Master Thesis in Geographical Information Science nr 158

Multi-scale Bark Beetle Predictions Using Machine Learning

Albert Wellendorf

2023

Department of

Physical Geography and Ecosystem Science

Centre for Geographical Information Systems

Lund University
Sölvegatan 12

Wellendorf, Albert (2023). Multi-scale Bark Beetle Predictions Using Machine Learning
Master degree thesis, 30 credits in Master in Geographical Information Science
Department of Physical Geography and Ecosystem Science, Lund University

Multi-scale Bark Beetle Predictions Using Machine Learning

Albert Wellendorf

Master thesis, 30 credits, in Geographical Information Sciences

Supervisors:

Ali Mansourian
Department of Physical Geography and Ecosystems Science, Lund University, Sweden

Pengxiang Zhao
Department of Physical Geography and Ecosystems Science, Lund University, Sweden

Acknowledgements

I want to thank both Ali Mansourian and Pengxiang Zhao for their patience and help during the project. Pengxiang Zhao was a great support during the modelling phase. I always felt that I could ask him when I had problems and he always took the time to answer me and try to understand my questions, although they were not always that clear.

I also want to thank Mitro Müller for his help with the data preparation and explanation of the process.

Abstract

Bark beetle attacks have led to widespread tree disturbance and deaths in many parts of the world, and thereby also economic and biodiversity losses. Forest-rich Sweden has experienced periodic attacks, latest in 2018. There is a great interest in identifying the most important explanatory features behind bark beetle attacks and making spatial predictions on where attacks might happen. This could limit the reliance on expensive ad-hoc measures to diminish the negative effects of bark beetle attacks. This is especially important in future years, as bark beetle attacks are expected to increase under climate change.

Machine learning is a family of algorithms that is capable of finding complex patterns in data and making predictions on unseen data. Earlier studies have already used different types of algorithms to predict bark beetle infestation spots and detect the most important features that characterise these spots. One problem with machine learning algorithms and the earlier studies in the field, is the lack of consideration of spatial autocorrelation and heterogeneity in the modelling. The current study aims to address these limitations by looking at the differences in prediction accuracy on different spatial scales. The study area (south-eastern Sweden) is divided into different numbers of zones (2, 4, 6, 8, 10, and 15) and the prediction accuracy and spatial distribution of feature importance are assessed and compared to that of the global model (full dataset). Furthermore, the results for a drought period (2018) and a normal period (2019-2020) are compared.

Different algorithms are assessed – Random Forest, Support Vector Machine and Logistic Regression. It is found that random forest performs best, albeit only marginally, compared to support vector machines on the global dataset (normal period). Random forest is therefore also used for the local modelling in the created zones.

The results from the local modelling indicate that zooming in to a more local scale (only considering the points in a zone) can result in better predictions both for the drought (year 2018) and normal period (year 2019 and 2020). Especially in areas with a relatively even number of infested and healthy records and also not too few points, the prediction accuracy is higher than for the global dataset. In the best performing local zones, the feature importance differs compared to the global model, and other features are generally most important here. This indicates that global modelling on the full dataset may mask the fact that some features are more important in different parts of the study area.

Multi-scale modelling can be beneficial for adaptation purposes and different factors can be prioritised in different areas depending on the local feature importance. Studies, like the current one, are important in the light of the future threat of an increase in bark beetle attacks. A more dynamic approach to the local modelling has been used in other fields where local machine learning models are created in all data records. It will be an interesting addition to current bark beetle research to make such dynamic studies in the future, but it is deemed out of the scope of the current study.

Keywords: Geographically weighted regression, bark beetle, GIS, machine learning, spatial prediction, Southern Sweden

Table of Contents

Overview of tables and figures

List of abbrevations

ML	Machine Learning
RF	Random Forest
SVM	Support Vector Machine
LR	Logistic Regression
GWR	Geographically Weighted Regression
DT	Decision Tree

1. Introduction

1.1 Bark beetle and forest disturbances

Forests suffer a range of different natural disturbances such as insect outbreaks that can lead to ecosystem changes, reduced tree growth, negative impacts on wildlife and biodiversity and ultimately large economic losses for forest owners (Hroššo et al., 2020; Kärvemo & Schroeder, 2010; Koreň et al., 2021; Olsson, 2016). Bark Beetle is a forest disturbance agent that primarily attacks spruce and pine trees and has resulted in a high number of tree deaths in many areas of the world (Hernandez, Saborio, Ramsey, & Rivera, 2012; Kärvemo & Schroeder, 2010; Olsson, 2016) – also in European forests, that have been disturbed periodically by the European spruce bark beetle (Ips typographus L.) (Hlásny et al., 2021; Olsson, 2016).

1.2 Triggers and drivers of bark beetle outbreaks

Wind-felled and dying trees may trigger bark beetle population growth as these trees offer optimal conditions for breeding in the beginning of an outbreak (Kärvemo & Schroeder, 2010; Lausch, Fahse, & Heurich, 2011). At this stage the bark population numbers are not high enough to overcome the living tree defences (Hroššo et al., 2020; Kärvemo & Schroeder, 2010).

Generally, higher temperatures support bark beetle development, both through beetle flights, that are initiated at around 20 ◦C and also through an increase in the number of generations per year (Lausch et al., 2011). In the warmer areas of the bark beetle range such as in the Central European lowlands, two generations are generally completed in a year, or even three in years with favourable weather. Further north one generation has generally been the rule (Lausch et al., 2011).

Climate change is expected to make the situation even more severe, thus increasing the need for adequate forest strategies and management. At the same time, a positive feedback between bark beetle outbreaks and climate change exist, as these lead to increased net carbon fluxes from the land to the atmosphere (Hlásny et al., 2021). Hernandez et al. (2012) found an increase in bark beetle attacks when climate change scenarios were included in the modelling. Higher temperatures and increased drought intensity will decrease tree vigour and increase the number of weakened and sensitive trees (Koreň et al., 2021). Also an increase in the number of bark beetle generations per year and the winter survival rate is expected (Lausch et al., 2011). There is an obvious increased risk of such situations in Northern Europe under global warming (Långström, Lindelöw, Schroeder, Björklund, & Öhrn, 2009).

Understanding the drivers behind bark beetle attacks is difficult and based on complex environmental processes that differ spatially and temporally (Koreň et al., 2021). Lausch et al. (2011) stated that the complexity of processes and features makes it nearly impossible to predict future outbreaks. A problem with some of the earlier studies in the field has been the lack of spatial and temporal dimensions, for instance including only one year of infestation data, and including only a limited number of explaining habitat factors, making it difficult to generalize the results to a larger scale (Lausch et al., 2011).

Different drivers have been included in previous studies to predict bark beetle infestations such as forest composition features, climatic features, and spatial-temporal features (Hernandez et al., 2012; Hroššo et al., 2020; Koreň et al., 2021; Lausch et al., 2011). Lausch et al. (2011) assessed the importance of different factors in the long-term outbreak in Bavaria National Forest starting in 1983-1984. The most important factors determining the sites of infestations were spatial-temporal factors such as distance to infestation sites from previous year and the area and perimeter of infestation sites from previous years. Elevation was found to be the most important abiotic factor, whereas slope and aspect played minor roles. Climatic factors such as potential solar radiation were not found to be important, thereby potentially explaining why aspect and slope were less important. Interestingly, temperature was, contrary to the results from other studies, not found to be linearly associated with an increase in outbreak probability. Hroššo et al. (2020) assessed the drivers of bark beetle infestations following a windthrow in the Slovakian Carpathians in 2014. Year of attack, solar radiation and tree dimensions were found to be positively correlated with infestation probability. Koreň et al. (2021) found some of the same factors to be the most important in the Horní Planá region in Czech Republic including potential solar radiation, spruce age, percentage of spruce in forest stand, volume per hectare and distance to actual forest damage. To underscore the complexity between drivers and the spread of bark beetle, Lausch et al. (2011) did not find any mono-causal correlations between individual factors and bark beetle infestations and suggested that this could be the reason for the different findings in earlier studies in the field. The choice of model, spatial scale and input factors all influence the results and the assessment of the relative importance of drivers.

1.3 Bark beetle outbreaks in Sweden

Sweden has suffered different bark beetle outbreaks since 1961 where huge storms initiated eleven years of outbreak (Kärvemo & Schroeder, 2010). A more recent outbreak occurred after the 2005 storm Gudrun that exemplified some of the typical dispersal and outbreak processes and resulted in a severe storm felling in southern Sweden and average volumes of damaged forests in the magnitude of 65 to 75 m^3 per ha due to the storm (Långström et al., 2009). Especially spruce trees were affected. Large volumes of damaged trees were still left in the forest in the summer of 2005 in spite of a large operation to clear the forests of these. The aim was to save timber and minimize economic losses but also to diminish the habitat for bark beetle development (Kärvemo & Schroeder, 2010; Långström et al., 2009). In the spring and summer of 2005 there was an abundance of dead trees in the south Swedish forests, but the bark beetle population numbers were low. Therefore, only a small percentage of these felled trees were attacked in the first period (Långström et al., 2009). If the population levels are able to sustain on the deadwood, they may eventually reach epidemic levels which usually takes one to three years (Hroššo et al., 2020). This can result in attacks and killings of standing trees at higher bark beetle population levels (Kärvemo & Schroeder, 2010).

An estimated 60% of the remaining wind-felled trees were attacked by bark beetle in Sweden in 2006 and in the end of the summer also 1.5 million m^3 of standing trees were killed (Långström et al., 2009). The reasons for this were the high temperatures in 2006 that led to intense beetle flight at the end of the summer and the completion of a second generation (Långström et al., 2009). In the following year, colder and wetter weather at the end of the summer led to decreased beetle flights and thus less attacks (Långström et al., 2009).

This example, following the storm Gudrun, shows a typical bark beetle outbreak that has also been reported in other countries following big wind-felling events (Hroššo et al., 2020; Långström et al., 2009; Lausch et al., 2011). Both Bavaria and Sweden have experienced prolonged bark beetle attacks (Kärvemo & Schroeder, 2010; Lausch et al., 2011). Outbreaks lasting a decade or longer can happen when continued storms lead to wind-fells and dry weather increase the sensitivity and decrease the tree defence of standing trees. So in general the bark beetle outbreak risk is high when the population numbers are high and there is an abundance of weakened trees, for instance due to dry weather (Hroššo et al., 2020; Kärvemo & Schroeder, 2010; Långström et al., 2009). The biggest bark beetle outbreak in Sweden was recorded following the exceptionally dry and warm weather in the summer of 2018 that led to the killing of approximately 17 million m^3 of spruce forest from 2018 to 2020 (Schroeder, 2020).

1.4 Outbreak dynamics and methods to predict bark beetle attacks

There is a great interest in better understanding the dynamics, dispersal processes and spatial dimension of bark beetle outbreaks. This could lead to more suitable forest management strategies, which could reduce economic losses as well as potential negative effects on biodiversity and ecosystem changes (Hlásny et al., 2021; Hroššo et al., 2020; Koreň et al., 2021). An increased understanding of outbreak dynamics and the most important drivers could also help to make precise predictions of the areas, that are most sensitive to bark beetle attacks. Such predictions would be beneficial as bark beetle management measures traditionally have been employed ad-hoc after the occurrence of infestations (Rammer & Seidl, 2019). Measures such as chemical control and removal of wind-felled trees are expensive and may be disapproved by society (Valdez Vasquez et al., 2020). According to Hroššo et al. (2020), 80% of the bark beetles must be killed and 80% of the windthrow must be removed to substantially decrease the risk. This can be problematic, especially after a huge storm event such as Gudrun in Sweden due to the vast number of wind-felled trees in the ground (Långström et al., 2009). Långström et al., (2009) therefore also found that the evidence suggested that these outbreaks cannot really be controlled by humans. The emphasis should be on understanding population dynamics and identify sensitive areas before infestation occurs (Hroššo et al., 2020; Långström et al., 2009).

In general, there is a growing interest among ecosystem managers and policy makers in precise predictions since these can forecast ecosystem trajectories and help to sustain the provision of ecosystem services that are essential to human beings and biodiversity (Rammer & Seidl, 2019). Pest prediction models is a traditional method that has been employed to predict the probability of pine bark beetle attacks in the US, based on the number of bark beetle trappings per day. The accuracy of these is highly variable and range between 32-85% (Billings & Upton, 2010; Munro, Montes, & Gandhi). Other traditional methods, such as linear regression, that has often been used in geographical analyses, are not suitable to model complex, non-linear relationships between independent and dependent variables as in the case of bark beetle attacks (Brunsdon, Fotheringham, & Charlton, 1996; Luo, Yan, McClure, & Li, 2022).

Advances in technology and data availability due to increased computational power, remote sensing and interpolation techniques have made such predictions much more feasible today (Rammer & Seidl, 2019), especially through the use of machine learning (ML). ML has been used in recent bark beetle studies that have assessed the performance of various ML algorithms in predicting sensitive forest areas (Koreň et al., 2021; Munro et al.; Ramazi, Kunegel-Lion, Greiner, & Lewis, 2021; Rammer & Seidl, 2019; Valdez Vasquez et al., 2020). These studies have shown that ML can improve the prediction accuracy of bark beetle infestations compared to traditional models such as pest prediction models and be helpful guidelines for forest managers (Valdez Vasquez et al., 2020).

1.5 Machine Learning and geographically weighted methods

ML algorithms are flexible, data-driven, sophisticated and effective at extracting knowledge from data (Georganos et al., 2021; Nikparvar & Thill, 2021). These algorithms can learn from data and eventually be used to predict new data points and to explore the relationship between dependent and independent variables (VanderPlas, 2016). The models could satisfy the increasing interest in more computationally intensive and data-driven algorithms. They can be used for a range of different tasks, such as pattern recognition, clustering, regression, and classification. ML has been used in many different fields and applications such as land use and land cover classification, disaster management, forest dynamic drivers, crop productivity prediction, forest disturbance assessment and population modelling (Georganos et al., 2021; Koreň et al., 2021; Nikparvar & Thill, 2021; Santos, Graw, & Bonilla, 2019).

Non-parametric ML models such as Random Forest (RF) can learn non-linear relationships from complex, high-dimensional data (Luo, Yan, & McClure, 2021) and are not sensitive to outliers such as is the case for linear regression models (Quiñones, Goyal, & Ahmed, 2021). These models are thus suitable to model complex, non-linear phenomena such as bark beetle attacks (Koreň et al., 2021). The problem with ML algorithms is that they are basically 'aspatial' and assumes that the relationship between independent and dependent variables do not vary over space, thereby disregarding potential spatial heterogeneity (Georganos et al., 2021). Not only the temporal, but also the spatial dimension appears to be important to describe the relationship between input factors and bark beetle infestations. The transition from endemic bark beetle attacks to self-driven epidemics is likely driven by individual factors whose importance vary depending on local conditions (Hroššo et al., 2020).

Earlier studies have tried to incorporate the spatial autocorrelation in ML modelling by combining ML and geographically weighted regression (GWR), introduced by Brunsdon et al. (1996). The rationale behind GWR is Tobler's first law of geography that states that things near to each other are more related and their attributes more alike compared to things more distant from each other (Arabameri, Pradhan, & Rezaei, 2019; Luo et al., 2022). Instead of one global regression model incorporating the whole area under study, GWR is based on multiple local regression models (Luo et al., 2022). GWR is an extension of the ordinary least squares method and uses a weighted least squares approach for each location, giving higher weights to observations that are closer than those farther away (Hagenauer & Helbich, 2021; Luo et al., 2022). As GWR is based on local linear regression models, the approach is not appropriate when relationships between independent and dependent variables are non-linear and multicollinearity between the independent variables exists (Luo et al., 2022; Quiñones et al., 2021). Newer studies have used novel approaches, combining RF and GWR to overcome the limitations of either approach and be able to model complex, spatial relationships and improve model performance (Georganos et al., 2021; Luo et al., 2021; Quiñones et al., 2021). Studies report a higher prediction accuracy of local models compared to global ones (Arabameri et al., 2019; Luo et al., 2021). Furthermore, these models are better at explaining the spatial heterogeneity of the relationships between independent and dependent variables (Quiñones et al., 2021). By dividing the study area into different neighbourhoods, mapping the prediction performances can reveal in what areas the local models perform well. In the areas with worse prediction accuracies, adding more features could improve the model performance (Georganos et al., 2021; Quiñones et al., 2021). Local scale modelling can guide the formulation of local management strategies focused on the most important factors in a specific area (Luo et al., 2021).

1.6 Earlier bark beetle studies

Different methodologies and algorithms have been used in the previous ML and bark beetle studies. For instance, Koreň et al. (2021) predicted the spatial distribution of spruce bark beetle infestation spots in the Horní Planá region, Czech Republic, using different machine learning algorithms, both linear and more complex non-linear models. Rammer & Seidl (2019) showed the usability of using deep learning and neural networks to predict bark beetle outbreaks in Bavarian Forest National Park in Germany. Ramazi et al. (2021) assessed the performance of eight machine learning models in future bark beetle predictions in Cypress Hills, Canada.

These studies have shown that ML algorithms can lead to better prediction accuracies compared to more traditional methods and be helpful in the spatial prediction of sensitive areas through the use of geographical data. At the same time, ML can be used to perform feature importance analysis that can shed light on the most important drivers and their relationships to bark beetle infestations (Koreň et al., 2021; Rammer & Seidl, 2019; Valdez Vasquez et al., 2020). (M. Müller, Olsson, Eklundh, Jamali, & Ardö, 2022) used a ML algorithm to perform feature analysis in the same study area and based on the same data as the current study.

Earlier bark beetle studies have used spatial input data, for instance in the form of coordinates, and distance-based input features such as the distance to damage spots from previous year and distance to forest edges (Koreň et al., 2021; Munro et al.; Rammer & Seidl, 2019; Valdez Vasquez et al., 2020). That being said, these studies (Koreň et al., 2021; M. Müller et al., 2022) still relied on global models that were used for predictions over the whole case area, thus not considering spatial autocorrelation and spatial non-stationarity of the relationship between bark beetle presence and explanatory features.

Another problem with earlier bark beetle studies is their limitations in time or space, either only covering one year or only a limited scale, making it difficult to make comparisons between studies and see which drivers are most important on a larger scale (Lausch et al., 2011).

The limitations of earlier studies in the field are related to the limited temporal and spatial scales of the input data as well as a lack of consideration of the spatial non-stationarity between the independent and dependent variables.

1.7 Aim and research questions

This study tried to fill in some of the research gaps that are present in earlier bark beetle studies. ML algorithms in the form of Random Forest (RF), Support Vector Machine (SVM) and Logistic Regression (LR) were applied to a dataset from south-eastern Sweden, consisting of bark beetle presence and related, explanatory features. LR was included to assess the difference in prediction accuracies between linear and non-linear models.

The aim of the study was foremost to use ML algorithms as a predictive tool – to establish the model with the best performance when it comes to predicting bark beetle infestation areas. ML models with a high prediction accuracy can guide future forest management and potentially help minimizing problems in forest-rich countries such as Sweden, especially under climate change and the expected increase in forest disturbances. The dataset was split into a drought and a normal period based on the weather in the assessed years. Modelling was performed on both datasets to assess the importance of drought, which could be an indication of potential impacts of climate change in the study area.

Multi-scale modelling was performed to see whether predictions would be more accurate on a more local scale and also to see whether the relationship between explaining factors and bark beetle presence varied over the study area. Both global models that considered all the data points in the study area, and local models only considering (spatially contiguous) subsets of the data in the study area, were assessed. The local zones were made both randomly and based on GWR scores. The best models, on both global and local scales were also used as exploratory tools, since knowing the most important factors on a local scale is another aspect that can help guide local bark beetle management in the future.

Multi-scale modelling overcame some of the scale limitations mentioned in the bark beetle literature and made it possible to study outbreak dynamics on different scales.

Answering the following research questions should help to fulfil the overall aim.

- Which machine learning algorithm predicts bark beetle infestations most accurately on the global scale?
- Does local modelling result in an overall higher prediction accuracy?
- What characterizes the best and worst performing local models and how do their feature importance vary over space?
- How do the results differ between the normal and the drought period and what are the implications of these results in relation to climate change?

2. Methods and material

2.1 Study area

The study area is a 48,600 km^2 large area in south-eastern Sweden located at 56.10 °N, 13.44 °E to 58.53 °N, 16.53 °E. It is a typical Swedish landscape, with large forest lands, water bodies, and less agricultural and built-up land (Fig. 1). The forests are mainly coniferous forests, with Norway spruce as the dominant species, with a low number of deciduous forests. Most of the area covers the hemi-boreal zone which is the transition zone between temperate and boreal zones and a small part is covered by the nemoral vegetation zone (Jonsson et al., 2016). Most of the study area is also in the hemi-boreal climate zone characterized by a humid-continental climate with hot summers. That being said, the climate is relatively mild. During the normal period 1991-2020 the mean temperature was -1.3 °C in January and 17.1 °C in July in Tranås, located in the northern part of the study area. In Växjö located in the south-central part, the January temperature was -1.1 °C and the July temperature identical to Tranås (“Dataserier med normalvärden för perioden 1991-2020 | SMHI”).

