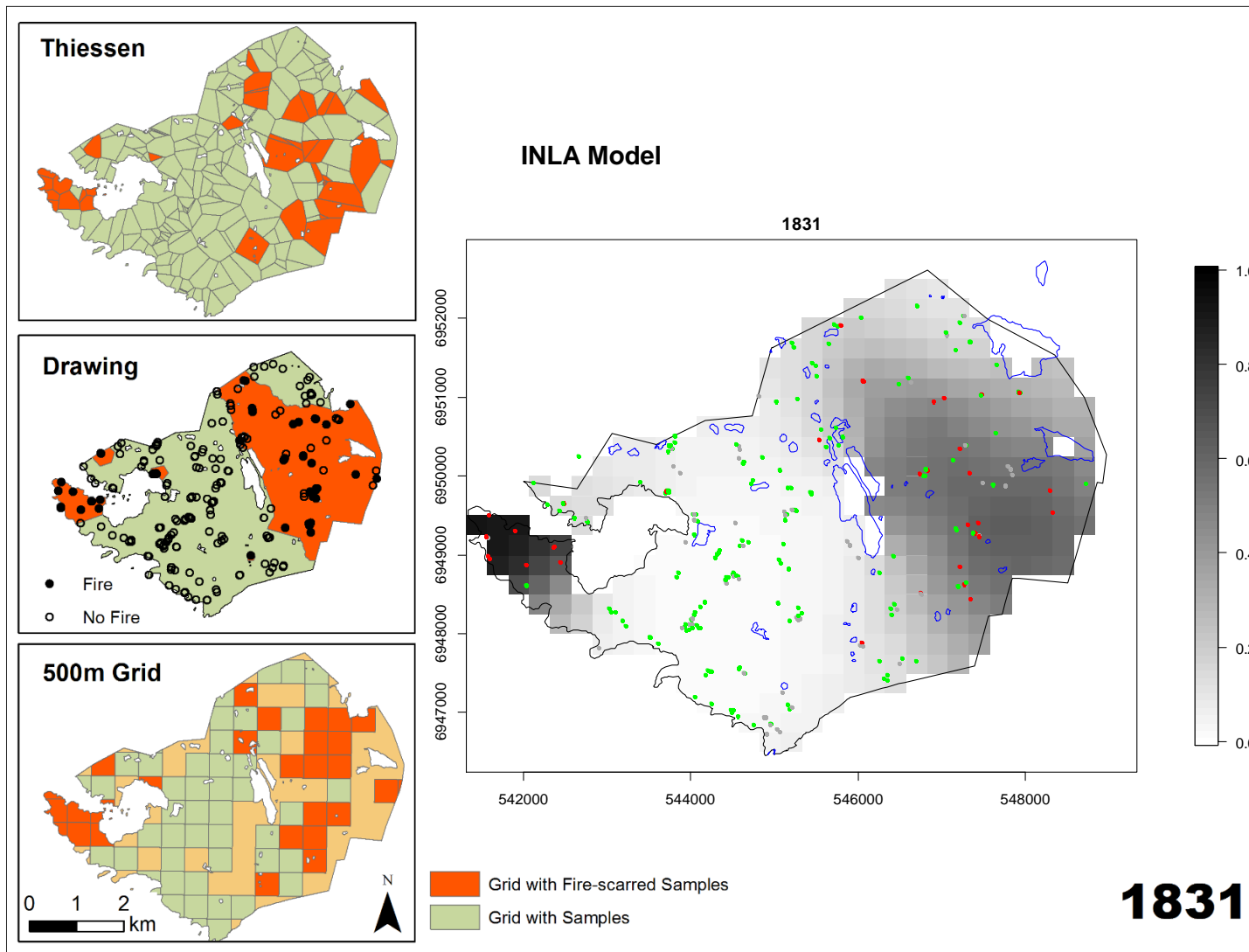


Figure 11.10: Method comparison between the different methods for the year 1831: Thiessen polygons, drawing an outline, grid method, and an INLA model.

