USE OF GEOSPATIAL METHODS TO CHARACTERIZE DISPERSION OF THE EMERALD ASH BORER IN SOUTHERN ONTARIO, CANADA

Farah Tasneem

A Thesis Submitted to the Faculty of Graduate Studies
in Partial Fulfillment of the Requirements for the Degree of
Master of Science

Graduate Program in Earth and Space Science
York University
Toronto, Ontario, Canada

© Farah Tasneem, 2019

ABSTRACT

Since the introduction of the Asian Emerald Ash Borer beetle (EAB, *Agrilus planipennis*) to Southern Ontario in 2002, the condition of all species of Ash trees *(Fraxinus)* in the province is currently at risk. Due to the aggressive nature of this beetle, early detection is critical in its eradication. Although species distribution modelling is not a new concept, several issues need to be addressed in order to increase its predictive accuracy. In this research, the effects of positive spatial autocorrelation as a result of sampling bias and data prevalence (i.e., proportion of absence to presence points) were investigated in an EAB dataset by applying a filtering distance threshold and employing various ratios of EAB presence to absence points during the modelling process. To analyze the impact of environmental and anthropogenic predictors on the distribution of the EAB, logistic regression, Random Forest (RF) and a hybrid of Random Forest and GLM known as the Random Generalized Linear Model (RGLM) were applied to EAB data from 2006-2012 across Ontario. Ultimately, three risk maps were created from the 2006-2012 EAB data by using the coefficients from logistic regression as weights and the creation of a risk map tool for RF and RGLM was used to validate the prediction dataset from 2013. High risk areas were identified from the risk maps for species prevalence and distribution monitoring. From these, precautionary measures can be implemented to stem the expansion of the beetle and thus reduce the destruction of the Ash tree species. All models identified June wind speed as the most important predictor variable followed by population centres. In terms of model transferability, logistic regression, Random Forest and RGLM achieved approximately 94% on the validation dataset. For the prediction dataset, RGLM had the best extrapolation accuracy (84%), followed by stepwise backward logistic regression (70%), and Random Forest (52%).

Acknowledgements

I would first like to sincerely thank my supervisor, Baoxin Hu. Her academic and moral guidance always pointed me in the right direction and helped me fill in my gaps where I lacked knowledge. She was very supportive about my ideas and aided in the completion of this research. In addition, I would like to thank Jianguo for diligently editing my drafts and providing helpful suggestions. Faizan and Wendy from my lab deserve recognition for their excellent expertise and support in some unfamiliar concepts encountered during my research. Cheryl and Ima provided very good company and moral support in the lab.

Next, I would like to thank Krista Amolins and Brent Hall from Esri Canada for their tremendous support and helpful suggestions for various research processes and also for Esri's financial contribution towards the research. In addition, I would like to extend gratitude to NSERC for their financial contribution as well. Haydn Lawrence at Esri deserves a huge shout out for his guidance in the creation of the risk map tool.

A special thanks goes out to the staff from the Canadian Food Inspection Agency (CFIA), Mireille Marcotte and Cameron Duff for their steadfast preparation of the EAB presence/absence points dataset which were the backbone of this research. Without access to the high quality sampled points spanning from several years in Ontario, the results of the model would certainly have been compromised.

Lastly, a huge thanks to my mother, Ayon and Rakin for their unconditional support and advice which motivated me to be complete my thesis sincerely and competently.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

The emerald ash borer (EAB, *Agrilus planipennis*) continues to severely impact ash trees since its invasion in 2002 in Windsor, Ontario. The stealthy nature of the beetle helped it stay undetected for approximately 10 years prior to its discovery in 2002 (de Groot et al., 2006; Fairmaire & Parsons, 2008). The EAB is an extremely aggressive pest as "in areas with a well established EAB population, ash trees can be mass attacked and killed in as little as two growing seasons" (Gaetz & Hildebrand, 2012, p. 1).

In the initial stages of the EAB outbreak in Canada, various US states including Ohio and Michigan relied heavily on surveys and ash-tree removal in an attempt to slow its spread. Michigan and Ohio had removed all ash trees within approximately 400 metres of the infested front. While the USA focused mainly on tree-removal strategies, Canada relied heavily on visual surveys during the early years of EAB infestation (Marchant, 2012). However, visual surveys are not the most reliable form of surveying as ash trees only exhibit visible symptoms until after a year following infestation (BenDor et al., 2006; Pontius et al., 2008).

Although surveying and detection methods may be effective in preparing conservation authorities for the arrival of the EAB in a particular area, the areas chosen to be surveyed are usually premeditated and spontaneous EAB emergence in the unregulated areas can go unnoticed. For this reason, predictive species distribution models using relevant anthropogenic and environmental factors can be used as a tool to draw attention to high-risk areas using risk maps derived from the models. BenDor et. al., (2006) modelled the impact of hypothetical anthropogenic enforcements such as firewood quarantines and EAB eradication programs on the attenuation of the EAB in an uninfected area in DuPage County, Illinois. The main inputs for the simulations were an EAB population model abstracted by the life cycle of the beetle and an ash tree population model which approximated the seeding and life span of the ash trees. Although BenDor et al. (2006)'s model served to be essentially hypothetical, it still required specified inputs such as an ash tree inventory dataset which may not be readily accessible for a larger study area.

Prasad et al. (2010) used a more practical approach to model the EAB in Ohio by combining historical EAB core zones (i.e., 1998-2006) and field sampled ash trees from 2004-2005 with relevant factors such as ash density, traffic density, distance to campgrounds and wood product industries. In addition, a predictor variable known as the "distance to the nearest EAB positive location from previous years" was included which added a spatio-temporal dimension to the model. In theory, Prasad et al. (2010)'s model calculated the probability of a cell becoming infested with the EAB and was validated by overlaying EAB points from 2007. Although the methodology used by Prasad et al. (2010) was pragmatic, the likelihood of positive autocorrelation present in the EAB dataset used to train the model was not addressed which violated the assumption of independent observations during modelling (Bebber, 1999; F. Dormann et al., 2007).

The most recent research on the EAB was conducted by Huset (2013) using similar anthropogenic variables as Prasad et al. (2010) but also incorporated various climactic variables such as precipitation and temperature in order to assess the impact of climate change. Huset (2013) used logistic regression and a popular presence-only machine learning model known as Maxent to identify the most important variables and visualize the spread of the EAB. Unlike Prasad et al. (2010), Huset (2013)'s methodology investigated spatial autocorrelation and addressed it by aggregating the presence points in a cell-wise basis on a symmetrical grid. However, Huset (2013)'s approach of addressing autocorrelation appeared exhaustive as it reduced the presence points by a factor of 10. Secondly, another drawback in Huset (2013)'s research was the ambiguous procedure used to create the risk map from the logistic regression results. For instance, Huset (2013) used the p-values of the significant variables identified by logistic regression as weights for the risk map of the study area. The more rational method to produce the risk map is to use the coefficients associated with the explanatory variables in logistic regression's equation as weights to produce EAB risk probability of each cell ranging from 0 to 1 (Mousavi et al., 2011; Ohlmacher & Davis, 2003). Although the risk map generation for logistic regression is fairly straight forward, the same cannot be concluded for the machine learning models where the risk probabilities are more difficult to display in a cartographic GIS platform (Holcombe et al., 2007).

While designing a species distribution model, the quality of the species and predictor variables can affect the model's authenticity which results in a poor transferability. For instance, since tree insect data is usually collected in premeditated areas such as near urban centres, campgrounds, and known areas of tree decline, the sampled points does not represent the true range of the species due to sampling bias. Secondly, temporal and scale discrepancies among the predictor variables raises questions about the integrity of the model. Lastly, the presence of multicollinearity amongst the predictor variables interferes with the variable importance rankings of the models (Hegyi & Laczi, 2015). Whereas excluding predictor variables that exhibit high correlation coefficients and VIF values is a way to handle multicollinearity prior to modelling for regression techniques, dealing with multicollinearity or correlated variables in machine learning models such as Random Forest lacks a breakthrough (Gregorutti et al., 2017).

With regards to the models used in this research, one of the most widely recognized statistical models used for species distribution purposes by ecologists is logistic regression, a type of generalized linear model (GLM), known for its ability to handle a binary distribution of species data such as presence and absence (Holcombe et al., 2007). Logistic regression is the most appropriate statistical model for this research because it takes a user-defined number of presence and absence species data, combines them with explanatory variables and derives an index of EAB risk in an study area ranging from 0 (low risk) to 1 (high risk). Aside from risk prediction, logistic regression also includes a variable importance ranking measure and regression coefficients which provide information about the positive or negative effects a predictor variable has on the response variable.

Aside from the conventional statistical models such as logistic regression, further advancements in analytical tools and software persuaded ecologists to explore alternate disciplines of species distribution such as machine learning, a type of algorithmic model that can be used to model a binary species distribution (Cushman & Huettmann, 2010). Given the complex nature of habitat selection for a species, the relationship between variables may be nonlinear or scale-dependent (Drew et al., 2011). For this reason, non-parametric machine learning approaches are attractive options that can overcome non-linear variable interactions and heterogeneity across spatial scales (Drew et al., 2011). Under the realm of machine learning, an algorithm known as Random Forest is gaining traction in the ecological community due to its brilliant predictive powers. Aside from its high prediction accuracy, Random Forest also includes variable importance measures (Archer & Kimes, 2008) unlike other machine learning methods such as k-nearest neighbours, support vector machines (SVM), and neural networks.

The third model that will be explored in this research is a hybrid between logistic regression and Random Forest known as random generalized linear model (RGLM) which aggregates the results from a number of logistic regression models contained in “bags” using randomly sampled data. In essence, RGLM combines the highly accurate ensemble classifier Random Forest with the interpretability of a forward logistic regression model (Song et al., 2013). That said, aside from classification accuracies, RGLM provides mean regression coefficients across all bags to assess the positive and negative effects that the predictor variables have on the response variable. RGLM was initially used for detecting cancer in gene expression datasets by Song et al. (2013) and has not been used in species distribution modelling. Although the three models did not include a dynamic component, in this research, a spatio-temporal dimension was added to the models by generating a variable known as “distance to the nearest EAB positive location from previous years". Moreover, this variable represented the significance of previously EAB-detected locations on sampled points from a current year.

Ultimately, the main research goals from this research is two-fold: address issues with data quality prior to modelling and devise a systematic framework to predict the dispersal of the EAB in Ontario. By addressing the problems associated with the species data and predictor variables, the transferability of the models can be increased and be effectively used as a template to predict EAB risk in a non-infested area. In order to achieve the research goals, four specific research objectives were devised:

1) Minimize high spatial autocorrelation present in the species data caused by sampling bias by applying a distance threshold to filter the data (Boria et al., 2014; Veloz, 2009) and produce spatially independent samples.

2) Assess multicollinearity among the predictor variables and determine ways to address it in the modelling process.

3) Investigate the performance of a statistical model (logistic regression) and two machine learning models (Random Forest and RGLM) on the predictive modelling of the EAB using various model performance indicators.

4) Design an automated risk map creation tool for the machine learning models to visualize areas prone to EAB infestation.

Furthermore, the layout of the thesis is as follows: following the introduction in Chapter 1, Chapter 2 discusses the origins of the EAB and its host species, detection methods, remediation strategies and a literature review on relevant research. Chapter 3 includes details about the study area, the species and predictor data used in this research and their limitations. Chapter 4 describes the methodology which includes details about data pre-processing and the species distribution models used. Chapter 5 includes the results and discussions on the findings. Lastly, Chapter 6 summarizes the main outcomes with regards to the research objectives and provides recommendations for future research. All in all, in the context of early detection of the EAB, the workflow devised from this research can be used to determine the relative EAB risk of a non-infested county or region.

CHAPTER 2: BACKGROUND & LITERATURE REVIEW

2.1 The EAB and Ash Species

The Forest Service defines an invasive species as “a significant environmental and economic threat to the Nation’s forests and rangelands” (Dix et al., 2010). The term “invasive” can be described as a foreign species that is not native to an area and become the dominant predators by consuming natural resources and outcompeting native species. That is to say, the emerald ash borer (EAB) is an invasive and exotic pest which arrived in Southern Ontario by attaching onto untreated ash wood used for packing material in ships.

The primary hosts of the EAB are ash trees belonging to the genus *Fraxinus* comprised of white ash (*F. americana*), green ash (*F. pennsylvanica*) and black ash (*F. nigra*) (McCullough and Katovich, 2004). Ash trees are an essential part of Southern Ontario’s native tree collection and make up a majority of woodlots, fence rows and trees surrounding water courses (“Emerald Ash Borer Management Plan update”, 2013). Ash trees are also a very popular choice in urban areas, both public and private due to their rapid growth and environmental adaptability. With regards to the economic impacts, green and white species are an important component of the hardwood forest industry as they are used to make cabinetry and sporting goods (Fairmaire & Parsons, 2008).

Upon landing on an ash tree, the EAB lays eggs approximately 0.6 to 1 mm in size under bark crevices (Appleton et al., 2017). When the eggs become larvae, they hatch within a few weeks and feed on vascular tissue of the tree creating S-shaped galleries. A majority of the larvae become pupae and remain in the tree's pupal chambers. When they become adults, the EAB emerges from distinct D-shaped exit holes.

2.2 Detection and Delimitation Efforts

The unprecedented incursion of the EAB has sparked the interest of conservation authorities and many precautionary measures have been taken to slow down its spread. Two of the native forms of invasive species detection are detection and delimitation surveys (Marchant, 2012). A detection survey essentially gathers data about the presence or absence of a pest in an area. However, it is are not necessarily designed to acquire the number of insects in the given area. It is usually conducted by a regulatory agency such as the Canadian Food Inspection Agency (CFIA) and aims to find physical evidence of the EAB in a host tree or nearby trap. Areas that are more prone to be inhabited by the EAB via human activities are given more priority such as campgrounds, trailer parks, sawmills and firewood suppliers, tree nurseries, rest-stops along major highways, and industrial areas which receive off-shore shipments (Marchant, 2012). On the other hand, a delimitation survey is conducted to quantify the density of the EAB found in a sampled tree. Delimitation surveys inform samplers about the age and severity of an EAB infestation but are often more labour-intensive and expensive than detection surveys. Currently, several municipalities in Ontario are conducting detection surveys along with tree protection programs.

Detection and delimitation surveys are conducted using four main methods: prism traps, visual examination, branch sampling and aerial and hyperspectral imaging. Prism traps have become a popular tool that is being incorporated into detection surveys by the CFIA. Prism traps are baited with a sticky green chemical known as Z-3-Hexenol to trap adult beetles and have proven to be quite effective in the early detection of the EAB (Marchant, 2012). As a reminder, prism traps used in detection surveys are not designed to quantitatively determine how many trees are infested, but rather whether trees at a particular location *are* infested. With this method in mind, there is a potential risk of false negative data for cases with low EAB population levels.

When the population of the EAB is significantly low and EAB adults are not detected, there is a chance of erroneously identifying a county as EAB-free when in fact it might be infested (Knight, et al., 2013; Pontius et al., 2008). On the other hand, false positive results might also be obtained if adult beetles are transported by wind or human vectors to a survey area.

The second method to detect and delimit the EAB is via visual survey. Visual surveys examine trees at the ground or canopy level. They are less accurate but also less invasive than similar surveying methods such as branch sampling. A caveat with visual surveying is that because signs and symptoms of EAB infestation do not appear until five years after the initial attack, the presence of the EAB might go undetected using visual surveys. Visual surveys are also very selective as regulatory agencies often select trees to perform visual surveys that are deemed to be higher risk based on their proximity to lumber yards, campgrounds, parks, sawmills and firewood processing facilities (Marchant, 2012).

The primary delimitation tool used in Canada is branch sampling and was recently developed by the Canadian Forest Service (CFS). It consists of sampling and the dissection of several branches of potentially infested ash trees delineated by their distances from the target facilities mentioned previously. Branch sampling is far more accurate than visual surveying because it is more effective at identifying infected ash trees at its initial stages with regards to signs and symptoms of the EAB. Although branch sampling is the most preferred method for early detection of the EAB, it is also costlier and more labour-intensive. However, to alleviate some of the costs and efforts associated with branch sampling, it can be integrated with other maintenance activities conducted by municipal forestry departments (Marchant, 2012).

The last more broad-scale method for detection and delimitation of the EAB uses Hyperspectral Imaging (HSI) to detect early EAB infection from tree canopy. The use of remote sensing technology such as hyperspectral imagery has the potential to identify signs of EAB infection before they become visible to the human eye (Marchant, 2012; Pontius et al., 2008). Hyperspectral imagery consists of hundreds of narrow adjoining spectral bands which allow a greater level of differentiation between objects of the same type (i.e., stressed vs. healthy tree) than multispectral imagery. In terms of tree physiology, stressed leaves of trees have reduced photosynthetic activity and chlorophyll content. As a result, subtle differences in chlorophyll content can be picked up by hyperspectral sensors in visible and near-infrared (NIR) bands (Pontius et al., 2008). Various wavelength indices can also be derived from hyperspectral imagery to optimize the detection of stressed ash trees.

In Canada, Hyperspectral imagery was tested in Oakville in 2010 by the USDA-Forest Service on the identification of ash trees and level of infestation. An accuracy of 80% in the identification of ash trees from other trees was achieved but an accuracy assessment on ash health was not performed because there were not enough spectral signatures from the field data (Hanou, 2011). While previous EAB detection procedures focused primarily on one method, an integrated assessment of visualizing the dispersal of an invasive species can be performed through predictive species modelling by combining historic species data with relevant explanatory variables.

2.3 Remediation Strategies

Although the ideal way to diminish the EAB is through the aforementioned “early detection”, it is not as straight forward as trees only show symptoms years after the attack (Marchant, 2012; Pontius et al., 2008). After the infestation of the EAB has taken place, remediation strategies can be applied to ensure the EAB does not spread such as implementing an ash free zone, quarantine of infected areas and injection of pesticides (Marchant, 2012).

An ash-free or a “firebreak” zone was implemented in the fall of 2003 by the CFIA west of the Chatham-Kent county in Southern Ontario. This approach included creating a barrier on the leading front of EAB spread by removing all ash trees in its path. This zone spanned 10 km and approximately 85, 000 ash trees were removed (Marchant, 2012). However, this method created a lot of controversy among residents and property rights activists and as a result was discontinued by 2005. The second remediation method includes establishing quarantines around EAB infested areas and restricting the movement of firewood. In 2005 and 2006, the CFIA regulated areas where the EAB has been spotted and restricted the movement of ash trees and all firewood from these zones (Gaetz & Hildebrand, 2012; Marchant, 2012). Within quarantined zones, individual infested trees were further quarantined using a radial zone of 5 kilometres. Quarantined specialists emphasized the effectiveness of quarantined zones in preventing new EAB outliers through human activities.

A remediation approach that affects ash trees at the individual level is the injection of an insecticide known as TreeAzin^{TM} (Azadirachtin) directly into the trees ("Emerald Ash Borer," 2012; Fairmaire & Parsons, 2008). TreeAzin's main ingredient is an extract from the seeds of the Indian neem tree. TreeAzin was first introduced in Canada in 2012 and is injected into the base of ash trees every two years until it effectively kills the EAB larvae by interrupting larval shedding (Fairmaire & Parsons, 2008; Gaetz & Hildebrand, 2012).

2.4 Species Distribution Models

As previously mentioned, the main objective of this research is to develop a species distribution model for the EAB in order to highlight the areas that may be potentially at risk in the foreseeable future based on the species' data from previous years. In order to develop the spread model, previous spread modelling methods on the EAB by several authors were explored. Existing EAB modelling papers by BenDor et al. (2006), Prasad et al. (2010) and Huset (2013) investigate natural and anthropogenic factors that influence the spread of the EAB using species distribution models. While BenDor et al. (2006), Prasad *et al.* (2010) used dynamic spatial models to establish the spread of the EAB, Huset (2013) used statistical and machine learning methods.

BenDor et al. (2006) simulated the effects of firewood quarantine and an EAB eradication program on the spread of the EAB in DuPage County, IL, a county in the Chicago metropolitan area not yet infested by the EAB at the time. In their research, BenDor et al. (2006) used a system dynamics model called STELLA and a Spatial Modeling Environment (SME) to test three different scenarios of EAB dispersal: the distribution of trees for different land uses, the establishment of firewood quarantine zones to limit anthropogenic influence and the implementation of an EAB eradication program. According to BenDor et al. (2006), incorporating a system dynamics model (STELLA) with a spatio-temporal model (SME) effectively captures the environmental heterogeneity presented across a study area. The SME used in BenDor et al. (2006)'s research incorporated the generic system dynamics model STELLA into a spatial array that is similar to cellular automata modelling. Within the SME, the features of the spatial data were used to create a matrix of independent spatially-specific system dynamics models.

STELLA was developed by using two sub-models: the EAB population model and the ash tree population model. The sub-models were characterized by the parasitic relationship between the EAB larvae and live ash bark. The models simulated the act of the EAB larvae consuming the live ash trees thereby reducing the amount of available bark area and propagating the EAB to farther distances. The EAB model was characterized by the life cycle of the beetle and model parameters corresponded to each life stage. With the study area organized as a matrix, it was assumed that the decision for an EAB to migrate to a cell is dependent on the density of the EAB adults in the cell and the total area of tree bark in the cell. For the purposes of modelling, an upper bound on EAB adult density was approximated as 4 adults per square metre of ash bark area. The total tree bark area was calculated by approximating the tree trunk as a cylinder and averaging different heights and diameters of various age cohorts of the ash trees.

Next, the ash tree population model was created by visualizing the seeding and life span properties of ash trees. This was done by taking the average of the number of seeds produced by each tree in an average year. Within the STELLA model, it was assumed that the ash trees were planted by urban foresters and in order to stratify the levels of germination, a seed bank with a land-use dependent germination rate was simulated. Lastly, using an ash tree inventory dataset and high-resolution land-use data, the ash tree density (number of ash trees per cell) was estimated for each land-use category. An algorithm was used to assign ash trees into different land-use classes in decreasing order of importance: uplands, floodplain forests, partial canopy/savanna and urban open space, low/medium density urban, and high-density urban.