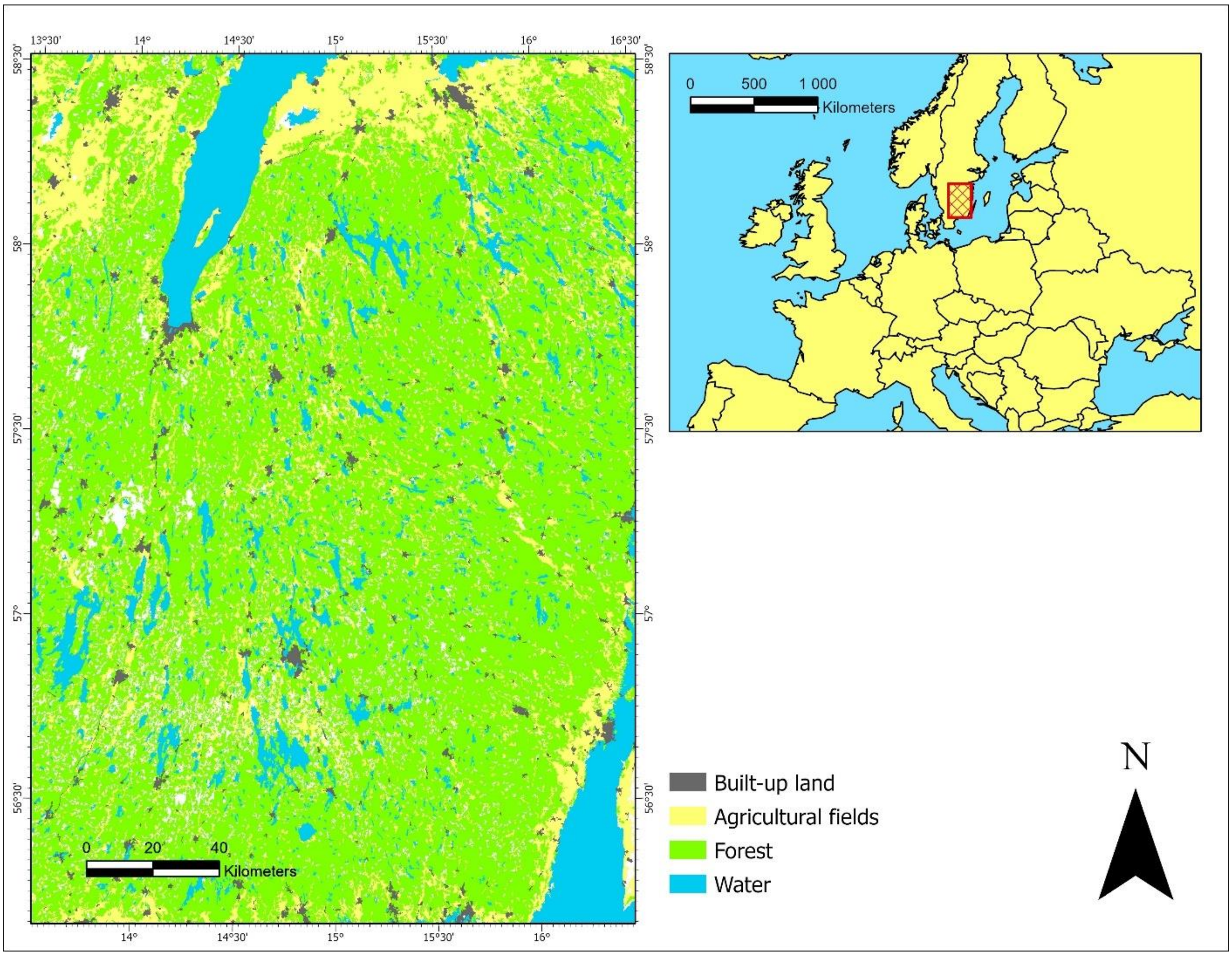


Figure 1. Map over study area

2.2 Data

2.2.1 Bark Beetle data

Harvested trees (infested with bark beetle) and their coordinate information were collected with harvester machines with global satellite navigation systems by Sveaskog and Södra in the years 2018 to 2020 (M. Müller et al., 2022). Data was removed if more than 10 coordinate duplicates were found. Since harvester machines potentially were harvesting more trees from the same position, this threshold was applied.

The data was structured in a 10x10 m grid covering the study area. Each pixel was characterized as infested if an infested tree was present here. Since the harvested data was only based on trees that were removed due to bark beetle infestations and surrounding trees, data about absence of bark beetle (healthy data) was needed (M. Müller et al., 2022). The chosen solution for this was to use landcover and property data, and defining healthy data as pixels that were not characterized as infested and at the same time located in estates that had harvested trees in other pixels (M. Müller et al., 2022).

Data from June to December 2019, from Sveaskog was discarded since there was uncertainty whether a part of the infestations from 2019 was actually from 2018. In the end since a substantial amount of data was discarded, instead of dividing it into annual parts, it was divided into a normal period covering the years 2019 and 2020 and a drought period covering 2018.

The end result was two grids based on healthy and infested pixels for both the normal (2019 and 2020) and drought period (2018). The number of healthy records far out-weighted infested records. To equalize the counts, a stratified random sampling was performed using the forest and soil type coverage percentages as grouping parameters (M. Müller et al., 2022). This could alleviate some of the problems with unbalanced datasets. In the end 24,433 records were included for the drought period, and 75,447 for the normal period. The final step was to combine the bark beetle presence data with the other data sources, that are described below.

2.2.2 Other data sources and explanatory features

Features that could help to explain bark beetle infestations were included through other data sources. All the data was included in the 10x10 m grid through resampling – nearest neighbour was used for discrete features and bilinear interpolation for continuous features (M. Müller et al., 2022). The list of data sources and the resultant features are included in Table 1.

Some variables had to be merged due to a high number of values. These were forest and soil type. Forest type was originally divided into: spruce forest (not on wetland), mixed coniferous forest (not on wetland), mixed forest (not on wetland), spruce forest (on wetland), mixed coniferous forest (on wetland), and mixed forest (on wetland). These were merged to the three types and information about position at wetlands omitted. The soil types were merged into eight types based on grain size (Table 1).

For the drought period, the features that included information about previous year such as *distPrevDmg* (see Table 1), were naturally not included. They were included in the normal period. Other features had NA-values due to deficiencies in geodata or other problems. Forest types were encoded with NA if the land cover class did not include spruce trees. Another reason for NA's for this feature was unsuccessful classification in the land cover data, most probably due to clouded remote sensing images. Some smaller islands were encoded with the value 'open water' in the soil wetness feature since resolution was too coarse to account for these. The 'open water' areas were removed from the data and then the soil wetness values were encoded as NA for these smaller islands.

Feature scaling was used, especially to prepare the data for the SVM algorithm as this model is sensitive to data on different scales. The VIF-score of features were used to assess the degree of multicollinearity, especially important for LR and its assumptions. Features with high VIF-scores were omitted for the LR algorithm.

Geographical coordinates were not used as a feature in the modelling as the geographical dimension was assumed to be included when going from global to more local models.

Table 1. Overview over data sources and variables in dataset

Data source	Variables	Units	Data
Harvester data	X Y	Meter, SWEREF99 TM Meter, SWEREF99 TM	Position Position
	bbPresence	Binary: 0= healthy, 1 = infested	Field records of Bark Beetle infested trees
Performed clearcuts [dataset] (Skogsstyrelsen, 2020) Raster	distToCC5	meter	Distance to clear cut conducted within 0 to 5 years.
	distToCC10	meter	Distance to clear cut conducted within 6 to 10 years.
Digital elevation model [dataset] (Lantm¨ateriet, 2020)	aspect	degrees	Sun exposition
	dem	meters	Elevation above sea level
	slope	degrees	Equal to the rise divided by the run.
	landforms	discrete classes	1=Depression, 2=Lower Slope, 3=Flat Area, 4=Middle Slope, 5=Upper Slope, 6=Upland
Soil wetness (Ågren et al., 2021)	soilWetness	Index 0-100	Relative soil-wetness in 'normal' circumstances
National landcover data [dataset] (Naturvårdsverket, 2018)	distForestEdge	meters	Distance to closest edge between forest and non-forest landcover class
	forestType	discrete classes	1 = Spruce forest, 2 = Mixed coniferous forest, 3 = Mixed Forest,
Harvester data	prevY_f11Sum	Total sum	Total sum of trees removed previous year in a 11×11 pixels neighborhood. Included only in normal period data.
	prevY_f21Sum	Total sum	Total sum of trees removed previous year in a 21×21 pixels neighborhood. Included only in normal period data.
	distPrevDmg	meter	Distance to closest pixel where infestations were recorded during drought period. Included only in normal period data.
Digital Forest Map [dataset] (Swedish University of Agricultural Sciences, 2015)	basalArea	Cubic meters/ha (m2/ha)	Area occupied by stems at height of 1.3 meters
	biomass	Tons of dry matter per hectare (t/ha)	Biomass of all vegetation. Deviating from volumetric biomass, branches are included. Does not include roots or tree stumps.
	canopyHeight	decimeters (dm)	Mean height of the canopy
	spruceVol	cubic meters per hectare (m3/ha)	Includes spruce stem volume over normal tree stump height, tree top and the bark.
Soil type [dataset] (Geological Survey of Sweden, 2015)	soilType	discrete classes	1=Organic, 2=Clay, 3=Silt, 4=Sand, 5=Gravel, 6=Moraine, 7=Rock, 8= Unspecified Sediment

2.3 Machine learning methods

The first step was to decide which type of ML task this was. Since the dependent variable was labelled and class-based – whether bark beetle is present or not – this was an example of a supervised classification task (Géron, 2019). Supervised methods are used when we have input/output pairs in the data and want to predict an output from a certain input (A. C. Müller & Guido, 2016). Model performance in this study was mainly defined as prediction accuracy – the number of samples that are correctly classified divided by total number of samples (Cracknell & Reading, 2014).

The non-parametric ML algorithms used in this study included RF and SVM. These are versatile and popular algorithms that can be used for both classification and regression tasks (Géron, 2019; A. C. Müller & Guido, 2016). These algorithms can learn from complex, non-linear data and should therefore be suited for the problem in this study. LR is a linear classification algorithm and very fast to work with. It was included to see how well it performed compared to more advanced, non-linear models. In the following each of these algorithms are presented but without putting too much emphasis on the mathematical part.

2.3.1 Random Forest

RF is an example of an ensemble algorithm that aggregates the results from multiple simpler estimators called decision trees (VanderPlas, 2016). A decision tree (DT) is a ML model on its own, but they often have problems with overfitting and therefore poor generalization performance. They tend to work well on the training data but worse on unseen test or validation data (A. C. Müller & Guido, 2016). DTs can model the data closely since they are non-parametric and therefore the number of parameters depends on the data, a posteriori (Géron, 2019). The strength of DT's is that they are very intuitive (VanderPlas, 2016). They arrive at their predicted output by establishing a series of if-else questions, also called tests, and answering these with the feature values (A. C. Müller & Guido, 2016). The crucial part of the training is to ask the right questions (VanderPlas, 2016). If the complexity of DTs is high and they contain pure leaves, meaning a 100% prediction accuracy on the training data, the final predictions are often more a result of noise or outliers than the true data distribution (VanderPlas, 2016). This will most definitely lead to overfitting when looking at unseen data, especially if outliers are present (A. C. Müller & Guido, 2016). To regularize DTs and diminish overfitting, different hyperparameters can be tweaked, such as restricting the depth of the tree (reducing the number of questions asked) or setting the minimum number of points in a node before splitting it, so data is not split into increasingly small leaves (Géron, 2019; A. C. Müller & Guido, 2016).

RF is designed to decrease the problem of overfitting by combining multiple DTs. They are trained on different random subsets of the data (Géron, 2019). Since only a subset of the training data is used for each tree, the bias is increased but aggregating the trees into a forest, reduces both variance and bias (Géron, 2019). Averaging the results in case of regression or using the majority vote in case of classification from slightly different DTs will diminish overfitting and result in more robust models that will work better on unseen data (Georganos et al., 2021; A. C. Müller & Guido, 2016). RF has both tree-specific hyperparameters which are the same as for the DTs and at the same time hyperparameters that are ensemble-specific (Géron, 2019). The most important parameters are the number of trees to grow, which should generally be set as high as computationally feasible to result in more robust results, and the number of features in each node of the tree (Georganos et al., 2021; A. C. Müller & Guido, 2016).

Even though RF seems superior to DT performance-wise, DTs are still used for easy visualization of the tree structure, showing how a prediction was made (A. C. Müller & Guido, 2016). DTs are white box models compared to the black box models of RFs. In general RF makes good predictions but it is much more difficult to explain how the model arrived at a result (Géron, 2019). RFs are easy to use and normally the hyperparameters do not need to be tuned to a very high degree (A. C. Müller & Guido, 2016).

2.3.2 Support Vector Machine

SVM is another type of powerful, supervised ML algorithm that can be used for both classification and regression tasks. In classification problems, SVMs use discriminative classification, meaning that a line, curve or manifold is used to divide the classes from each other in space (VanderPlas, 2016). SVMs are binary classifiers and therefore suited to the data in this study as the dependent variable only can take two values (Géron, 2019).

Together with LR, linear SVM is one of the most common linear classification algorithms (A. C. Müller & Guido, 2016). SVMs are maximum margin estimators, which means that in linear problems, a line is drawn that maximises the margin between the two classes (VanderPlas, 2016). The key here is the support vectors that lie on this margin. All the points further away from the margin do not influence the algorithm as long as they are placed on the correct side of the line (VanderPlas, 2016). SVMs can be extended to more efficiently handle non-linear data by projecting the data into a higher dimension (VanderPlas, 2016), thereby making the classes linearly separable in the higher dimension.

For linear SVMs, the hyperparameter that determines the degree of regularization is C. For low values of C, the model generalizes better and puts more emphasis on the majority of the points, whereas a high values of C puts more emphasis on the single points, thus trying to fit the training data as well as possible (A. C. Müller & Guido, 2016). This can lead to overfitting and make the model perform worse on unseen data. Other hyperparameters that are important for non-linear SVMs, are the gamma hyperparameter for the gaussian kernel, which determines the radius of the kernel and the influence of nearby points. If the gamma is set low, the radius is high and many points are considered important for building the decision boundaries, often leading to a more general model, prone to underfitting (A. C. Müller & Guido, 2016)

SVMs have potentially more weaknesses than RFs. Working on big datasets, the computation time, especially for non-linear SVMs, can be very high (A. C. Müller & Guido, 2016; VanderPlas, 2016). Furthermore, the data must be scaled, which is not a prerequisite of RF.

SVMs are also very sensitive to choice of hyperparameters, and cross-validation or grid search could be performed, which would also increase the cost and time of the modelling (Géron, 2019; VanderPlas, 2016). That being said, a SVM was chosen here, because they are very versatile and once trained the prediction phase can be very fast, making them suitable for predictions on unseen data (VanderPlas, 2016).

2.3.3 Logistic Regression

LR was chosen to be able to compare a simpler, linear model to the non-parametric models described above. LR is despite its name a classification algorithm, that tries to estimate the probability that an instance belongs to one of two classes, thus making it a binary classifier (Géron, 2019). It can be extended to a multinomial LR where the dependent variable can fall in more than two classes (Ae, 2013). The probability is found through the fitting of a logistic sigmoid function to the relationship between independent and dependent variables (Ae, 2013). This S-shaped curve has values between 0 and 1 and the outputted number indicates the probability of a positive event (Ae, 2013). The parameter C determines the regularization of the model, just as in the case of SVM. So, if the model tends to overfit, a good procedure would be to lower this parameter (A. C. Müller & Guido, 2016). Especially in the case of multivariable analysis with many features in high dimensions such as in this study, being aware of overfitting is important (A. C. Müller & Guido, 2016).

LR has been included in this study since they are fast to train and use for prediction. They are especially suited for large datasets, that would otherwise often lead to very long computation times for non-parametric models. Sometimes linear models are simply used due to computation and time constraints, even though they expectedly result in worse prediction accuracies on non-linear problems (A. C. Müller & Guido, 2016). And of course, this is the big problem with these types of models – that they will often not result in very accurate predictions on real-world, complex, and non-linear data.

LR makes fewer assumptions about the data than linear regression. No assumptions are made that the relationships between dependent and independent variables are linear (Ae, 2013). LR can furthermore handle both continuous and categorical input data (Ae, 2013). That said, there are other assumptions that are problematic in relation to the data in this study, including independence of samples and multicollinearity between features. The latter can be checked with Pearson's correlation coefficient or preferably variance inflation factor (VIF) that shows how much the variance of the estimated coefficient is influenced by multicollinearity (Senaviratna & Cooray, 2019). If multicollinearity is indicated, omitting or combining correlated features or increasing sample size are possible solutions (Senaviratna & Cooray, 2019).

2.4 Modelling working steps

A flowchart of the working steps can be seen in Fig. 2.

2.4.1 Global modelling

The first step was to run the global models for the normal period data (2019 and 2020). The global models were run on the entire dataset. The models were initially run ‘quick and dirty’, meaning without fine-tuning the hyperparameters, just to get a feeling which model might yield the best results. Hereafter the hyperparameters leading to the best results were found through cross-validation and grid-search. The ML model that performed best for the normal period was also used for the drought period. The global models (both normal and drought) were assessed through cross-validated scores based on five splits as well as confusion matrices.

Furthermore, partial dependence plots were established for the most important features in the best model to get an understanding of the relationship between these independent variables and the dependent variable. Feature importance for the model with highest accuracy was noted and compared to the results for the local models.

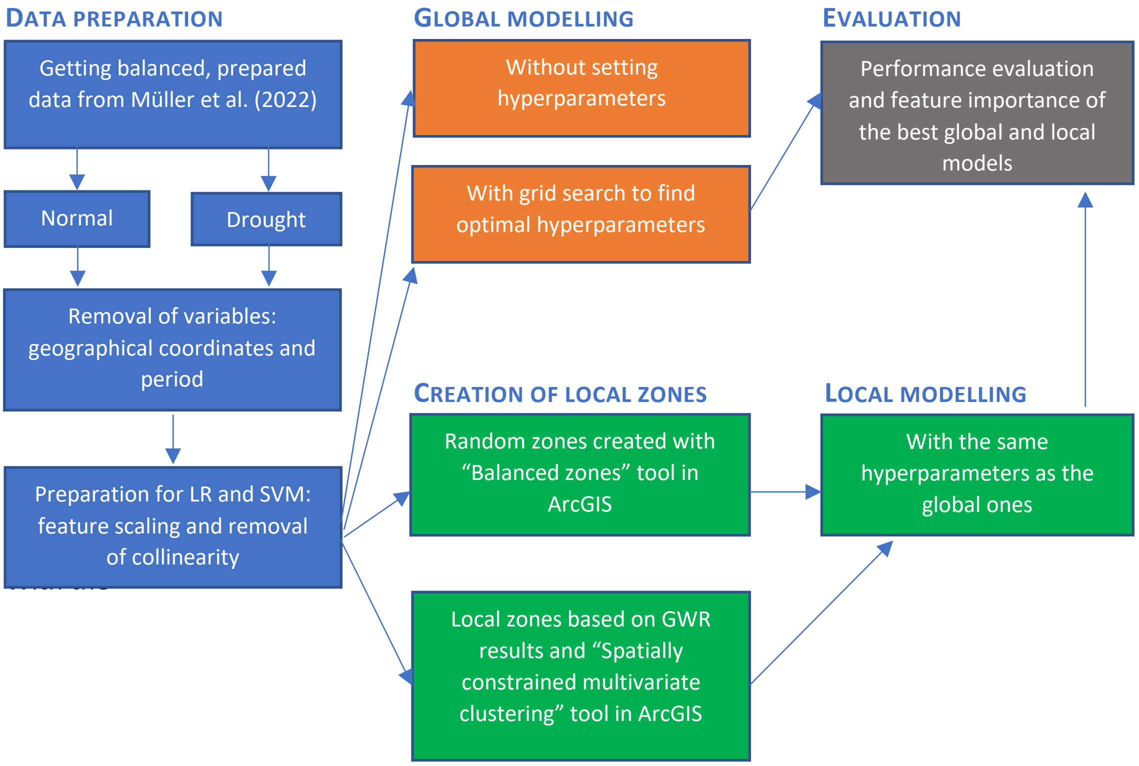

DATA PREPARATION
Getting balanced, prepared data from Müller et al. (2022)
Normal
Drought
Removal of variables: geographical coordinates and period
Preparation for LR and SVM: feature scaling and removal of collinearity
GLOBAL MODELLING
Without setting hyperparameters
With grid search to find optimal hyperparameters
EVALUATION
Performance evaluation and feature importance of the best global and local models
CREATION OF LOCAL ZONES
Random zones created with “Balanced zones” tool in ArcGIS
Local zones based on GWR results and “Spatially constrained multivariate clustering” tool in ArcGIS
LOCAL MODELLING
With the same hyperparameters as the global ones

Figure 2. Flowchart over methodology. Data preparation was not included to a high degree, since the prepared data from M. Müller et al. (2022) was used. The local modelling was only performed with the ML algorithm that resulted in the best global results. Blue boxes = data preparation (for both global and local models), orange boxes = global modelling, green boxes = local modelling, grey box = evaluation.

2.4.2 Local modelling

As an initial step to see whether local models could be expected to perform better, a spatial autocorrelation analysis was performed in ArcGIS Pro through the *Moran's I global test.* To run it more smoothly the point data was collapsed into a 400x200 pixel fishnet and aggregated, so the proportion of bark beetle presence in each polygon was calculated. The size of the fishnet was determined through trial and error, both bigger and smaller polygons resulted in error and therefore 400x200 was used. The distance method was set to Euclidean and the conceptualization of spatial relationships to inverse distance ("Spatial Autocorrelation (Global Moran's I) (Spatial Statistics)—ArcGIS Pro | Documentation").

Initially the idea was to perform the local modelling in the same way as in earlier spatial ML studies (Georganos et al., 2021; Quiñones et al., 2021; Santos et al., 2019) where a dynamic approach was used to combine GWR with RF. These studies were based on raster or polygon-data, for instance counties in the US (Quiñones et al., 2021) with fewer data records compared to the current study. The more dynamic approach creates local models for each data record only including neighbours to this specific data record. This can result in very long computing times on big datasets, since the same number of models as data points in the dataset basically have to be created. The simpler approach used in the current study was to divide the full dataset into geographical entities (spatially contiguous zones) in two different ways in ArcGIS: random zones and GWR zones.