11.11 Year 1844

75

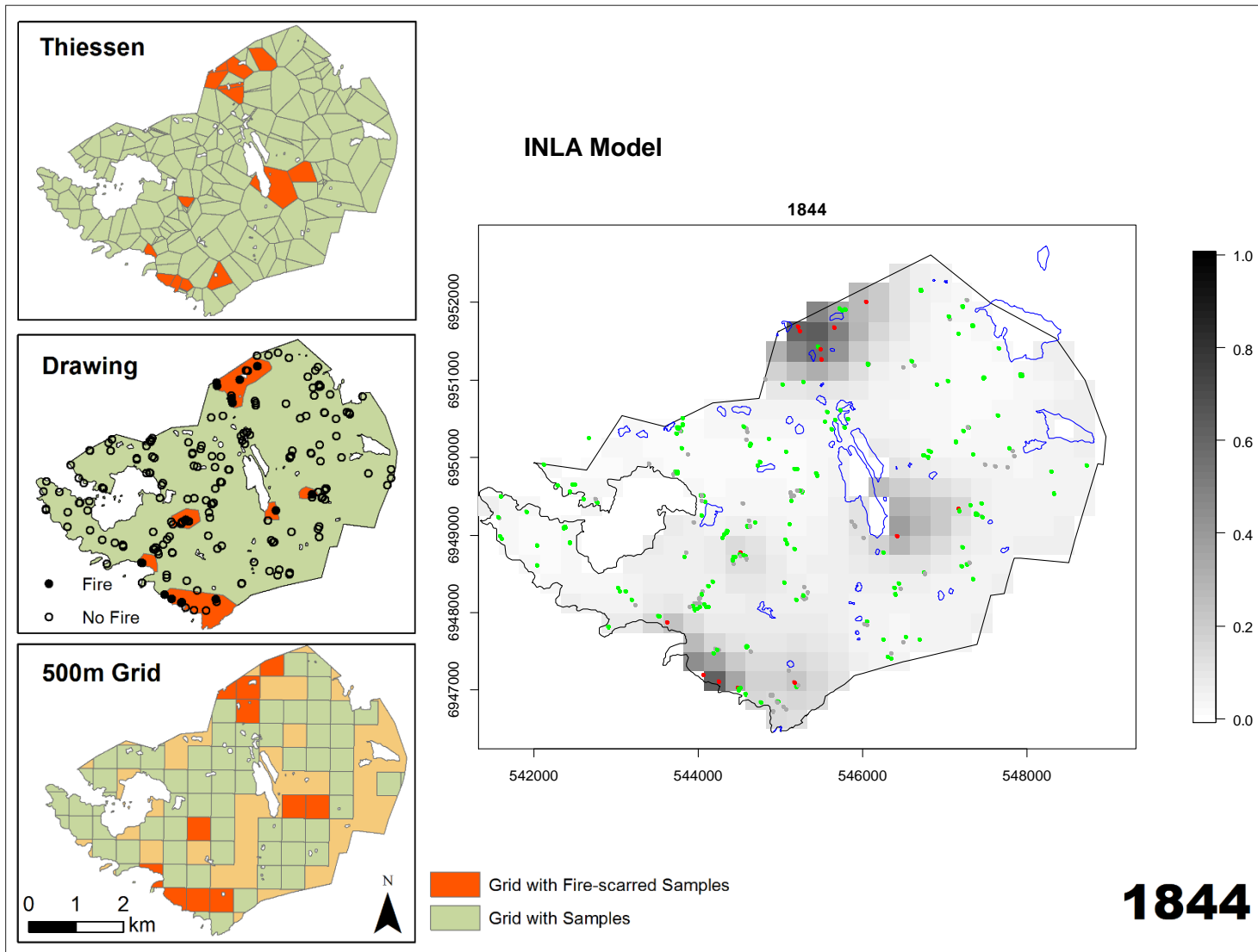


Figure 11.11: Method comparison between the different methods for the year 1844: Thiessen polygons, drawing an outline, grid method, and an INLA model.