Following the implementation of the two sub-models into the SME framework, three different factors potentially affecting the spread of the EAB were simulated: the distribution of the ash trees and land-use, the ability of a county-wide firewood quarantine program to limit anthropogenic influence and the implementation of an EAB eradication program. BenDor et al. (2006)'s findings show that implementing a firebreak zone (a buffer around known EAB infested locations) and limiting firewood movement from infested sites attenuated EAB larvae spread and adult population (BenDor et al., 2006). All in all, although BenDor et al. (2006)'s research provided some realistic simulations of primarily anthropogenic influences, two of the model inputs were based on theoretical estimations of EAB density and ash germination rates which cannot be approximated at alternate spatial scales. In addition, BenDor et al. (2006)'s research requires many detailed parameters such as an abstraction of the biological characteristics of the EAB and ash trees which requires extensive research for valid results. While being aware of the strictly theoretical nature of the research, BenDor et al. (2006) concludes that using scenario-driven models can build on empirical information on the EAB population and spread dynamics in order to identify knowledge gaps and provide a framework for field surveys.

Furthermore, BenDor et al. (2006) simulated the potential spread of the EAB to a new area using various anthropogenic constraints while Prasad et al. (2010) used known EAB presence and absence points from 2003-2004, historic EAB maps and various relevant explanatory variables to simulate EAB spread in Ohio. In addition to obtaining known EAB presence points from Michigan's conservation services, Prasad and co-authors also conducted their own field visits and surveyed infested ash trees. In the research, a spatially explicit cellular spread model called SHIFT was used to combine sub-models of the insect's short-distance dispersal (insect flight model (IFM)) with human-mediated long-distance dispersal mechanisms (insect ride model (IRM)) to create a hybrid spread model. In a spatially explicit cellular model, a transition state model is fitted to a real system where the transition from one step to the next depends on the empirical relationship between the target cell and its neighbours.

While the only explanatory variable used by BenDor et al. (2006) was ash density, Prasad used Random Forest to identify the most important variables from variables such as ash density, human population density, traffic density, campground size and usage, and wood products industries. Using the variables previously mentioned, Prasad attempted to incorporate landscape heterogeneity in the SHIFT model. The first component of SHIFT, the insect flight model (IFM), calculates the probability of infestation of a future year based on EAB infestation of the previous year using an empirical spread rate of 20 km/year. In this research, the insect flight model was modified to a single predictor variable (distance to the nearest EAB positive location from previous years) in order to incorporate spatio-temporal aspects of EAB dispersal.

The second component of SHIFT, the insect ride model (IRM) uses the same mechanism as the insect flight model but using a search radius of 400 km to accommodate long distance dispersal. Two variables of the insect ride model (roads and campgrounds) were incorporated using a gravity model which argues that the movement of a test subject being modelled is motivated by the attractiveness of its destination. For instance, the gravity model calculated the number of campers that travel between two locations which was used to estimate the potential risk of the destination. The distance between a location and a campground was calculated as the road network distance between the location's zip code and the campground. Next, weights were assigned based on the density of people at a camp and its proximity to core zones. The variables population centres and wood product industries were categorized into classes according to population density and added to the SHIFT model.

In theory, the SHIFT model calculates the probability of a cell becoming infested with the EAB based on its distance to a nearby EAB-infested cell. The infestation probability for each unoccupied cell is a value from 0 to 1 and is characterized by their proximity to infested cells. SHIFT advances the EAB infested front based on its current location, the abundance of EAB behind the front, and the amount of ash trees ahead of the front.

In conclusion, the SHIFT model was able to correctly identify 97% of the known outlier EAB positive points that fell in medium to high risk zones on the risk map generated. Aside from the high accuracy of Prasad et al., (2010)'s model, the execution of independent field work on ash trees served as an asset to the validation process. Ultimately, Prasad et al., (2010) concluded that the distance from current EAB centres was the most important variable, followed by the distance to roads, population density and the influence from quarantined counties (Prasad et al., 2010). Regardless of the methods used, the work of Prasad et al., (2010) served as a great insight into the potential anthropogenic factors that should be included in this research. In addition, Prasad tested the transferability of the model using verified EAB presence points from a previous year (i.e., 2005) which is coincidentally one of the objectives of this research.

Pontius et al. (2008) shifted the focus from predictive spread modelling of the EAB to using purely remote sensing techniques to visualize the decline of ash trees. Pontius and co-authors used hyperspectral imagery collected in Michigan and Ohio to classify healthy and thinning ash canopies using various vegetation indices which examined chlorophyll and moisture content. Similar to Prasad et al. (2010), independent fieldwork was also employed in this research to collect ground-truth infested vs. non-infested ash trees with the assistance of a Trimble GPS unit. Unlike the research mentioned previously which incorporated primarily anthropogenic variables, Pontius performed stepwise linear regression on only biotic variables derived from vegetation bands such as a greenness index (GI), a water band ratio (WBI) and four chlorophyll-related indices. The risk map created by Pontius achieved a very high accuracy of 97% in terms of separating the five categories of ash decline. Although the classification results achieved by Pontius were informative, the usage of hyperspectral imagery to examine ash health for the large scale of Ontario would be far too costly and impractical. As a result, the more readily available multispectral imagery was used.

The most recent study on EAB spread by Huset in 2013 used logistic regression and maximum entropy modelling (Maxent) to identify the factors most associated with the presence of the EAB in New York. Previously, Prasad et al. (2010) and BenDor et al. (2006) examined the spread of the EAB using only ash density and anthropogenic sources. On the other hand, Huset (2013) incorporated climactic variables along with anthropogenic and land-use data. The 17 variables used by Huset (2013) included distance to campgrounds, wood product industries, water, known EAB locations, wind power, human population, percent forested, percent developed, NDVI, elevation, slope, aspect, precipitation, and temperature. The logistic regression model used in Huset (2013)'s research essentially predicted the probability of the outcome of a dependent variable based on the values of various independent variables. Since logistic regression requires a binary dependent variable, pseudo-absence points were generated as Huset (2013) only had access to EAB presence points from 2009-2011. In contrast, since both EAB presence and true absence points in Ontario were obtained for this research, the usage of logistic regression seems plausible.

On the other hand, the machine learning classifier Maxent requires presence-only species data. In theory, Maxent estimates the probability distribution of a target species by determining the probability distribution of maximum entropy that is the most uniform. In addition, a set of constraints is enforced in Maxent that represents incomplete information about the target distribution. For instance, if the sampled data exhibits sampling bias, the resulting predictions can be erroneous since the probability distribution is dependent on the observed presence data (Cushman & Huettmann, 2010). In such a case, the Maxent model explicitly assumes that the sample locations used in the model are compared to a sample of available locations across the study area. Within the Maxent model, the relationships between the response and the predictor variables can take on a variety of forms such as linear, quadratic, threshold and piecewise. In addition, the relationship between the response and predictor variables is assessed at different ranges or scales. For instance, a predictor variable could be modelled against the response variable using a linear function at its lowest range, using an interaction with another predictor in its middle range and using a threshold function in its upper range.

The outputs of the Maxent model include variable importance ranking and a validation scheme referred to as a "random test percentage" which uses a quarter of the dataset as testing data. The most visual output of Maxent is a risk map which represents the probability distribution of the species across the study area by finding the combination of predictor variables which maximizes the log-likelihood of the model. In order to prevent overfitting, the log-likelihood is penalized in Maxent by a regularization parameter which increases in accordance to the complexity of the model (Phillips & Dudík, 2008). Huset (2013)'s variable importance conclusions were consistent with Prasad et al. (2010)'s findings such that the distance to known locations of the EAB served as the most important variable (Huset, 2013). All in all, the incorporation of climactic variables with anthropogenic variables in Huset (2013)'s research was also implemented in this research along with some methodologies such as addressing multicollinearity of predictor variables, reducing positive autocorrelation of the sampled EAB points, determining the optimal ratio of EAB presence to absence points and creating the risk maps. That said, the descriptions of the three spread models used in this research are outlined below:

2.4.1 Logistic Regression

2.4.1.1 Model Description

Logistic regression belongs to the family of generalized linear models and is one of the most widely used species distribution modelling methods due to its ability to predict the distribution of a dichotomous response variable based on a number of predictor variables (Peng et al., 2002). Generalized linear models are comprised of a large category of regression models such as linear regression, logistic regression, multinomial regression and Poisson regression (McCullagh & Nelder, 1989).

Logistic regression is more appropriate for this research over linear regression due of its ability to only model binary dependent variables, rather than continuous ones used in linear regression. Another problematic aspect of linear regression models is that the residuals are assumed to be normally distributed with a constant variance, an assumption that cannot be made for all types of datasets. For example, if the response variable can only take on the value of 0 or a 1, the data cannot be normally distributed. In that case, logistic regression is used because it generalizes statistical assumptions by allowing other types of distributions. Logistic regression was implemented as a Microsoft Excel-add on statistical software known as XLSTAT (version 2017).

To begin, the main category of binary response variable being analyzed in this research (EAB presence/absence), can be represented by the random variable Y_i with its realization of y_i. Since y_i is a binary variable, it can take on two values: $y_i = 1$ (presence), with probability p_i, or $y_i = 0$ (absence) with probability $1 - p_i$. Hypothetically, if y_i is treated as a random variable Y_i, it takes on a Bernoulli distribution according to (2.1).

$$P_r\{Y_i = y_i\} = p_i^{y_i}(1 - p_i)^{1 - y_i} \qquad (2.1)$$

If this Bernoulli distribution is repeated n times, the joint distribution of Y_i follows a binomial distribution. Suppose that n_i observations in group i are independent and share the same probability p_i and y_i is the number of units associated with the attribute of interest in group i. Thus, the probability distribution function of Y_i can be given by (2.2).

$$P_r\{Y_i = y_i\} = \binom{n_i}{y_i} p_i^{y_i}(1 - p_i)^{n_i - y_i} \qquad (2.2)$$

For $y_i = 0,1,2, \dots, n_i$ where $\binom{n}{y_i} = \frac{n!}{y_i!(n-y_i)!}$ is called the binomial coefficient, $p_i^{y_i}(1 - p_i)^{1-\,p_i}$ is the probability of obtaining y_i successes and $n_i - y_i$ failures at the same time in a specified order. Although (2.2) addresses the response (or dependent) variable (presence/absence), it does not account for the explanatory (predictor or independent) variables. The next step is to introduce the explanatory variables by having the probability p_i associated with $x_i \ = (1 \ x_{i1} \ \dots \ x_{im})'$, a ($m$ + 1) dimensional vector of covariates and $\beta = (\beta_o \ \ \beta_1 \dots \beta_m)'$, a ($m$ + 1) dimensional vector of regression coefficients.

The linear probability model derived from the dimensional vector of covariates and regression coefficients can be condensed as $p_i = x_i{}'\beta = \ \beta_o + \ \beta_1 x_{i1} + \ \beta_2 x_{i2} + \ \dots + \beta_m x_{im}$. To emphasize, the goal of a logistic regression model is to identify the best fitting model between the dependent and independent variables by generating the coefficients, standard errors and significance levels in order to predict a logit transformation of the probability of presence of the characteristic of interest. The logit link function in (2.3) models the log odds of the probability of the presence of the dependent variable as a function of the given independent variables. The logistic regression model focuses on choosing and estimating parameters that maximize the likelihood of the observed samples, rather than the ones that minimize the sum of squared errors in ordinary least squares regression.

$$logit(p_i) = \ x_i{}'\beta = \ \beta_o + \ \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_m x_{im} \qquad (2.3)$$

In consideration of its advantages, a major caveat with (2.3) is that the linear predictor $x_i{}'\beta$ can represent any real value but the probabilities for each value are to be between 0 and 1 for the purposes of this research. In order to limit the predicted values to the correct range, complex restrictions need to be imposed. A way to circumvent this issue is to transform the probability p_i to eliminate the range restrictions and model the transformation as a linear function of the covariates. This is achieved by a two-step procedure. First, the probability p_i is converted to the odds as the ratio of the probability to its complement. Although the odds ratio and probability are similar, the odds ratio is preferred over the probability because it can take on any positive value and has no ceiling restriction (Rodriguez, 2007). Next, the logarithm is taken to the logit or log-odds of the odds ratio as in (2.4) to remove the floor restriction and map probabilities within 0 and 1.

$$n_i = logit(p_i) = log\,\frac{p_i}{1-\,p_i} \qquad (2.4)$$

Furthermore, ceiling and floor effects are measurement errors preventing the distinction of values at the upper and lower regions of a scale where negatively skewed values experience ceiling effects and positively skewed values experience floor effects (Koedel & Betts, 2010; Yu, 2000). The next transformation involves incorporating the explanatory variables in the link logit function in (2.3), which needs to be set equal to (2.4) as the right sides of both equations are equal to the probability, p_i. Exponentiating the resulting equation defines a multiplicative model (2.5) for the odds for the i-th unit.

$$\frac{p_i}{1-\,p_i} = \exp\{x_i{}'\beta\} \qquad (2.5)$$

Finally, solving for the probability p_i to link the logit model in (2.3) results in (2.6) for logistic regression.

$$p_i = \frac{\exp\{x_i{}'\beta\}}{1+\exp\{x_i{}'\beta\}} = \frac{\exp(b_0+b_1x_{i1}+b_2x_{i2}+\cdots+\,b_mx_{im})}{1+\,\exp(b_0+b_1x_{i1}+b_2x_{i2}+\cdots+\,b_mx_{im})} \qquad (2.6)$$

The predicted values of the logit are converted back into the predicted odds using the logistic regression function. For linear models, the β parameters can be directly estimated using the maximum likelihood function whereas for logistic regression, an exact analytical solution does not exist ("XLSTAT," 2017). As a result, an iterative algorithm such as the Newton-Raphson algorithm is required to approximate a solution. In the context of EAB modelling, p_i is given as the probability of EAB presence on a scale of 0 to 1 based on the values of explanatory variables $x_i{}'$for a given area.

2.4.1.2 Model Selection Methods and Variable Importance Measures

There are various variable selection methods available for logistic regression such as forward, backward, forward (step-wise) or backward (step-wise). The standard forward selection procedure starts with an intercept only model and adds the first predictor that makes the largest contribution to the model based on its Wald or likelihood ratio statistic. Next, the remaining variables are sequentially added to the model if their associated p-values is less than the specified entry p-value. Once a variable enters the model, it is never removed. This process continues until none of the variables meet the specified significance level for entry in the model (Bursac et al., 2008). Conversely, the backward elimination starts with a full model containing all the variables and at each step, the likelihood ratio or the Wald statistic is examined for each variable. If a variable exceeds the removal significance threshold, it is removed. The process is repeated until all the variables that fail to meet the specified removal threshold are removed.

The dilemma with forward and backward variable selection methods is that they do not consider the contributions of other variables to the model while adding or eliminating a variable. As a result, the step-wise variable selection overcomes the risk of adding an insignificant variable or eliminating a significant variable by reaffirming the importance of previously added variables at each step. In the forward step-wise variable selection, the model initially includes only the intercept and the remaining variables are added in a step-wise fashion according to the guidelines followed by the standard forward variable selection procedure. However the main difference with the standard forward variable selection procedure and forward stepwise variable selection is that after the addition of a variable using forward stepwise variable selection, backward variable selection is performed subsequently and the predictor variable with the smallest value of the removal significance threshold is removed (Wang, Zhang, & Bakhai, 2004).

On the other hand, backward step-wise variable selection starts with a full model containing all the variables and standard forward and backward variable selection procedures are performed at each step until the procedure stops and no variable can be added nor removed. Although the step-wise strategy is very popular, it still has its limitations. The apparent issue with the stepwise process is that during the initial selection of variables at the bi-variant level (i.e., a comparison of an independent variable with the response variable), only partial information about the relationship between the independent variables and the dependent variable can be utilized (Akinwande et al., 2015). As a result, when a predictor variable appears uncorrelated to the dependent variable and is consequently eliminated, the eliminated variable may actually significantly improve the model's performance (Akinwande et al., 2015; Courville & Thompson, 2015). Moreover, the stepwise elimination of uncorrelated variables using stepwise cause an underestimation of some of the parameters and compromises the predictive power of the model (Lutz, 1983; Harrell, F. E., 2015).

Prior to discussing the variable importance measures in logistic regression, it is important to discuss a challenge posed by the training dataset which interferes with the significance rankings of the predictor variables. Complete separation of the training dataset by a logistic regression model occurs when a predictor variable has the ability to separate the two classes of the response variable completely. In such a case, the maximum likelihood estimate for the predictor variable does not exist as it approaches infinity and the resulting estimated regression coefficient is not reliable. For quasi-separation, the predictor variable has the ability to separate the response variable to a high degree but the maximum likelihood estimate is not infinite. In the case of complete or quasi-separation, XLSTAT recommends using the "Firth's method" (Firth, 1993) which is a penalized likelihood approach that produces finite estimates of regression parameters when maximum likelihood estimates are infinite (Williams & Dame, 2018). Using the Firth's method, a corrective procedure is applied to the maximum likelihood estimate in order to reduce bias before the variable's coefficient is calculated. As for variable importance measures, XLSTAT uses the likelihood ratio as the criterion to test the statistical significance of each independent variable in the model. The likelihood ratio test involves comparing a full model to an alternate model with the exclusion or inclusion of a particular variable thereby making it analogous to the F-test in linear regression. If the difference between the full and alternate model is statistically significant, then the full model fits the data significantly better than the reduced model. The likelihood ratio is calculated by (2.7):

$$LR = -2(l(\hat{B}^0) - l(\hat{B})) \tag{2.7}$$

where $l(\hat{B})$ represents the log likelihood of the full model and $l(\hat{B}^0)$ represents the log likelihood of the reduced model evaluated at the maximum likelihood estimates specified by the null hypothesis ("The Pennsylvania State University", 2018). The test statistic for the likelihood ratio test follows a Chi-squared distribution with $p - r$ degrees of freedom where p represents the total number of variables in the full model and r represents the number of variables in the reduced model. If the p-value associated with the test statistic is lower than the specified significance level, then the change due to the excluded variable in the reduced model is not significant.

The second variable importance test is known as the Wald test which approximates the likelihood ratio test except that it only requires estimating one model and examines individual regression coefficients. Much like the likelihood ratio test, the Wald test also follows a Chi-squared distribution with the degrees of freedom of 1 to indicate each individual regression coefficient. To that end, the Wald test is similar to the *t*-test in linear regression. The Wald test statistic is calculated using (2.8):

$$Z^2 = \left(\frac{\widehat{\beta}_i}{se(\widehat{\beta}_i)}\right)^2 \tag{2.8}$$

where $\hat{\beta}_i$ represents the regression coefficient and se represents the standard error of the coefficient. All in all, the Wald test tests that the explanatory variables chosen by a variable selection procedure in the model are simultaneously equal to 0. The hypothesis for the Wald test are stated below:

H_o: The regression coefficient of the variable β_i is equal to 0
H_a: The regression coefficient of the variable β_i is not equal to 0

Aside from the p-values indicating the significance of each variable, logistic regression includes coefficients indicating the degree of the effect the predictor variables have on the response variable. There are two kinds of coefficients provided by logistic regression: unstandardized and standardized. Unstandardized coefficients provide the change in the response variable for 1 unit increase in the log odds of the predictor variable whereas the standardized coefficients provide the change in the response variable given one standard deviation increase in the log odds of the response variable.

2.4.2 Random Forest

2.4.2.1 Model Description

Random Forest (Breiman, 2001) is an ensemble machine learning method that is gaining popularity in the ecological community due to its strong predictive power (Cushman & Huettmann, 2010). The usage of multiple decision trees to build a forest of predictions (i.e., an ensemble learning method) gives Random Forest its distinct name. In theory, Random Forest is a classification and regression trees (CART) based method that uses bagging, or bootstrap aggregation to independently construct a bootstrapped sample of the dataset for each decision tree followed by an aggregation of the predictions at the bottom of each tree. On its own, CART is sensitive to outliers in the training sample and is therefore unstable whereas Random Forest stabilizes any variations in the samples by acquiring bootstrapped samples from the original dataset (Breiman, 1996). Instability in a model can reduce the overall accuracy and discrepancies in variable importance. In essence, bootstrapping is defined as creating multiple versions of the original dataset by taking random subsets of samples from the training dataset with replacement of the samples (Song et al., 2013). The random subsets account for roughly 2/3 of the training data for each tree. The remaining 1/3 of the training data is referred to as out-of-bag (OOB) and is used to calculate the misclassification error of each tree (Bento et al., 2013; Breiman, 2001) for cross-validation purposes. Lastly, much like CART, Random Forest can be used for both regression and classification tasks to predict a phenomenon based on a training dataset (Louppe, 2014).

In Random Forest, the replacement of the observations in the dataset is a key component in building a bootstrapped sample because it mimics the statistical properties of the original data by allowing repetitions. For the classification mode in Random Forest, a majority vote is taken at the bottom of each tree across all terminal nodes for the prediction of the outcome (Bento et al., 2013). Unlike standard decision trees, where each node is split using the best split among all variables, Random Forest splits each node using the best predictor among a sub-set of predictors randomly chosen at that node. This random selection prevents overfitting by minimizing the correlation between the trees thereby outperforming other classifiers such as discriminant analysis, support vector machines, and neural networks (Archer & Kimes, 2008; Breiman, 2001). The same algorithm is used for classification and regression within Random Forest. Classification is used for a categorical response variable whereas regression is used for a continuous response variable. Random Forest was implemented in R, version 3.4.0 (R Development Core Team, 2011) as the randomForest package. The Random Forest algorithm is explained by the following steps in detail:

Step 1: Collect a number of randomly selected bootstrap samples from the original dataset specified by the parameter “ntree”. The usage of randomly selected subsets of the dataset prevents the need for increasing the sample size and is particularly useful for smaller datasets.