2.4.2.1 Random zones

The first approach was based on the *Balanced zones* tool in ArcGIS, that divides the study area into spatially contiguous zones based on the data in the area. The criteria used here was 'defined number of zones', with 2, 4, 6, 8, 10 and 15 zones created. This option is useful if the aim is to make zones with the same number of records ("How Build Balanced Zones works—ArcGIS Pro | Documentation"). The bark beetle presence variable could have been considered to try to keep a relative balance between infested and non-infested records inside each zone. In the end it was decided to keep things simple and not use the dependent variable for the zone-building. Including attribute criteria would make it much more difficult to create zones with approximately equal number of records, and this was the aim here.

2.4.2.2 GWR zones

The second method was generally more sophisticated and based on GWR results. This one was also performed in ArcGIS where GWR can be calculated. Since GWR is a linear model, problems can exist in case of multicollinearity, and this proved to be the case here. After removing problematic features (canopy height, basal area, and biomass for drought period and *distToCC10*, aspect, slope, *prevY_f11sum*, *prevY_f21sum*, basal area, biomass, canopy height and soil type for normal period), a distance-based kernel of 50 km, also called a fixed kernel, was used. The model type was specified as binary. The weighting scheme used was set to bisquare, gradually decreasing the influence of nearby-points based on distance in the local models (“How Geographically Weighted Regression (GWR) works—ArcGIS Pro | Documentation”). The output included the percentage of deviance explained by both the global and local model (“How Geographically Weighted Regression (GWR) works—ArcGIS Pro | Documentation”). A higher deviance explained by the local model(s) indicates spatial autocorrelation and the appropriateness of using local models for the problem in question.

The tool *Spatially constrained multivariate clustering* in ArcGIS was used to create the local zones based on the GWR results. This tool creates spatially contiguous zones based on some attribute values (“Spatially Constrained Multivariate Clustering (Spatial Statistics)—ArcGIS Pro | Documentation”), in this instance the local percent deviance scores. No cluster size constraints were used, and the number of clusters was set to respectively 2, 4, 6, 8, 10 and 15, the same as for random zones. This meant that, in opposite to the random zones, the local zones would generally include different number of records. By not setting any constraints (for instance the minimum number of records in each zone), the local zone-building was based solely on geographic contiguity as well as the local percent deviance scores, making it possible to create distinct zones where the scores were relatively similar.

2.4.2.3 The modelling of the local models

The final step was to run the models in the created zones. As already said, It is important to acknowledge that the term local models differ here compared to the other studies already mentioned (Georganos et al., 2021; Quiñones et al., 2021; Santos et al., 2019). Here it is more understood as local zones where one model was run for the entire zone.

The same hyperparameters were used in the local zones as in the global model, meaning if the best global model was RF, the local models were run with RF with the same hyper-parameters as in the global model. Also, five-folded cross-validation was used for the accuracy assessments as well as confusion matrices. The final and overall accuracy assessment for the local models was performed by weighing the cross-validation accuracy scores by the size of the zones (number of records in each). The best results for the normal and drought period were mapped, including the feature importance for each of these zones.

3. Results

3.1 Data exploration

The spatial proportions of bark beetle infestations in the normal and drought period are included (Fig. 3 and 4). These maps are based on the full data set (rather than the balanced, stratified dataset) to get an overview on the true extent of the problem and the spatial diversion. The identical maps based on the stratified and balanced data are included in the appendix (Fig. 17). It should be noted here that this is based on the available data only, and there would be bark beetle attacked trees, that we do not know about in the study area.

Instead of showing the full data set (2,659,545 records in normal period, and 1,356,357 records in drought period), a fishnet was created (100x50 pixels) and the proportion of bark beetle inside each polygon calculated. This fishnet size was deemed best for visualization purposes. Since the proportion of bark beetle is limited in both the normal and drought datasets, the majority of polygons were characterized by a low proportion (below 7%) in both maps (Fig. 3 and 4). The majority of bark beetle presence in the normal period was found in the south-eastern part of the study area, whereas some clusters were found in the northern and southern part in the drought period. More single polygons with higher bark beetle presence were found in the normal period compared to the drought where more clusters were found.

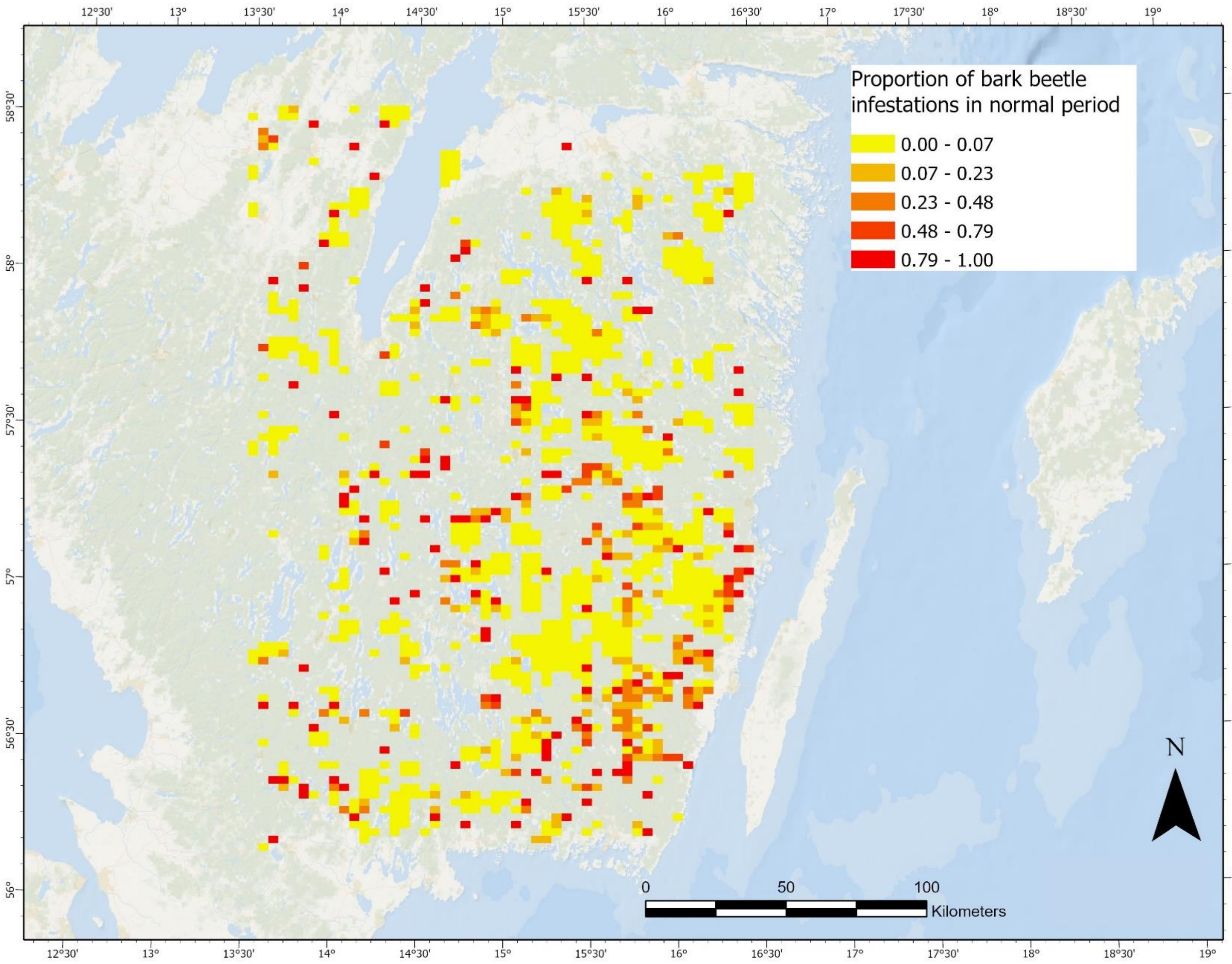


Figure 3. Proportion of infested trees from bark beetle (based on the full dataset) for the normal period. Classification based on Natural Breaks (Jenks).

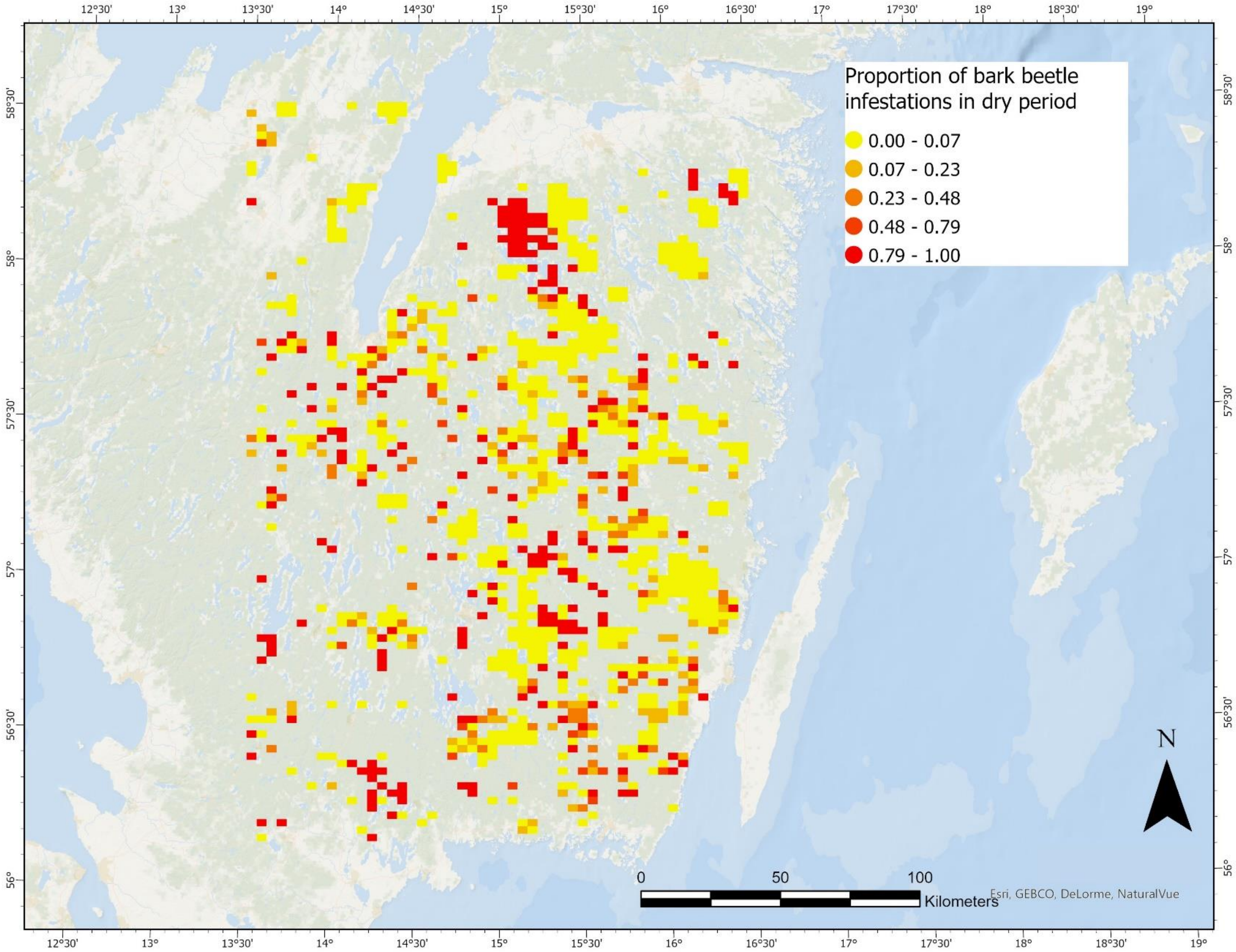


Figure 4. Proportion of infested trees from bark beetle (based on the full dataset) for the drought period. Classification based on Natural Breaks (Jenks).

3.1.1 Correlations

Fig. 5 and 6 show the correlations between the numerical variables in the two balanced data sets (normal and drought period). In the rest of the study the balanced datasets were used since the modelling was based on these. It was most interesting to look at the correlations between bark beetle infestations and the explanatory features, but intercorrelations between the features could show important properties such as multicollinearity.

Generally, the correlations did not differ to a high degree between the two periods (normal and drought). The tree characteristic features (basal area, biomass, canopy height, spruce volume) had a positive relationship with bark beetle presence. The correlations between these were very high, and multicollinearity must be considered. Omitting one or more of these features would probably not decrease model performance. The tree characteristic features generally had a negative relationship with the other features, such as elevation and soil wetness, indicating that in drier and lower-lying areas the tree characteristic features could be higher.

The elevation related features generally had a negative relationship with bark beetle presence, but the effect was higher in the normal period, especially for elevation. The relationship to the tree characteristic features could also partly explain this.

Soil wetness was negatively correlated to bark beetle presence in both periods, indicating that soil dryness could be a problem in both normal and drought periods. Wetness in the soils has a mitigating effect, lowering the amount of bark beetle presence. Distance to previously clear cut areas, *distToCC5* and *distToCC10* (see Table 1), had a negative effect on bark beetle especially in the drought period where both these features had a relatively strong negative correlation.

To sum up, the same correlation patterns were seen between the two periods, but elevation showed a more remarked negative effect on bark beetle presence in the normal period. It should be noted here that the main aim of this study was not to make a thorough feature analysis per se, but instead to assess model performance and the importance of modelling scale. Therefore, no figures or maps showing the extent of the features were included here. On the other hand, M. Müller et al. (2022) was more focused on analysing the explanatory features for bark beetle attacks, in the study area, based on the same data.

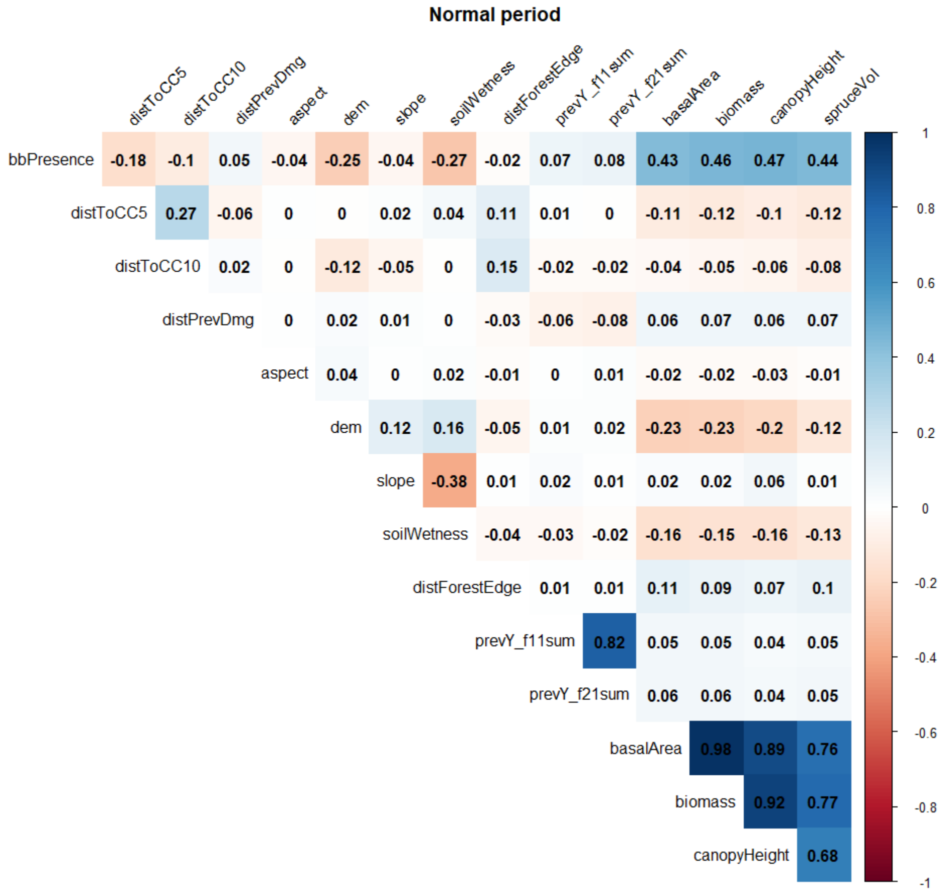


Figure 5. Correlations between numerical variables, normal period

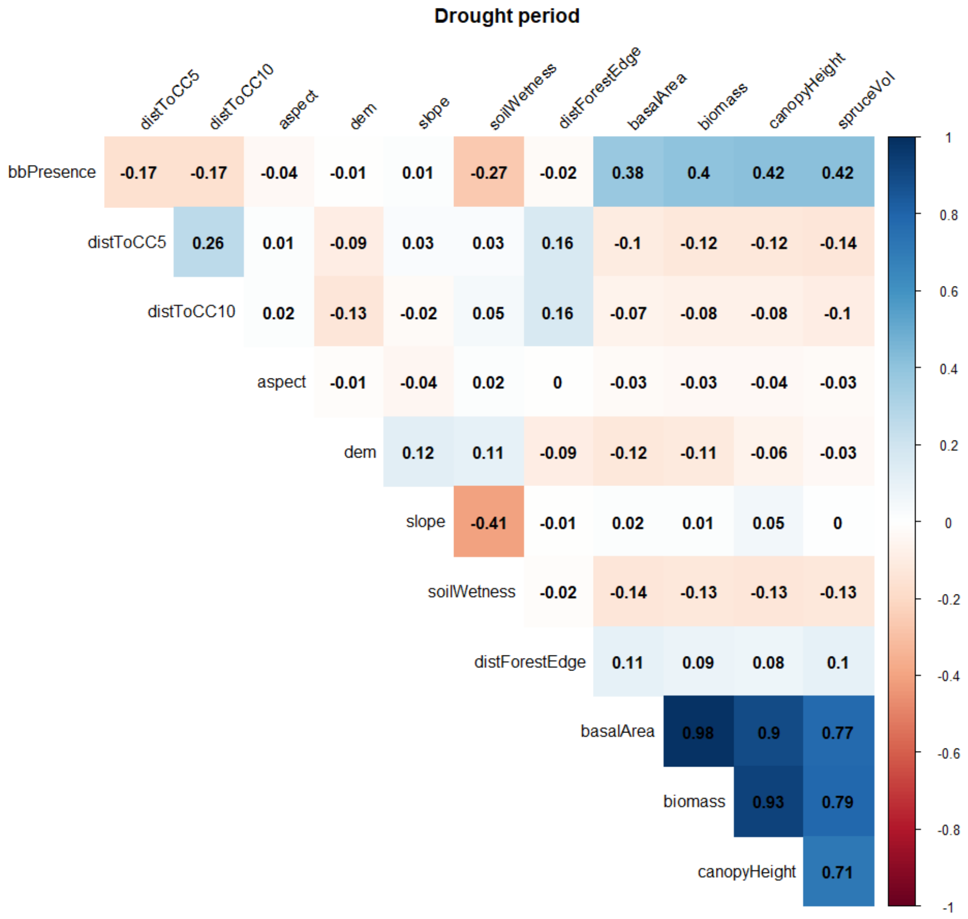


Figure 6. Correlations between numerical variables, drought period

3.1.2 Spatial autocorrelation and GWR results

Spatial autocorrelation was calculated both through the *Moran's I global* test and the subsequent GWR results. Both were calculated in ArcGIS Pro. For both the normal and drought period, Moran's I showed significant spatial autocorrelations ($p < 0.001$) with positive Moran's I-scores, indicating that bark beetle presence was clustered (Fig. 18 and 19 in the appendix). Based on these results, a local modelling approach that better incorporate spatial autocorrelation could be expected to perform better than a global model.

For both periods, the local models performed better than the global ones and explained more of the deviance, which also indicated spatial autocorrelation (Table 2). The results also showed that the effect was more pronounced for the drought period, but only marginally. By comparing the residual sum of squares, the improvement in performance when going from global to local was 19.2% for normal period and 21.6% for drought period ("How Geographically Weighted Regression (GWR) works—ArcGIS Pro | Documentation"). The reason for this could also be explained by the fact that more features were included in the drought model, as multicollinearity was more of a problem in the normal period dataset. For instance, the soil type variable was omitted in the normal period model (Table 2).

The GWR score is based on a linear (OLS) model. A complex problem like the one assessed here consisting of a large number of records and many different environmental and geospatial features is not necessarily well explained by a linear model. None the less, this served as a good starting point to divide the study area into zones with different deviance scores – areas with higher local deviance scores should perform better when modelling on a more local scale.

Table 2. GWR outputs for normal and drought period, calculated in ArcGIS.

For both models	
Distance band	50 km
Dependent variable	Bark beetle presence
Model for normal period	
Explanatory variables	• DistToCC5 • DistPrevDmg • Dem • Landforms • Soil wetness • DistForestEdge • Forest type • Spruce vol
Deviance explained by the global model (non-spatial):	0.27
Deviance explained by the local model:	0.41
Deviance explained by the local model vs. global model:	0.19
Model for drought period	
Explanatory variables	• DistToCC5 • DistToCC10 • Aspect • Dem • Landforms • Slope • Soil wetness • DistForestEdge • Forest type • Spruce vol • Soil type
Deviance explained by the global model (non-spatial):	0.22
Deviance explained by the local model:	0.39
Deviance explained by the local model vs. global model:	0.22

3.2 Global modelling results

3.2.1 Model performance and hyperparameters

Model runs were made with the three ML models on the normal period data and hyperparameters were tuned with the help of grid search. The model performance was assessed through five-fold cross validation. In general, all the models performed quite well, with the non-linear models performing better than the linear ones (Table 3). The best model performance was found with RF, although SVM and RF performed similarly after hyperparameter tuning. The improvement from hyperparameter tuning was only marginal for RF whereas it was more pronounced for SVM. In the end, the RF was deemed to be the best performing model for the normal period and used both for the global drought modelling and the subsequent local modelling. RF performed worse in the drought period compared to the normal period but still had a rather high prediction accuracy.