11.12 Year 1868

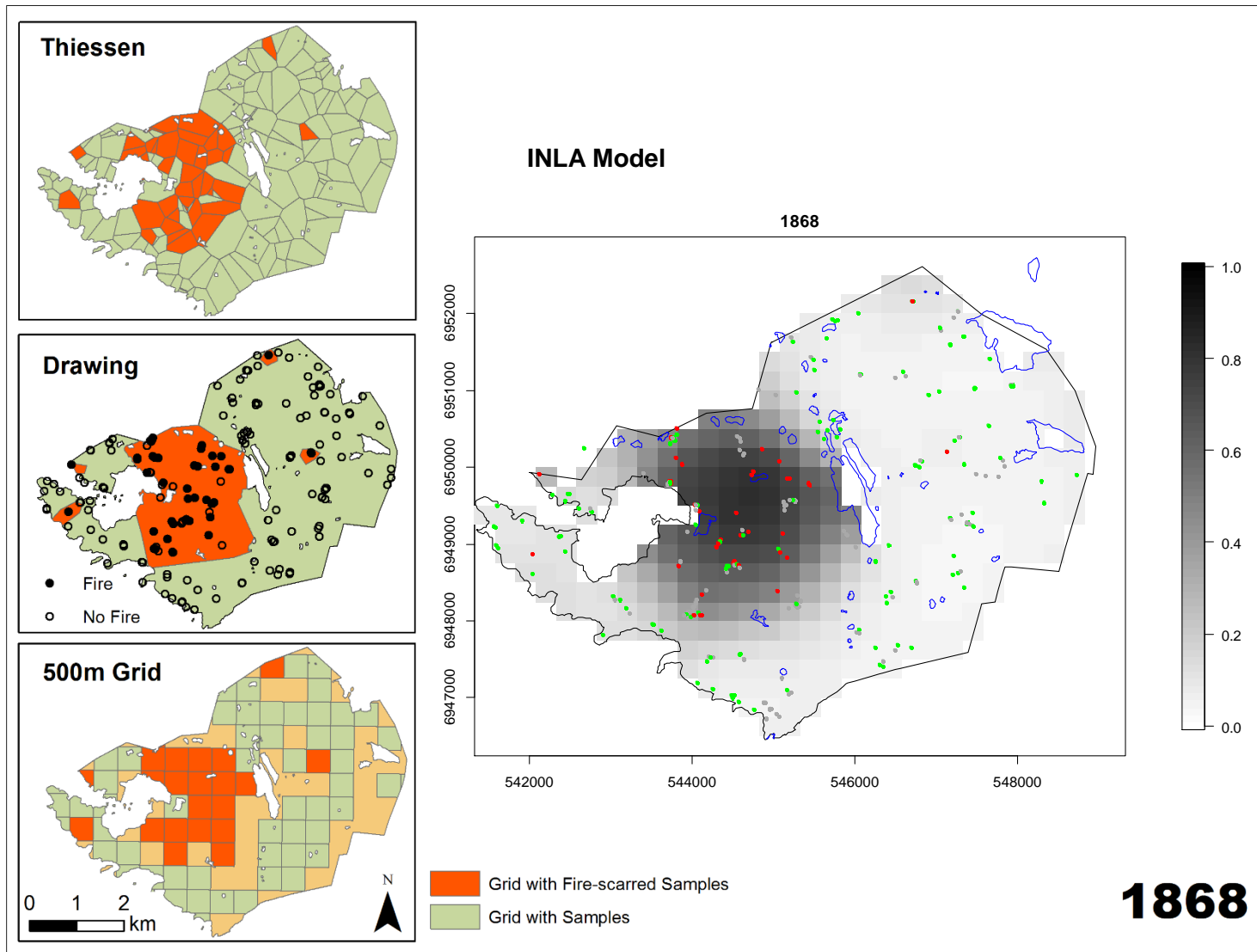


Figure 11.12: Method comparison between the different methods for the year 1868: Thiessen polygons, drawing an outline, grid method, and an INLA model.