Step 2: For each bootstrap sample, a classification tree is grown un-pruned which contains all sections of the tree. At each node of the tree, a user-specified number of predictors is used to choose the best split and this process continues until the bottom of the tree is reached.

Step 3: New data such as a testing or prediction dataset is predicted by aggregating the predictions of the original bootstrapped samples (i.e., the majority votes for classification).

Step 4: At each bootstrap iteration, the out-of-bag (OOB) data is predicted using the tree grown with the bootstrapped samples and the error rate is computed for each tree.

2.4.2.2 Variable Importance Measures of Random Forest

Useful outputs from the Random Forest model include the OOB misclassification rate, variable importance rankings, and a confusion matrix for the testing or prediction samples. The number of trees to be constructed for each bootstrapped sample is arbitrary and depends on the user. The importance of a variable in Random Forest is usually measured in two ways: the mean decrease in impurity importance (MDI) and the mean decrease in accuracy (MDA) (Hur, et al., 2017; Louppe et al., 2013).

The mean decrease in impurity importance (MDI) was used in this research. To begin, the impurity index is used to decide where to make the split in a tree and the variable(s) that are used to make the split. The mean decrease in impurity importance (MDI) for a variable is calculated by adding up the weighted impurity decreases for all nodes where the variable makes a split and is divided by the number of trees in the forest. As a result, the sum of the impurity reductions in all the trees is calculated as the importance of the variable (Hur et al., 2017) in (2.9):

$$Imp(X_m) = \frac{1}{N_T} \sum_{T} \sum_{t \in T : v(s_{t)} = X_m} p(t)\Delta i(s_t, t) \qquad (2.9)$$

where X_m is the variable in question, $p(t)\Delta i(s_t, t)$ refers to the portion of weighted impurity decreases for all nodes (t) and splits (s_t) where X_m is used and averaged over all trees (N_T) (Louppe et al., 2013). In Random Forest, the index for impurity reduction (MDI) is known as the Gini coefficient. On the other hand, MDA is based on a permutation test which measures the accuracy decreases across all OOB predictions when the variable in question is permuted while all other variables are left unchanged (Bento et al., 2013).

2.4.3 Random Generalized Linear Model (RGLM)

2.4.3.1 Model Description

Logistic regression models that use the forward stepwise variable selection method are useful because their results are easy to interpret. However, the disadvantage is the prediction accuracy is compromised because it may lead to an overfitting and result in unstable predictors (Song et al., 2013). On the other hand, Random Forest has superior prediction abilities but is difficult to interpret (Cushman & Huettmann, 2010). As a result, a hybrid of a generalized linear model (GLM) and Random Forest known as RGLM (Song et al., 2013) is explored in this section which takes advantage of the excellent prediction capabilities of Random Forest and the palatability of a forward selection generalized linear model. Under these circumstances, RGLM can be used to predict binary, continuous and count outcomes.

In order to improve the prediction capabilities of a logistic regression model, the concept of generating various subsets of the training dataset and aggregating the predictions is a plausible method. Interestingly, the idea of a bagged generalized linear model(GLM) has been proposed by Breiman (1996) but RGLM extends the proposition even further by incorporating an element of randomness through randomly selecting a specified number of variables in each bag. In fact, Song et al. (2013) compared the RGLM with other prediction models and found that it achieved the highest mean accuracy amongst gene expression datasets.

In brief, much like Random Forest, RGLM is an ensemble predictor based on non-parametric bootstrap aggregation (bagging) of several logistic regression models (i.e., bags) where the key features (explanatory variables) are selected using forward variable selection according to the AIC criterion. RGLM arrives at the final prediction of the outcome (i.e., binary response variable) by aggregating the predictions using the selected explanatory variables across all the bags. Similar to Random Forest, the number of variables selected for forward regression for each bag is user-specified. In conclusion, the randomness incorporated in RGLM stems from creating several bootstrap samples of the training dataset like bagged GLMs and by selecting a random subset of explanatory variables for each bootstrap sample.

RGLM was implemented as the randomGLM package in R, version 3.4.0 (Development Core Team, R., 2011) and realized through 5 major steps as follows:

Step 1: Create a number of bootstrapped samples specified by the parameter "nBags" and for each bag, a number of samples equal to the number of samples of the original dataset is generated with replacement. In the case that a bag contains less than the number of samples indicated by the "minInBagObs" parameter, it is discarded and resampled.

Step 2: The parameter "nFeaturesInBag" specifies the number of predictor variables to randomly select for each bag. This parameter is usually included if the number of predictor variables is quite large and a reduction of the number of variables is required. The variables are ranked according to their individual association with the response variable tested using the Wald or likelihood ratio test in each bag using a univariate GLM model.

Step 3: The cut-off for the number of highest ranked variables in each bag (i.e., the variables with the most significant univariate significance levels) indicated by the parameter "nCandidateCovariates" are chosen by forward selection based on the "stepAIC R function in the MASS R library" (Song et al*.,* 2013, p. 3) to build a multivariate generalized linear model.

Step 4: The predictions from the multivariate generalized linear model are aggregated across all bags to provide a final ensemble prediction. Similar to Random Forest, if the classification mode is used, a majority vote known as the adjusted Majority Vote (aMV) strategy is used to average the predicted probabilities across all bags (Prinzie & Van den Poel, 2008; Song et al., 2013). Lastly, a binary prediction is obtained using the predicted probabilities by choosing a threshold of 0.05.

2.4.3.2 Variable Importance Measures of RGLM

RGLM has three main variable importance measures. The first measure, and the most intuitive, is the number of times that a variable is selected by forward regression across all bags. A more specific measure is the number of times that a variable is selected as the candidate covariate for forward regression. The third measure is the sum of the absolute regression coefficient values for each variable. Among the three variable importance measures, Song et al. (2013) recommended to use the number of times that a variable is selected through forward regression as it provides a direct association with the outcome variable.

CHAPTER 3: STUDY AREA AND DATA

This chapter consists of descriptions of the study area, the species data and the predictor datasets used in the distribution models. The species and predictor data required extensive processing in order to improve their quality and maintain consistency with the spatial and temporal scales of the datasets. Lastly, the average values for the predictor variables for the EAB presence and absence points were examined for preliminary insights into the significant variables.

3.1 Study Area

The study area in this study is Southern Ontario, Canada (43° 10' 26.40" N, -81° 18' 57.60" W), approximately 136, 907 km^2 in size and represents the extent of the EAB dataset provided by the CFIA as shown in Figure 3.

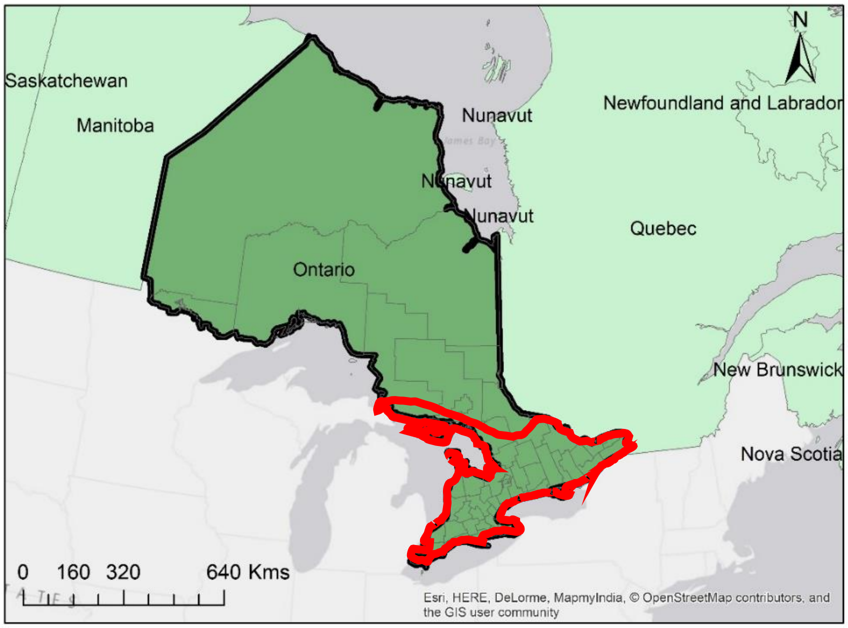

N
Saskatchewan
Manitoba
Nunavut
Newfoundland and Labrador
Nunavut
Nunavut
Quebec
Ontario
New Brunswick
Nova Scotia
0 160 320 640 Kms
Esri, HERE, DeLorme, MapmyIndia, © OpenStreetMap contributors, and the GIS user community

Figure 3. Study area of Southern Ontario, Canada displayed in a red outline

3.2 Data: Response and Predictor Variables

The two main types of variables required for this research are locational data of the EAB (presence/absence) and various environmental abiotic, biotic and anthropogenic variables. In this research, the EAB data is referred to as the response/dependent variable and the environmental and anthropogenic variables take on various interchangeable terms such as explanatory, predictor, and independent variables. Mac Nally (2000) argued that the variable selection process and subsequently, statistical inferences of a distribution model could be substantially improved if it builds on existing knowledge and theory. As a result, the explanatory variables were selected based on an *a priori* influence on the spread of the EAB. The datasets are listed in Table 3 followed by their descriptions of the response and explanatory variables.

Table 3. Descriptions of the data layers acquired for data collection

Dataset Title	Source	Format	Features Extracted	Resolution	Date Revised/Coverage
1. Emerald Ash Borer data	Canadian Food Inspection Agency (CFIA)	Excel file containing geographic coordinates	Presence/absence EAB points from Ontario 2006-2013	-	2014
2a. Landsat 5 TM scenes	USGS Earth Explorer	Raster (TIF)	NDVI	30 m	Acquired from 2006-2012 (May-August)
2b. Landsat 7 ETM+ scenes	USGS Earth Explorer	Raster (TIF)	NDVI	30 m	Acquired from 2013 (May-August)
3. Population Centres	Statistics Canada	Vector (points)	Medium and large population classes	-	2011
4. Accommodations Point	DMTI Spatial Inc.	Vector (points)	Campgrounds	-	January 09, 2015
5. Forest Processing Facilities	Ontario Ministry of Natural Resources via Geoportal	Vector (points)	-	-	February 15, 2008
6. Ontario Sea Ports	SeaRates - Retrieved from https://www.searates.com/maritime/canada.html	Longitude/ Latitude coordinates	-	-	2017
7. Ontario Road Network (ORN)	Land Information Ontario (LIO)	Vector (lines)	Freeways	-	February 01, 2010
8. Provincial DEM (South) Version 3.0	Ministry of Natural Resources via Geoportal	Raster	Elevation, slope and Aspect	30 m	February 2014
9. Ontario Wind Resources Information	Ontario Ministry of Natural Resources	NAD83	Southern Ontario	Vector (points) separated by 1 km	Monthly (June) speed data covered during 20 years at 30 m height
10. WorldClim Version 2	WorldClim – Global Climate Data	Raster (TIF)	Precipitation and solar radiation in June	1 km	Covered during 1970-2000
11. MODIS/Terra Land Surface Temperature/Emissivity	Land Processes Distributed Active Archive Center (LP DAAC)	Raster (TIF)	Land surface temperature (LST) daytime products	1 km	8-day composite products from March 2000-present

Dataset 1: Emerald Ash Borer Data

The CFIA provided sampled points of EAB presence and absence data from 2002 to 2013 in Ontario. In terms of sampling, the types of sampling activity conducted by the CFIA includes green prism traps and visual surveys. All baited traps were installed on June 1st and taken down on August 31st (Appleton et al., 2017). Visual surveys were conducted during late August which marks the period of time when signs and symptoms of the EAB is most prominent. The visual surveys were conducted in premeditated areas where the EAB could potentially have been introduced through human activities. These areas include: areas with ash decline, urban centres, provincial parks, campgrounds, rest stops along major transportation corridors, ash nursery stocks, and other areas identified by the public (Appleton et al., 2017).

For green prism traps' placement and density, a set of guidelines were followed by the CFIA. First of all, traps were deployed in urban centres using a triangular grid system. Traps were only placed in trees that were located along a forest edge, in an open area or open stand of trees. One trap was placed per chosen site and situated as high as possible within the canopy. In terms of orientation of the traps, they were placed on the south or southwest side of the tree in the middle of a branch. It should be noted that the GPS coordinates of the trees were truncated (i.e., rounded to three decimal places) for privacy and confidentiality reasons.

With regards to the general shift of survey locations from a year to the next, each time a county had confirmed EAB sightings, it was declared regulated and sampling did not occur in the upcoming year within the same county. The presence and absence data collected by the CFIA from 2002-2013 are displayed in Figures 3.1 and 3.2 and the counts of the points from are provided in Table 3.1.

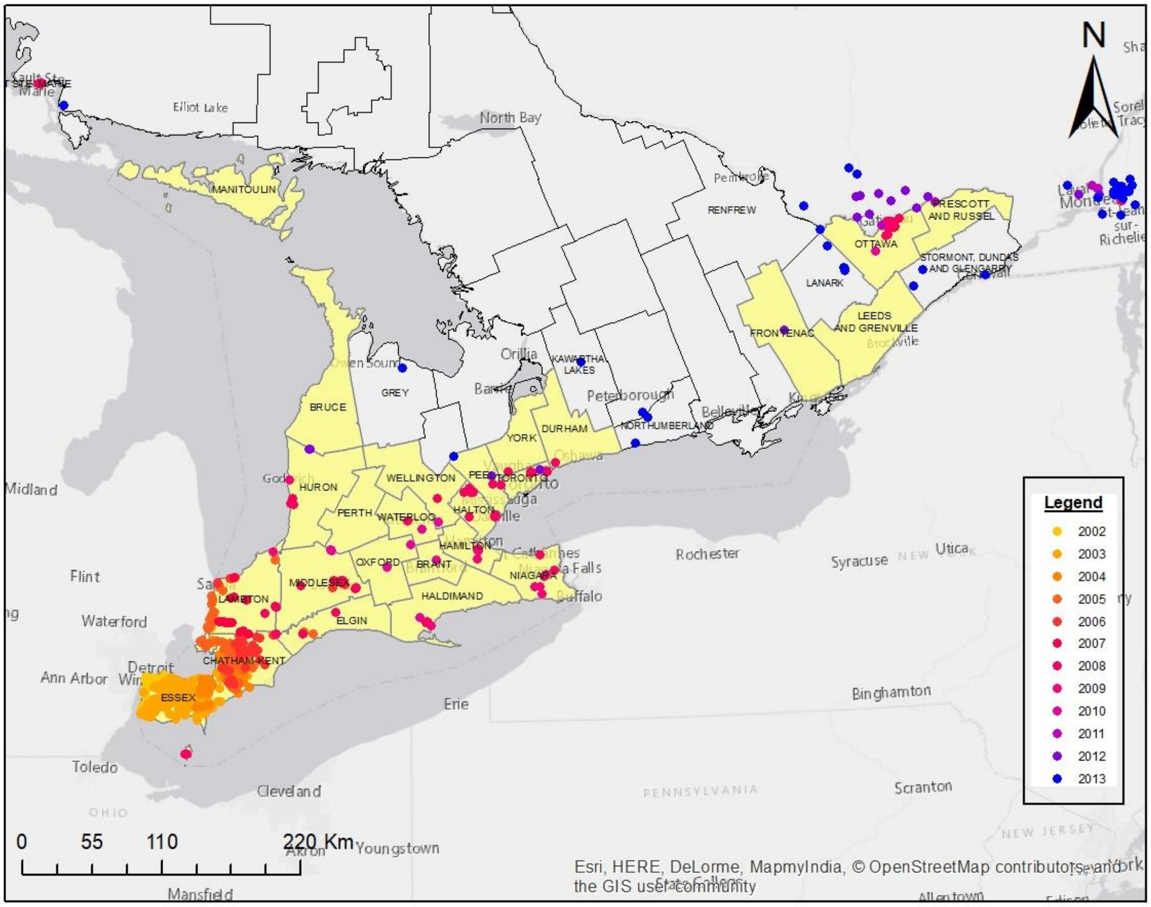


Figure 3.1. Study area of Southern Ontario overlain by EAB presence points collected from 2002-2013. The counties highlighted in yellow were regulated as of 2013

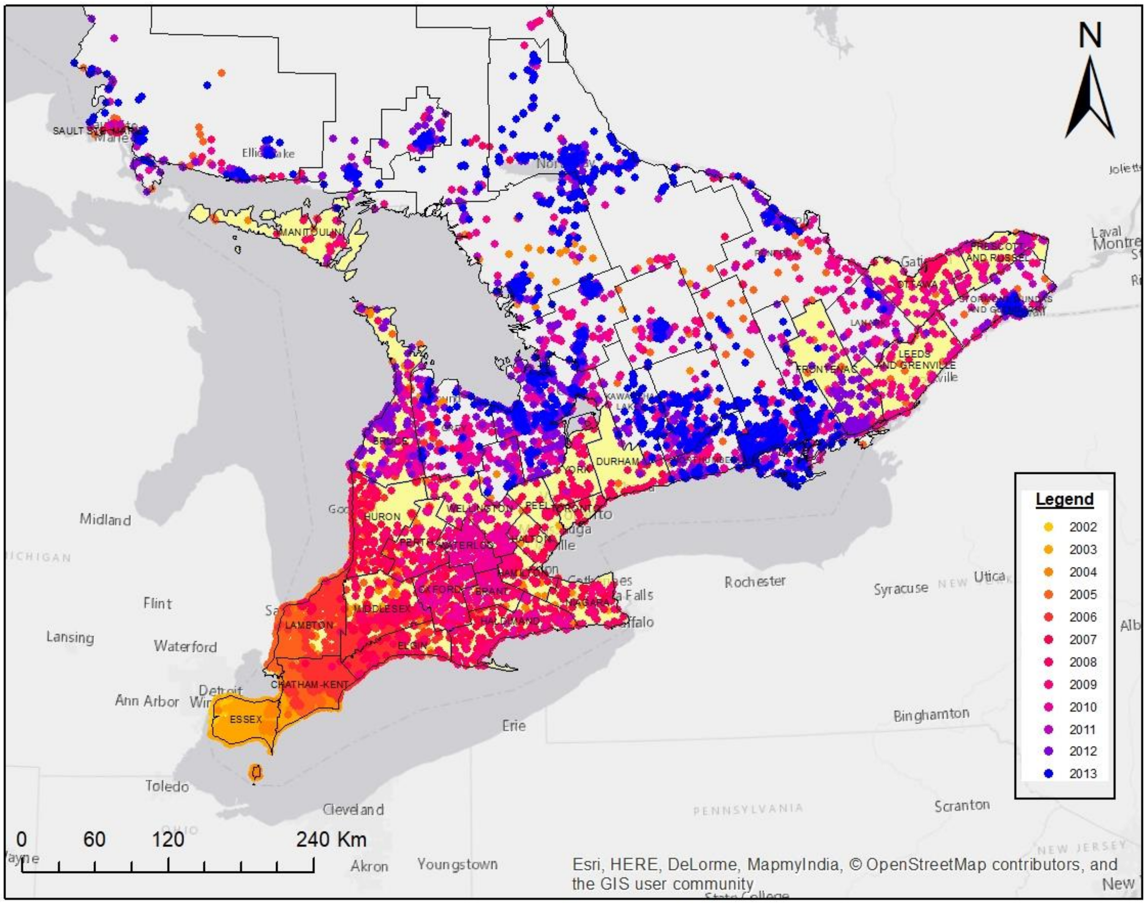


Figure 3.2 Study area of Southern Ontario overlain by EAB absence points collected from 2002-2013. The counties highlighted in yellow were regulated as of 2013

Table 3.1. Number of sampled EAB presence and absence points from their respective years in Southern Ontario. *Note** only the rows outlined in bold were used as inputs into the models

Year	Number of EAB presence points	Number of EAB absence points
2002	36	440
2003	356	3912
2004	163	6573
2005	146	7952
2006	**59**	**6706**
2007	**69**	**1417**
2008	**124**	**1189**
2009	**17**	**907**
2010	**11**	**986**
2011	**1**	**521**
2012	**6**	**981**
2013	22	1253

From the Figures 3.1 and 3.2, it should be noted that the counties highlighted in yellow were regulated by the CFIA as of 2013. Since the new EAB sightings were observed in 2013 in unregulated areas, the CFIA has declared the entire Southern Ontario regulated as of April 1, 2014 as shown in Figure 3.3.