Table 3. Global model performance (prediction accuracy) with and without hyperparameter tuning (HPT)

Normal period	RF	LR	Linear SVM	SVM
Without HPT	0.87	0.76	0.79	0.83
With HPT	0.89 {'bootstrap': False, 'max_depth': 60, 'max_features': 'auto', 'min_samples_leaf': 1, 'min_samples_split': 3, 'n_estimators': 200}	0.76 {'C': 1.0, 'penalty': 'l1', 'solver': 'liblinear'}	0.79 {'C': 1.0, 'penalty': 'l2'}	0.88 {'C': 10, 'gamma': 1, 'kernel': 'rbf'}
Drought period	RF			
Without HPT	0.83			
With HPT	0.84 {'bootstrap': False, 'max_depth': 70, 'max_features': 'auto', 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 200}			

3.2.2 Feature importance for global models

Feature importance was calculated for both the normal and drought period to assess whether there were differences and to see what features could be more important in future drought years compared to wetter years (Fig. 7 and 8). Partial dependence plots were made for the six most important features in each period to get a deeper understanding of the relationship between these features and bark beetle presence (Fig. 9 and 10). Only two tree characteristic features were included in the partial dependence plots as these features were expected to have a rather identical relationship to bark beetle presence and therefore it was more interesting to include other features.

The most important features included tree characteristics (only 2 included here but more were actually in top six), elevation, soil wetness, and *distToCC5* (see Table 1) for both models. So generally, the most important features did not differ remarkably between the two periods. In the normal period also the *distPrevDmg* (see Table 1) was included.

3.2.2.1 Tree characteristic features and elevation

As for the correlations, the importance of the tree characteristic features was generally similar for the two periods (Fig. 7 and 8). These features had high importance, especially canopy height. As can be seen on the partial dependence plots (Fig. 9 and 10) the relationship was positive for canopy height and spruce volume, but the positive effect seemed to weaver off for both features in both periods. Canopy height had a very steep increase from around 170 dm, resulting in an approximately doubling of bark beetle presence at canopy heights around 220 dm in the normal period. The relationship between spruce volume and bark beetle presence was more linear but the marginal increase was higher in the drought period at higher spruce volume values whereas the plateau was reached earlier in the normal period (Fig. 9 and 10).

Elevation was important in both periods, but especially in the normal period (Fig. 9 and 10). The lowest areas in the normal period had the highest bark beetle presence after which a steep decrease happened around 50 m.a.s.l. From around 80 m.a.s.l. the rate of change was lower with the overall trend being a slight decrease. A trend was not obvious in the drought period where an increase was followed by a decrease and the partial dependence returned to almost the same value.

3.2.2.2 The other features

Soil wetness was more important in the drought period, but still included as one of the top six features in the normal period (Fig. 9 and 10). This was not surprising and in line with the correlation values. At dry soils the bark beetle presence was generally higher in the drought period. The marginal effect of soil wetness was higher for low increases in the drought period, indicating that even a slight increase in soil wetness can have an effective mitigating effect. The partial dependence decreases more steadily in the normal period.

The distance to previous attack (*distPrevDmg* – see Table 1), only included in the normal period, had a low correlation with bark beetle presence but relatively high importance in the model (Fig. 7). The reason could be that this feature was not correlated to the other explanatory features and therefore had a unique relationship to bark beetle presence. The partial dependence plot showed a rather trendless relationship between the feature and bark beetle presence, but at lower levels there was more going on with a decrease and subsequent increase in partial dependence (Fig. 9). Since most of the records had rather low values of this variable and 60% of the records were found in this interval (approx. 0-2000 m) this could explain the big importance in the model. With increasing distance, the partial dependence dropped in the beginning and subsequently increased. The distance features – *distToCC5* and *distToCC10* (see Table 1) were less important and decreased rather linearly in both periods (Fig. 9 and 10).

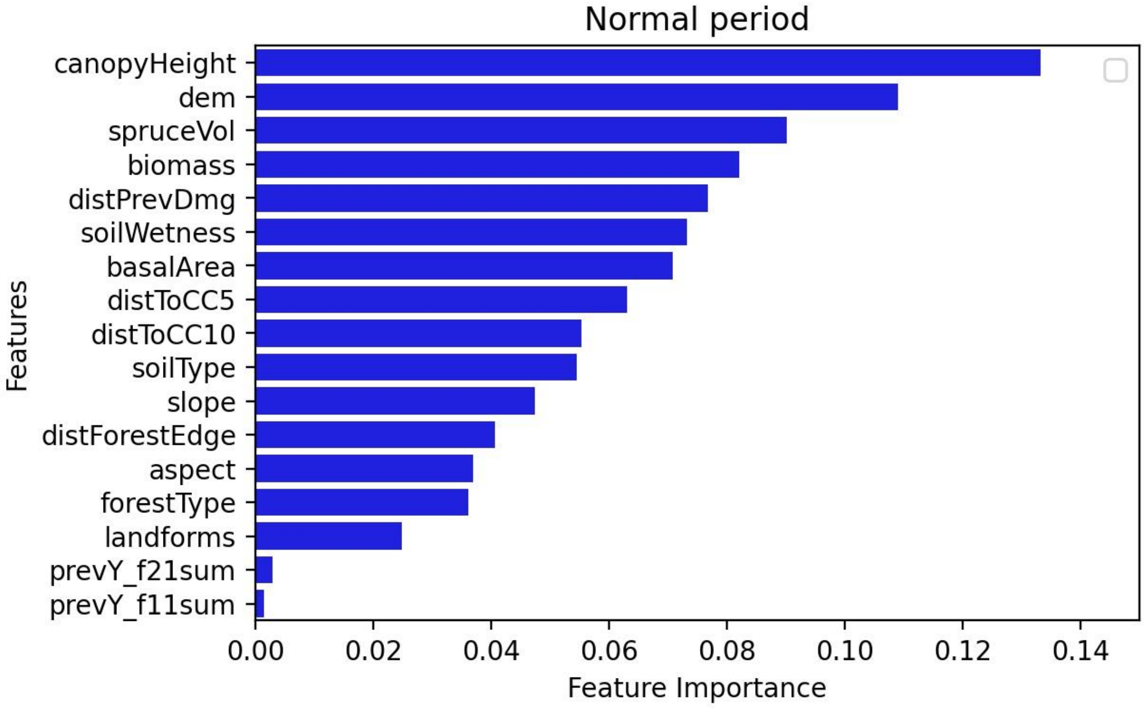


Figure 7. Feature importance for normal period (global model, RF).

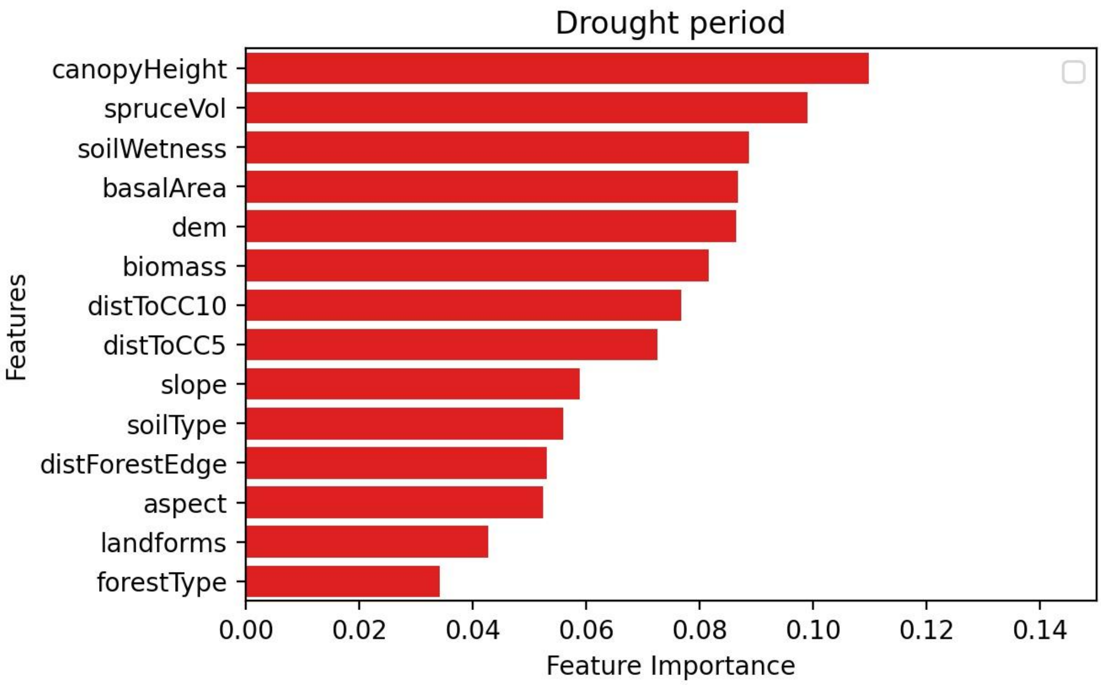


Figure 8. Feature importance for drought period (global model, RF).

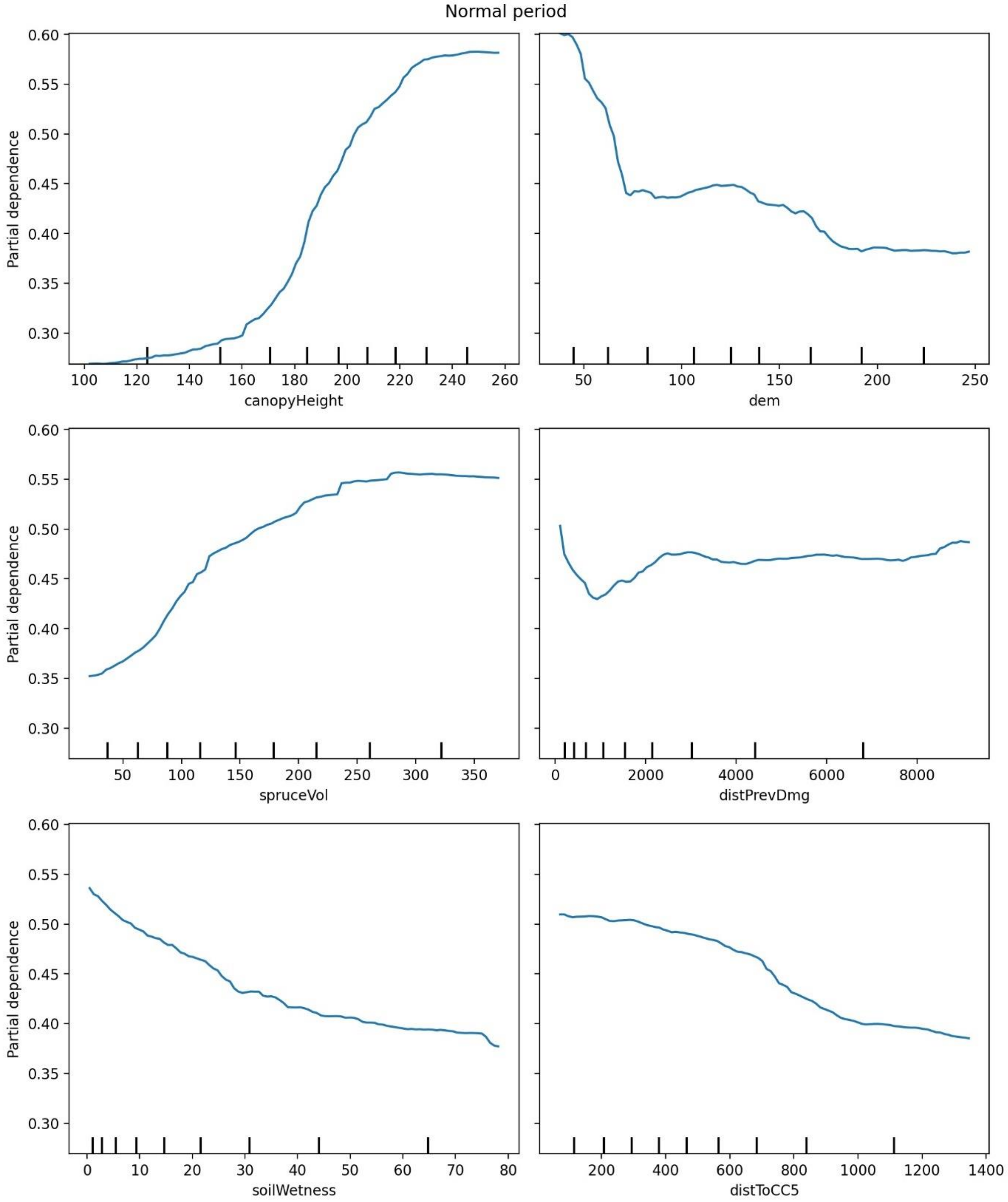


Figure 9. Partial dependence plots for the 6 most important features (organized after importance), normal period (global model, RF). Only two tree characteristic features included.

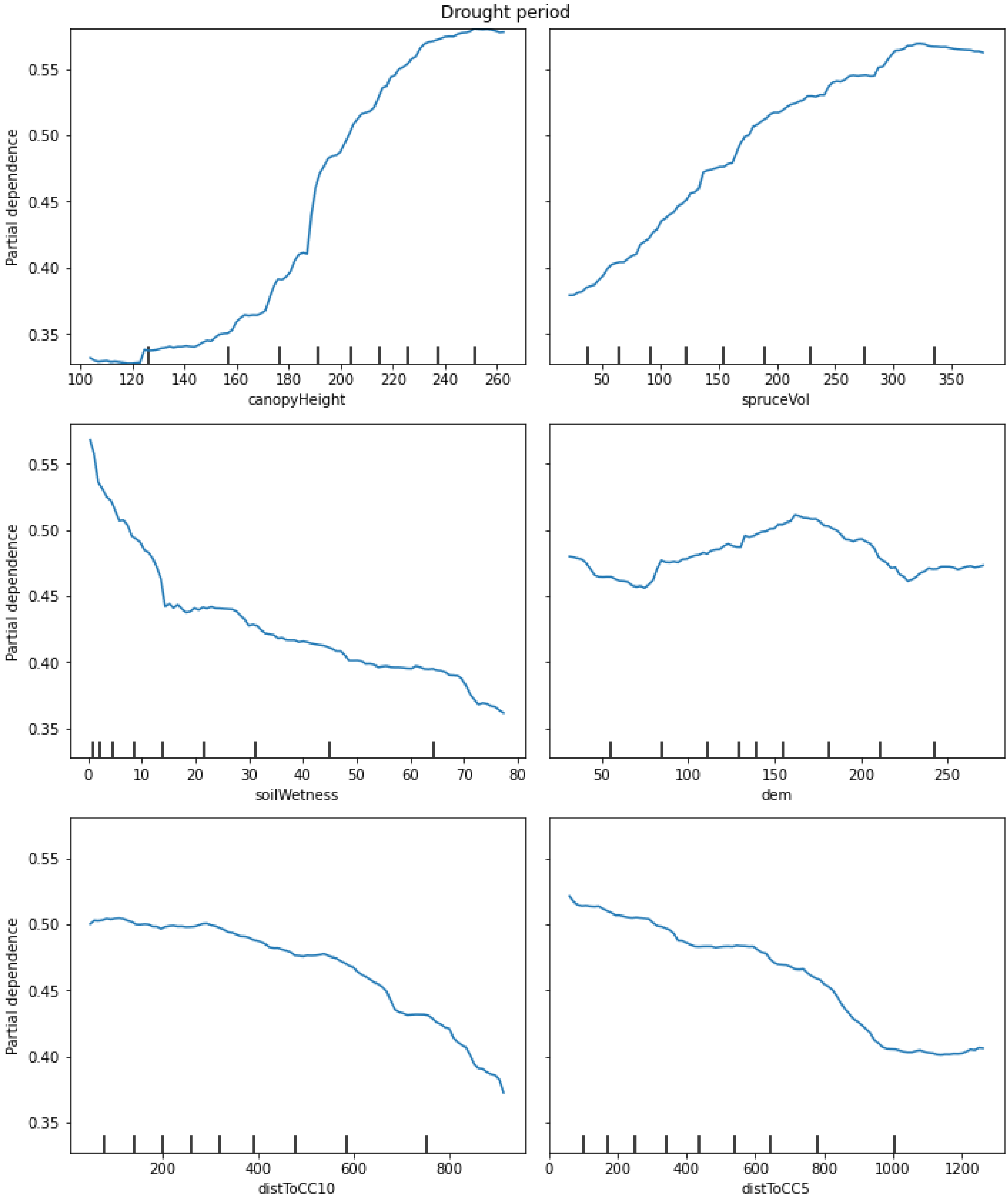


Figure 10. Partial dependence plots for the 6 most important features (organized after importance), drought period (global model, RF). Only two tree characteristic features included.

3.3 Local modelling results

3.3.1 Overall local model performance

The general picture was the same for the normal and drought local models with overall, local models performing better than the global counterparts. With an increasing number of local zones, the overall performance (the average of the performances inside each zone weighted after the number of records in the zones) generally increased for both the random zones and the zones based on GWR results (Fig. 11 and 12). For the normal period, all the local models based on random zones performed better than the models based on GWR zones. The local models based on 10 zones actually performed worse than the ones based on 8 zones, but accuracy increased again with 15 zones (Fig. 11). For the drought period, the random local models performed best at lower number of zones but GWR performed better at 15 zones (Fig. 12). It must be noted, that even though an increase in performance was seen with more zones, the absolute increase was generally marginal.

The local models were based on RF. Due to the vast number of different results, only the best performing local models for both the normal and drought period and their characteristics are presented here under. The best performing local models were the random model based on 15 zones for the normal period (Fig. 11), and the GWR model based on 15 zones for the drought period (Fig. 12). From now on, these two will be described as *Random 15 normal* and *GWR 15 drought* for simplicity. Once again, it should be noted that the performances were in general similar, especially from 8 zones and more. The full results for each zone in *Random 15 normal* and *GWR 15 drought*, including cross-validated performance and feature importance are included in Table 4 and Table 5 in the appendix.

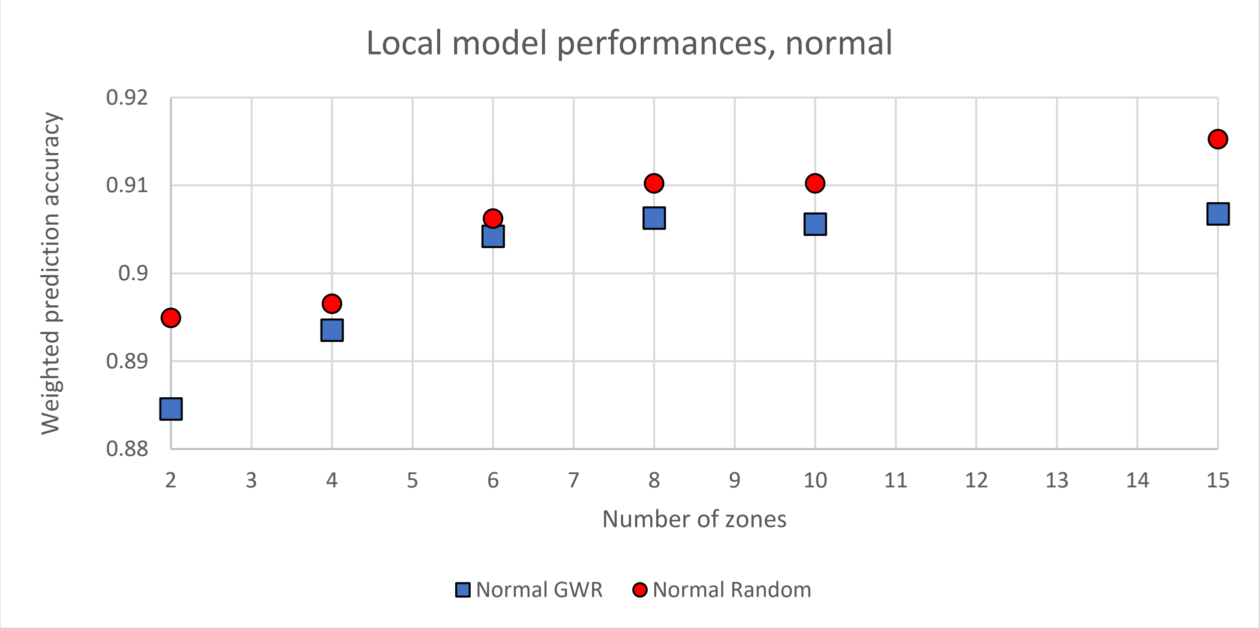


Figure 11. Local model performances, normal period. The figure is based on the number of zones – only 2, 4, 6, 8, 10, and 15 zones were assessed. The prediction accuracy is the average of the accuracy inside each zone weighted after number of records inside each zone.