Department of Physical Geography and Ecosystem Science

Master Thesis in Geographical Information Science

1. *Anthony Lawther*: The application of GIS-based binary logistic regression for slope failure susceptibility mapping in the Western Grampian Mountains, Scotland (2008).
2. *Rickard Hansen*: Daily mobility in Grenoble Metropolitan Region, France. Applied GIS methods in time geographical research (2008).
3. *Emil Bayramov*: Environmental monitoring of bio-restoration activities using GIS and Remote Sensing (2009).
4. *Rafael Villarreal Pacheco*: Applications of Geographic Information Systems as an analytical and visualization tool for mass real estate valuation: a case study of Fontibon District, Bogota, Columbia (2009).
5. *Siri Oestreich Waage*: a case study of route solving for oversized transport: The use of GIS functionalities in transport of transformers, as part of maintaining a reliable power infrastructure (2010).
6. *Edgar Pimiento*: Shallow landslide susceptibility – Modelling and validation (2010).
7. *Martina Schäfer*: Near real-time mapping of floodwater mosquito breeding sites using aerial photographs (2010).
8. *August Pieter van Waarden-Nagel*: Land use evaluation to assess the outcome of the programme of rehabilitation measures for the river Rhine in the Netherlands (2010).
9. *Samira Muhammad*: Development and implementation of air quality data mart for Ontario, Canada: A case study of air quality in Ontario using OLAP tool. (2010).
10. *Fredros Oketch Okumu*: Using remotely sensed data to explore spatial and temporal relationships between photosynthetic productivity of vegetation and malaria transmission intensities in selected parts of Africa (2011).
11. *Svajunas Plunge*: Advanced decision support methods for solving diffuse water pollution problems (2011).
12. *Jonathan Higgins*: Monitoring urban growth in greater Lagos: A case study using GIS to monitor the urban growth of Lagos 1990 - 2008 and produce future growth prospects for the city (2011).
13. *Mårten Karlberg*: Mobile Map Client API: Design and Implementation for Android (2011).
14. *Jeanette McBride*: Mapping Chicago area urban tree canopy using color infrared imagery (2011).

15. *Andrew Farina*: Exploring the relationship between land surface temperature and vegetation abundance for urban heat island mitigation in Seville, Spain (2011).
16. *David Kanyari*: Nairobi City Journey Planner: An online and a Mobile Application (2011).
17. *Laura V. Drews*: Multi-criteria GIS analysis for siting of small wind power plants - A case study from Berlin (2012).
18. *Qaisar Nadeem*: Best living neighborhood in the city - A GIS based multi criteria evaluation of ArRiyadh City (2012).
19. *Ahmed Mohamed El Saeid Mustafa*: Development of a photo voltaic building rooftop integration analysis tool for GIS for Dokki District, Cairo, Egypt (2012).
20. *Daniel Patrick Taylor*: Eastern Oyster Aquaculture: Estuarine Remediation via Site Suitability and Spatially Explicit Carrying Capacity Modeling in Virginia's Chesapeake Bay (2013).
21. *Angeleta Oveta Wilson*: A Participatory GIS approach to *unearthing* Manchester's Cultural Heritage 'gold mine' (2013).
22. *Ola Svensson*: Visibility and Tholos Tombs in the Messenian Landscape: A Comparative Case Study of the Pylian Hinterlands and the Soulima Valley (2013).
23. *Monika Ogden*: Land use impact on water quality in two river systems in South Africa (2013).
24. *Stefan Rova*: A GIS based approach assessing phosphorus load impact on Lake Flaten in Salem, Sweden (2013).
25. *Yann Buhot*: Analysis of the history of landscape changes over a period of 200 years. How can we predict past landscape pattern scenario and the impact on habitat diversity? (2013).
26. *Christina Fotiou*: Evaluating habitat suitability and spectral heterogeneity models to predict weed species presence (2014).
27. *Inese Linuza*: Accuracy Assessment in Glacier Change Analysis (2014).
28. *Agnieszka Griffin*: Domestic energy consumption and social living standards: a GIS analysis within the Greater London Authority area (2014).
29. *Brynja Guðmundsdóttir*: Detection of potential arable land with remote sensing and GIS - A Case Study for Kjósarhreppur (2014).
30. *Oleksandr Nekrasov*: Processing of MODIS Vegetation Indices for analysis of agricultural droughts in the southern Ukraine between the years 2000-2012 (2014).
31. *Sarah Tressel*: Recommendations for a polar Earth science portal in the context of Arctic Spatial Data Infrastructure (2014).
32. *Caroline Gevaert*: Combining Hyperspectral UAV and Multispectral Formosat-2 Imagery for Precision Agriculture Applications (2014).
33. *Salem Jamal-Uddeen*: Using GeoTools to implement the multi-criteria evaluation analysis - weighted linear combination model (2014).