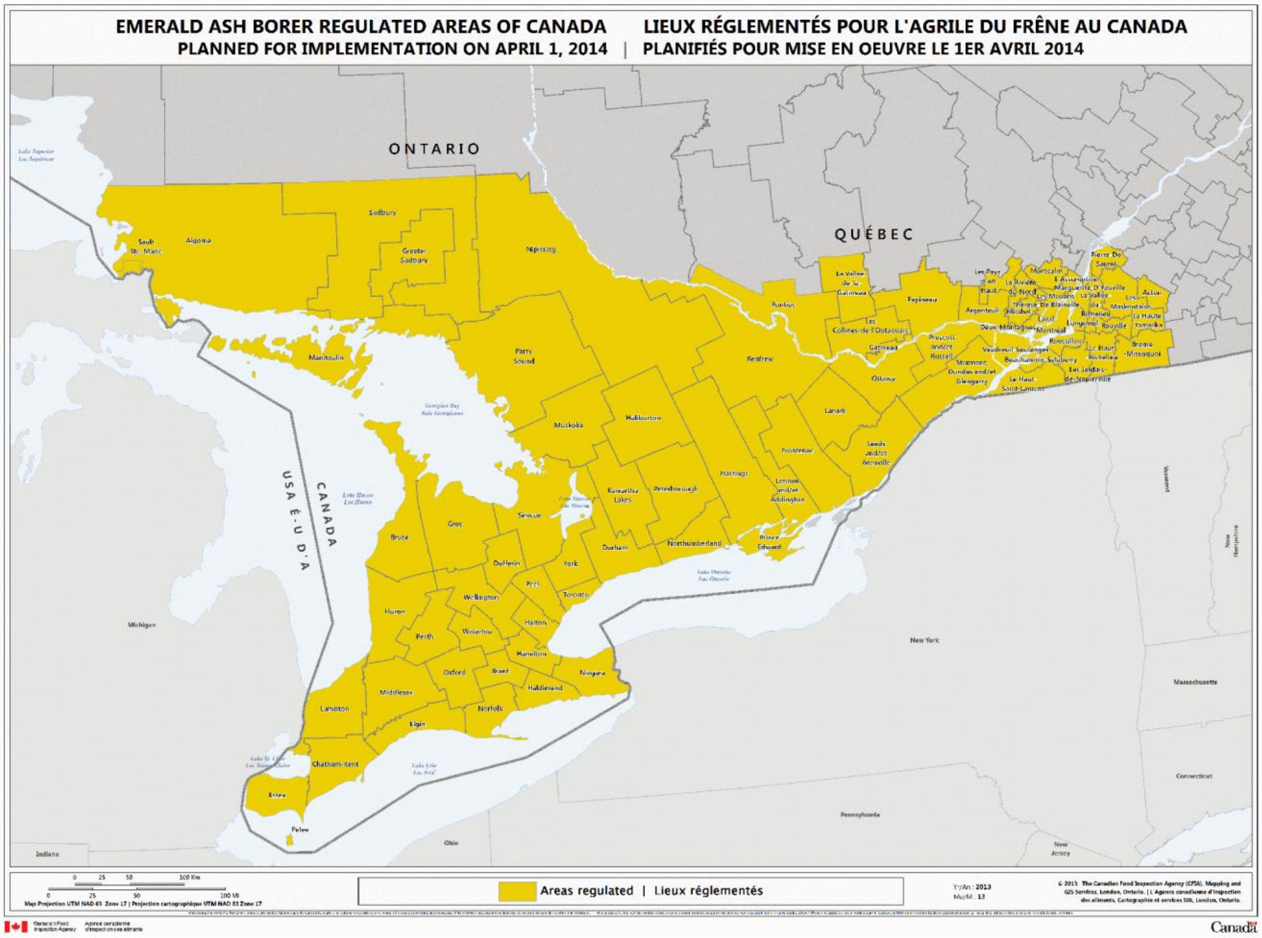


Figure 3.3. EAB-regulated counties in Ontario and Quebec since April 2014. Source: (Canadian Food Inspection Agency, 2014)

As evident in Figure 3.3, all the counties in Southern Ontario became regulated since 2014 and the EAB data acquired prior to 2014 is the only insight into the historical trends of the EAB since its introduction and further substantiates the importance of this research. The confirmed EAB points from 2002-2005 during its introduction were omitted from the model because they represented EAB infestation from the initial point of outbreak during which EAB appeared in dense clusters. An inclusion of these points in the model would provide inaccurate results because upon introduction into Canada, the EAB appeared to be in a state of frenzy and occupied as many ash trees in sight. This phenomenon defeats one of the main objectives of this research – to identify the areas the EAB prefer to inhabit without the influence of an introduction hotspot.

Datasets 2a and 2b: Landsat 5 TM and Landsat 7 ETM+ scenes

A total of eight Landsat 5 “Collection 1 Higher-Level” scenes were downloaded from the USGS Earth Explorer website from May-August pertaining to the years of EAB coverage over Southern Ontario from 2006-2012. The months May to August were chosen because they coincide with the months when the EAB traps were deployed and collected and coincidently, it is also the growing season for ash trees in a given year (Royo & Knight, 2012). The range of four months allowed for the collection of an adequate number of scenes that covered the entire extent of EAB coverage in Ontario from 2006-2012 which would not be possible if the collection of scenes were from one particular month (i.e., the month of June). Having said that, the scenes were collected according to the extent of the EAB coverage for each consecutive sampling year (i.e., 2006-2012).

The images were chosen to exclude as much cloud cover as possible and were already atmospherically corrected, avoiding the burden of post-production processing. As a result, the output products provided atmospherically corrected surface reflectance bands. The USGS mainly produced higher-level Landsat data products to facilitate land surface change studies. Surface reflectance data products simulate information that would be received if the sensor was just above the earth’s surface without the inclusion of any artifacts which often decrease consistency. Aside from the Landsat 5 scenes, sets of eleven scenes from Landsat 7 were downloaded for the year of 2013 to be used as prediction.

Although the images were atmospherically corrected, there were two problematic pixel types: fill and saturate. The fill pixels indicated an absence of data and the saturate pixels indicated high brightness observed from clouds, white sandy deserts, and other bright surfaces. As a result, these pixel types were converted to “No Data” and interpolated using a 12 by 12 rectangle to calculate the average value of the cells in the neighbourhood. Lastly, the scale factor “0.0001” was applied to all bands (“Landsat 4-7 Surface Reflectance (LEDAPS) Product Guide,” 2018).

Furthermore, the modified Landsat images were used to compute a normalized difference vegetation index (NDVI) layer to analyze the health of the trees that the EAB invade. The red band (Band 3) and near infrared band (NIR) (Band 4) from the Landsat images were used in the NDVI equation: reflectance in NIR band – reflectance in red band / reflectance in NIR band + reflectance in red band. The NDVI is a commonly used vegetation index designed to study the health of vegetation using the red band in which chlorophyll in plants absorbs radiation and the NIR band in which chlorophyll reflects vegetation. The NDVI ranges from -1 to 1 where values closer to 1 indicate greener, more healthy vegetation.

Dataset 3: Population Centres

The population centres polygon layer was used as a proxy for urban land cover in that it generalizes the density of humans within a given area. It contains the boundaries of all population centres defined for the 2011 Ontario census. By definition, a population centre has a minimum of 1000 people and a population density of 400 people per square kilometer (Statistics Canada, 2011). The population centre size class groups include small, medium and large urban population centres. For the purposes of this research, only the medium population centres (30,000-99,999 population) and large population centres (>100,000 population) were chosen because due to their size, they have a potentially greater influence over smaller urban centres.

Dataset 4: Accommodations Points (Campgrounds)

A major potential anthropogenic influence on the spread of the EAB is the distance to campgrounds. Campgrounds were extracted from the accommodation points layer which included locations of all housing facilities such as hotels, motels, campgrounds, inns, hostels and resorts in Ontario to assess their impact. The summer peak season for camping coincides with the peak time of EAB maturation from May and thus can be considered a vector of anthropogenic spread (BenDor et al., 2006). In addition, campgrounds were used as a predictor variable in the research of Prasad et al. (2010) for estimating EAB abundance in Ohio.

Dataset 5: Forest Processing Facilities

Much like the campgrounds layer, the forest processing facilities is another anthropogenic pathway by which the EAB can spread. All types of forest processing facilities (pulp/paper/paperboard, sawmill, veneer and composite/panel) were included because they are in direct contact with raw wood material (i.e., logs). To emphasize, Prasad et al. (2010) had selected wood product industries as one of the variables to be included in their insect ride model (IRM).

Dataset 6: Ontario Sea Ports

Since the initial outbreak of the EAB occurred due to importing infected ash wood via ships into North America, it is highly likely that this same transport mechanism could be used to spread the EAB across counties close to large water bodies. The wood contained in these ships are used for stabilizing cargo and packing heavy consumer products. SeaRates (www.searates.com/maritime/canada.html) is an online shipping company that arranges cargo delivery in shiploads in Canada. Geographic coordinates of the ship docking stations from SeaRates in Southern Ontario were collected and converted into a points shapefile.

Dataset 7: Ontario Road Network (ORN)

The Ontario road network (ORN) contains all types of roads in Ontario with the positional accuracy of 10 metres or greater (“User Guide for ORN Segment with Address,” 2016). The road classes include alleyway, arterial, collector, expressway/highway, freeway, local/strata, local/street, and local/unknown. To clarify, Ontario roads was included as an anthropogenic variable to analyze long-distance dispersal of the EAB. Roads act as a long-distance vector for the EAB due to insect-hitchhiking and firewood transported in trucks. Since roads with a higher volume of traffic alludes to a greater EAB risk, only the expressways/highways were included as they carry more than 40,000 vehicles per day. A buffer of 1 kilometer was applied to the highways to account for the “increased probability of insects attached to windshields, radiators, or the vehicle itself” (Prasad et al., 2010, p. 359).

Dataset 8: Provincial Digital Elevation Model (DEM) (South) Version 3.0

The provincial DEM was used as an explanatory factor in addition to the slope and aspect. The DEM surface represents true ground elevations in Southern Ontario and was generated using Ontario Radar DSM, OBM, DTM points and 2002 GTA Ortho contours (“Provincial Digital Elevation Model Technical Specifications, v3.0,” 2013). Notably, the elevation, slope and aspect serve as indirect environmental gradients which rarely affect the distribution of species but are included due to their correlation with more relevant predictors such as temperature, precipitation and solar radiation (Elith & Leathwick, 2009). In the case of the ash trees, the slope was used as a proxy for the species type and stress level of the ash trees that are found on the terrain. Research performed by Royo and Knight (2012) suggest that areas with steeper slopes (> 45 degrees) and a history of defoliation which indicates stripping of leaves contain stressed ash trees (Royo & Knight, 2012). Distinctly, the EAB is known to invade healthy ash trees but stressed ash trees are even more susceptible to invasion due to an increase in vulnerability (McCullough et al., 2009). The aspect was used to indicate the direction of the slope and analyze its relationship to the EAB points. For example, the south facing slopes are known to be warmer and drier than the north facing slopes (Chatt & Walberg, 2005). The values of the aspect correspond to a direction accompanying the slope. For instance, 0 corresponds to north, 90 corresponds to east, 180 corresponds to south, and 270 corresponds to west.

Dataset 9: Ontario Wind Resources Information

The abiotic factor wind speed is a prominent factor in contributing to the spread of the EAB across short and long distances. The dataset provided by the Ministry of Natural Resources contained long term average monthly wind speed data in the form of points covering Southern Ontario at 30 and 80 metres above ground level. The month to use for the wind speed data needed to coincide with the time period that EAB traps were employed and removed (i.e., June to August). That said, June was the most appropriate month to use as it represented the peak emergence of EAB adults in Canada which occurs mid-to late June (Appleton et al., 2017; Marchant, 2012) and the height chosen was 30 m because ash trees, particularly green ash trees grow to a maximum height of about 30 metres ("Emerald Ash Borer," 2012).

Dataset 10: WorldClim Version 2 – Global climate data

WorldClim (Fick & Hijmans, 2016) provided various climactic layers such as June precipitation and solar radiation. These variables represented the effects of climate on the presence and absence of EAB points. They were obtained at a spatial resolution of 30 seconds (1 km^2) between the years 1970-2000. The month of June was selected again because it represents the month during which adult EABs emerge from ash trees and the precipitation and solar radiation are potential variables that could affect the survival of the EAB in preceding years.

Dataset 11: MOD21A2: MODIS/Terra Land Surface Temperature

Another abiotic variable that is potentially conducive to the dispersal of the EAB is land surface temperature which a prime indicator of global warming facilitated by anthropogenic activities such concentrated human activities and the establishment of paved land cover or barren lands (Settur et al., 2013). As a side note, temperature was also included as one of the explanatory variables in Huset (2013)'s research on the EAB to investigate one of the objectives of the research – how anthropogenic climate change affects EAB risk. Although surface temperature/emissivity could have been calculated from the thermal band of the Landsat TM images used to calculate NDVI, it was not recommended for a large heterogenous study area (Vlassova et al., 2014). Since Landsat images have a high spatial resolution, they are more sensitive to the thermal contrast in land surface temperatures (LSTs) between landcover features such as tree canopy during the summer months. Research conducted by Vlassova et al. (2014) support this theory where LSTs retrieved from Landsat images using the single-channel (SC) method produced the greatest variances (above 6 °C) in the summer and lower variances in the winter months compared to LST values from MODIS. Coincidentally, since the month of EAB emergence (June) falls during the summer months, the usage of the MODIS LST products seemed more practical. In addition, the coarse resolution (1 km^2) of the MODIS LST products was consistent with the other climactic variables.

The land processes distributed active archive center (LP DAAC) website offers various MODIS land surface temperature products from the Terra and Aqua satellites. The Terra satellite was chosen over the Aqua because it is timed to cross the equator from north to south in a descending mode rather than the Aqua's ascending mode from south to north. The MODIS21 product uses the Temperature Emissivity Separation (EST) algorithm to retrieve the LST and emissivity dynamically from the three MODIS thermal infrared bands 21, 31, and 32 (Hulley & Hook, 2017). The MOD21A2 dataset is essentially an 8-day composite LST product that uses a simple averaging method to calculate the average from all the cloud free MOD21A1D (day) and MOD21AN (night) products from the 8-day period. Moreover, much like the NDVI dataset, land surface temperatures for all the EAB presence and absence points from the month of June were obtained for the corresponding years. Lastly, a scale factor of 0.02 was applied.

3.3 Descriptive Statistics of the EAB Data

Before conducting the spread modelling, an insight needs to be made into the statistics of the EAB presence and absence data by analyzing their descriptive statistics such as mean, median, standard deviation (STD), minimum and maximum for each of the 14 explanatory variables using the XLSTAT software. For a fair comparison of the EAB presence and absence data, the original EAB presence points and the same number of randomly selected absence points were analyzed. The descriptive statistics for the presence and absence data are summarized in Tables 3.2 and 3.3. In addition, a visual depiction of the average values of the explanatory variables is displayed in Figure 3.4 as bar plots.

Table 3.2. Descriptive statistics such as minimum, maximum, median, mean and standard deviation using quantitative explanatory variables for the EAB presence data

Statistic	**Minimum**	**Maximum**	**Median**	**Mean**	**Standard Deviation**
Population Centres (m)	0	68578.35	1555.96	11090.06	15852.74
NDVI	0.10	0.92	0.55	0.55	0.22
Elevation (m)	51.48	380.53	185.67	188.27	63.52
Aspect (Direction)	1.42	359.37	193.70	194.50	92.45
Slope (°)	0.00	9.20	0.50	0.84	1.14
Ports (m)	1693.07	83230.06	45559.03	45777.19	25568.89
June Precipitation (mm)	68.00	94.00	85.00	83.63	5.94
Forest Processing Facilities (m)	3750.00	77762.18	39600.92	40326.24	21634.19
June Wind Speed (m/s)	2.60	3.50	3.30	3.34	0.13
Nearest EAB Positive Location (m)	111.05	34613.16	468.79	2343.66	4974.10
Highways (m)	0.00	34780.78	10852.95	10050.17	6510.89
Camps (m)	1351.33	57937.22	10802.08	16329.93	16466.11
June Land Surface Temperature (Kelvin)	263.32	308.32	299.38	298.14	8.04
June Solar Radiation (kJ/m^{2}day)	20613.00	21822.00	21320.00	21317.64	154.09

Table 3.3. Descriptive statistics such as minimum, maximum, median, mean and standard deviation using quantitative explanatory variables for the EAB absence data

Statistic	Minimum	Maximum	Median	Mean	Standard Deviation
Population Centres (m)	0.00	197916.03	21161.66	25905.63	25984.87
NDVI	0.12	0.92	0.67	0.65	0.16
Elevation (m)	55.77	361.43	214.03	216.55	37.64
Aspect (Direction)	3.08	359.95	195.71	189.69	103.95
Slope (°)	0.00	10.82	0.54	1.09	1.52
Ports (m)	241.87	81729.22	34205.86	37037.08	18964.50
June Precipitation (mm)	61.00	109.00	85.00	84.25	5.31
Forest Processing Facilities (m)	1594.24	74813.92	23635.22	25057.93	15928.96
June Wind Speed (m/s)	2.70	5.43	3.97	3.94	0.37
Nearest EAB Positive Location (m)	82.29	9947.05	4129.07	4503.09	2665.12
Highways (m)	0.00	40616.28	11782.45	14107.51	11664.53
Camps (m)	30.00	99536.00	27152.50	28095.70	16699.83
June Land Surface Temperature (Kelvin)	321.18	345.92	339.14	338.33	3.13
June Solar Radiation (kJ/m²day)	20613.00	21822.00	21356.50	21353.62	185.61

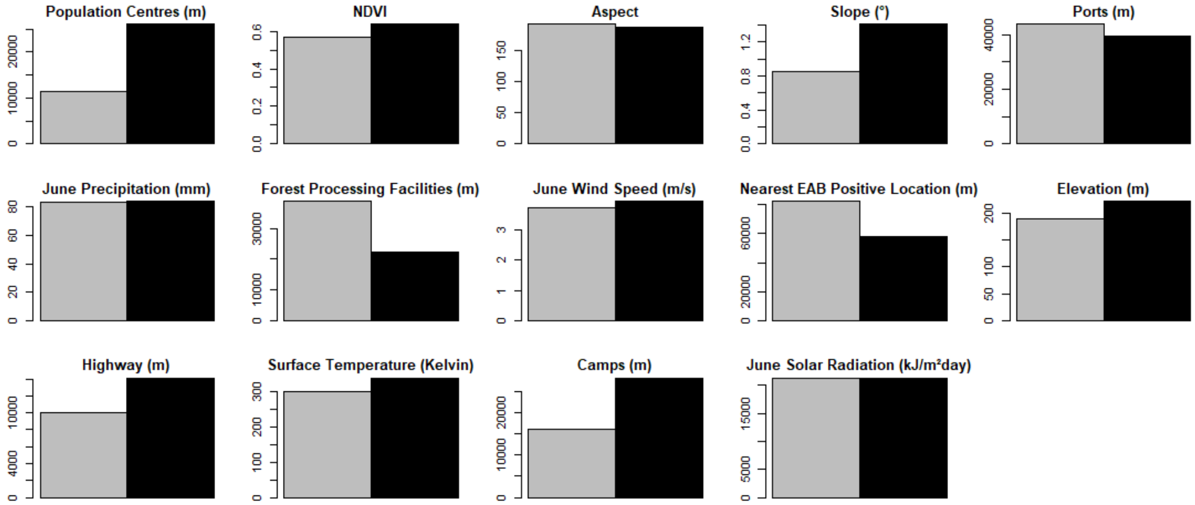


Figure 3.4. Bar plot of average values of explanatory variables for EAB presence points (grey) and EAB absence points (black)

First of all, it is quite evident from the descriptive statistics in Tables 3.2 and 3.3 that the distance from population centres is shorter to the presence points than the absence points. The mean distance from the population centres to the presence points is approximately 11,090 m whereas the mean distance from the population centres to the absence points is 25,906 m. However, the standard deviation of the mean distance from the population centres to the absence points is substantially large (25,984 m) compared to the one from the presence points (15,852 m), suggesting that there is a large variation in the distances of the absence points. Nonetheless, the distance from population centres to the presence points is most likely an important variable, which will be confirmed by the variable importance tests in Chapter 5.

Another anthropogenic factor, forest processing facilities, showed some unexpected results. The mean distance from a forest processing site to presence points is 40,326 m whereas the mean distance to the absence points is 25,057 m. This correlation is unusual because of the implication that it is about twice as likely for a location to be infested by the EAB the farther away it is from a forest processing site. However, the standard deviation values of the distance from forest processing facilities for the presence and absence points tell a different story. For instance, the standard deviation of the presence points was 21,634 m whereas the standard deviation of the absence points was 15,929 m which implies perhaps there might be some presence points that are closer to the forest processing facilities. However, based on the mean values, this phenomenon is the opposite for Huset (2013) as her research concluded that the closer a location is to a wood product industry, the greater the risk of EAB infestation (Huset, 2013). Similarly, the mean distance from sea ports to the EAB presence points was 45, 777 m while the mean value to the absence points was 37, 037 m. The mean distance from campgrounds to the presence points was 16, 330 m and the absence points was 28, 096 m. Moreover, from the descriptive statistics, the distances from the two anthropogenic variables (sea ports and forest processing facilities) had a negative relationship with the presence points such that the greater the distance from an anthropogenic facility, the greater the chance of EAB presence.

The last anthropogenic layer, distances to highways showed that the median distance to the presence points from highways (10,050 m) was shorter than the average distance to the absence points (14,108 m) which suggested that a majority of EAB presence points are found near highways. The standard deviations support this theory as the standard deviation for the EAB presence points is lower (6511 m) than the one (11, 665 m) for the EAB absence points. Furthermore, the mean and standard deviation values conclude that there is less spatial variation among the EAB presence points from the highways than among the absence points which makes this variable a potential significant variable.

Next, for the NDVI data for the presence points, the mean value was 0.55 whereas the mean NDVI absence value was approximately 0.67. This suggests there may be a correlation between the health of the trees and EAB infestation where the EAB is more likely to infest an unhealthier tree rather than a healthy tree. For the abiotic variables such as the elevation, aspect, and the slope, the presence and absence points showed similar trends. The mean elevation for the absence points was 217 m whereas the mean elevation for the presence points was 188 m. The standard deviation was about two times greater for the presence points (63.519 m) than the absence points (37.636 m). For the aspect, both averaged presence and absence points fell in the 190 ° - 195 ° range and according to Burrough's aspect guideline, this range suggests that ash trees generally face South (Burrough, 1998). The findings are consistent with the EAB surveying protocols for placing traps on the south side of trees.

Lastly, the average slope was 0.84 for the presence points and 1.09 for the absence points, respectively. Although their distinction was not significant, this relationship opposes Royo & Knight's (2012) theory that steeper slopes contain stressed trees thus making them more prone to attacks by the EAB (Royo & Knight, 2012). Although it cannot be confirmed whether the sampled EAB-absent ash trees were in fact stressed, according to the descriptive statistics, the EAB infected trees are more likely to exist at the lower angles. When the standard deviation values were examined, the relationship was not as clear as the values were similar for both presence (1.14) and absence points (1.52).