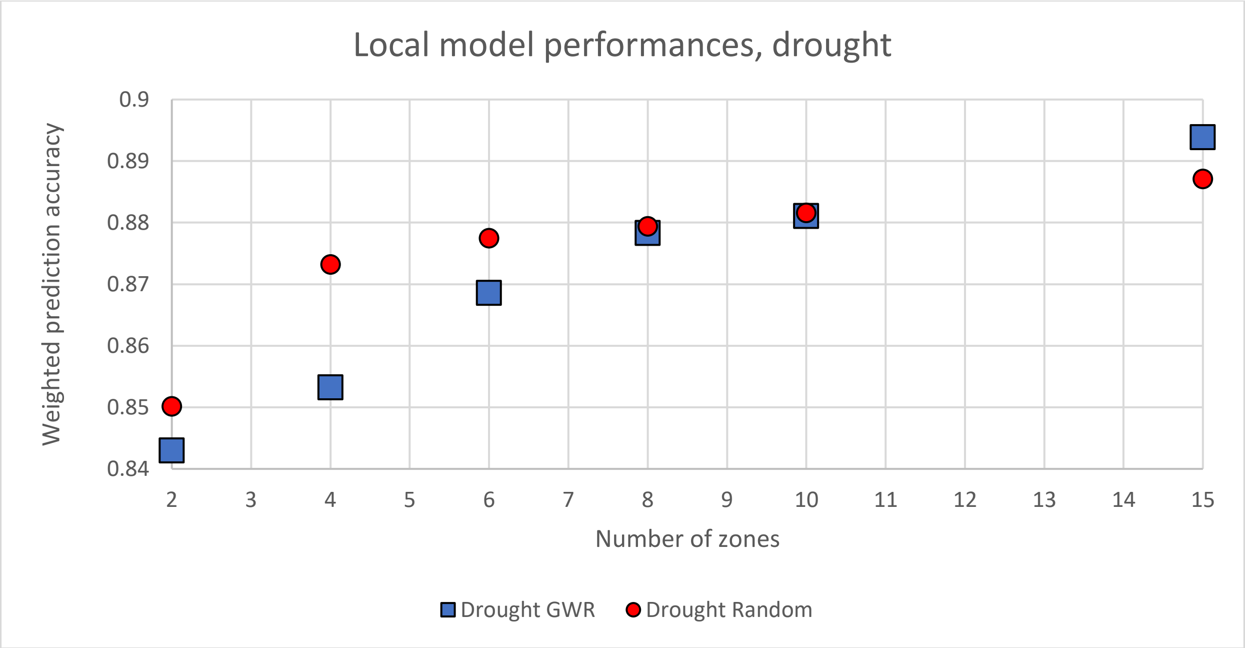


Figure 12. Local model performances, drought period. The figure is based on the number of zones – only 2, 4, 6, 8, 10, and 15 zones were assessed. The prediction accuracy is the average of the accuracy inside each zone weighted after number of records inside each zone.

3.3.1.1 Spatial model performance

The spatial distribution of local model performance in *Random 15 normal* and *GWR 15 drought* is seen in Fig. 13. For both periods, the same trends were seen with the best performing zones located in the north-western and mid-eastern part of the study area. Once again it should be noted that performance did not differ much between the different zones on an absolute scale, but it was interesting that the zones in the western and eastern part (located under the mid-eastern good performing part) were performing even worse than the global model for the normal period.

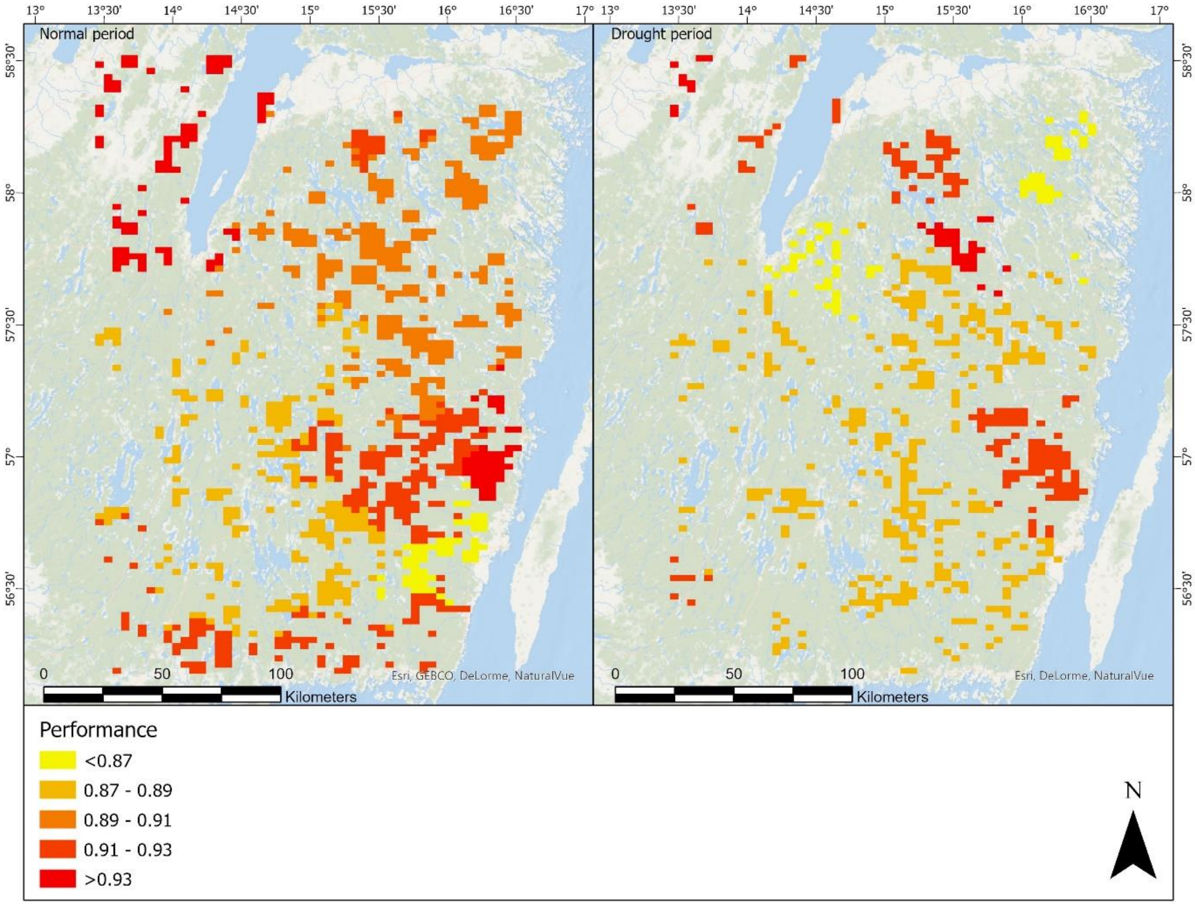


Figure 13. The spatial (local) model performances for *Random 15 normal* (left) and *GWR 15 drought* (right). Classification made with Natural Breaks (Jenks) based on the data distribution of *Random 15 normal* (also used for *GWR 15 drought*).

3.3.2 Local feature importance

Fig. 14 and 15 show the local feature importance of the five most important features for both models (these were defined as the features most often found in top 5 in each of the 15 zones). Since the tree characteristic features were assumed to be correlated as in the global model, only two of these features were included here, so other types could be assessed. The feature importance was generally higher in *Random 15 normal* compared to *GWR 15 drought*. The five most important features inside each zone in the two local models can be seen in Table 4 and Table 5 in the appendix.

3.3.2.2 Local feature importance for the normal model/zones

When looking at the best performing zones in *Random 15 normal*, the tree characteristic features seemed to be less important than they were in the relatively worse performing zones. The elevation, *distPrevDmg*, and *distToCC10* (see Table 1) were more important in these areas (Table 4). The feature importance of *distPrevDmg*'s was high in the north-western and mid-eastern part (Fig. 14). Elevation had a high importance, especially in the middle and southern part of the study area, two areas defined by a relatively high performance.

The worse performing zones in *Random 15 normal* were located in north-east, mid-west, and south-east (Fig. 13). These generally had tree characteristic features as the most important feature, such as canopy height and spruce volume. Spruce volume for instance had a relatively high feature importance in the north-eastern part, whereas canopy height had a high importance in the middle part of the study area (Fig. 14). The patterns are naturally not that clearly defined, spruce volume for instance also showed a high importance in the north-western part – an area defined by high model performance. Soil wetness was generally less important than the other mentioned features with lower importance in areas that performed better such as the mid-eastern part, and a higher importance in areas that performed worse such as the north-eastern part.

3.3.2.2 Local feature importance for the drought model/zones

The best performing zones in *GWR 15 drought* were generally characterized by a high importance for forest type, especially in the north-western and northern part of the study area (Fig. 15).

Tree characteristic features seemed to have a more varied relationship with model performance. Both of these (canopy height and spruce volume) were important in the mid-eastern part, and spruce volume also in the north-western part of the study area. Canopy height on the other hand had low importance in the north-eastern, well-performing part of the study area. There were indications that elevation had a more complicated relationship with model performance in the *GWR 15 drought* compared to *Random 15 normal* with low importance in the north-western part but higher importance in the mid-eastern part (Fig. 15). Elevation therefore also seemed to have a relatively high importance in zones that did not perform that well.

DisttoCC5 (see Table 1) and canopy height also had a higher importance in worse-performing zones – in the middle part of the study area. Even forest type also had a high importance in the middle part (under the lake Vättern, located in the mid northern part of the study area) that performed relatively bad as well as the southern part that also did not perform as well as the best areas. In general, it seemed more difficult to make some general statements about the local feature importance in the drought model.

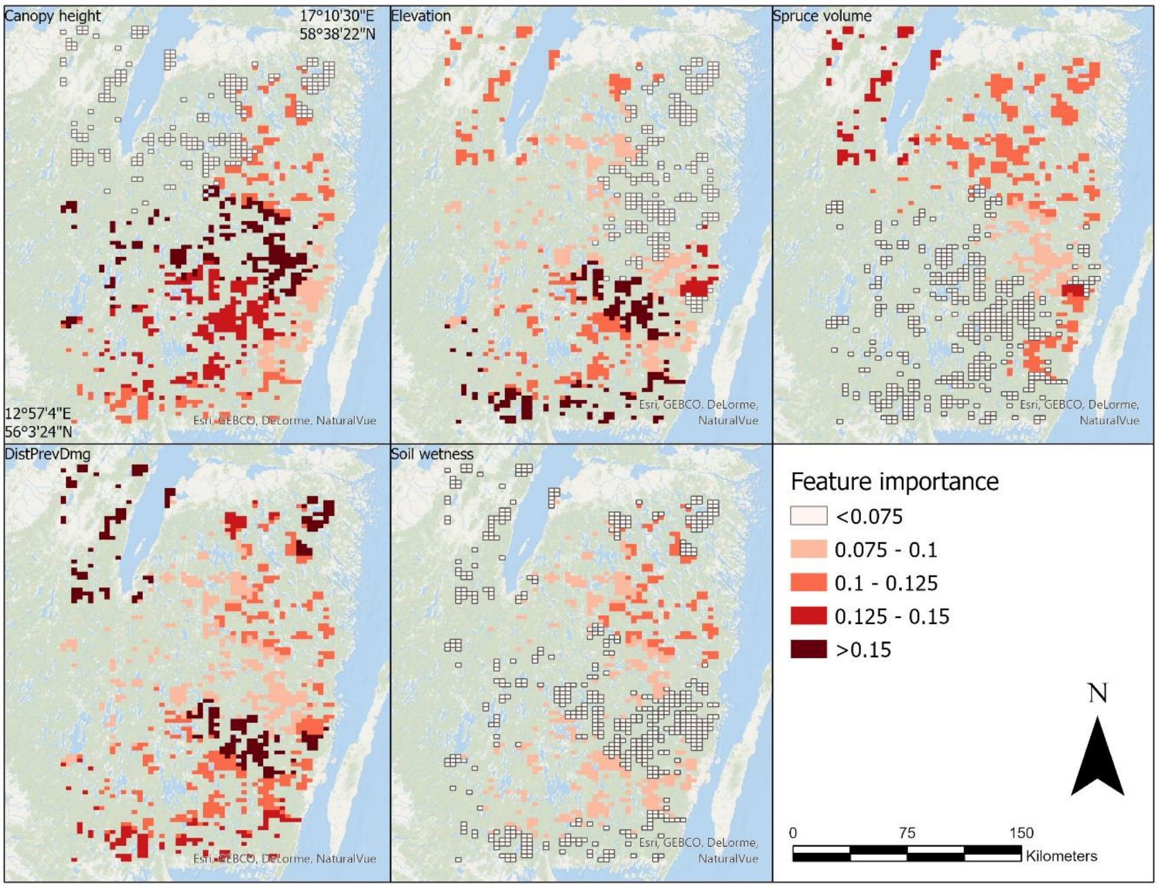


Figure 14. Spatial distribution of local feature importance in *Random 15 normal*. The five most important features included here (defined as the features most often found in top 5 in each of the 15 zones). Only two tree characteristic features included. Classification based on manual interval for easier comparison.

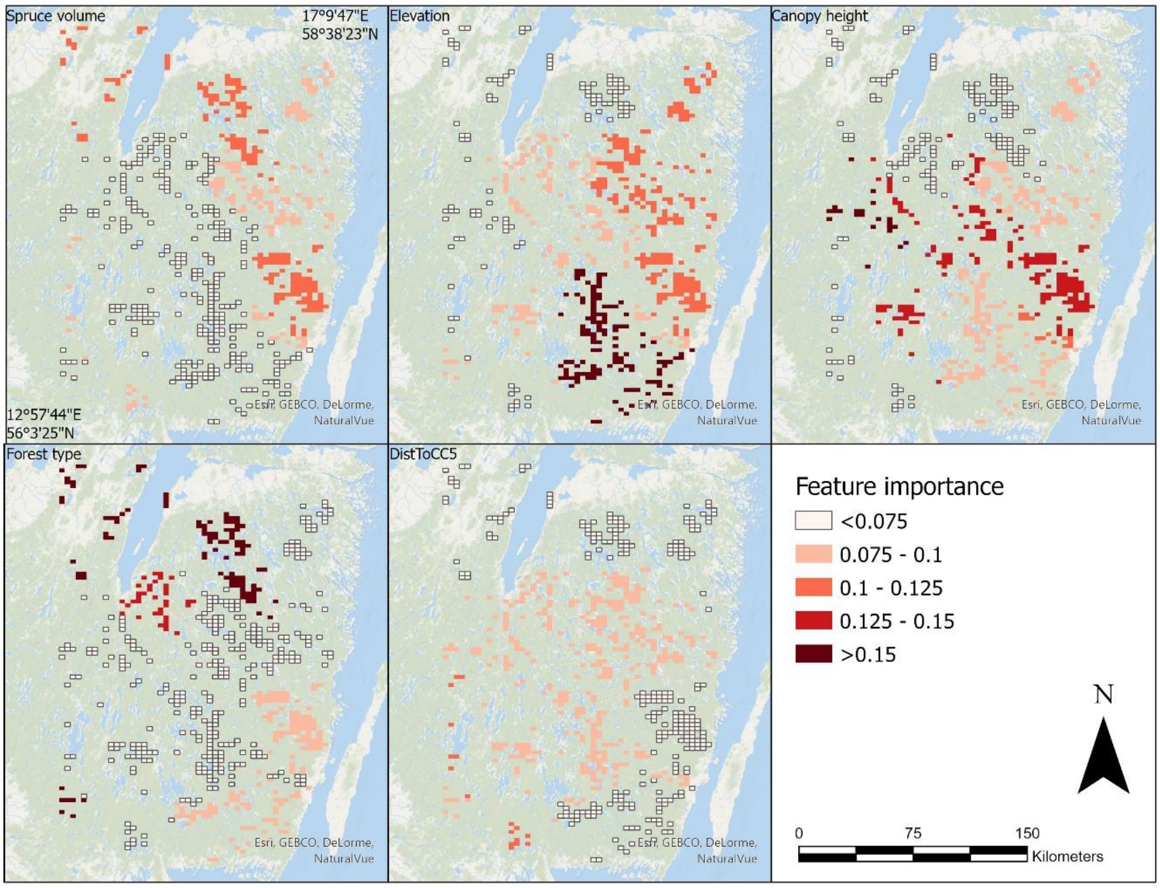


Figure 15. Spatial distribution of local feature importance in *GWR 15 drought*. The five most important features included here (defined as the features most often found in top 5 in each of the 15 zones). Only two tree characteristic features included. Classification based on manual interval for easier comparison.

3.3.3 Characteristics of zones based on their performance

One of the aims of this study was to characterise the best performing local zones and to assess the potential of using more local models to predict bark beetle presence. In the appendix, Table 4 and Table *5* show a detailed account of the different zone results for the two best performing local models – *GWR 15 drought* and *Random 15 normal*.

The number of zones clearly had a big influence, indicating the degree of locality. Both local models based on the greatest number of zones (15) performed best. The number of records in each zone differed based on how the study area was split. The random zones were characterized by an approximately equal number of data records in each zone, whereas the GWR-based zones were characterized by an uneven number of records. The correlation between model performance and the number of records inside the zone was assessed for the two best performing local models. Both correlation coefficients were negative indicating that fewer number of records could result in better model performance. That said, it was difficult to make conclusions about the relationship between the number of records and performance, since none of these tests were significant (p-value > 0.05). It should of course be noted that it was expected that the random model would not result in significant correlations due to the approx. even number of records in each zone.

Performance was highly dependent on the distribution of bark beetle presence inside each zone, with an equal distribution generally resulting in the best local performance. Zone 3 and 9 in the *GWR 15 drought* (Table 5) were examples where local models did not perform very well as data in the zone was unbalanced resulting in a low prediction accuracy for healthy records in zone 3 and infested records in zone 9.

The local deviance score was found to be significantly correlated (p-value < 0.05) to model performance for *GWR 15 drought* with a correlation value of 0.63. The areas with high GWR scores were characterized by high spatial autocorrelation and therefore using more local input data resulted in higher model performance. Figure 16 shows the local deviance scores for the drought dataset. The north-western, south-western, and mid-eastern parts of the study area had high local deviance scores, which coincided to a high degree with high model performance (Figure 13).

The feature importance in the best performing zones in both the normal and drought models generally differed from the zones that did not perform as well. These had certain features with higher importance, especially *distPrevDmg* (see Table 1) and elevation in *Random 15 normal*, and forest type in *GWR 15 drought*. The relatively worse-performing zones were more associated with high feature importance of the tree characteristic features, such as in the global models. As said, it was a bit more difficult to make general statements about local feature importance in the drought model, but the general conclusion must be that the best performing zones differ from the global ones in their feature importance.

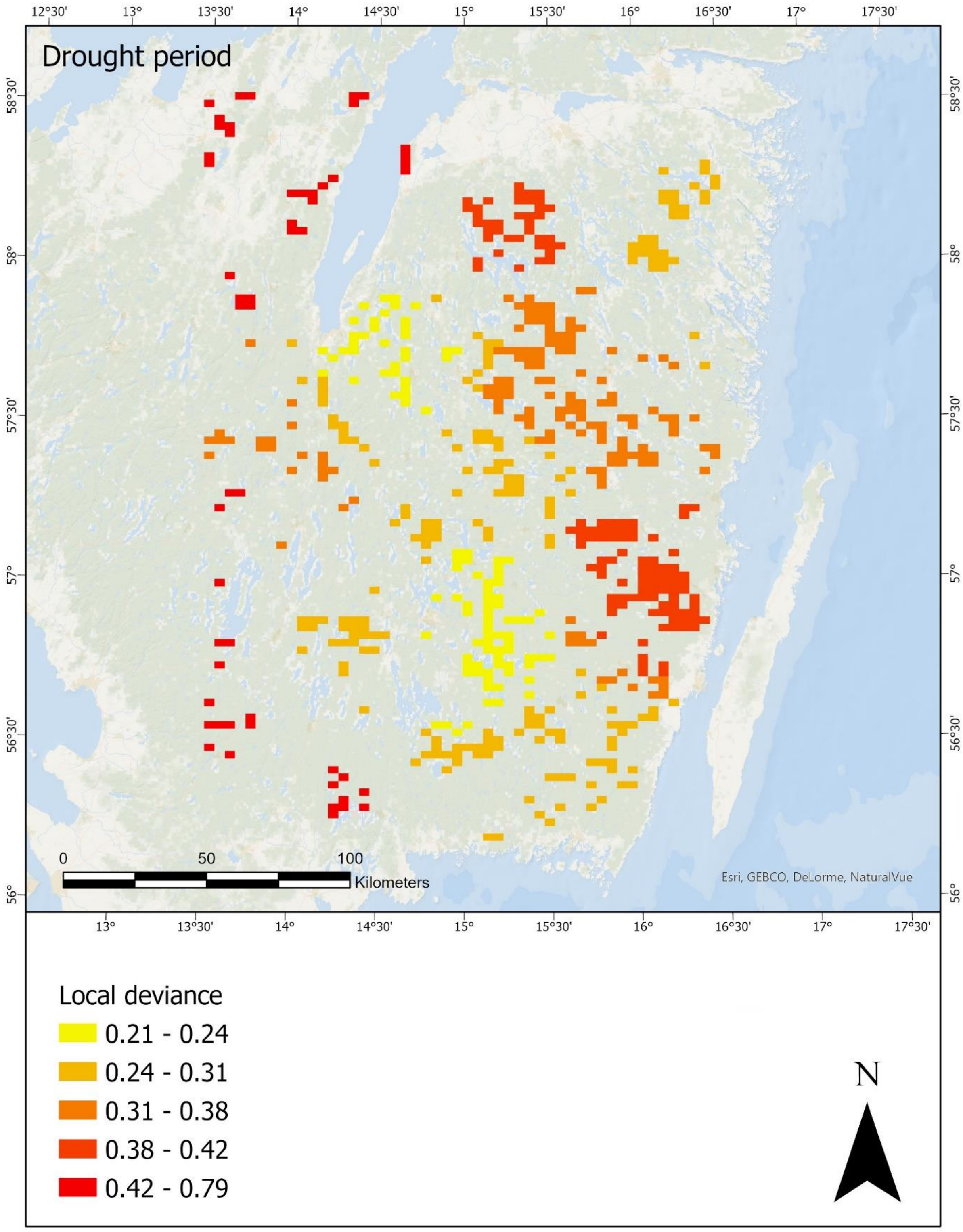


Figure 16. Spatial distribution of local deviance scores, *GWR 15 drought*. Classification made with Natural Breaks (Jenks).

4. Discussion

In this study, an assessment of using ML models to predict bark beetle attacks was performed on different scales. Many methodological choices had to be made that could potentially influence the results and make them less reliable.

4.1 Global results

Only a handful of features and ML models were assessed. The features included were varied and consisted of both temporal and spatial features as well as more static, landscape features. One can always consider whether the chosen features are appropriate for a specific application, but the high model performances obtained in the study indicate that the chosen features were good at estimating and predicting future bark beetle attacks in the case area.