34. *Samanah Seyedi-Shandiz*: Schematic representation of geographical railway network at the Swedish Transport Administration (2014).
35. *Kazi Masel Ullah*: Urban Land-use planning using Geographical Information System and analytical hierarchy process: case study Dhaka City (2014).
36. *Alexia Chang-Wailing Spitteler*: Development of a web application based on MCDA and GIS for the decision support of river and floodplain rehabilitation projects (2014).
37. *Alessandro De Martino*: Geographic accessibility analysis and evaluation of potential changes to the public transportation system in the City of Milan (2014).
38. *Alireza Mollasalehi*: GIS Based Modelling for Fuel Reduction Using Controlled Burn in Australia. Case Study: Logan City, QLD (2015).
39. *Negin A. Sanati*: Chronic Kidney Disease Mortality in Costa Rica; Geographical Distribution, Spatial Analysis and Non-traditional Risk Factors (2015).
40. *Karen McIntyre*: Benthic mapping of the Bluefields Bay fish sanctuary, Jamaica (2015).
41. *Kees van Duijvendijk*: Feasibility of a low-cost weather sensor network for agricultural purposes: A preliminary assessment (2015).
42. *Sebastian Andersson Hylander*: Evaluation of cultural ecosystem services using GIS (2015).
43. *Deborah Bowyer*: Measuring Urban Growth, Urban Form and Accessibility as Indicators of Urban Sprawl in Hamilton, New Zealand (2015).
44. *Stefan Arvidsson*: Relationship between tree species composition and phenology extracted from satellite data in Swedish forests (2015).
45. *Damián Giménez Cruz*: GIS-based optimal localisation of beekeeping in rural Kenya (2016).
46. *Alejandra Narváez Vallejo*: Can the introduction of the topographic indices in LPJ-GUESS improve the spatial representation of environmental variables? (2016).
47. *Anna Lundgren*: Development of a method for mapping the highest coastline in Sweden using breaklines extracted from high resolution digital elevation models (2016).
48. *Oluwatomi Esther Adejoro*: Does location also matter? A spatial analysis of social achievements of young South Australians (2016).
49. *Hristo Dobrev Tomov*: Automated temporal NDVI analysis over the Middle East for the period 1982 - 2010 (2016).
50. *Vincent Muller*: Impact of Security Context on Mobile Clinic Activities A GIS Multi Criteria Evaluation based on an MSF Humanitarian Mission in Cameroon (2016).
51. *Gezahagn Negash Seboka*: Spatial Assessment of NDVI as an Indicator of Desertification in Ethiopia using Remote Sensing and GIS (2016).
52. *Holly Buhler*: Evaluation of Interfacility Medical Transport Journey Times in Southeastern British Columbia. (2016).

53. *Lars Ole Grottenberg*: Assessing the ability to share spatial data between emergency management organisations in the High North (2016).
54. *Sean Grant*: The Right Tree in the Right Place: Using GIS to Maximize the Net Benefits from Urban Forests (2016).
55. *Irshad Jamal*: Multi-Criteria GIS Analysis for School Site Selection in Gorno-Badakhshan Autonomous Oblast, Tajikistan (2016).
56. *Fulgencio Sanmartín*: Wisdom-volcano: A novel tool based on open GIS and time-series visualization to analyse and share volcanic data (2016).
57. *Nezha Acil*: Remote sensing-based monitoring of snow cover dynamics and its influence on vegetation growth in the Middle Atlas Mountains (2016).