The biotic variables, both with the presence and absence points had similar values as the mean precipitation was approximately 84 mm and the mean solar radiation was 21,300 kJ/m^{2}day. The mean June wind speed was 3.34 m/s for the presence points and 3.94 m/s for the absence points, which suggests that EAB infestation decreases at locations with strong winds. The mean and standard deviations of the land surface temperatures for the EAB presence points were 298 and 8.0 kelvins whereas the mean and standard deviations of the land surface temperatures for the EAB absence points were 338 and 3.13 kelvins. The mean surface temperatures for the EAB absence points were approximately 40 kelvins greater than the EAB presence points suggesting that the EAB prefer cooler areas.

Lastly, the mean distance to the nearest EAB positive location was examined. The distinction between the EAB presence and absence points based on the mean, standard deviation and median values could not be made with confidence. For instance, the mean distance to the nearest EAB positive location from the presence points was 2344 m whereas the mean distance to the nearest EAB positive location from the absence points was 4503 m. This relationship was inconsistent with the standard deviations where the standard deviation for the EAB presence points was 4974 m whereas it was 2665 m for the absence points. When the median values were examined, it was 469 m for the EAB presence points and 4129 m for the EAB absence points. Based on these relationships, it is difficult to predict a relationship between the distance to the nearest EAB positive location for EAB presence and absence points. However, Huset (2013)'s results concluded the average distance from a given location to a known EAB location was shorter for the EAB-containing quadrants than for the non-EAB quadrants.

CHAPTER 4: METHODOLOGY

The methodology includes the complete workflow of pre-processing the response and predictor variables and analyzing the data using three distribution models (i.e., logistic regression, Random Forest and RGLM). To overcome the issue of high positive spatial autocorrelation, the sampled EAB presence and absence points were filtered using a distance threshold corresponding to the distance between two points at which maximum clustering occurred. Next, the models were trained using a 1:1 ratio of prevalence as it provided an equal representation of the samples from each class. Despite this, the effects of using a greater proportion of absence points was tested by gradually increasing the number of absence points using logistic regression and examining the misclassification rates.

Next, a correlation coefficient matrix, significance testing and VIFs (variance inflation factors) of the predictor variables were assessed to determine the variables that were required to be removed prior to modelling due to exhibiting multicollinearity. In addition to addressing multicollinearity before modeling, a method known as recursive backward elimination (RFE) was used in Random Forest to address multicollinear variables during the modeling process. Following the pre-processing of the species data and predictor variables, the transferability of the three distribution models (i.e., logistic regression, Random Forest and RGLM) was tested by assessing the classification accuracies, ROC curves and AUC values, and risk maps of the models. An automated risk map tool was generated for the machine learning models (Random Forest and RGLM) to visualize the outputted probabilities from the models. A flow chart of the methodology is displayed in Figure 4.

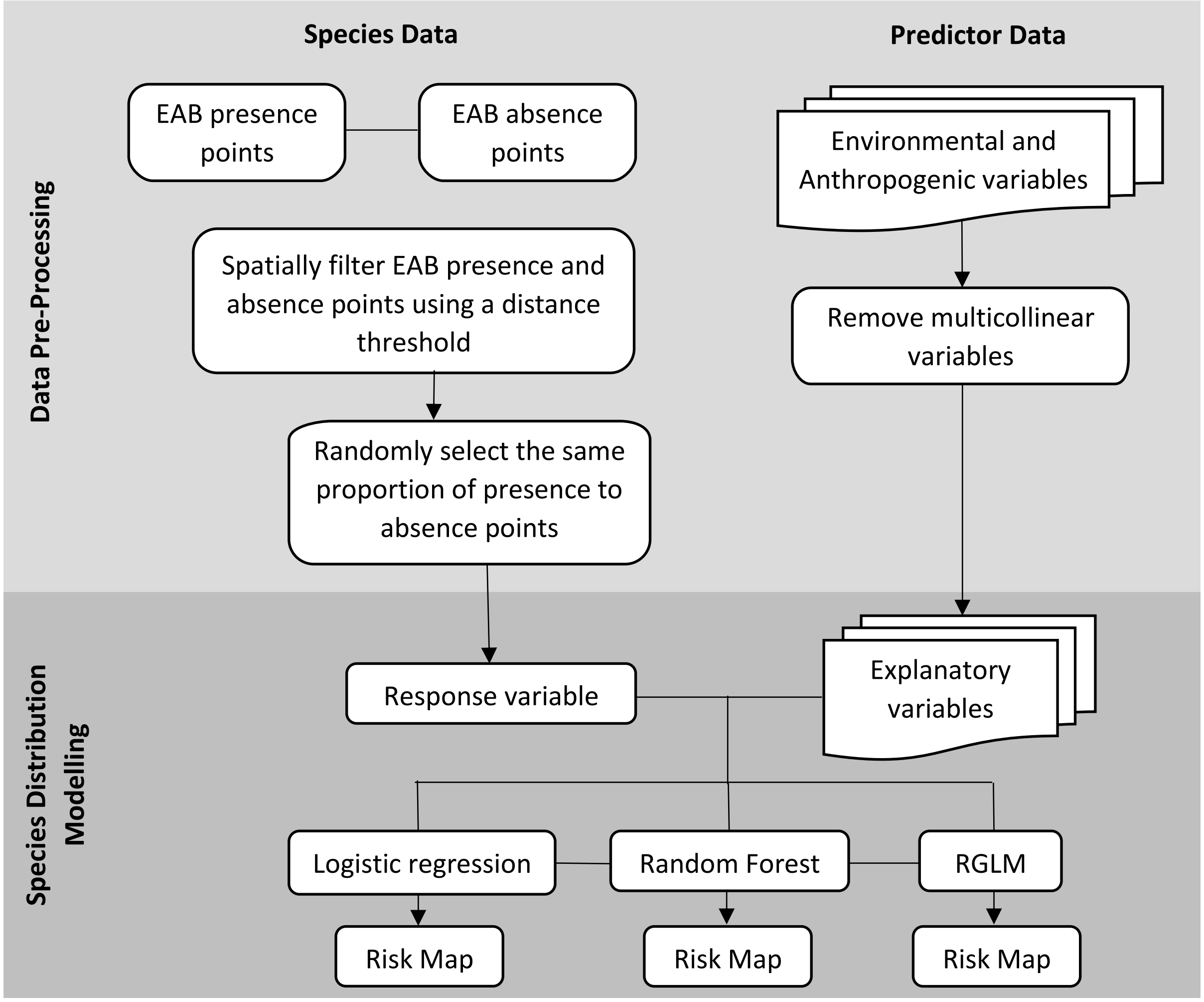

Species Data
Predictor Data
Data Pre-Processing
EAB presence points
EAB absence points
Environmental and Anthropogenic variables
Spatially filter EAB presence and absence points using a distance threshold
Remove multicollinear variables
Randomly select the same proportion of presence to absence points
Species Distribution Modelling
Response variable
Explanatory variables
Logistic regression
Random Forest
RGLM
Risk Map
Risk Map
Risk Map

Figure 4. Flow diagram of the methodology of the EAB species distribution model

4.1 Data Pre-Processing

In order to achieve an ideal species distribution model, the quality of the species data is an essential component as it forms the foundation of the model. However, there are various roadblocks and limitations that affect the quality of the species data in a species distribution model. Among the factors, two of the most prominent ones are the scale of the study area and species sampling bias.

Although the province-wide scale of the EAB data allows for a broad understanding of macro-level EAB interactions, the study area consequently also represents significant spatial heterogeneity across the landscape. Since the main goal of a species distribution model is to predict the suitability of a landscape for a species, the chosen scale for modelling should illustrate a strong interaction between the species and the limiting resources in its environment (Cushman & Huettmann, 2010). However, the optimal spatial scale that maximizes the relationship between the species and its surroundings is unknown to researchers and is somewhat restricted based on the species data acquired. In the case of the EAB, because the data was collected in counties of various sizes, there was a discrepancy between the scale of the species data and the predictor variables which were generated at a small scale for the anthropogenic variables and a coarse scale for the climactic variables. In order to maintain consistency, all the absence points that fell within a 1 km radius of the presence points were eliminated to ensure that each cell of the coarsest predictor variables, which also had a resolution of 1 km^2, did not have overlapping EAB presence and absence points. This was done to address the scale discrepancy between the EAB data and the predictor variables.

Aside from the spatial scale discrepancy, because the EAB data was collected over several years, there was also a discrepancy with the temporal scale of the species data and the predictor variables. For instance, most of the anthropogenic variables (i.e., campgrounds, sea ports, forest processing facilities, etc.) were collected during one specific year. However, since these anthropogenic features are static and do not typically change positions overtime, their effects would be similar during each year of EAB sampling. However, for the variables or layers that were available for multiple years such as NDVI and land surface temperature, each sampled point collected during a given year was matched with the variable from the same year. Lastly, the disjoint NDVI and surface land temperature predictor layers from several years were mosaicked as a single layer to be used as inputs for generating the risk maps from the models. With regards to modelling, a layer was generated known as the “distance to the nearest EAB positive location from previous years” which indicated the distance from a presence or absence point from a particular year to the nearest EAB presence point of previous years in order to account for the spatio-temporal variability in the dataset. This was done using the near tool in ArcMap 10.6 (Environmental Systems Research Institute (ESRI), 2016) which takes a point dataset and computes the Euclidean distance to another set of features. For instance, for the year 2008, the distances from all the presence and absence points to the nearest EAB presence point from the years 2002-2007 was calculated. After the distances for all the presence and absence points were calculated, the presence and absence points from the years 2006-2012 were merged as a separate points shapefile and used as the input data. Moreover, this variable indicated the likelihood for an arbitrary point to be infested by the EAB based on its proximity to locations of infested trees from past years.

Secondly, during field sampling, sampling bias presented in species data induces positive spatial autocorrelation (i.e., a cluster of EAB points in close proximity to one another). The sampling bias is broken down into two categories: geographic and climactic. Geographic sampling bias occurs when data acquisition is performed within a specified area such as along main roads, highways, or within high risk areas (Barbet-Massin et al., 2012; Syfert et al., 2013). In the case of EAB sampling performed by the CFIA, the data acquisition was performed in premeditated areas such as the areas along highways, near urban centres, provincial parks, campgrounds and ash nursery. Geographic sampling bias is demonstrated by the EAB presence points found serially along an agricultural road as shown in Figure 4.1.

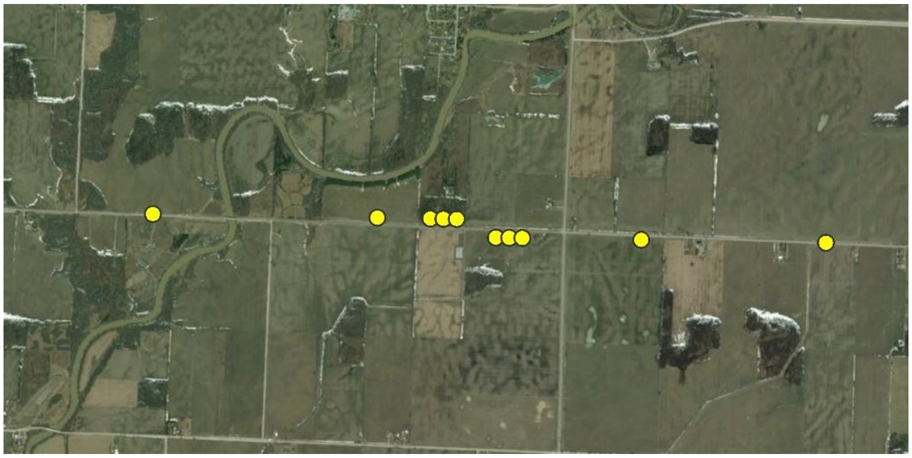


Figure 4.1. Example of geographic sampling bias of EAB samples taken along roads

If these areas were to be analyzed according to an explanatory variable such as land cover, it would appear that some EAB prefer to inhabit agricultural areas when in reality sampling was just performed on trees alongside agricultural fields. As a result, in order to avoid potential inconsistent modelling results due to the sampling bias, a land use dataset will not be included as one of the explanatory variables.

Likewise, climatic sampling bias occurs when the sampling isn't carried out over an entire environmental range, such as sampling only at low altitudes due to a lack of accessibility to higher altitudes (Barbet-Massin et al., 2012). In both sampling cases, a bias is present in the acquired species samples due to a locational accessibility of the samplers. When sampling bias is present in the sample, it does not adequately represent the true distribution of the species. Climatic sampling bias does not directly apply to the sampling design of the EAB as ash trees are found on relatively low laying surfaces which are easily accessible to samplers. In the context of the EAB data, if neighbouring sampled points share the same attributes for an explanatory variable (i.e., similar elevation values), the resulting model will be overfitted and model performance values will appear inflated (Boria et al., 2014; Hijmans & Elith, 2013; Veloz, 2009). An overfitted model is a model that fits too closely to the calibration data and limits the model's ability to predict an independent set of testing data (Boria et al., 2014).

In a broader sense, an inflated test statistic increases the chance of the Type I error or an incorrect rejection of the null hypothesis which was demonstrated by Veloz (2009). Veloz (2009) concluded that spatially autocorrelated occurrence data of an invasive plant species known as Asteraceae led to a lack of independence between the training and testing datasets resulting in a bias in model predictions. Veloz (2009)'s research demonstrated that accounting for spatial autocorrelation significantly improved the prediction accuracies of the GARP model as indicated by the similarity statistic, which was 0.719 before filtering the occurrence points and 0.828 after spatial filtering.

Similarly, Václavík and co-authors (2012) used various species distribution models to predict the spread of an invasive plant pathogen known as P. *ramorum* and have concluded that accounting for multi-scale structure of spatial autocorrelation by using spatial eigenvector mapping (SEVM) significantly enhanced the predictive capability of the models. The presence of autocorrelation in a dataset can be detected using the Moran's I, Geary's c (Elith & Leathwick, 2009; F. Dormann et al., 2007) and the nearest neighbour index (Fisher et al., 2007) which compares the average distance from a point's centroid to its nearest neighbour's centroid with the hypothetical average distance if the points were randomly distributed. The average nearest neighbour index, or ratio is calculated as the ratio of the observed average distance to the expected average distance for a random distribution (Environmental Systems Research Institute (ESRI), 2016). If the average nearest neighbour index is less than 1, the spatial distribution of the points is clustered whereas when the spatial distribution of the points is dispersed, the index is greater than 1. Autocorrelation of the EAB points was assessed using the nearest neighbour index where the observed average distance ($\overline{D}_o$) between each point and its nearest neighbour is equal to the sum of the nearest distances for each point and divided by the number of points. On the other hand, the expected mean distance of the points for a random distribution ($\overline{D}_E$) is calculated by (4.1).

$$\overline{D}_E = \frac{0.5}{\sqrt{n/A}} \qquad (4.1)$$

where n is the number of points and A is the area of the rectangle surrounding the points. The average nearest neighbour z-score statistic is obtained by subtracting $\overline{D}_E$ from $\overline{D}_o$ and further dividing by the standard error (SE), which is equal to $0.26136/\sqrt{n^2/A}$.

The EAB presence points achieved an observed mean distance of 2349.93 metres whereas the expected mean distance was 19031.80 metres resulting in a nearest neighbour ratio of 0.123. Similarly, the EAB absence points achieved an observed mean distance of 752.48 metres with an expected mean distance of 2882.03 metres, resulting in a nearest neighbour ratio of 0.261.

The statistical testing of the average nearest neighbour analysis assigns a significance level p-value for specified ranges of the critical value (z-scores) achieved by the analysis. For instance, within the critical value range between -2.58 to -1.65, the significance level is at 0.05 and with the critical values beyond -2.58, the significance level is 0.1. The z-score achieved by the EAB presence and absence points were -29.43 and -151.07, respectively which suggests a less than 1% chance that the clustered distribution was the result of random chance.

The presence of autocorrelation in the species data can be addressed during modelling using methods such as autocovariate regression, spatial eigenvector mapping, generalized least squares and generalized linear mixed models (F. Dormann et al., 2007). Where the aforementioned methods analyzes spatial autocorrelation during modelling, the effects autocorrelation was addressed pre-modelling in this research by filtering the species data using a specified distance tolerance (Boria et al., 2014; Veloz, 2009). By filtering the clusters of species localities to a single point, the resulting points become spatially independent which is essential for model calibration and evaluation (Brown, 2014). However, the distance to spatially filter species data is arbitrary but should be selected with a valid justification. For instance, Boria et al. (2014) spatially filtered occurrence data of a shrew species known as Microgale cowani by removing localities that were within 10 km of one another. The distance 10 km was chosen based on the high spatial heterogeneity of the mountains in Madagascar, the habitat of the Microgale cowani.

Since the primary focus of this research is to understand the nature of EAB dispersal, intuitively the distance between neighbouring EAB presence points should hold more importance as it determines the beetle's dispersal range. With this in mind, the distance at which EAB clustering resulted in positive autocorrelation was determined using a heuristic approach rather than a mathematical one. A technique proposed by Kramer-Schadt et al. (2013) to reduce autocorrelation included filtering Malay civets occurrence points within a radius of 10 kilometers. This radius was chosen because it represented the home range distance of individual Malay civets in Borneo, Southeast Asia. Similarly, the natural spread of the EAB was simulated by USDA's Northern Research Station where the EAB beetles were released and captured at outlier sites originated 1 and 3 years earlier from infested nursery trees. Conclusively, EAB-colonized trees were found 638 and 540 metres from the epicentres at the 1 year and 3-year sites, respectively (Northern Research Station, 2016). This suggested that the extent of the EAB spread after a year was approximately 540 metres from its epicentre and clusters of the EAB were likely to be found along this radius.

Although USDA's research provided a great insight into the realized dispersal extent of the EAB, it was merely a simulation and cannot be approximated for the EAB samples used in this research. As a result, the shortest distance from an EAB-infested ash tree to nearby infested ash trees as an approximation of the realized dispersal extent of the EAB was determined using a histogram in Figure 4.2.

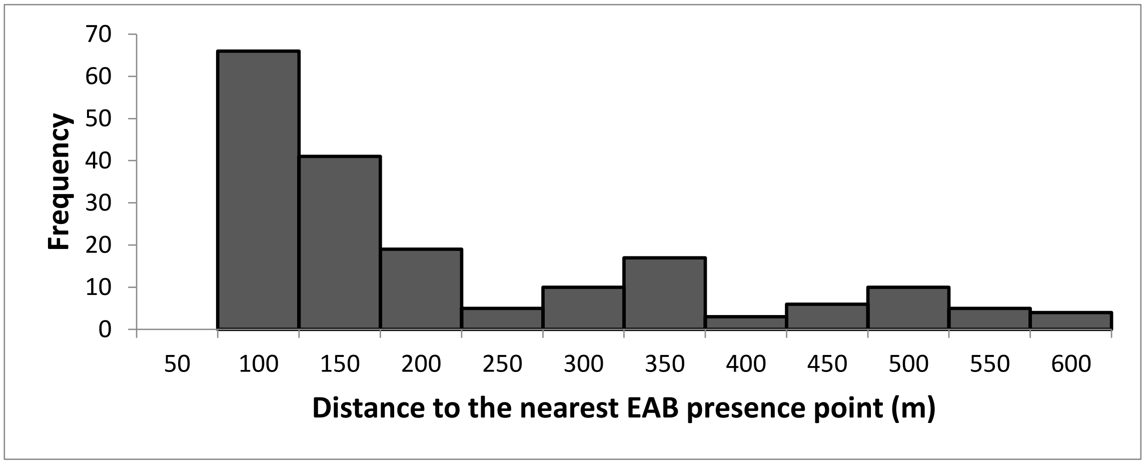


Figure 4.2. Histogram of the frequency of distances between two neighbouring EAB presence points

According to the histogram in Figure 4.2, maximum clustering of the EAB presence points occurred at a radius of 100 metres and decreased gradually as the distance increased. Furthermore, to filter the autocorrelated EAB presence points, a filtering distance of 100 metres was chosen. Similarly, since the EAB absence points were treated as a separate group, the minimum distance between two EAB absence points at which clustering was most prominent was also determined using a histogram in Figure 4.3.

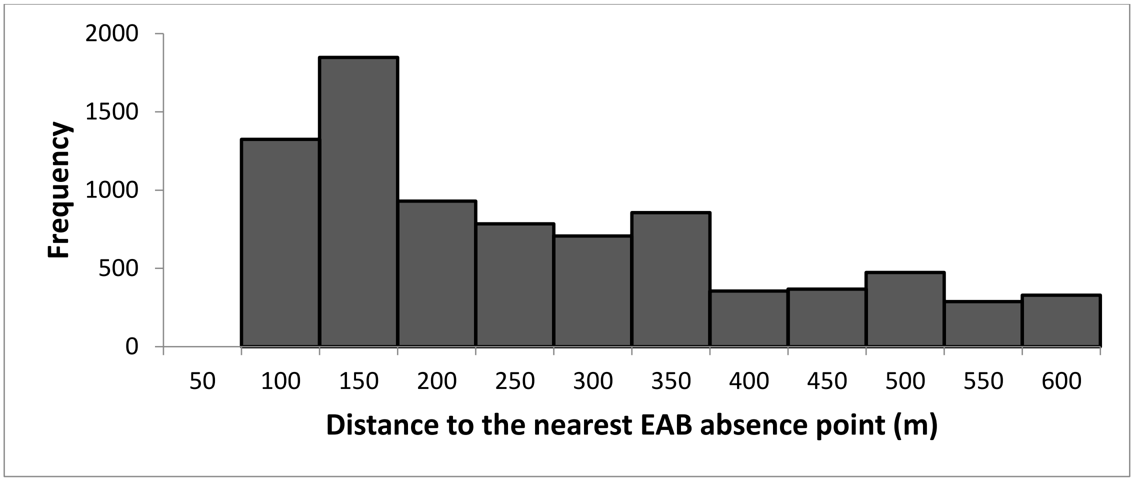


Figure 4.3. Histogram of the frequency of distances between two neighbouring EAB absence points

After the optimal distances for the EAB presence and absence points were determined, a filtering tool known as the "Spatially Rarefy Occurrence Data for SDMs" created by Jason L. Brown (Brown, 2014) was used to filter multiple sampled records to a single record within the two specified distances. This tool is essentially used to reduce the bias of predictor variables resulting from spatially autocorrelated occurrence data, which otherwise would compromise a distribution model's ability to predict spatially independent data. The tool functions by calculating a distance matrix for all the points and then systematically removing points that are closer than the specified search distance. It is a non-random process where the closest cluster is removed first, then the table of distances is re-evaluated until all the points are removed at the specified distance. This tool was used on both the EAB presence and absence data to reduce the points to their respective distance thresholds.