Of the ML models assessed, RF was found to perform (marginally) better on the global dataset. This is in line with the results from Koreň et al. (2021) that found tree based models to perform best in the spatial prediction of bark beetle presence. After some hyperparameter tuning, SVM performed almost as good as RF. But there are other reasons, apart from pure performance, to prefer RF – for instance due to its fast computation time, the fact that standardization and scaling of data is not required, and also its ability to handle collinearity between features (M. Müller et al., 2022; Quiñones et al., 2021) – the first two are often seen as disadvantages of the SVM model. Cracknell & Reading (2014) compared RF and SVM among other ML models and reported that RF performed better in different aspects compared to SVM, here under stability and processing time. SVM on the other hand showed a higher computation time and instability over the cross-validated scores. In the current study the stability was identical between SVM and RF – meaning there was no significant difference in variance between the cross-validated performance scores for the two models. For some of the aforementioned reasons, RF has been chosen in different spatial studies as the only ML model (Georganos et al., 2021; M. Müller et al., 2022; Santos et al., 2019).

Since RF performed best for the normal period, it was also used for the global drought modelling. The prediction accuracy was lower compared to the normal dataset but still performed adequately. This could indicate a more complex relationship between the explaining features and bark beetle presence under drought – feature importance was generally lower. But as already mentioned in M. Müller et al. (2022) it could also be due to the lower number of features included in the drought dataset.

Tree characteristic features such as canopy height and spruce volume were found to be highly indicative of sensitive areas, with higher values increasing the risk of bark beetle presence. Both in the normal and drought period, these factors were found to be highly important (Figure 7 and Figure 8). This was in accordance with other studies such as Koreň et al. (2021) who found percentage of spruce and spruce forest age to be important features. Topographic features have also been noted as important features in other studies (M. Müller et al., 2022) and elevation was found to be the most important abiotic factor predicting bark beetle attacks in Lausch et al. (2011). The same was found in this study, in both the normal and drought period. The feature importance was especially high in the normal period, where higher elevation was associated with lower bark beetle presence. The negative correlation between tree characteristic features and elevation in the normal period can partly explain this (Figure 5). The relationship between bark beetle presence and elevation seemed more diffuse in the drought period with a correlation value very close to zero.

Lausch et al. (2011) found temporal and distance-based features to be highly indicative of bark beetle risk, such as distance from the site of infestation from previous year. This makes sense as bark beetles will typically attack trees located close to trees recently attacked (M. Müller et al., 2022). The distPrevDmg feature (see Table 1) included in the normal period, was found to be the most important feature after the already mentioned (tree characteristics and elevation), even though it was only marginally, positively correlated to bark beetle presence. According to M. Müller et al. (2022) this distance-based feature could explain the better performance in the normal global model compared to the drought one, and this is confirmed here by the high importance of this feature in the normal model. In Koreň et al. (2021), distance to actual forest damage was one of the features included in the best performing model.

4.2 Local results

Local zones, defined as spatially contiguous zones, were created based on both GWR scores and random selection, and modelling was performed inside these. In general, local modelling resulted in better model performance, especially when a sufficient number of zones were created – best results were obtained with 15 zones. Quiñones et al. (2021) found that geographically weighted RF had a higher potential for accurate diabetes predictions compared to traditional and global methods. Georganos et al. (2021) examined the spatial scales at which local models could perform better than global models, by considering different kernel sizes (number of neighbours), and found that if a suitable geographical scale is chosen, local RF can perform better than global RF.

One thing that must be noted here, is that the term ‘local model’ differs between this study and earlier studies (Georganos et al., 2021; Quiñones et al., 2021; Santos et al., 2019). The approach in this study was more static, meaning a comparison was made between a global model that used data from the whole study area and ‘local’ models that only used data from one zone. The more dynamic approach employed in earlier studies is based on kernels and bandwidths where a model is performed in each record, and only based on nearby points, thus constantly changing the inputs in each model. One potential problem with the more static approach used in this study is the fact that models could perform worse at the zone edges, and there will naturally be more of these when more zones and models are considered compared to one global. Since many of the nearest data records will be located in other zones for these edge records, the accuracy could be worse. The more dynamic approach comes with a more intense computing cost, due to the vast number of models that potentially must be established. In this case, it becomes even more important to consider the choice of ML model. As reported in Cracknell & Reading (2014) using multiple training samples to account for spatial variation in a study area can result in a large increase in SVM processing times.

Most of the geographically weighted ML studies have understandably used RF (Georganos et al., 2021; Quiñones et al., 2021; Santos et al., 2019).

In the aforementioned spatial ML studies, the data mainly consisted of polygons, for instance counties in the US (Quiñones et al., 2021**).** The point data in this study with a high number of records, even after balancing the dataset, would be a big problem computation-wise if dynamic, local modelling should be employed. The data could have been generalized from point to polygon, for instance by finding the proportion of bark beetle presence inside polygons. But in the end, the more static approach was chosen since the urge was to use the point data, and not lose information by generalizing the points into polygons. Another point was that most of these earlier spatial ML studies were based on a continuous dependent variable (opposite this study that had a binary dependent variable), and therefore local regressions were performed. Georganos et al. (2021) developed and used *SpatialML*, a R-package that performs local RF regressions. Writing an own program to perform a local classification was deemed out of the scope of the project, due to time constraints. That being said, performing some dynamic, local modelling, in different zones in *Random 15 normal* and *GWR 15 drought* could have been implemented to see whether performance would improve inside these zones. The computation costs would not be that high in this case due to the lower number of records inside the zones compared to the global dataset. This can be employed in future bark beetle studies. Once again, since the performance was indeed quite good, the potential for improvement was minor.

The overall, local model performances were more similar in the normal and drought period, when dividing the area into more zones (Table 4 and Table 5), which indicated that the drought model would generally benefit more from zooming in than the normal model. When the study area was divided into more zones, very well-performing zones were found for both the normal and drought period and generally most local models performed better than their global counterparts. Generally, the best-performing zones had a different feature importance compared to the worst-performing ones that had a feature importance that more resembled the global models with tree characteristic features being the most important. Elevation and *distPrevDmg* (see Table 1) were important in the best-performing zones in the normal period and forest type was important in the best-performing zones in the drought period.

As noted in Quiñones et al. (2021) being able to map out model performance spatially makes it possible to find areas with worse performances and where additional features could be needed to increase prediction accuracy (Georganos et al., 2021). What characterizes good and bad zones was therefore an important assessment in this study and this included feature importance.

4.3 Limitations and uncertainties

Different uncertainties surrounded the data, modelling, and choice of features in this study. One of the big insecurities with the data was the establishment of healthy records. Infested trees were collected through harvester data, whereas a proxy for healthy trees was made – defined as pixels not characterized as infested and located in estates that had harvested trees in other pixels. This is indeed only a proxy and could hide the fact that many of these areas had infested trees. In the end this was deemed the best option, and since the performance was high, this approach could be said to be acceptable.

Other studies have compared different ML models, for instance Koreň et al. (2021) who assessed the spatial distribution of bark beetle presence in Czech Republic. In the current study, fewer models were compared. Initially, other complex models, such as neural networks were considered. In the end, fewer models were included, due to time constraints and the fact that the aim was more focused on comparing global and local models. A more thorough comparison of global and local models would have included using more ML algorithms for the local models and not just RF. SVM and RF performed almost equally good on the global dataset. Making local models with SVM would have been a great addition since many of the earlier geographically weighted ML studies are based on RF (Georganos et al., 2021; Quiñones et al., 2021; Santos et al., 2019). This should be implemented in future studies where time and computational constraints are less of an issue.

The model hyperparameters were tuned using grid search. The linear models and RF did not improve to a high degree. RF often performs well even without much hyperparameter tuning and this can also be considered one of the strengths of the model. The non-linear version of SVM improved more markedly due to hyperparameter tuning. Only the RBF-kernel was assessed and including other kernels such as polynomial in the assessment could have further improved model performance (Géron, 2019). Even though there are many reasons to prefer RF, a more intense and broader grid search could potentially have improved model performance and given the edge to other models. A trade-off between computation time and broadness of grid search must of course be considered (Cracknell & Reading, 2014). Another limitation concerning hyperparameters was the fact that the tuned global hyperparameters were also used in the local models. This was also chosen due to computation and time constraints. Making local hyperparameter tuning in each zone could have resulted in higher local model performance. This improvement would likely have been higher if SVM had been used for the local performance due to the lower effect of hyperparameter tuning for RF.

Following the choice of the more static approach used in this study, one thing that had to be considered was how to divide the study area into different zones. A random procedure, that kept the number of data records in each zone fairly constant as well as a GWR-based approach based on the spatial autocorrelation in each area were implemented. When creating zones, it must be decided what should be considered. The proportion of bark beetle presence (trying to balance the proportion between infested and healthy records inside each zone) was not considered to a high degree since this would make it difficult to make spatially contiguous zones. The number of records in each zone was the most important factor in the random approach, and the local deviance score was the most important in the GWR-based approach. Using a more dynamic procedure would have avoided many of these choices, as a constant number of input points could have been used, but for the reasons already mentioned, this method was not implemented.

The results indicate that it was better to divide the area into a higher number of zones, and therefore it would be central to ask, whether overall performance could be improved by dividing the study area into an even higher number of zones – for instance 20 or even 30 zones. This could make sense if the random procedure is followed. But if the GWR method is used with different zone sizes, this would increase the chance of having very small zones where an unbalanced data distribution and low number of records could result in bad performances. Already at 15 zones, some of the zones get so small that performance is severely affected. Unbalanced data can decrease model performance – Hernandez et al. (2012) for instance used unbalanced data to model bark beetle attacks and acknowledged that the relatively low accuracy could be improved with more balanced data. Another problem using GWR is the fact that it is based on local linear regression models and therefore not really suitable for complex, non-linear problems like bark beetle presence (Luo et al., 2022). Still, it was clear that high local deviance scores found in the GWR-output was associated with better local model performance, and the GWR-approach had its merit.

Deciding which features to include and omit is always one of the most important aspects and considerations of a ML study. Features that could potentially have increased performance in this study were coordinates and meteorological. The coordinate features were left out of the final models, although other similar studies have included these. (Georganos et al., 2021) argued that including coordinates as input features is good practice when working with spatial data. If training data is of high quality, using only coordinates as explaining features can be sufficient to achieve an acceptable model performance (Cracknell & Reading, 2014). That being said, using coordinates can be problematic. Training a ML model in a restricted study area using coordinates, can make the model less able to generalize outside of the study area. In studies, like this one, where data is partly based on field collection, spatial clustering of input data due to time constraints and accessibility is often the case, and this can lead to poor generalization outside of these clusters (Cracknell & Reading, 2014). Since the aim was to assess spatial autocorrelation and the use of more local models, it was decided not to use coordinates in the modelling. Instead, the spatial dimension was included in the GWR models that were used to create the local division. Geographical coordinates were thus omitted in both the global and local models.

Another group of explaining features that were not directly considered in this study were meteorological features. Other bark beetle studies have included solar radiation, temperature and precipitation in their models (Hernandez et al., 2012; Koreň et al., 2021; Lausch et al., 2011). Whereas Lausch et al. (2011) did not find solar radiation to contribute to higher model performance, solar radiation was one of the features that were included in the best performing model in Koreň et al. (2021). A big insecurity regarding these meteorological features, and the features in general, is their temporal importance during a bark beetle outbreak. Lausch et al. (2011) found that bark beetle preferred different temperatures at different times in the outbreak, making the relationship even more complex. Looking at the spatial importance of meteorological features such as air temperature and precipitation can be problematic and unreliable on a high spatial resolution as the data in this study (M. Müller et al., 2022). Due to these spatial and temporal uncertainties, these features were not included directly. Comparing data from a drought and normal period, was a way to include these features indirectly and in a more reliable way.

The most important features were assessed for both the global and local models, but a more sophisticated feature selection was not implemented. The full set of features were included in all the models, but the feature importance was often low for several of these. Koreň et al. (2021) reported that the number of features could be reduced without a significant reduction in model performance. Fewer features even improved prediction accuracy of diabetes prevalence in Quiñones et al. (2021). Using a smaller number of features could be beneficial if for instance the more dynamic, local modelling approach was implemented, as computation time would be lower – something that would especially be beneficial with SVM. Fewer input features would also make the model more interpretable (Quiñones et al., 2021) and make it easier to make an in-depth survey of the relationship between the dependent variable and the independent variables, for instance through the use of partial dependence plots. In M. Müller et al. (2022) the focus was on feature importance and bark beetle presence and a pruning of the included features in the models was performed to decrease collinearity between these. This pruning led to higher model performances, indicating the potential for even increasing model performance in the current study. Local feature selection is another topic where more research could be implemented. Using the same features for each local model is a simplistic approach and could be made more sophisticated by selecting different features in each locality (Georganos et al., 2021). Adding features in localities with a relatively low prediction accuracy could potentially result in better model performance in these zones (Quiñones et al., 2021).

As a final note on limitations, it could be argued that the spatial non-stationarity between independent and dependent variables where not really considered, as one could argue that the local models are still global in nature, as we do not use a dynamic approach. Even though this spatial non-stationarity is not considered directly, by using GWR in the zone creation and comparing results and feature importance between the global and local models, it is included at least indirectly.

4.4 Implications of the results

We must differ between overall local model performance (weighted mean of the performance inside each zone) and performance inside each zone. One of the clear advantages of using a GWR-approach is that it shows in which areas it might be beneficial to zone in on a smaller scale and perform local modelling. If a forest entity wants to assess the risk of bark beetle presence in an area and make a feature analysis, it could be beneficial to zoom in and use more local data as this could be helpful in understanding the features that are most important inside this zone and which should be considered in future management schemes.

It is important to consider the underlying data, and make sure that the number of records in the local model is still relatively high and the distribution of infested and healthy records not too unbalanced. If one must use unbalanced data, a higher number of infested trees in the input data would generally be preferred as this could mean, the model would at least not be worse at predicting infested trees compared to healthy ones. Generally the true positive rate (sensitivity) would be satisfactory in that case (Koreň et al., 2021). An infested tree can lead to infestations on nearby trees and detection of these is therefore more important than detection and prediction of healthy trees. Harvesting a healthy tree is better than not harvesting an infested one, and this is a trade-off that must be considered when establishing local models. It is still important to acknowledge that reliable ML models must be characterized by both high sensitivity and specificity (Koreň et al., 2021) – something that can be an issue depending on the local model creation and the number of zones assessed.

Global warming and climate change is expected to lead to larger bark beetle habitats and also making the risk of a second beetle generation in one season higher (in areas where this typically has not been possible) (M. Müller et al., 2022). Higher temperatures and resulting drought can decrease tree defence and leave more trees weakened and increase their sensitivity to bark beetle attacks (Koreň et al., 2021). Another possible consequence is that longer frost-free periods under climate change might result in less stable trees that are more sensitive to wind damage (M. Müller et al., 2022) – one of the predispositions of bark beetle attacks, for instance in Sweden. Generally model performance was worse under drought than normal but could, as already mentioned, be explained by the extra features included in the normal dataset. To assess complex spatial phenomena such as forest damage from insects, advanced methods are needed, and our results indicate, that zooming in and performing more local modelling had more potential in the drought period with almost all zones performing better than the global model. In the future under climate change, there will be an increased focus on prediction and detection of areas sensitive to bark beetle attacks. This will help to effectively select areas where protective measures should be taken, such as increasing tree drought resistance (M. Müller et al., 2022). Zooming in on a finer scale, and even using a more dynamic approach in the creation of local models could be highly beneficial to increase model performance and help with more effective local management, if the crucial features differ between areas, as the results of this study indicated. Other advanced techniques that can be used together with ML in the future is remote sensing data that can help to more easily identify past and current infestations (Koreň et al., 2021) and make the input data more reliable than in this study.

There are still many unknowns when it comes to the spatial and temporal dimensions of bark beetle outbreaks due to the high complexity. As already mentioned, predicting future outbreaks can be almost impossible due to these complexities and as reported in Lausch et al. (2011), only a combination of factors could determine the spread of bark beetle. As said in M. Müller et al. (2022), since features are, in most circumstances, not expected to be independent in natural and complex matters such as bark beetle outbreaks, methods assessing partial importance should be implemented.

The temporal and spatial dimensions are important to consider in bark beetle studies and have often been rather limited in earlier studies, focusing on relatively small study areas and a limited time period (Lausch et al., 2011; M. Müller et al., 2022). One aspect that should be examined in future studies is the use of ML on longer time scales – to assess how the predictive and exploratory dimensions of the model changes during an outbreak and also look at the spatial dimension of these. A longer time scale than 3 years, as in the current study, should be assessed. Ramazi et al. (2021) made a temporal split of the data into different years and thereby could assess how different ML algorithms could predict mountain pine beetle attacks both short- and long-term. Defining in what year a bark beetle attack occurred can be complex, as seen in the current study. But considering the temporal dimension and continuously looking at different features and feature combinations will be important to improve our understanding of the dynamics of bark beetle attacks.

In the end, using ML in combination with remote sensing and other advanced techniques to detect infested spots and using this in different areas, can give us important insights of the spatial dimension of bark beetle attacks, and at the same time be a valuable tool in future forest management under climate change. This can result in a higher potential for detecting sensitive areas before infestation occurs, thereby decrease the dependency on expensive and substantial ad-hoc measures in the forest industry. Instead of control actions, focus should be on cheaper and effective prevention measures such as removing infested trees and laying out pheromone traps (Rammer & Seidl, 2019; Valdez Vasquez et al., 2020).

5. Conclusion

The ML model that most accurately predicted bark beetle presence in this study, based on the whole dataset (global model) was RF, followed by non-linear SVM. Both models performed very well on the data, with the chosen features. They performed almost identically due to a higher model improvement from hyperparameter tuning for SVM, but RF ultimately performed marginally better. In a real-world complex problem, not having to do intensive hyperparameter tuning and grid search could be preferred to keep computation costs down – a reason for choosing RF, based on the results of this study.

Splitting the study area into smaller entities (local modelling) resulted in generally higher prediction accuracy, especially when a higher number of local zones were considered. The highest number of assessed zones, 15, led to the best overall performance. The models in the local zones generally performed better than the global ones, especially in zones with a balanced distribution of healthy and infested records, not too few records, and a high local deviance score for the GWR-based local zones. Generally, feature importance was different in the best performing zones compared to the zones that performed worse. The worst-performing zones resembled the global ones, with many tree-characteristic features being the most important. Other features, such as forest type in the drought period, and *distPrevDmg* (see Table 1) in the normal period, were important in many well-performing zones, indicating that these zones could perform well because other features, than the globally most important, were important inside these. By only using local datapoints, the model was better able to include this information.

Increasing temperatures and drought instances under climate change is expected to lead to a higher risk of bark beetle attacks in many areas. The models generally performed better on the data from the normal period compared to the drought period, indicating a more complex relationship between the independent and dependent features under drought and climate change. The additional number of features in the normal models could also explain this difference. The local zones improved model performance in the drought period more than in normal period, indicating that the local models were better able to explain the potentially more complex feature relationship. Using more local and potentially more dynamic (and computationally intensive) models could improve prediction accuracy of bark beetle presence, and at the same time give valuable insights to local feature importance. This could guide future forest management on a local scale – something that will be crucial under climate change due to the spatial dimension and expected increase in insect damage in Sweden and other forest-rich countries.

References

Ae, H. (2013). An Introduction to Logistic Regression : From Basic Concepts to Interpretation with Particular Attention to Nursing Domain, *43*(2), 154–164.

Arabameri, A., Pradhan, B., & Rezaei, K. (2019). Gully erosion zonation mapping using integrated geographically weighted regression with certainty factor and random forest models in GIS. *Journal of Environmental Management*, *232*(September 2018), 928–942. https://doi.org/10.1016/j.jenvman.2018.11.110

Billings, R. F., & Upton, W. W. (2010). A methodology for assessing annual risk of southern pine beetle outbreaks across the southern region using pheromone traps. *Advances in Threat Assessment and Their Application to Forest and Rangeland Management*, 73–85. Retrieved from http://www.treesearch.fs.fed.us/pubs/37019#.VgsGsisRCKg.mendeley

Brunsdon, C., Fotheringham, A. S., & Charlton, M. E. (1996). Geographically Weighted Regression:, *28*(4).

Cracknell, M. J., & Reading, A. M. (2014). Geological mapping using remote sensing data: A comparison of five machine learning algorithms, their response to variations in the spatial distribution of training data and the use of explicit spatial information. *Computers and Geosciences*, *63*, 22–33. https://doi.org/10.1016/j.cageo.2013.10.008

Dataserier med normalvärden för perioden 1991-2020 | SMHI. Retrieved 04/21/2022 from https://www.smhi.se/data/meteorologi/dataserier-med-normalvarden-for-perioden-1991-2020-1.167775

Georganos, S., Grippa, T., Niang Gadiaga, A., Linard, C., Lennert, M., Vanhuysse, S., … Kalogirou, S. (2021). Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling. *Geocarto International*, *36*(2), 121–136. https://doi.org/10.1080/10106049.2019.1595177

Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. " O'Reilly Media, Inc."