An overview of the tool suggests that spatially filtering species data at 5 km^2, 10 km^2, and 30 km^2 in areas of high, medium, and low environmental heterogeneity (Brown, 2014). Given the large scale of Southern Ontario and the range of elevation values (21.5 to 655 m), the study area fell under the category of high environmental heterogeneity and the distances derived (i.e., 100 and 150 metres) for filtering the EAB presence and absence points were confirmed as appropriate. When the filtering thresholds of 100 and 150 metres were used on the presence and absence points, respectively, the spatially rarefy occurrence tool reduced 269 presence points to 250 points and 11,422 absence points to 9525 points. The pictorial depiction of the spatially rarefy tool is displayed in Figure 4.4.

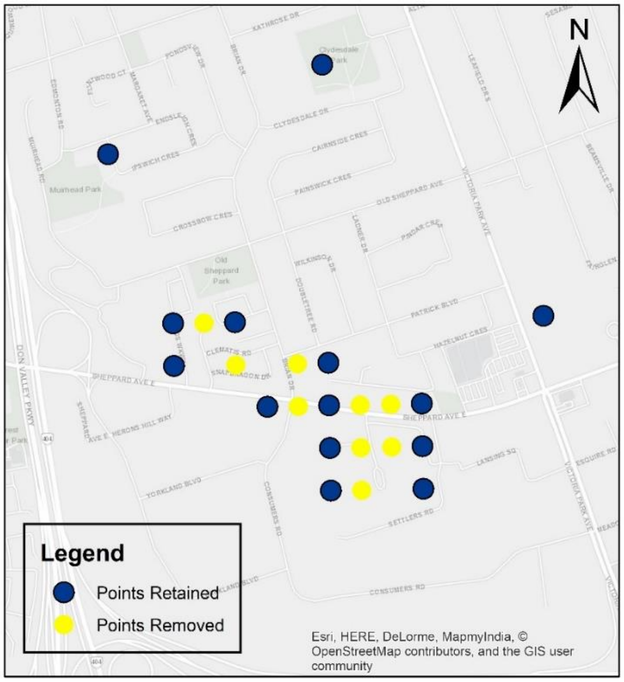


Figure 4.4. Usage of the spatially rarefy occurrence data tool using EAB presence points. The yellow points represent the points that were removed, and the blue points represent points retained by the tool

Next, 20% of the presence points from 2006-2012 were randomly selected for the validation dataset including an equal number of absence points. Subsequently, increasing numbers of absence points were randomly selected to assess the effects of prevalence on the logistic regression model. The points were selected using the "Sampling Design" tool developed by the National Oceanic and Atmospheric Administration (NOAA) (NOAA Biogeography Branch, 2013). This tool is designed to exercise sampling procedures under a GIS framework, and derive information about population metrics. NOAA's main usage of this tool was focused on marine habitats; however, its use can be extended to any type of population spread over physical space. There are two sampling selection procedures in this tool: simple random and stratified random. The simple random procedure randomly selects a user-defined number or percentage of samples whereas the stratified random procedure selects samples from various levels of strata, which is specified as an attribute. The absence points were chosen using the simple random procedure which ensures an unbiased selection of absence points from all the counties regardless of their concentration of absence points.

4.2 The Removal of Multicollinearity in Predictor Variables

Prior to building a spread model, the issue of multicollinearity in the predictor data should be analyzed thoroughly. Multi-collinearity is a phenomenon that leads to a deficiency in the proposed models, with which two or more predictor variables are linearly correlated such that one or more predictor variables can be derived from the others. A perfect collinearity exists between two independent variables if the correlation is equal to 1 or -1 (Akinwande et al., 2015). In theory, it is rare for ecological datasets not to exhibit multicollinearity, especially when climactic variables are modelled. For the explanatory variables used in this research, it is possible for the climactic variables such as June temperature, wind speeds, precipitation, and solar radiation to exhibit multicollinearity as they are driven by similar atmospheric circulation processes (Braunisch et al., 2013). If the purpose of a species distribution model is to predict within the range of the training dataset for interpolation purposes, then it can be assumed that the collinearity between variables will remain constant but for extrapolation purposes, consistent collinearity patterns cannot be presumed (Dormann et al., 2013; Werkowska et al., 2017). Since the models were tested on their interpolation and extrapolation abilities, the likelihood of multicollinearity was investigated. The presence of multicollinearity in a dataset can be detected using a correlation coefficient matrix where the variables with absolute correlation coefficients close to 1 indicate a strong correlation. Although a strong correlation does not necessary translate to collinearity, high correlation coefficients can usually be used to approximate linear relatedness, or collinearity (Dormann et al., 2013).

There are two methods to address multicollinearity in ecological data: cluster-dependent and cluster-independent. Cluster-dependent methods such as principle component analysis (PCA) and cluster analysis identifies the predictor variables that form clusters and creates a proxy set of the variables. On the other hand, cluster-independent methods such a correlation coefficient matrix and variance inflation factors (VIFs) bypasses the creation of clusters by identifying predictor variables that exhibits collinearity (Dormann et al., 2013) prior to modelling. Since cluster-dependent methods such as PCA (Lee et al., 2012) and cluster analysis (Wille, 2004, p. 273) are typically used for high dimensional data (i.e., numerous predictor variables) (Werkowska et al., 2017), the cluster-independent methods were more appropriate for this research as there were fifteen predictor variables and multicollinearity was a suspicion amongst only the climactic variables.

The correlation coefficient matrix of the predictor variables determines which variables are strongly correlated based on their pair-wise correlation coefficient (r). However, the appropriate cut-off value for declaring a variable as strongly multicollinear is subjective. As a result, a hypothesis test can be performed to test the significance of the correlation coefficients of the variables which takes into consideration the number of observations in the dataset. The p-value is calculated using a t distribution with $n-2$ degrees of freedom evaluated at the significance level of 0.05 (4.2):

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \tag{4.2}$$

where r is the correlation coefficient and the hypothesis statements are provided below:

H_o: The correlation coefficient is equal to 0

H_a: The correlation coefficient is different from 0

Although the correlation matrix and p-values provide insights about each pair of predictors, this may be limiting as even if pairwise correlations are small for two variables, there may be a linear dependence among three or more variables (The Pennsylvania State University, 2018). A more robust test for multicollinearity is the variance inflation factor (VIF) which summarizes the variables that are correlated to one or more variables by a single value (Kgosiesele, 2010; Midi et al., 2010). The VIF for an explanatory variable is obtained by running an ordinary least square regression as a function of the remaining variables as seen in (4.3).

$$VIF_k = \frac{1}{1-{R_k}^2} \tag{4.3}$$

In (4.3), ${R_k}^2$ is the coefficient of multiple determination of x_k with the other variables. When ${R_k}^2$ = 0, the VIF is equal to 1, and x_k is not correlated with other variables which indicates an absence of multicollinearity between predictors. When the VIF value is between 5-10, it indicates a high correlation (Akinwande et al., 2015) with one or more variables. The test for multicollinearity was performed in XLSTAT and the results are summarized in chapter 5.

When multicollinearity exists among the predictors, it affects each model type (statistical vs. machine learning) in a different manner. In logistic regression, multicollinearity inflates the variances of the estimated parameters and therefore provides incorrect inferences of the relationships between the explanatory and response variables (Midi et al., 2010; P. Vatcheva & Lee, 2016) based on the unstable p-values for the predictors. As a result, the inclusion of linearly correlated variables in the model can lead to an erroneous identification of the significant variables. For moderate and large sample sizes, it is recommended to eliminate one of the highly correlated variables in order to reduce multicollinearity (Midi et al., 2010). The upside to removing multicollinear variables is that it does not reduce the predictive powers of the model, but rather affects the calculations for individual predictors.

The Random Forest model also experiences negative effects of multicollinearity on its variable importance measures such as impurity-based ranking (i.e., Gini index). When a multicollinear variable is included in the model, it compromises the importance of all the variables that interact with the multicollinear variable such that the total importance of a given variable is either extremely low or suspiciously high because similar information is spread within the multicollinear variable(s) (Toloşi & Lengauer, 2011). Consequently, if the multicollinear variable(s) are considered several times for computing the total importance (Louppe et al., 2013), the importance of other variables will have a lower reported importance since a large portion of the impurity is already removed by the multicollinear variable. A strategy devised by Guyon *et. al.,* (2002) on handling correlated predictor variables in Random Forest is known as recursive feature elimination (RFE), which re-examines ranked variables using a permutation importance measure at each step of a backward elimination algorithm. Using this method, RFE can effectively create models with fewer, significant variables that remain during the last steps of the backward elimination procedure even with the presence of correlated variables (Gregorutti et al., 2017). The RFE algorithm was accessed through the caret package in R (Kuhn et al., 2018). The model was evaluated with a 10-fold cross validation scheme by first computing the variable importance for all the variables and then recursively re-evaluating the importance using each subset of variables $S_i, i = 1 \dots S$, where S represents the total number of variables. A similar approach by Strobl *et. al.,* (2008) used a conditional importance measure which consisted of conditionally permuting the variables to correlated variables. However, this method is only feasible for a small number of predictor variables as it is computationally intensive.

As for RGLM, since each bag of the model is constructed using an individual logistic regression model and forward variable selection is performed on each set of observations using a specified number of random variables for each bag, it can be assumed that the inclusion of multicollinear variables in RGLM will also increase the likelihood of inflated standard errors of the regression coefficients. In addition, during the forward variable selection process, if one collinear predictor is selected rather than another, then the selection process could take very different paths (Dormann et al., 2013). Nonetheless, the multicollinearity effect in RGLM is not known to its full extent.

4.3 Species Distribution Models Specifications

The species distribution models were developed by adjusting specific parameters to accommodate the response and explanatory variables. However, in order to obtain accurate information about the significance of the variables, the existence of complete or quasi-separation caused by the predictor variables was assessed using the “brglm2” package in R and using a contingency table. With regards to the models, in logistic regression, first a backward then forward stepwise variable selection method was used to effectively identify the most important variables. For starters, backward variable selection is preferred by many statisticians (Shtatland et al., 2001) as it starts with a full model and eliminates variables at each step that fail to meet the critical p-value (Haque et al., 2018). However, the backward variable selection tends to over-estimate the number of significant variables in a dataset compared to forward or step-wise variable selection methods as demonstrated by Austin & Tu (2004), Haque et al. (2018) and Iwundu & Efezino, (2015). For this reason, the predictor variables identified as important using the backward variable selection procedure were compared to forward and backward stepwise variable selection procedures which applies multiple removal and addition steps on the variables. Moreover, given the caveats of stepwise variable selection such as biases in parameters and over-fitting (Sainani, 2013; Whittingham et al., 2006), it was justified to perform backward variable selection prior to stepwise for a fair comparison of the results.

In XLSTAT, the logit link function was used according to the likelihood criterion. This way, the change in log likelihoods between each step is tested to determine which variables should be excluded from the model. If the overall fit of the model increases following the elimination of a variable, it is removed (Sarkar et al., 2010). The tolerance was set to 0.001, below which a variable is automatically ignored. The entry probability for an explanatory variable into the model was set 0.15 while the probability for the removal of a variable was 0.2 as recommended by Hosmer & Lemeshow (2000). Under the specifications for stop conditions, the number of iterations performed was set as 100 and a value of 0.00001 for the convergence. The confidence interval was set to be associated with the probability of 95%, indicating a significance level (α) of 0.05.

Generally speaking, since logistic regression focuses on arriving at a single best model using multiple iterations of the same dataset, it can be argued that standard logistic regression is not a robust method for predicting new datasets. The concept of bagging or bootstrap aggregation can be exercised to avoid the instability of automatic variable selection methods of logistic regression (Sainani, 2013) by taking random subsets of replicates of the training dataset with replacement to establish relationships with the predictor variables. By using the out-of-bag (OOB) dataset to calculate the misclassification rate of the bootstrapped samples, cross-validation is performed internally in Random Forest. However, since random variability in model fitting is exhibited by Random Forest (Wenger & Olden, 2012), 100 iterations were performed.

The two main parameters in Random Forest that governs the construction of trees are the number of trees constructed and the number of predictor variables selected for splitting the nodes of each tree. Although Random Forest provides a guideline on how to select the number of predictor variables to be randomly chosen at each node of the tree (i.e., $\sqrt{predictor\ variables}$), the number of trees to be constructed is subjective. As a result, a sensitivity analysis was performed which analyzed the OOB error rates of the model by increasing the number of trees. Using this method, the optimal number of trees that produced the lowest OOB error rate was selected. Aside from Random Forest's bagging feature which averages the results of multiple bootstrapped samples, its high predictive power is also achieved by a randomized selection of the predictor variables. The randomization process essentially de-correlates the trees making them less variable (James et al., 2013). However, since the variables are chosen randomly and not according to an advanced variable selection scheme, this also means insignificant variables have an equal chance of being selected as significant variables for the root node (Song et al., 2013). With this in mind, RGLM, a statistical-machine learning hybrid model combines the advantages of bagging samples with forward variable selection in each bag in order to maximize the selection of significant variables.

Another stipulation that RGLM enforces on the variable selection process is that prior to performing forward variable selection on a specified number of variables, it ranks all the variables according to their individual association with the response variable. The variables for ranking are selected from the entire pool of variables according to the “nFeaturesInBag” parameter. According to RGLM’s guidelines, the value of this parameter depends on the total number of variables. If the number of predictor variables is between 11-300 (Song et al., 2013), then the value for this parameter is calculated by N(1.0276 – 0.00276N), where N is the number of total variables. Using a value of 13 as N for the equation above provided the value of 12.89. As a result, the rounded-up value of 13 was used. Using this method, the predictor variables are subjected to two rounds of importance ranking and uses fewer variables for prediction than Random Forest (Song et al., 2013).

Next, much like Random Forest, a sensitivity analysis was performed for RGLM to determine the optimal number of candidate explanatory variables to be considered for forward regression in each bag that resulted in the lowest OOB error rate. As has been noted, RGLM includes a range of variable selection parameters that helps capitalize on the selection of important variables. A valuable parameter of RGLM that should be mentioned is “mandatoryCovariates” which has the ability to force a variable into the bags even if it is not selected for forward variable selection. This can be useful to researchers when assessing the impact of a variable of interest on the model performance. However, since there were no proposed target variables for the EAB, this parameter was not used.

4.4 Assessment Methods of Model Performance

After the development of distribution models using species data, it is informative to compare their performance using a collection of assessment methods. In this research, three different approaches were used: classification tables, receiver operating characteristic (ROC) curves and risk maps. The standard classification tables, or confusion matrix cross-tabulates the actual versus the predicted presence and absence records. The accuracies are commonly displayed as percentages. However, classification tables are threshold dependent such that by convention, a probability of 0.5 or greater predicted by the model is classified as presence (1) and probabilities below 0.5 are classified as absence (0). That is to say, the threshold is user-enforced and its selection can have a significant impact on the model accuracy and the predicted prevalence (Freeman & Moisen, 2008).

The usage of ROC plots has been adopted by ecological studies due to its threshold-independence in evaluating presence-absence models. The ROC curve essentially plots the true positive rate (sensitivity) against the false positive rate (1 – specificity) of a model across various probability thresholds ranging from 0 to 1. In order to construct the ROC curve, the probabilities generated for each observation of the prediction dataset are ordered from lowest to highest and each probability is sequentially used as a threshold to classify the points. The points for the curve start at the origin and increases by one unit for every positive outcome and increases one unit to the right for every negative outcome. Lastly, the area under the ROC curve, referred to the AUC is determined by the trapezoidal rule. The AUC index provides a discrimination measure of model performance (Lobo et al., 2008) where effective models exhibit AUC values near 1 and poor models exhibit AUC values close to 0.5. The ROC curves and AUC values were obtained using the “pROC” package (Xavier *et. al.,* 2011) in R.

The last method of validation of the distribution models is performed visually through risk map validation. The risk maps reflect the level of risk associated with each cell in the study area using the same classification scheme of the probabilities in ArcMap 10.6. For instance, the probabilities associated with each risk level were reclassified such that the lowest risk corresponded to EAB presence probabilities between 0 – 0.22, low risk corresponded to probabilities between 0.22 – 0.40, moderate risk between 0.40 – 0.50, high risk between 0.50 – 0.72 and highest risk between 0.72 – 1. The range for the highest risk was liberal due to the possibility of areas exhibiting latent signs of EAB infestation. The resolution of the risk maps was set as 1 km to maintain consistency with the resolution of the coarsest variables and are displayed in the Appendix section.

The risk maps were validated by determining the proportion of EAB presence and absence points from 2013 that fell under each category of risk. Although the classification scheme of the three risk maps outputted by the SDMs were similar, the process that involved the creation of the risk maps differed. To begin, to create the risk map for 2013, some layers were required to be generated and updated. For instance, the NDVI layer was updated to a 2013 version and a Euclidean distance layer was included which displayed the relative distances from each presence point throughout the years 2002-2012. For the logistic regression model, only the variables identified as significant were included in the risk map. The unstandardized coefficients associated with the predictor variables were used as weights. The equation was directly inserted into ArcMap's "Raster Calculator" tool to create the risk map for 2013. Furthermore, for each cell in the study area, the values for the predictor variables were used as input data to derive an EAB presence probability index between 0 to 1.

The risk maps for RGLM and Random Forest were generated using a different approach as unlike logistic regression, the probabilities could not be directly imported into a GIS platform for visualization. First, a lattice of equally spaced points (1000 m by 1000 m) was created across the study area. This resolution was chosen to approximate the large scale of Southern Ontario and maintain consistency with the climactic variables. The points were then projected to the spatial reference system "NAD 83 UTM Zone 17" and the values for each explanatory variable were extracted for each point. The R-ArcGIS bridge was used to obtain the EAB presence probabilities from the Random Forest and RGLM models and assign the probabilities to the sampled points. Lastly, the "Inverse distance weighting" tool was used to convert the points into a surface. In theory, each pixel of the risk map is a value from 0 to 1 which represents EAB risk, where 0 is the lowest EAB risk and 1 is the highest EAB risk. The risk maps for the logistic regression model, random forest and RGLM are displayed in the Appendix.

CHAPTER 5: RESULTS & DISCUSSION

Chapter 3 highlighted the data layers used in this research and the descriptive statistics provided preliminary insights of the EAB presence and absence data using minimum, maximum, median, mean and standard deviation values. The methodology in Chapter 4 featured a workflow of processing the response variable and explanatory variables and details about their inclusion into the three species distribution models (i.e., logistic regression, Random Forest and RGLM). Lastly, Chapter 5 will summarize the results of the multicollinearity test of the explanatory variables, the classification accuracies of the models and the variable importance rankings.

5.1 Multicollinearity Results of Explanatory Variables

To reflect, multicollinearity is a phenomenon that arises when there is an approximately linear relationship between two or more predictor variables resulting in an unfit model. Hence, it is important to exclude them from the input variables (Akinwande et al., 2015; Hegyi & Laczi, 2015). Multicollinearity of the variables was tested using the Pearson coefficient correlation matrix and a multicollinearity test. According to Evans (1996), correlation coefficient values between 0.40-0.59 indicate a moderate correlation, values between 0.60-0.79 indicate a strong correlation and values between 0.80-1.0 indicates a very strong correlation. A correlation matrix of the predictor variables used in this research is summarized in Table 5.

Table 5. Correlation coefficient matrix of the predictor variables

Variables	Population Centres	NDVI	Aspect	Slope	Ports	June Precipitation	Forest Processing Facilities	June Wind Speed	Nearest EAB Positive Location	Elevation	Camps	Highways	Surface Temperature	June Solar Radiation
Population Centres	**1.000**	0.110	0.076	-0.183	0.182	0.077	-0.439	0.381	0.265	0.251	0.786	0.086	-0.247	0.377
NDVI		**1.000**	0.020	0.095	0.123	-0.028	-0.010	0.087	0.024	0.036	-0.004	0.006	-0.082	-0.017
Aspect			**1.000**	-0.113	-0.025	0.081	-0.048	0.121	-0.026	0.019	0.014	0.009	0.001	-0.051
Slope				**1.000**	-0.089	-0.156	0.149	-0.380	-0.149	-0.108	-0.222	0.003	0.034	0.017
Ports					**1.000**	0.578	0.168	0.128	0.134	-0.325	0.171	0.194	-0.031	-0.521
June Precipitation						**1.000**	-0.154	-0.073	0.094	-0.251	0.180	0.040	-0.050	-0.522
Forest Processing Facilities							**1.000**	-0.276	-0.255	-0.649	-0.613	0.144	0.170	-0.600
June Wind Speed								**1.000**	0.173	0.320	0.357	-0.107	-0.192	0.287
Nearest EAB Positive Location									**1.000**	0.080	0.384	0.320	0.144	0.149
Elevation										**1.000**	0.258	-0.259	0.008	0.627
Camps											**1.000**	0.066	-0.227	0.455
Highways												**1.000**	0.081	-0.217
Surface Temperature													**1.000**	-0.049
June Solar Radiation														**1.000**

According to Table 5, four variables exhibited strong correlations with two or more variables. For instance, June solar radiation had a strong correlation with the variables forest processing facilities (0.6) and elevation (0.627). The variable camps experienced strong correlations with population centres (0.786) and forest processing facilities (0.613). Elevation had a strong correlation with June solar radiation (0.627) and forest processing facilities (0.649). Lastly, forest processing facilities was strongly correlated with elevation (0.649), camps (0.613), and June solar radiation (0.6). In addition, the variables that exhibited significant p-values indicating strong correlations with other variables are as follows: camps, June solar radiation, forest processing facilities, surface temperature, highways, elevation, nearest EAB positive location, and June wind speed.

The issue with the correlation coefficient matrix and p-values are that only pair-wise relationships are assessed and therefore relationships with other interacting variables are not considered. As a result, the variance inflation factors (VIF) of all the variables is summarized in Table 5.1 which provides a single value for the degree of multicollinearity of a variable against all the other variables. Although the threshold VIF value for considering a variable as multicollinear is arbitrary, according to existing literature, VIF scores between 5 to 10 indicates a high correlation (Akinwande et al., 2015).

Table 5.1. Multicollinearity test using all explanatory variables

Variable	Population Centres	NDVI	Aspect	Slope	Ports	June Precipitation	Forest Processing Facilities	June Wind Speed	Nearest EAB Positive Location	Elevation	Camps	Highways	Surface Temperature	June Solar Radiation
R^2	0.667	0.087	0.070	0.239	0.598	0.731	0.793	0.402	0.311	0.627	0.787	0.273	0.209	0.825
VIF	3.007	1.095	1.075	1.314	2.485	3.717	4.827	1.673	1.452	2.684	4.687	1.375	1.264	5.714

From Table 5.1, it is evident that aside from elevation, the predictor variables that were correlated with two or more variables in Table 5 also exhibited high VIF values, except for camps. For instance, the calculated VIF values for the variables June solar radiation, distance from camps and forest processing facilities are greater than or close to 5. Because June solar radiation had the greatest VIF value (5.714) compared to the rest of the variables, it was removed from the list of input predictor variables. For the variables camps and forest processing facilities, because they shared similar second highest VIF values (4.687, 4.827), it was more difficult to omit one. In addition, according to literature (Huset, 2013; Prasad et al., 2010), the two variables shared a known contribution in the dispersal of the EAB. As a result, they were retained in the model until further analysis.

5.2 Logistic Regression

As mentioned previously, before analyzing the results of logistic regression, the predictor variables' abilities to separate the response variable into two mutually exclusive groups was assessed using the "detect separation" method of the brglm2 package. According to the results, separation was not detected by the variables as the maximum likelihood estimates of all the variables were all finite. However, contingency tables were created for all the variables to assess the likelihood of quasi-separation. The results indicated that the variable June wind speed was able to separate the two classes of EAB presence and absence the most effectively. The result of the contingency table is provided in Table 5.2.

Table 5.2. Contingency table of June wind speed values of the EAB presence and absence points

June Wind Speed Ranges (m/s)	**Presence**	**Absence**
2.50 - 2.94	35	5
2.95 - 3.38	88	14
3.39 - 3.82	127	67
3.83 - 4.26	0	125
4.27 - 4.70	0	30
4.71 - 5.14	0	9

According to the contingency table of the EAB presence and absence training points in Table 5.2, although it is certain that the EAB do not inhabit areas with wind speeds greater than 3.83, for wind speeds lower than 3.83, there is an overlap of the two classes suggesting a case of quasi-separation. Furthermore, in order to ensure finite maximum likelihood estimates, the Firth's method was applied to the stepwise models.

5.2.1 Backward Variable Selection Method excluding Multicollinear Variable(s)

As mentioned in the methodology, after eliminating the variable June solar radiation due to its exhibition of high multicollinearity, the standard backward variable selection method was used to eliminate insignificant variables based on the probability of the likelihood-ratio statistic. The results of the backward variable selection process are summarized in Table 5.3.

Table 5.3. Results of the backward variable selection procedure for logistic regression

Number of Variables	Variables	Variable IN/OUT	Status	-2 Log(Likelihood)	Pr > LR
13	Population Centres / NDVI / Aspect / Slope / Ports / June Precipitation / Forest Processing Facilities / June Wind Speed / Nearest EAB Positive Location / Elevation / Camps / Highway / Surface Temperature			174.718	0.00
12	Population Centres / NDVI / Aspect / Slope / Ports / June Precipitation / Forest Processing Facilities / June Wind Speed / Nearest EAB Positive Location / Elevation / Camps / Surface Temperature	Highway	OUT	174.742	0.00
11	Population Centres / NDVI / Aspect / Slope / Ports / June Precipitation / Forest Processing Facilities / June Wind Speed / Nearest EAB Positive Location / Elevation / Camps	Surface Temperatu re	OUT	174.756	0.00
10	Population Centres / NDVI / Slope / Ports / June Precipitation / Forest Processing Facilities / June Wind Speed / Nearest EAB Positive Location / Elevation / Camps	Aspect	OUT	174.981	0.00
9	Population Centres / Slope / Ports / June Precipitation / Forest Processing Facilities / June Wind Speed / Nearest EAB Positive Location / Elevation / Camps	NDVI	OUT	175.582	0.00
8	Population Centres / Slope / Ports / June Precipitation / Forest Processing Facilities / June Wind Speed / Nearest EAB Positive Location / Elevation	Camps	OUT	176.397	0.00

According to Table 5.3, the variables highways, June surface temperature, aspect, NDVI and camps were eliminated from the full model using the backward variable selection process as the -2 Log(Likelihood) values increased at each step following the removal of the variables. A noteworthy observation about the camps variable is that in addition to being excluded from the model, it also exhibited a high VIF value from the multicollinearity test (Table 5.1).

5.2.2 Stepwise Variable Selection Method excluding Multicollinear Variable(s)

After an insight was made into the variables eliminated by the standard backward variable selection process, forward and backward stepwise variable selection methods were used by applying the Firth's method to prevent a quasi-separation of the data by the June wind speed variable. The estimated Chi-Squared (Wald) test statistic for each variable was calculated as the squared ratio of the coefficient to the standard error and the p-value was used to test the null hypothesis (H_o). If the significance of the p-value was lower than the specified significance level of α (i.e., 0.05), the null hypothesis was rejected. As a side note, the cut-off p-value is rounded to the tenth decimal place, thus a value of p-value = 0.049 is rounded to 0.05. The two stepwise variable selection methods were compared using three criteria: the variables identified by the methods as significant (i.e., p-value less than 0.05), the variables that were considered insignificant (i.e., p-value greater than 0.05), and the presence and absence classification accuracies of the validation and prediction datasets. The results of the two stepwise variable selection methods are compared in Table 5.4.

Table 5.4. Variable selection results and classification accuracies of the forward and backward stepwise methods of logistic regression

Stepwise Method	**Significant Variables**	**Insignificant Variables**	**Validation Accuracy (%)**	**Prediction Accuracy (%)**
Forward	June Wind Speed Slope Nearest EAB Positive Location Population Centres June Precipitation Elevation	Forest Processing Facilities Ports Highway Surface Temperature Camps NDVI Aspect	93%	52.38%
Backward	June Wind Speed Slope Nearest EAB Positive Location Elevation	Forest Processing Facilities Ports Highway Surface Temperature Camps NDVI Aspect Population Centres June Precipitation	93%	70.45%

From Table 5.4, it is evident that after eliminating the variables June precipitation and population centres by the stepwise backward selection method, it achieved a prediction accuracy 18% greater than the stepwise forward selection method. As a result, although the validation accuracies of both the selection methods (i.e, forward and backward) were similar, the stepwise backward variable selection method produced a more parsimonious model which had greater extrapolation capabilities. Overall, the accuracies for the validation dataset exceeded the accuracies of the prediction dataset. An insight into the lower accuracy of the prediction dataset from 2013 could be attributed to the spatio-temporal variation of the dataset compared to the validation dataset.

An important observation regarding the variable camps is that since it was eliminated by the standard backward variable selection and the stepwise methods in addition to exhibiting multicollinearity, the inclusion of camps in subsequent models can potentially undermine the models' predictive powers. As a result, the variable camps was excluded from the Random Forest and RGLM models. A closer inspection of the variable importance rankings of the stepwise backward method was assessed in Table 5.5 using the unstandardized and standardized coefficients.

Table 5.5. Variable importance ranking of the backward stepwise process excluding multicollinear explanatory variables

Variable	DF	Unstandardized Coefficients	Chi-Square (Wald)	Pr > Wald	Significance at α = 0.05	Standardized Coefficients
June Wind Speed	1	-9.02E+00	67.47	< 0.0001	Significant	-2.314
Elevation	1	1.19E-02	5.09	0.019	Significant	0.444
Slope	1	-5.31E-01	12.07	0.001	Significant	-0.426
Nearest EAB Positive Location	1	-5.20E-06	5.47	0.024	Significant	-0.317

In Table 5.5, the degrees of freedom (DF) corresponds to each variable estimated in the model. For each variable, one DF is required to define the Chi-Squared distribution to infer whether the unstandardized coefficient for the variable in question is 0 given the remaining variables are in the model. The unstandardized coefficients for each variable indicate the amount of change in EAB presence given by a one-unit increase in the log-odds of the variable. However, it should be noted that with the unstandardized coefficients, the difference in the units of the explanatory variables is not taken into account. For instance, the unstandardized coefficient of the anthropogenic variable (nearest EAB positive location) is considerably lower than the remaining variables because it is expressed in metres as opposed to the climactic variables which have less variation in the units of measurement. As a result, the magnitude of the relationship between the explanatory variables and the response variable should not be assessed using the unstandardized coefficients but rather the standardized coefficients which takes into consideration the units of measurement of the explanatory variables. The standardized coefficients are measured in units of standard deviations and indicates the change in the standard deviation of the dependent variable given one standard deviation increase in the log odds of an independent variable. The absolute value of the standardized coefficients, which are adjusted to the units of measurement for each variable are used to rank the variables in order of importance. A discussion regarding the change in the response variable (EAB presence/absence) caused by the predictor variables is as follows:

a. June wind speed

An increase of 1 m/s in wind speed decreased the log odds of EAB presence probability by 9.02. This relationship is in contrast with the theory that stronger winds assist EAB migration to farther distances (Stohlgren et al., 2010). However, due to the large temporal and spatial scale of the EAB data, it can be said that the propagation methods of the EAB has potentially changed over the years 2002-2012 and that the EAB now require stagnant winds for survival. June wind speed was ranked as the most significant variable as indicated by its absolute standardized coefficient which is reflected by its ability to quasi-separate the EAB presence and absence points from Table 5.5.

a. Elevation

The elevation exhibited a positive relationship with EAB risk such that for every 1 unit increase in the elevation of an area, the log odds of EAB presence probability increased by 0.0119 suggesting that the EAB prefer to inhabit trees on higher grounds.

b. Slope

Since the elevation possessed a positive relationship with EAB risk, it was expected that the slope would also exhibit a negative relationship as the slope is a bi-product of elevation. The slope had a negative relationship to the EAB risk as for one degree increase in slope, there is a 0.531 decrease in the logit of EAB presence probability. This relationship does not hold true for research conducted by Royo et al. (2012) where ash trees faced greater dieback and reduced crown conditions on upper slopes than ash populations in lower slopes. Although this is inherent for ash trees, the same conclusion cannot be made about the ash trees chosen by the EAB dataset in this research as according to logistic regression's results, the EAB prefer to inhabit ash trees on lower slopes.

c. Distance to the nearest EAB positive location

Perhaps one of the most important explanatory variables in this research is the distance from a presence or absence point to presence points from past years. It is expected from existing literature (Huset, 2013; Prasad et al., 2010) that it is more likely for a location to be infested by the EAB based on its proximity to past EAB-infested locations. According to Table 5.5, the results corroborate conclusions made by past literature where for every metre increase in the distance from a sampled point of a current year to the presence points from previous years, the log odds of the probability of EAB presence decreased by 0.0000052. This relationship, although very small, coincides with findings by Huset (2013) where for every metre increase in the distance from known EAB locations, the log odds of EAB presence also decreased. In addition, it was overwhelmingly the most significant variable as identified by logistic regression and Maxent models' low p-values compared to other variables by Huset (2013). However, according to the unstandardized coefficients and Pr > Wald values in Table 5.5, although the variable distance to the nearest EAB positive location proved to be significant, it did not provide an overwhelmingly substantial predictive power over June wind speed based on its unstandardized coefficient value.

Above all, it should be noted that Huset (2013) used only 3 years of fine-scaled EAB presence data in New York (2009-2011) where a majority of the EAB points appeared in clusters. However, the EAB data used in this research was collected over 11 years (2002-2012) across the province of Ontario. Following the outbreak of the EAB in 2002, the sightings appeared in close range until 2008. The EAB sightings appeared sporadic, especially for 2009 and 2010 when the EAB was discovered in isolated counties as seen in Figure 3.1. As a result, the distance to the nearest EAB positive location differed greatly from a current year to past years which could have contributed to the weaker relationship.

5.2.3 Assessing the Effects of Prevalence using Logistic Regression

As a reminder, although there was a greater proportion of absence points to presence points available, the logistic regression model was trained using a 1:1 prevalence to avoid a class imbalance issue (Cushman & Huettmann, 2010). Nonetheless, the effects of increasing the size of the absence points was further investigated in Table 5.6 by assessing the misclassification rates of the stepwise backward selection models.

Table 5.6. Misclassification rates of the training dataset using logistic regression

Number of Absence Points	**Misclassification Rate (%)**
200	7.00%
1200	3.64%
2200	4.17%
3200	3.74%
4200	3.11%
5200	3.00%
6200	2.64%
7200	2.30%
8200	2.14%

From the misclassification rates in Table 5.6, it is evident that as the number of absence points was increased from 200, the misclassification rate decreased gradually from 7% to 2.14% when 8200 absence points was used. This suggests that the logistic model performed the best when a 1:41 ratio of EAB presence to absence points was used. By the same token, it is expected that the high accuracy would be transferred to the validation and prediction datasets. To test this theory, the accuracies were examined using a 1:41 ratio (i.e., 8200 absence points) in Tables 5.7 and 5.8.

Table 5.7. Classification table of the validation dataset with 8200 absence points using logistic regression

	Absence	Presence	Total	% Correct
Absence	49	1	50	98.00%
Presence	33	17	50	34.00%
Total	82	18	100	66.00%

Table 5.8. Classification table of the prediction dataset with 8200 absence points using logistic regression

	Absence	Presence	Total	% Correct
Absence	3	19	22	13.64%
Presence	2	20	22	90.91%
Total	5	39	44	52.27%

According to the testing and prediction accuracies in Table 5.7 and 5.8, the results refute Barbet-Massin et al. (2012)'s findings that using a higher ratio of species presence to absence points achieves greater transferability for GLMs such as logistic regression. Compared to Table 5.4 where the models were trained using 200 absence points, the validation accuracy decreased by 27% whereas the prediction accuracy decreased by approximately 18% from the stepwise backward selection method. However, the key difference between Barbet-Massin et al. (2012)'s research and this research is the usage of pseudo-absence points whereas true absence points were used in the latter and thus is subjected to sampling bias. Furthermore, all subsequent models were trained with a 1:1 prevalence.

5.3 Random Forest

Among the various specifications of Random Forest, the two main parameters that influences the performance of the model are the number of trees constructed and the number of predictor variables randomly chosen at each node of the tree. Although a guideline to specify the number of variables to be randomly chosen for each node is determined by the value of the square root of the total number of predictors, the total number of trees to be constructed is arbitrary. A pragmatic solution is to conduct a sensitivity test by increasing the number of trees and comparing the output OOB error rates. The number of variables to be randomly chosen was rounded down to "3" from a calculated value of 3.46 by taking the square root of the total number of variables (i.e., 12). Next, the OOB error rate for the training sample for each step increase in the number of trees (n_{tree}) is displayed in Table 5.9.

Table 5.9. n_{tree} vs. OOB estimate of error rates for Random Forest

N_{tree}	**OOB Estimate of Error Rate (%)**
20	6.25%
50	5.75%
100	5.50%
200	5.25%
500	5.25%
1000	5.25%

It is evident from Table 5.9 that for the EAB training dataset, the OOB error rate is not overly sensitive to the number of trees used to construct the random forest and the rate remains constant at 5.25% beyond 200 trees. As a result, 200 was used as the number of trees. Although the OOB rates for the training sample were relatively low, the robustness of Random Forest is truly tested by its performance on novel datasets. As a reminder, one of the main objectives of this research is to assess the transferability of distribution models in two scenarios: unsampled areas within the same time frame (i.e., validation data) and areas in a future time frame (i.e., prediction data). From the logistic regression results in section 5.2, since it was concluded that the models with the same proportion of EAB presence and absence points had the best performance, the results were compared with Random Forest in Tables 5.10 and 5.11, respectively.

Table 5.10. Classification table of the validation dataset using Random Forest

	Absence	**Presence**	**Total**	**% Correct**
Absence	45	5	50	90.00%
Presence	0	50	50	100.00%
Total	45	55	100	95.00%

Table 5.11. Classification table of the prediction dataset from 2013 using Random Forest

	Absence	**Presence**	**Total**	**% Correct**
Absence	4	18	22	18.18%
Presence	3	19	22	86.36%
Total	7	37	44	52.27%

It is evident from Tables 5.10 and 5.11 that Random Forest had an excellent classification accuracy (95%) for the validation dataset but a poor prediction accuracy (52%). This suggests evidence of overfitting by Random Forest where the high accuracy achieved by the validation dataset was not reflected in the prediction dataset. Furthermore, much like logistic regression's results, the inability to transfer in novel datasets that vary spatio-temporally is a proven shortcoming of Random Forest.

5.3.1 Variable Importance of Random Forest

The variable importance chart for Random Forest is summarized in Table 5.12 using the mean decrease Gini which is also referred to as the mean decrease impurity (MDI) (Louppe, 2014).

Table 5.12. The mean decrease Gini (MDI) for the predictor variables in Random Forest

Variable	Mean Decrease Gini (MDI)
June Wind Speed	14.96
Population Centres	2.25
Highway	1.84
Land Surface Temperature	1.69
Forest Processing Facilities	1.67
Nearest EAB Positive Location	1.51
NDVI	1.20
Elevation	1.10
Slope	1.10
Ports	1.07
Aspect	0.62
June Precipitation	0.52

As a reminder, for an arbitrary tree in Random Forest, each time a parent node makes a split using a variable, the Gini impurity index for the two preceding children nodes is lesser in value than the parent node. As a result, by summing up the Gini decreases for each variable between the parent and children nodes across all trees in the forest, a variable importance score can be obtained. Accordingly, Table 5.12 shows that the greatest decrease in Gini was achieved by the variable June wind speed (MDI = 14.96) as it was the most significant variable. The remaining variables' decreases in Gini showed insignificant differences between two successive variables suggesting that their importance was not as prominent compared to the decrease in Gini of the first variable. As a reminder, it was established that the inclusion of multicollinear variables in Random Forest interferes with variable importance and thus were eliminated by examining their VIF values prior to modelling. A second method to handle multicollinearity was by employing the RFE algorithm in Random Forest. A comparison of the Gini index values for the two methods (i.e., VIF-guided vs. RFE) and a standard Random Forest model with the inclusion of all the variables is displayed in Figure 5.

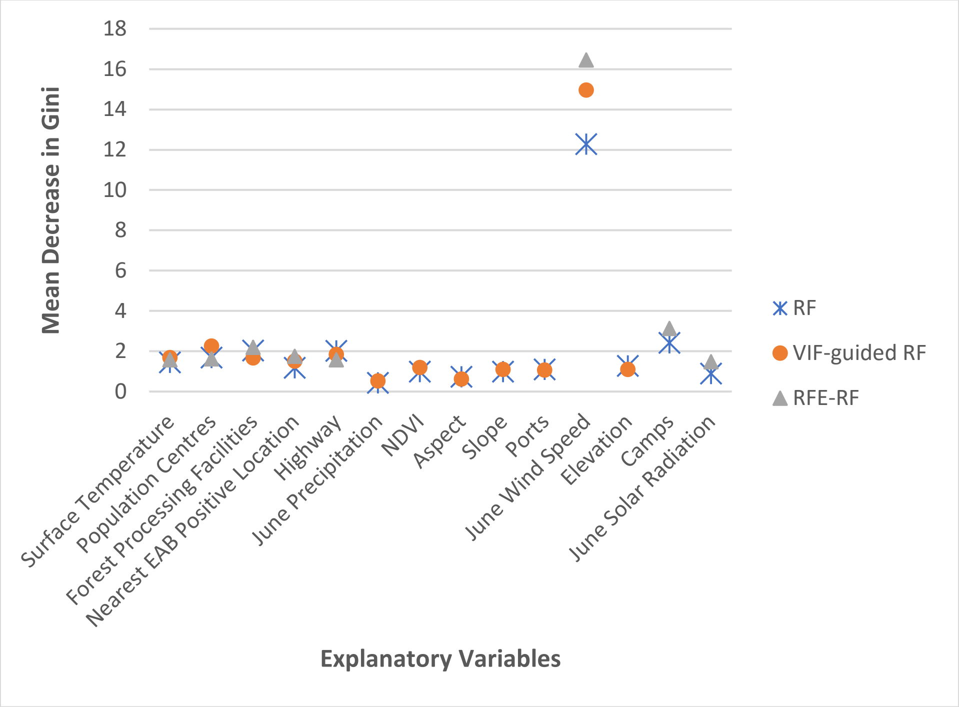


Figure 5. Comparison of the mean decrease Gini of the RF, VIF-guided RF, and RFE-RF models. *Note* The explanatory variables without values for the VIF-guided RF or RFE-RF models suggest that they were eliminated by the methods*

An examination of Figure 5 reveals that the inclusion of all the variables in the RF model achieved similar variable importance rankings as the VIF-guided RF model and the RFE-RF method. The variable June wind speed remained the top most important variable identified by all methods. For the RFE-RF method, the importance of June wind speed was the most substantial compared to the other models following the elimination of the variables NDVI, aspect, slope, ports, June precipitation. A noteworthy observation about the variables camps and June solar radiation, which exhibited multicollinearity and was consequently eliminated by the VIF-guided RF method, were included in the model by the RFE-RF method. A study conducted by Darst *et. al.,* (2018) concluded that in the presence of many highly correlated variables, the standard random forest model outperformed the RFE-RF model in terms of highly ranking the causal variables. Although there was not an abundance of the number of correlated variables in this research and there is mainly one causal variable, a reflection of the ability of the methods to effectively transfer to novel datasets was assessed by their accuracies. While all three models scored similar accuracies for the validation dataset (96%), the accuracies for the prediction dataset achieved by the standard RF and VIF-guided RF were 52.27% whereas the accuracy of the RFE-RF method was 47.73% which suggests the RFE-RF method eliminated too many variables which were necessary for calibration of the model. Furthermore, since the RF model achieved the same prediction accuracy as the VIF-guided RF method, it supports the theory that RF is robust against multicollinear variables (Matsuki et al., 2016).