Hagenauer, J., & Helbich, M. (2021). A geographically weighted artificial neural network. *International Journal of Geographical Information Science*, *00*(00), 1–21. https://doi.org/10.1080/13658816.2021.1871618

Hernandez, A. J., Saborio, J., Ramsey, R. D., & Rivera, S. (2012). Likelihood of occurrence of bark beetle attacks on conifer forests in Honduras under normal and climate change scenarios. *Geocarto International*, *27*(7), 581–592. https://doi.org/10.1080/10106049.2011.650652

Hlásny, T., König, L., Krokene, P., Lindner, M., Montagné-Huck, C., Müller, J., … Seidl, R. (2021). Bark Beetle Outbreaks in Europe: State of Knowledge and Ways Forward for Management. *Current Forestry Reports*, *7*(3), 138–165. https://doi.org/10.1007/s40725-021-00142-x

How Build Balanced Zones works—ArcGIS Pro | Documentation. Retrieved 10/19/2022 from https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/learnmore-buildbalancedzones.htm

How Geographically Weighted Regression (GWR) works—ArcGIS Pro | Documentation. Retrieved 10/19/2022 from https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/how-geographicallyweightedregression-works.htm

Hroššo, B., Mezei, P., Potterf, M., Majdák, A., Blaženec, M., Korolyova, N., & Jakuš, R. (2020). Drivers of spruce bark beetle (Ips typographus) infestations on downed trees after severe windthrow. *Forests*, *11*(12), 1–15. https://doi.org/10.3390/f11121290

Jonsson, B. G., Ekström, M., Esseen, P. A., Grafström, A., Ståhl, G., & Westerlund, B. (2016). Dead wood availability in managed Swedish forests - Policy outcomes and implications for biodiversity. *Forest Ecology and Management*, *376*(September), 174–182. https://doi.org/10.1016/j.foreco.2016.06.017

Kärvemo, S., & Schroeder, L. (2010). A comparison of outbreak dynamics of the spruce bark beetle in Sweden and the mountain pine beetle in Canada (Curculionidae : Scolytinae). *Entomologisk Tidskrift*, *131*(3), 215–224. Retrieved from http://www.sef.nu/

Koreň, M., Jakuš, R., Zápotocký, M., Barka, I., Holuša, J., Ďuračiová, R., & Blaženec, M. (2021). Assessment of machine learning algorithms for modeling the spatial distribution of bark beetle infestation. *Forests*, *12*(4). https://doi.org/10.3390/f12040395

Långström, B., Lindelöw, Å., Schroeder, M., Björklund, N., & Öhrn, P. (2009). The spruce bark beetle outbreak in Sweden following the January-storms in 2005 and 2007. *IUFRO Forest Insect and Disease Survey in Central Europe, September 15-19 2008*, (May 2006), 1–8.

Lausch, A., Fahse, L., & Heurich, M. (2011). Factors affecting the spatio-temporal dispersion of Ips typographus (L.) in Bavarian Forest National Park: A long-term quantitative landscape-level analysis. *Forest Ecology and Management*, *261*(2), 233–245. https://doi.org/10.1016/j.foreco.2010.10.012

Luo, Y., Yan, J., & McClure, S. (2021). Distribution of the environmental and socioeconomic risk factors on COVID-19 death rate across continental USA: a spatial nonlinear analysis. *Environmental Science and Pollution Research*, *28*(6), 6587–6599. https://doi.org/10.1007/s11356-020-10962-2

Luo, Y., Yan, J., McClure, S. C., & Li, F. (2022). Socioeconomic and environmental factors of poverty in China using geographically weighted random forest regression model. *Environmental Science and Pollution Research*, (0123456789). https://doi.org/10.1007/s11356-021-17513-3

Müller, A. C., & Guido, S. (2016). *Introduction to machine learning with Python: a guide for data scientists*. " O'Reilly Media, Inc."

Müller, M., Olsson, P. O., Eklundh, L., Jamali, S., & Ardö, J. (2022). Features predisposing forest to bark beetle outbreaks and their dynamics during drought. *Forest Ecology and Management*, *523*(June). https://doi.org/10.1016/j.foreco.2022.120480

Munro, H. L., Montes, C. R., & Gandhi, K. J. K. A New Approach to Evaluate the Risk of Bark Beetle Outbreaks Using Machine Learning Methods. *Available at SSRN 4043457*.

Nikparvar, B., & Thill, J. C. (2021). Machine learning of spatial data. *ISPRS International Journal of Geo-Information*, *10*(9), 1–32. https://doi.org/10.3390/ijgi10090600

Olsson, P.-O. (2016). *Monitoring insect defoliation in forests with time-series of satellite based remote sensing data-near real-time methods and impact on the carbon balance*

Olsson, Per-Ola.

Quiñones, S., Goyal, A., & Ahmed, Z. U. (2021). Geographically weighted machine learning model for untangling spatial heterogeneity of type 2 diabetes mellitus (T2D) prevalence in the USA. *Scientific Reports*, 1–13. https://doi.org/10.1038/s41598-021-85381-5

Ramazi, P., Kunegel-Lion, M., Greiner, R., & Lewis, M. A. (2021). Predicting insect outbreaks using machine learning: A mountain pine beetle case study. *Ecology and Evolution*, *11*(19), 13014–13028. https://doi.org/10.1002/ece3.7921

Rammer, W., & Seidl, R. (2019). Harnessing Deep Learning in Ecology: An Example Predicting Bark Beetle Outbreaks. *Frontiers in Plant Science*, *10*(October), 1–9. https://doi.org/10.3389/fpls.2019.01327

Santos, F., Graw, V., & Bonilla, S. (2019). *A geographically weighted random forest approach for evaluate forest change drivers in the Northern Ecuadorian Amazon*. *PLoS ONE* (Vol. 14). https://doi.org/10.1371/journal.pone.0226224

Schroeder, M. (2020). Granbarkborrens förökningsframgång i dödade träd under sommaren 2020 i sydöstra Småland , Värmland och Uppland / Västmanland Sammanfattning Bakgrund.

Senaviratna, N., & Cooray, T. (2019). Diagnosing Multicollinearity of Logistic Regression Model. *Asian Journal of Probability and Statistics*, 1–9. https://doi.org/10.9734/ajpas/2019/v5i230132

Spatial Autocorrelation (Global Moran's I) (Spatial Statistics)—ArcGIS Pro | Documentation. Retrieved 12/03/2022 from https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/spatial-autocorrelation.htm

Spatially Constrained Multivariate Clustering (Spatial Statistics)—ArcGIS Pro | Documentation. Retrieved 10/19/2022 from https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/spatially-constrained-multivariate-clustering.htm

Valdez Vasquez, M. C., Chen, C. F., Lin, Y. J., Kuo, Y. C., Chen, Y. Y., Medina, D., & Diaz, K. (2020). Characterizing spatial patterns of pine bark beetle outbreaks during the dry and rainy season's in Honduras with the aid of geographic information systems and remote sensing data. *Forest Ecology and Management*, *467*(December 2019). https://doi.org/10.1016/j.foreco.2020.118162

VanderPlas, J. (2016). *Python data science handbook: Essential tools for working with data*. " O'Reilly Media, Inc."

Appendix

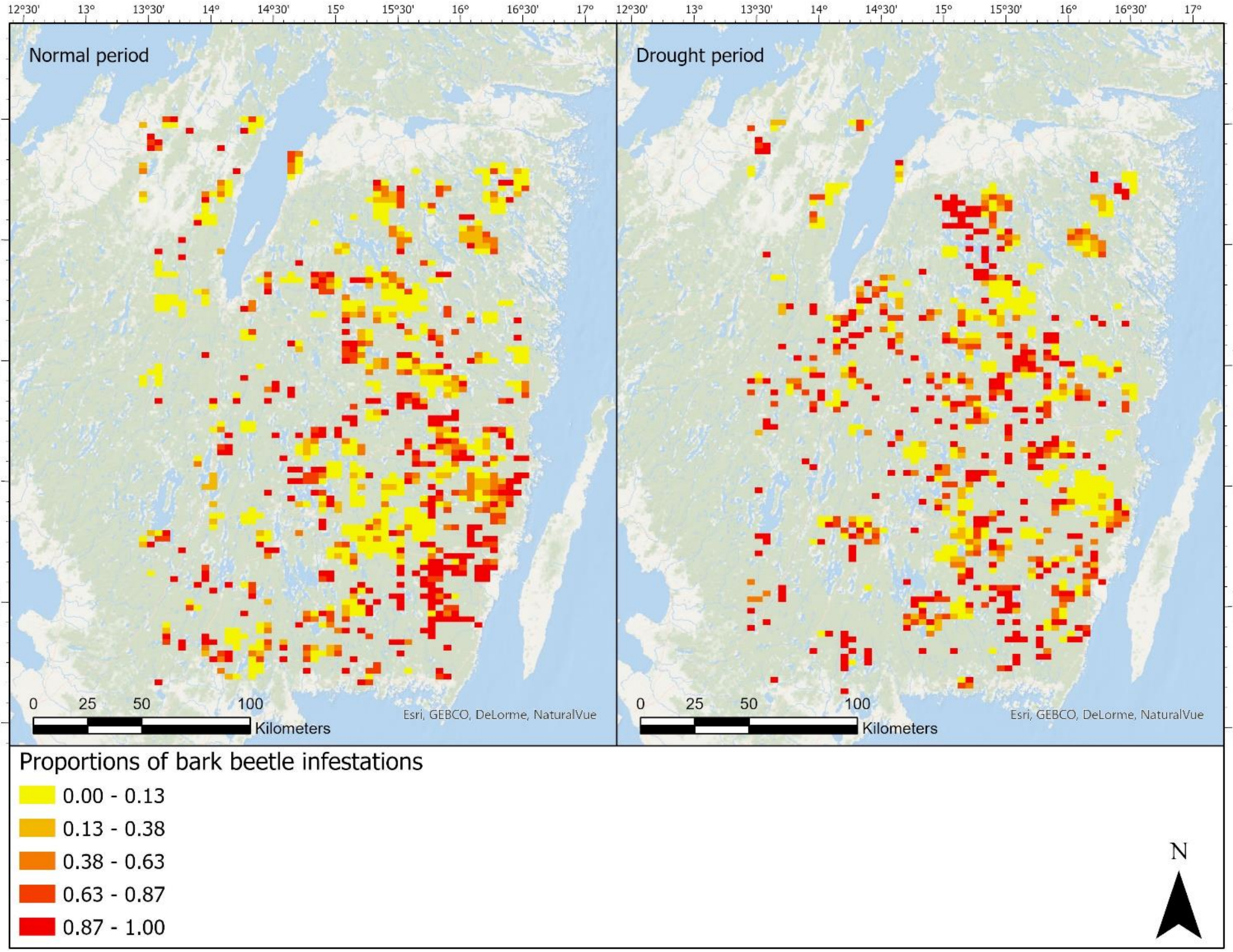
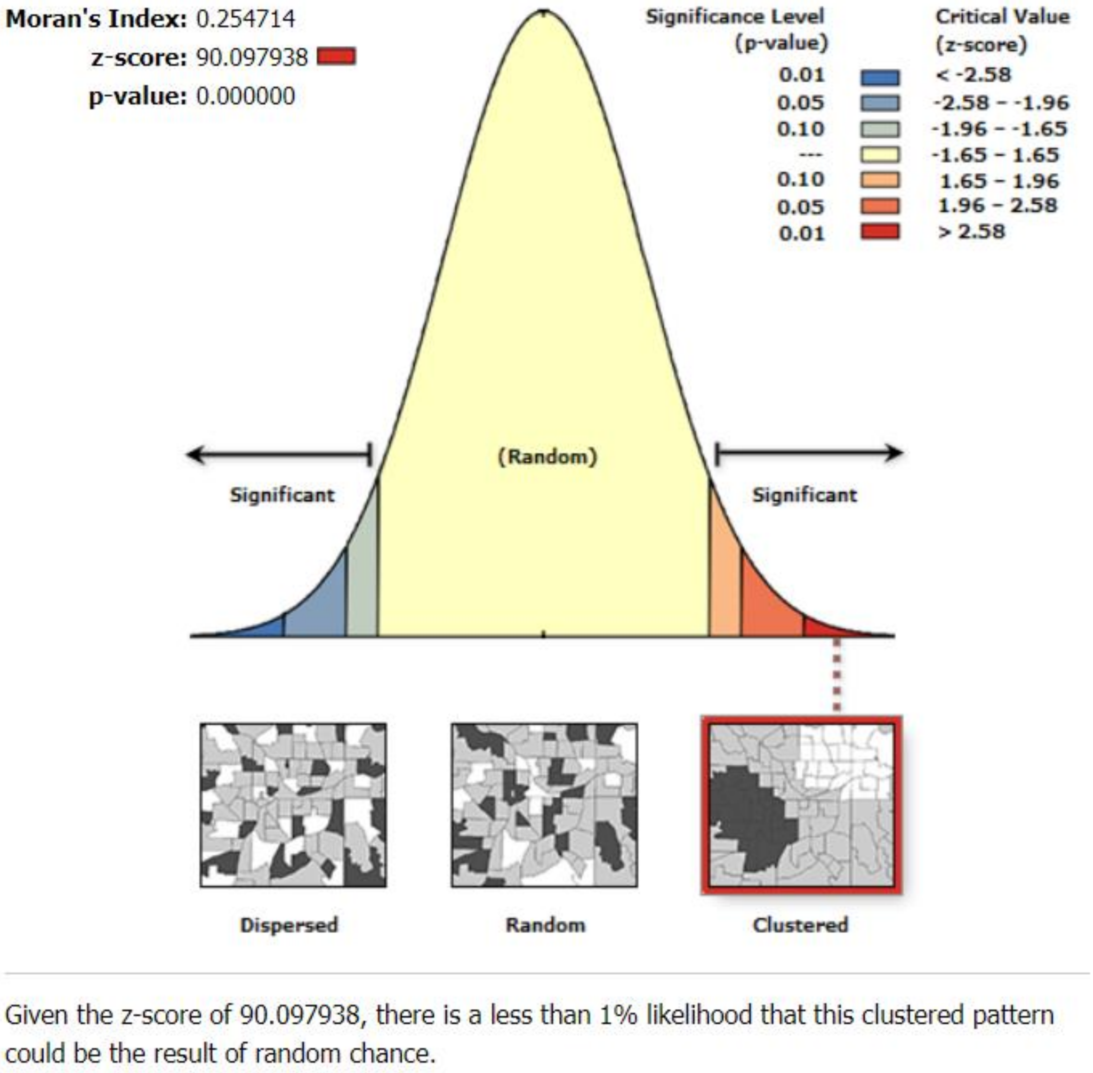


Figure 17. Proportion of infested trees from bark beetle (based on the balanced dataset) for the normal and drought period. Classification based on Natural Breaks (Jenks) on the data distribution in the normal period.



Moran's Index: 0.254714

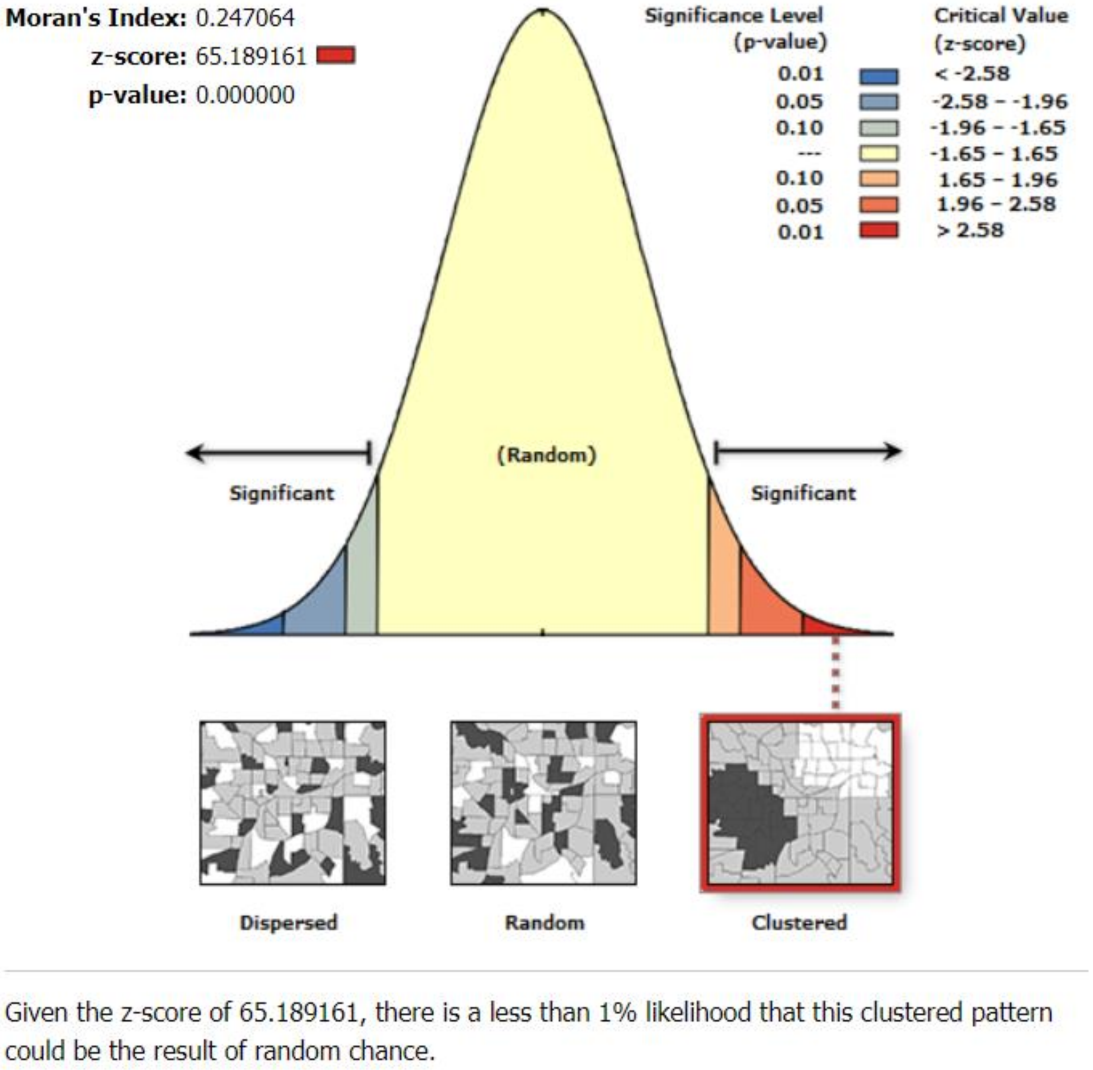
z-score: 90.097938 🟥

p-value: 0.000000

Significance Level (p-value)		Critical Value (z-score)
0.01	🟦	< -2.58
0.05	🟦	-2.58 - -1.96
0.10	🟩	-1.96 - -1.65
---	🟨	-1.65 - 1.65
0.10	🟧	1.65 - 1.96
0.05	🟧	1.96 - 2.58
0.01	🟥	> 2.58

Given the z-score of 90.097938, there is a less than 1% likelihood that this clustered pattern could be the result of random chance.

Figure 18. ArcGIS output from Moran's I test on the global dataset from the normal period.



Moran's Index: 0.247064

z-score: 65.189161 🟥

p-value: 0.000000

Significance Level (p-value)		Critical Value (z-score)
0.01	🟦	< -2.58
0.05	🟦	-2.58 - -1.96
0.10	⬜	-1.96 - -1.65
---	🟨	-1.65 - 1.65
0.10	🟧	1.65 - 1.96
0.05	🟧	1.96 - 2.58
0.01	🟥	> 2.58

Given the z-score of 65.189161, there is a less than 1% likelihood that this clustered pattern could be the result of random chance.

Figure 19. ArcGIS output from Moran's I test on the global dataset from the drought period.

Table 4. Results from the RF modelling in the local zones in *Random 15 normal*. Canopy = canopy height. Other feature explanations in Table 1.

Results from local models (based on random)

Zone	Number of records	CV mean score, accuracy	f1-score, healthy (number of points)	f1-score, infested (number of points)	Correlations	Feature importance
1	5756	0.92	0.93 (2822)	0.93 (2934)	canopy 0.583597 biomass 0.504479 spruce vol 0.496253 basal area 0.458528 forest type -0.397257	canopy 0.1783 biomass 0.0878 basal area 0.0873 elevation 0.0855 distprevdmg 0.0849
2	4947	0.91	0.87 (1729)	0.93 (3218)	canopy 0.549142 biomass 0.453942 spruce vol 0.433183 basal area 0.410896 soil wetness -0.399094	canopy 0.1603 soil wetness 0.0984 spruce vol 0.0946 distprevdmg 0.0884 distToCC5 0.0781
3	6224	0.9	0.91 (4303)	0.85 (1921)	spruce vol 0.460787 biomass 0.399738 basal area 0.398165 canopy 0.394555 soil wetness -0.288430	spruce vol 0.1145 canopy 0.1083 soil wetness 0.1081 distprevdmg 0.1072 soilType 0.0754
4	5611	0.90	0.93 (3384)	0.91 (2227)	spruce vol 0.397342 soil wetness -0.348596 basal area 0.319154 biomass 0.318797 canopy 0.278216	spruce vol 0.1054 soilType 0.1024 soil wetness 0.0981 distprevdmg 0.0960 elevation 0.0804
5	4877	0.95	0.95 (2542)	0.95 (2335)	spruce vol 0.543003 biomass 0.521588 basal area 0.501124 canopy 0.492405 elevation -0.271071	distprevdmg 0.1688 spruce vol 0.1320 elevation 0.1059 biomass 0.0887 basal area 0.0831
6	5767	0.89	0.89 (3324)	0.88 (2443)	biomass 0.475540 canopy 0.473179 basal area 0.444799 spruce vol 0.435291 elevation - 0.289420	canopy 0.1322 elevation 0.1154 distprevdmg 0.1142 soil wetness 0.0950 biomass 0.0860
7	5231	0.93	0.96 (3567)	0.92 (1664)	canopy 0.482379 elevation - 0.459498 biomass 0.455405 spruce vol 0.419550 basal area 0.417332	elevation 0.1668 distprevdmg 0.1600 canopy 0.1394 biomass 0.0772 distToCC10 0.0707
8	4856	0.89	0.91 (3486)	0.82 (1370)	canopy 0.426670 biomass 0.392014 basal area 0.355379 spruce vol 0.332656 landforms 0.197715	canopy 0.1548 biomass 0.1030 basal area 0.0950 distprevdmg 0.0890 elevation 0.0860
9	4815	0.92	0.93 (2387)	0.93 (2428)	canopy 0.438828 biomass 0.407843 elevation -0.390659 basal area 0.357678 spruce vol 0.266670	elevation 0.1978 distprevdmg 0.1400 canopy 0.1122 distToCC5 0.0823 biomass 0.0692
10	4751	0.87	0.57 (432)	0.92 (4319)	canopy 0.301523 biomass 0.293823 basal area 0.280703 spruce vol 0.269991 soil wetness -0.143569	spruce vol 0.1111 canopy 0.1005 distprevdmg 0.0976 soil wetness 0.0961 elevation 0.0952

11	4801	0.95	0.92 (1665)	0.95 (3136)	spruce vol 0.472286 forest type -0.410608 canopy 0.400753 biomass 0.396443 basal area 0.384877	distprevdmg 0.2626 spruce vol 0.1209 soil type 0.0792 canopy 0.0777 forest type 0.0705
12	4855	0.94	0.93 (1845)	0.95 (3010)	forest type -0.517837 spruce vol 0.514797 canopy 0.475814 elevation -0.466554 biomass 0.450284	elevation 0.1464 spruce vol 0.1457 distprevdmg 0.1178 forest type 0.0906 disttoCC10 0.0896
13	4691	0.94	0.90 (1018)	0.97 (3673)	disttoCC10 -0.575986 elevation -0.438623 canopy 0.419563 disttoCC5 -0.416691 forest type -0.405900	disttoCC10 0.2315 elevation 0.1398 distprevdmg 0.1131 canopy 0.0978 forest type 0.0648
14	4909	0.91	0.94 (3482)	0.87 (1427)	spruce vol 0.467468 basal area 0.432490 biomass 0.426014 canopy 0.390151 soil type 0.273944	distprevdmg 0.1492 spruce vol 0.1133 biomass 0.0947 basal area 0.0851 soil type 0.0825
15	3355	0.92	0.91 (1737)	0.91 (1618)	spruce vol 0.556518 biomass 0.520893 basal area 0.511248 forest type -0.454068 canopy 0.453409	distprevdmg 0.1500 spruce vol 0.1232 elevation 0.1115 forest type 0.1060 biomass 0.0738

Overall assessment:
Weighted accuracy (cross-validated): 0.915

Table 5. Results from the RF modelling in the local zones in *GWR 15 drought*. Canopy = canopy height. Other feature explanations in Table 1.

Results from local models (based on GWR)

Zone	Local Deviance (GWR score)	Number of records	CV mean score, accuracy	f1-score, healthy (number of points)	f1-score, infested (number of points)	Correlations	Feature importance
1	0.38	806	0.97	0.98 (514)	0.96 (292)	forest type -0.653165 spruce vol 0.490541 soil wetness -0.452312 landforms 0.301733 canopy 0.291320	forest type 0.2040 elevation 0.1351 soil wetness 0.1342 spruce vol 0.1197 disttoCC10 0.0790
2	0.55	711	0.93	0.95 (353)	0.94 (358)	spruce vol 0.597040 biomass 0.591876 basal area 0.568361 forest type -0.544246 canopy 0.525821	forest type 0.1732 spruce vol 0.1238 biomass 0.1158 basal area 0.1060 canopy 0.0796
3	0.55	284	0.89	0.46 (28)	0.93 (256)	soil type 0.353618 basal area 0.297425 soil wetness -0.295767 biomass 0.282939 disttoCC5 0.233201	basal area 0.2014 soil type 0.1630 disttoCC5 0.1455 biomass 0.0983 spruce vol 0.0892
4	0.41	4342	0.93	0.94 (2381)	0.93 (1961)	spruce vol 0.525608 canopy 0.517333 forest type -0.484912 biomass 0.464548 basal area 0.431727	elevation 0.1284 canopy 0.1236 spruce vol 0.1019 disttoCC10 0.0952 forest type 0.0920
5	0.34	3195	0.88	0.89 (1495)	0.90 (1700)	biomass 0.411299 canopy 0.408229 basal area 0.404059 spruce vol 0.400820 soil wetness -0.308871	elevation 0.1238 soil wetness 0.1025 basal area 0.0988 spruce vol 0.0984 disttoCC5 0.0893
6	0.79	276	0.95	1 (58)	1 (218)	forest type -0.602606 biomass 0.569137 spruce vol 0.563673 basal area 0.547598 canopy 0.524032	forest type 0.3351 basal area 0.1388 spruce vol 0.1123 disttoCC10 0.0748 biomass 0.0706
7	0.21	1503	0.87	0.86 (695)	0.88 (808)	forest type 0.347213 canopy 0.299949 biomass 0.279729 basal area 0.262846 spruce vol 0.255361	forest type 0.1271 disttoCC5 0.0986 elevation 0.0947 soil wetness 0.0930 disttoCC10 0.0909
8	0.24	1433	0.88	0.92 (821)	0.88 (612)	biomass 0.424999 canopy 0.420471 elevation -0.382045 basal area 0.381743 spruce vol 0.349669	elevation 0.2274 canopy 0.0924 disttoCC5 0.0887 basal area 0.0855 biomass 0.0834
9	0.26	2557	0.84	0.91 (1840)	0.77 (717)	soil type 0.276547 spruce vol 0.267380 landforms 0.262529 biomass 0.248889 basal area 0.246669	soil type 0.1054 elevation 0.1029 soil wetness 0.0977 spruce vol 0.0869 disttoCC10 0.0854
10	0.76	74	0.93	0.91 (28)	0.95 (46)	forest type 0.467625 canopy 0.411538 distforestedge 0.397547 slope -0.380454 biomass 0.357218	distforestedge 0.1891 slope 0.1354 forest type 0.1253 disttoCC5 0.0977 canopy 0.0729

11	0.28	4065	0.89	0.86 (1614)	0.9 (2451)	canopy biomass basal area spruce vol soil wetness	0.390827 0.372103 0.347565 0.319847 -0.304831	elevation soil wetness canopy forest type disttoCC5	0.1710 0.0986 0.0950 0.0806 0.0800
12	0.38	598	0.88	0.88 (313)	0.85 (285)	canopy biomass basal area spruce vol soil type	0.541835 0.484620 0.436271 0.384780 0.288572	canopy biomass disttoCC5 basal area elevation	0.1762 0.1053 0.0879 0.0858 0.0760
13	0.42	1183	0.93	0.94 (469)	0.96 (714)	forest type spruce vol canopy soil wetness biomass	-0.754661 0.513449 0.397888 -0.369100 0.350702	forest type spruce vol soil wetness basal area biomass	0.4396 0.1108 0.0565 0.0470 0.0464
14	0.34	522	0.89	0.88 (242)	0.89 (280)	spruce vol biomass elevation canopy basal area	0.515619 0.512159 -0.511675 0.501779 0.477932	canopy soil wetness disttoCC10 biomass spruce vol	0.1347 0.1259 0.1159 0.0937 0.0886
15	0.31	2891	0.88	0.88 (1369)	0.89 (1522)	canopy biomass basal area spruce vol disttoCC5	0.461871 0.430926 0.397680 0.382887 -0.250006	canopy disttoCC5 biomass basal area elevation	0.1312 0.1011 0.0982 0.0919 0.0869

Overall assessment:
Weighted accuracy (cross-validated): 0.894

Series from Lund University

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-volkano: 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).

58. *Julia Hjalmarsson:* A Weighty Issue: Estimation of Fire Size with Geographically Weighted Logistic Regression (2016).

59. *Mathewos Tamiru Amato:* Using multi-criteria evaluation and GIS for chronic food and nutrition insecurity indicators analysis in Ethiopia (2016).

60. *Karim Alaa El Din Mohamed Soliman El Attar:* Bicycling Suitability in Downtown, Cairo, Egypt (2016).

61. *Gilbert Akol Echelai:* Asset Management: Integrating GIS as a Decision Support Tool in Meter Management in National Water and Sewerage Corporation (2016).

62. *Terje Slinning:* Analytic comparison of multibeam echo soundings (2016).

63. *Gréta Hlín Sveinsdóttir:* GIS-based MCDA for decision support: A framework for wind farm siting in Iceland (2017).

64. *Jonas Sjögren:* Consequences of a flood in Kristianstad, Sweden: A GIS-based analysis of impacts on important societal functions (2017).

65. *Nadine Raska:* 3D geologic subsurface modelling within the Mackenzie Plain, Northwest Territories, Canada (2017).

66. *Panagiotis Symeonidis*: Study of spatial and temporal variation of atmospheric optical parameters and their relation with PM 2.5 concentration over Europe using GIS technologies (2017).

67. *Michaela Bobeck:* A GIS-based Multi-Criteria Decision Analysis of Wind Farm Site Suitability in New South Wales, Australia, from a Sustainable Development Perspective (2017).

68. *Raghdaa Eissa*: Developing a GIS Model for the Assessment of Outdoor Recreational Facilities in New Cities Case Study: Tenth of Ramadan City, Egypt (2017).

69. *Zahra Khais Shahid*: Biofuel plantations and isoprene emissions in Svea and Götaland (2017).

70. *Mirza Amir Liaquat Baig*: Using geographical information systems in epidemiology: Mapping and analyzing occurrence of diarrhea in urban - residential area of Islamabad, Pakistan (2017).

71. *Joakim Jörwall*: Quantitative model of Present and Future well-being in the EU-28: A spatial Multi-Criteria Evaluation of socioeconomic and climatic comfort factors (2017).

72. *Elin Haettner*: Energy Poverty in the Dublin Region: Modelling Geographies of Risk (2017).

73. *Harry Eriksson*: Geochemistry of stream plants and its statistical relations to soil- and bedrock geology, slope directions and till geochemistry. A GIS-analysis of small catchments in northern Sweden (2017).

74. *Daniel Gardevärn:* PPGIS and Public meetings – An evaluation of public participation methods for urban planning (2017).

75. *Kim Friberg:* Sensitivity Analysis and Calibration of Multi Energy Balance Land Surface Model Parameters (2017).

76. *Viktor Svanerud:* Taking the bus to the park? A study of accessibility to green areas in Gothenburg through different modes of transport (2017).

77. *Lisa-Gaye Greene*: Deadly Designs: The Impact of Road Design on Road Crash Patterns along Jamaica's North Coast Highway (2017).

78. *Katarina Jemec Parker*: Spatial and temporal analysis of fecal indicator bacteria concentrations in beach water in San Diego, California (2017).

79. *Angela Kabiru*: An Exploratory Study of Middle Stone Age and Later Stone Age Site Locations in Kenya's Central Rift Valley Using Landscape Analysis: A GIS Approach (2017).

80. *Kristean Björkmann*: Subjective Well-Being and Environment: A GIS-Based Analysis (2018).

81. *Williams Erhunmonmen Ojo*: Measuring spatial accessibility to healthcare for people living with HIV-AIDS in southern Nigeria (2018).

82. *Daniel Assefa*: Developing Data Extraction and Dynamic Data Visualization (Styling) Modules for Web GIS Risk Assessment System (WGRAS). (2018).

83. *Adela Nistora*: Inundation scenarios in a changing climate: assessing potential impacts of sea-level rise on the coast of South-East England (2018).

84. *Marc Seliger*: Thirsty landscapes - Investigating growing irrigation water consumption and potential conservation measures within Utah's largest master-planned community: Daybreak (2018).

85. *Luka Jovičić*: Spatial Data Harmonisation in Regional Context in Accordance with INSPIRE Implementing Rules (2018).

86. *Christina Kourdounouli*: Analysis of Urban Ecosystem Condition Indicators for the Large Urban Zones and City Cores in EU (2018).

87. *Jeremy Azzopardi*: Effect of distance measures and feature representations on distance-based accessibility measures (2018).

88. *Patrick Kabatha*: An open source web GIS tool for analysis and visualization of elephant GPS telemetry data, alongside environmental and anthropogenic variables (2018).

89. *Richard Alphonce Giliba*: Effects of Climate Change on Potential Geographical Distribution of Prunus africana (African cherry) in the Eastern Arc Mountain Forests of Tanzania (2018).

90. *Eiður Kristinn Eiðsson*: Transformation and linking of authoritative multi-scale geodata for the Semantic Web: A case study of Swedish national building data sets (2018).

91. *Niamh Harty*: HOP!: a PGIS and citizen science approach to monitoring the condition of upland paths (2018).

92. *José Estuardo Jara Alvear*: Solar photovoltaic potential to complement hydropower in Ecuador: A GIS-based framework of analysis (2018).

93. *Brendan O'Neill*: Multicriteria Site Suitability for Algal Biofuel Production Facilities (2018).

94. *Roman Spataru*: Spatial-temporal GIS analysis in public health – a case study of polio disease (2018).

95. *Alicja Miodońska*: Assessing evolution of ice caps in Suðurland, Iceland, in years 1986 - 2014, using multispectral satellite imagery (2019).

96. *Dennis Lindell Schettini*: A Spatial Analysis of Homicide Crime's Distribution and Association with Deprivation in Stockholm Between 2010-2017 (2019).

97. *Damiano Vesentini*: The Po Delta Biosphere Reserve: Management challenges and priorities deriving from anthropogenic pressure and sea level rise (2019).

98. *Emilie Arnesten*: Impacts of future sea level rise and high water on roads, railways and environmental objects: a GIS analysis of the potential effects of increasing sea levels and highest projected high water in Scania, Sweden (2019).

99. *Syed Muhammad Amir Raza*: Comparison of geospatial support in RDF stores: Evaluation for ICOS Carbon Portal metadata (2019).

100. *Hemin Tofiq*: Investigating the accuracy of Digital Elevation Models from UAV images in areas with low contrast: A sandy beach as a case study (2019).

101. *Evangelos Vafeiadis*: Exploring the distribution of accessibility by public transport using spatial analysis. A case study for retail concentrations and public hospitals in Athens (2019).

102. *Milan Sekulic*: Multi-Criteria GIS modelling for optimal alignment of roadway by-passes in the Tlokweng Planning Area, Botswana (2019).

103. *Ingrid Piirisaar*: A multi-criteria GIS analysis for siting of utility-scale photovoltaic solar plants in county Kilkenny, Ireland (2019).

104. *Nigel Fox*: Plant phenology and climate change: possible effect on the onset of various wild plant species' first flowering day in the UK (2019).

105. *Gunnar Hesch*: Linking conflict events and cropland development in Afghanistan, 2001 to 2011, using MODIS land cover data and Uppsala Conflict Data Programme (2019).

106. *Elijah Njoku*: Analysis of spatial-temporal pattern of Land Surface Temperature (LST) due to NDVI and elevation in Ilorin, Nigeria (2019).

107. *Katalin Bunyevácz*: Development of a GIS methodology to evaluate informal urban green areas for inclusion in a community governance program (2019).

108. *Paul dos Santos*: Automating synthetic trip data generation for an agent-based simulation of urban mobility (2019).

109. *Robert O' Dwyer*: Land cover changes in Southern Sweden from the mid-Holocene to present day: Insights for ecosystem service assessments (2019).

110. *Daniel Klingmyr*: Global scale patterns and trends in tropospheric NO2 concentrations (2019).

111. *Marwa Farouk Elkabbany*: Sea Level Rise Vulnerability Assessment for Abu Dhabi, United Arab Emirates (2019).

112. *Jip Jan van Zoonen*: Aspects of Error Quantification and Evaluation in Digital Elevation Models for Glacier Surfaces (2020).

113. *Georgios Efthymiou*: The use of bicycles in a mid-sized city – benefits and obstacles identified using a questionnaire and GIS (2020).

114. *Haruna Olayiwola Jimoh*: Assessment of Urban Sprawl in MOWE/IBAFO Axis of Ogun State using GIS Capabilities (2020).

115. *Nikolaos Barmpas Zachariadis*: Development of an iOS, Augmented Reality for disaster management (2020).

116. *Ida Storm*: ICOS Atmospheric Stations: Spatial Characterization of CO2 Footprint Areas and Evaluating the Uncertainties of Modelled CO2 Concentrations (2020).

117. *Alon Zuta*: Evaluation of water stress mapping methods in vineyards using airborne thermal imaging (2020).

118. *Marcus Eriksson*: Evaluating structural landscape development in the municipality Upplands-Bro, using landscape metrics indices (2020).

119. *Ane Rahbek Vierø*: Connectivity for Cyclists? A Network Analysis of Copenhagen's Bike Lanes (2020).

120. *Cecilia Baggini*: Changes in habitat suitability for three declining Anatidae species in saltmarshes on the Mersey estuary, North-West England (2020).

121. *Bakrad Balabanian*: Transportation and Its Effect on Student Performance (2020).

122. *Ali Al Farid*: Knowledge and Data Driven Approaches for Hydrocarbon Microseepage Characterizations: An Application of Satellite Remote Sensing (2020).

123. *Bartlomiej Kolodziejczyk*: Distribution Modelling of Gene Drive-Modified Mosquitoes and Their Effects on Wild Populations (2020).

124. *Alexis Cazorla*: Decreasing organic nitrogen concentrations in European water bodies - links to organic carbon trends and land cover (2020).

125. *Kharid Mwakoba*: Remote sensing analysis of land cover/use conditions of community-based wildlife conservation areas in Tanzania (2021).

126. *Chinatsu Endo*: Remote Sensing Based Pre-Season Yellow Rust Early Warning in Oromia, Ethiopia (2021).

127. *Berit Mohr*: Using remote sensing and land abandonment as a proxy for long-term human out-migration. A Case Study: Al-Hassakeh Governorate, Syria (2021).

128. *Kanchana Nirmali Bandaranayake*: Considering future precipitation in delineation locations for water storage systems - Case study Sri Lanka (2021).

129. *Emma Bylund*: Dynamics of net primary production and food availability in the aftermath of the 2004 and 2007 desert locust outbreaks in Niger and Yemen (2021).

130. *Shawn Pace*: Urban infrastructure inundation risk from permanent sea-level rise scenarios in London (UK), Bangkok (Thailand) and Mumbai (India): A comparative analysis (2021).

131. *Oskar Evert Johansson*: The hydrodynamic impacts of Estuarine Oyster reefs, and the application of drone technology to this study (2021).

132. *Pritam Kumarsingh*: A Case Study to develop and test GIS/SDSS methods to assess the production capacity of a Cocoa Site in Trinidad and Tobago (2021).

133. *Muhammad Imran Khan*: Property Tax Mapping and Assessment using GIS (2021).

134. *Domna Kanari*: Mining geosocial data from Flickr to explore tourism patterns: The case study of Athens (2021).

135. *Mona Tykesson Klubien*: Livestock-MRSA in Danish pig farms (2021).

136. *Ove Njøten*: Comparing radar satellites. Use of Sentinel-1 leads to an increase in oil spill alerts in Norwegian waters (2021).

137. *Panagiotis Patrinos*: Change of heating fuel consumption patterns produced by the economic crisis in Greece (2021).

138. *Lukasz Langowski*: Assessing the suitability of using Sentinel-1A SAR multi-temporal imagery to detect fallow periods between rice crops (2021).

139. *Jonas Tillman*: Perception accuracy and user acceptance of legend designs for opacity data mapping in GIS (2022).

140. *Gabriela Olekszyk*: ALS (Airborne LIDAR) accuracy: Can potential low data quality of ground points be modelled/detected? Case study of 2016 LIDAR capture over Auckland, New Zealand (2022).

141. *Luke Aspland*: Weights of Evidence Predictive Modelling in Archaeology (2022).

142. *Luís Fareleira Gomes*: The influence of climate, population density, tree species and land cover on fire pattern in mainland Portugal (2022).

143. *Andreas Eriksson*: Mapping Fire Salamander (Salamandra salamandra) Habitat Suitability in Baden-Württemberg with Multi-Temporal Sentinel-1 and Sentinel-2 Imagery (2022).

144. *Lisbet Hougaard Baklid*: Geographical expansion rate of a brown bear population in Fennoscandia and the factors explaining the directional variations (2022).

145. *Victoria Persson*: Mussels in deep water with climate change: Spatial distribution of mussel (Mytilus galloprovincialis) growth offshore in the French Mediterranean with respect to climate change scenario RCP 8.5 Long Term and Integrated Multi-Trophic Aquaculture (IMTA) using Dynamic Energy Budget (DEB) modelling (2022).

146. *Benjamin Bernard Fabien Gérard Borgeais*: Implementing a multi-criteria GIS analysis and predictive modelling to locate Upper Palaeolithic decorated caves in the Périgord noir, France (2022).

147. *Bernat Dorado-Guerrero*: Assessing the impact of post-fire restoration interventions using spectral vegetation indices: A case study in El Bruc, Spain (2022).

148. *Ignatius Gabriel Aloysius Maria Perera*: The Influence of Natural Radon Occurrence on the Severity of the COVID-19 Pandemic in Germany: A Spatial Analysis (2022).

149. *Mark Overton*: An Analysis of Spatially-enabled Mobile Decision Support Systems in a Collaborative Decision-Making Environment (2022).

150. *Viggo Lunde*: Analysing methods for visualizing time-series datasets in open-source web mapping (2022).

151. *Johan Viscarra Hansson*: Distribution Analysis of Impatiens glandulifera in Kronoberg County and a Pest Risk Map for Alvesta Municipality (2022).

152. *Vincenzo Poppiti*: GIS and Tourism: Developing strategies for new touristic flows after the Covid-19 pandemic (2022).

153. *Henrik Hagelin*: Wildfire growth modelling in Sweden - A suitability assessment of available data (2023).

154. *Gabriel Romeo Ferriols Pavico*: Where there is road, there is fire (influence): An exploratory study on the influence of roads in the spatial patterns of Swedish wildfires of 2018 (2023).

155. *Colin Robert Potter*: Using a GIS to enable an economic, land use and energy output comparison between small wind powered turbines and large-scale wind farms: the case of Oslo, Norway (2023).

156. *Krystyna Muszel*: Impact of Sea Surface Temperature and Salinity on Phytoplankton blooms phenology in the North Sea (2023).

157. *Tobias Rydlinge*: Urban tree canopy mapping - an open source deep learning approach (2023).

158. *Albert Wellendorf*: Multi-scale Bark Beetle Predictions Using Machine Learning (2023).