5.4 Random Generalized Linear Model (RGLM)

In terms of transferability, both logistic regression and Random Forest shared similar results in that they performed relatively well on the validation dataset and poorly on the prediction dataset. Comparatively, logistic regression performed better than Random Forest on the prediction dataset (16% increase) which suggests superiority of a statistical model over the machine learning model for transferability purposes using species data. Since RGLM combines the concept of bootstrapped GLMs and a randomness aspect from Random Forest, it is expected that RGLM should have the best performance. As mentioned before, although a guideline for the number of variables randomly selected for each node is specified by Random Forest, RGLM does not provide such recommendations. As a result, while using the same number of trees in Random Forest as the total number of bags, the number of variables required for forward regression in each bag was determined by assessing the OOB error rates in a sensitivity analysis in Table 5.13.

Table 5.13. nCandidateCovariates vs. the out-of-bag (OOB) estimate of error rates for RGLM

nCandidateCovariates	OOB Estimate of Error Rate (%)
12	7.50%
11	7.50%
10	7.00%
9	7.00%
8	7.25%
7	7.00%
6	7.00%
5	6.50%
4	6.50%
3	6.50%
2	6.75%
1	7.25%

According to Table 5.13, it is evident that using 200 bags, the ideal number of candidate variables is difficult to assess as there is no single number of candidate predictor variables greater than which the OOB rate stabilizes at a constant value. As a result, the validation and prediction accuracies were analyzed for each number chosen for the candidate covariates. The results indicated that using a "nCandidateCovariates" value of 7 achieved the greatest transferability as summarized in Table 5.14 and 5.15, respectively.

Table 5.14. Classification tables of the validation dataset using RGLM

	Absence	Presence	Total	% Correct
Absence	46	4	50	92.00%
Presence	1	49	50	98.00%
Total	47	53	100	95.00%

Table 5.15. Classification tables of the prediction dataset from 2013 using RGLM

	Absence	Presence	Total	% Correct
Absence	20	2	22	90.91%
Presence	5	17	22	77.27%
Total	25	19	44	84.09%

From the classification results in Tables 5.14 and 5.15, while the validation accuracy achieved by RGLM were consistent with logistic regression and Random Forest, its prediction accuracy significantly exceeded both models. Furthermore, by abstracting the RGLM as a bagged regression model as opposed to an individual regression model, it was expected that its accuracies would surpass logistic regression's and by combining the concept of randomly selecting a specified number of predictor variables for each bag like Random Forest, RGLM proved to the most robust model for both interpolation and extrapolation purposes.

5.4.1 Variable Importance of RGLM

Next, the variable importance measure of RGLM was investigated. In RGLM, a command used as a proxy for variable importance provides the variables that are repeatedly chosen in each bag for forward regression using the AIC criterion. Using this command, the variables that are chosen the most times across all bags are considered the most significant ones. Accordingly, the variable importance ranking displaying the variables and the number of times they were chosen in each bag is given in Figure 5.1.

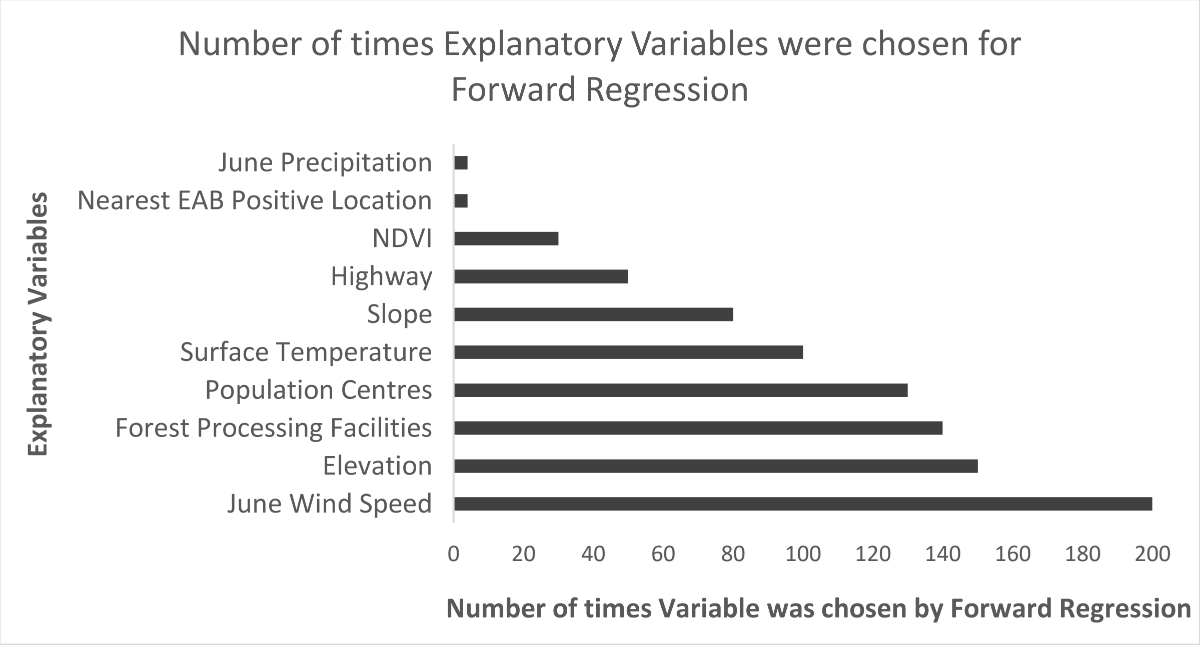


Figure 5.1. Most important variables identified by the RGLM model determined by the number of times the variable was chosen for forward regression in each bag using RGLM

According to Figure 5.1, the single most important variable selected in forward regression in all 200 bags was June wind speed. The rest of the variables were chosen by forward selection in 150 or lesser number of bags. A noteworthy observation is that the variable aspect was not chosen to be included in any bag as it was considered the least significant variable and both logistic regression and Random Forest support the same conclusion.

5.5 Comparison of the Three Model Performances

Before discussing the results of the three models, it is important to discuss various factors of the species and predictor data that could have affected model performance. For starters, the two main issues associated with the species data were sampling bias and a greater proportion of EAB absence data compared to the presence data. Inconsistencies in the collection of the species data misrepresents the true range of the species and therefore makes it challenging for distribution models to predict in a new area. By filtering the species data using a threshold and determining the optimal ratio of presence to absence points required to achieve the highest classification accuracies, the quality of the species data can be improved.

Secondly, inconsistencies between the scale and acquisition periods of the predictor variables and the species data undermines the authenticity of the models. For instance, due to the high costs associated with acquiring climactic predictor variables such as rainfall and wind speed, they are only available as a long-run average over several years at a coarse scale (i.e., 1 km^2). Having said that, if the scale of the species data is greater than the predictor variables, each cell of the predictor variables can contain both presence and absence points. In order to overcome this issue, eliminating the absence points that fell within 1 kilometer of the presence points ensured that each cell of the coarser scale predictor variables represented one sampled point. Notably, a previous trial without filtering the absence points achieved lower model accuracies (approximately 15% lower) for the validation dataset and slightly lower model accuracies for the prediction dataset. However, one of the drawbacks of filtering the species data in order to account for the scale of the predictor variables is that it caused some variables to achieve biased importance rankings. For instance, after filtering the absence points, the variable June wind speed was overwhelmingly ranked the most important variable by the three models which was not the case prior to filtering. Consequently, whether this relationship is entirely valid or as a result of the filtering is uncertain. That said, species distribution models built for large scale studies should carefully inspect the species data to ensure it is consistent with the scale of the predictor variables. With regards to the models, evaluating a species distribution model based on two aspects of transferability (i.e., interpolation and extrapolation) provides a means of “identifying relationships between the species and predictor variables that are truly general thereby reducing the risk of overfitting” (Wenger & Olden, 2012, p. 262). That said, a robust species distribution model is one that can predict both validation data with similar spatio-temporal characteristics and a novel dataset successfully with varying conditions. In terms of improving model performance, machine learning methods such as support vector machines (Drake et al., 2006), classification and regression trees, and Random Forest (Breiman, 2001) are known to match non-linear complex relationships in a dataset more efficiently than traditional generalized linear modelling (Wenger & Olden, 2012). As a result, both generalized linear modelling and machine learning methods were exercised in this research to test the transferability capabilities of the models.

With regards to fine-tuning the models, adjustments of model parameters were essential in achieving high classification accuracies. For instance, logistic regression’s backward stepwise variable selection method achieved the highest prediction accuracy compared to the forward stepwise variable selection methods as it used fewer variables to effectively predict EAB risk. For the machine learning models, although it is not certain how sensitive Random Forest is to the number of predictor variables randomly selected for each tree because the default value was utilized, the number of variables selected for forward regression in each bag for RGLM was quite sensitive to the classification accuracies. As a result, a sensitivity analysis was performed by increasing the number of predictor variables for each tree or bag and analyzing the validation and prediction accuracies in Figure 5.2.

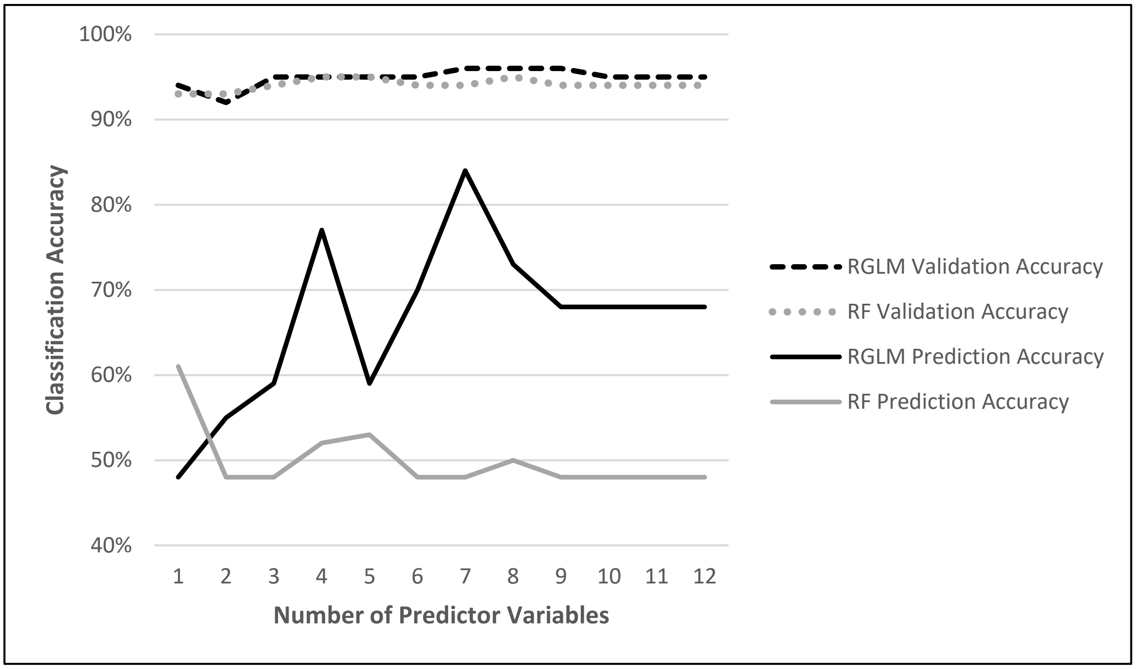


Figure 5.2. Sensitivity analysis of the classification accuracies vs. varying numbers of predictor variables for RGLM and Random Forest

As evident in Figure 5.2, when the number of predictor variables increased, the validation accuracies of RGLM and RF stayed relatively constant beyond an increase of three variables but the prediction accuracy changed drastically. RGLM was more sensitive to the number of predictor variables chosen to be selected for forward regression in each bag compared to RF, where the accuracy significantly dropped and plateaued after using just one variable. In general, as the number of predictor variables increased, the classification accuracy increased for RGLM with two distinct hikes in accuracy using four and seven variables. Moreover, the sensitivity analysis exemplified the superiority of RGLM over RF in predicting a novel dataset and an adjustment in the number of variables made a huge impact on the classification accuracy.

According to the classification accuracies, all models achieved high validation accuracies, stepwise backward logistic regression (93%), Random Forest (95%), and RGLM (95%), respectively. For the prediction dataset, RGLM outperformed the two models by achieving 84% whereas logistic regression achieved 70% and Random Forest performed the poorest (52%).

Nevertheless, since the classification accuracies were obtained by enforcing a threshold of 0.5 on the probabilities, the results could potentially be subjective. Furthermore, a more robust form of assessing model performance is by using ROC curves and corresponding AUC values which is independent of the probability threshold. The ROC curves are displayed in Figure 5.3 and the corresponding AUC values and standard deviations are summarized in Table 5.16.

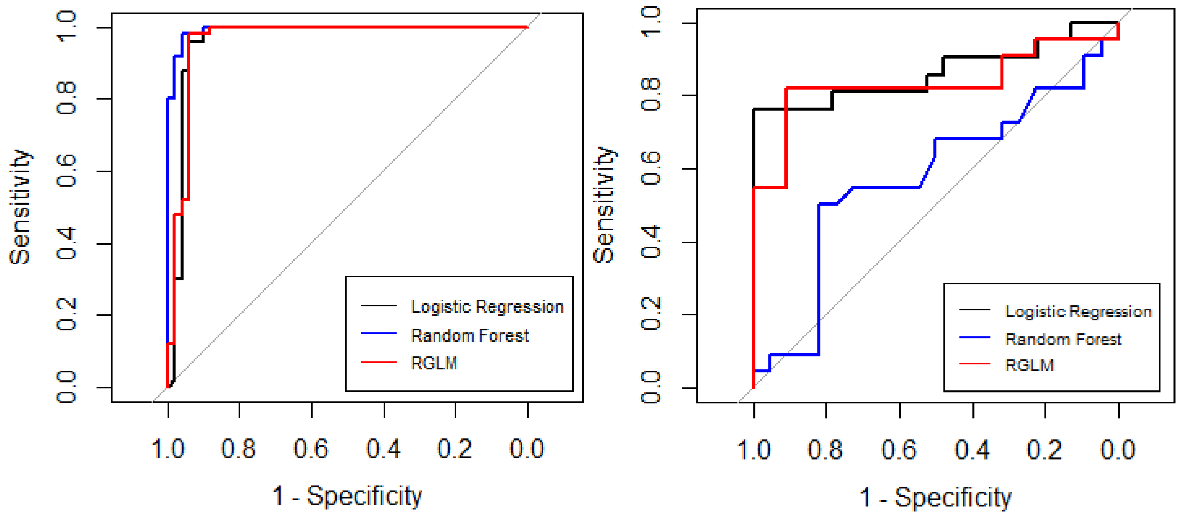


Figure 5.3. A comparison of the ROC curves of the three models (logistic regression=black line, random forest=blue line, RGLM=red line) for the validation dataset (left) and prediction dataset (right)

Table 5.16. Average AUC and standard deviation values for the three models (logistic regression, Random Forest and RGLM)

Model	Validation Dataset		Prediction Dataset	
	AUC	SD	AUC	SD
Logistic Regression	0.96	0.02	0.86	0.06
Random Forest	0.99	0.006	0.58	0.09
RGLM	0.96	0.02	0.83	0.07

From Table 5.16, it is evident that the AUC values of logistic regression, Random Forest, and RGLM are very close to 1 with Random Forest attaining an almost perfect AUC value for the validation dataset. On the other hand, for the prediction dataset, the AUC values of logistic regression (0.86) and RGLM (0.83) were similar compared to Random Forest (0.58). In the same way, the ROC curve of Random Forest in Figure 5.3 gravitates towards the diagonal where the true positive rate is equal to the false positive rate (AUC = 0.5) for all thresholds, suggesting that the predictive ability of the model is equivalent to random assignment (Freeman & Moisen, 2008). Overall, based on the results of the ROC curves and AUC values, RGLM and logistic regression had similar predictive powers compared to the classification accuracies which established RGLM as the most robust model. Next, the results of the risk maps were validated (Table 5.17) by calculating the percentage of EAB presence and absence points that fell under the five categories of risk.

Table 5.17. Risk map validation results of the EAB prediction dataset from 2013

	Logistic Regression		**Random Forest**		**RGLM**	
	Presence Points (%)	Absence Points (%)	Presence Points (%)	Absence Points (%)	Presence Points (%)	Absence Points (%)
Lowest Risk	23%	13%	59%	36%	23%	23%
Low Risk	41%	41%	41%	59%	40%	50%
Moderate Risk	4.5%	23%	0%	5%	23%	9%
High Risk	18%	13%	0%	0%	0%	18%
Highest Risk	13.5%	10%	0%	0%	14%	0%

According to Table 5.17, none of the models were able to successfully divide the EAB detected and non-detected ash trees into appropriate categories based on risk. However, by examining the two extreme risk levels (lowest risk and highest risk), some important insights can be made. For logistic regression, the second greatest proportion of presence points (23%) fell in the lowest risk category compared to a very low proportion of points in the highest risk category (13.5%). Similar percentages are obtained for RGLM with the exception of a greater proportion of points in the moderate risk category than the high-risk category compared to logistic regression. Conversely, Random Forest incorrectly predicted 59% of the EAB presence points in the lowest risk category and 0% presence points were predicted in the high or highest risk category suggesting substandard prediction abilities. The classification accuracies of the absence points were relatively high for all three models with a large proportion of absence points falling in the low and lowest risk categories.

Based on the results of Random Forest, research performed by Wenger & Olden (2012) on the occurrence of the brook and brown trout in the western United States concluded similar findings where Random Forest outperformed all models using a random cross-validation dataset but performed poorly on a dataset that varied spatially. On the other hand, a generalized linear mixed model (GLMM) displayed greater transferability than Random Forest. It is proposed by the authors that the GLMM performed better than Random Forest due to its inclusion of random effects and correlated errors in the dataset during model calibration (Wolfinger & O'connell, 1993). To be specific, in the GLMM, the variable temperature was represented by a quadratic relationship due to an *a priori* association with the species niche (Magnuson et al., 1979). Conversely, Random Forest models the response variable against the predictor variables empirically without making assumptions about the true form of the data (Wenger & Olden, 2012). The authors established this phenomenon by constructing a partial dependence plot of relative occurrence vs. response to temperature for Random Forest and compared it to a plot of the occurrence probability vs. temperature of the best-supported GLMM. The results showed that the response curve for Random Forest appeared very jagged whereas the response curve for GLMM was smoother. Furthermore, the results indicated that unlike GLMM, the response curve for Random Forest matched the training data so closely that it failed to transfer to new datasets.

In conclusion, the performance of the models in this research coincided with the findings by Wenger & Olden (2012). Although a standard logistic GLM was used in this research as opposed to a GLMM used in the latter, the fact remains that logistic regression also scrutinizes the selection of predictor variables into the model by offering various variable selection methods (i.e., forward, backward, forward step-wise, backward step-wise, etc.) in order to maximize the selection of only significant variables. By the same token, RGLM capitalizes on the forward variable selection method by incorporating the concept of bootstrapped samples and ranking the predictor variables prior to being selected for forward selection.

On the contrary, Random Forest does not provide the user with many model specifications that pertain to variable selection. As a result, although Random Forest is known to model complex relationships between a response variable and predictor variables without any *a priori* knowledge, it fails to detect conditional dependencies among the predictor variables in a new dataset if the training dataset does not exhibit such a relationship (Prajwala, 2015; Robnik-Šikonja, 2004). This limitation of Random Forest can be attributed to its variable importance measure, the Gini index as it measures the impurity before and after a split is made by the candidate variable at the node level. In this way, Random Forest assumes a conditional independence of the predictor variables as the Gini index evaluates each variable individually and does not take into consideration other contributary variables (Robnik-Šikonja, 2004) and is biased to the order of variables at each node of the tree (Hur et al., 2017). With this in mind, the likelihood ratio test and AIC criterion used in logistic regression and RGLM compares competing models with different combinations of predictor variables in each model and selects the best model while preserving conditional dependencies between the variables thereby providing greater a transferability than Random Forest.

In summary, when comparing the three models, they all performed well on the validation dataset but relatively poorly on the prediction set from 2013. A logical explanation would be that the environmental conditions had changed from 2012 in 2013 thereby affecting the explanatory variables, especially the values of the environmental variables. This change was assessed by comparing the average values of the explanatory variables of the EAB presence and absence points in 2006-2012 to the EAB points in 2013 using bar plots in Figures 5.4 and 5.5.

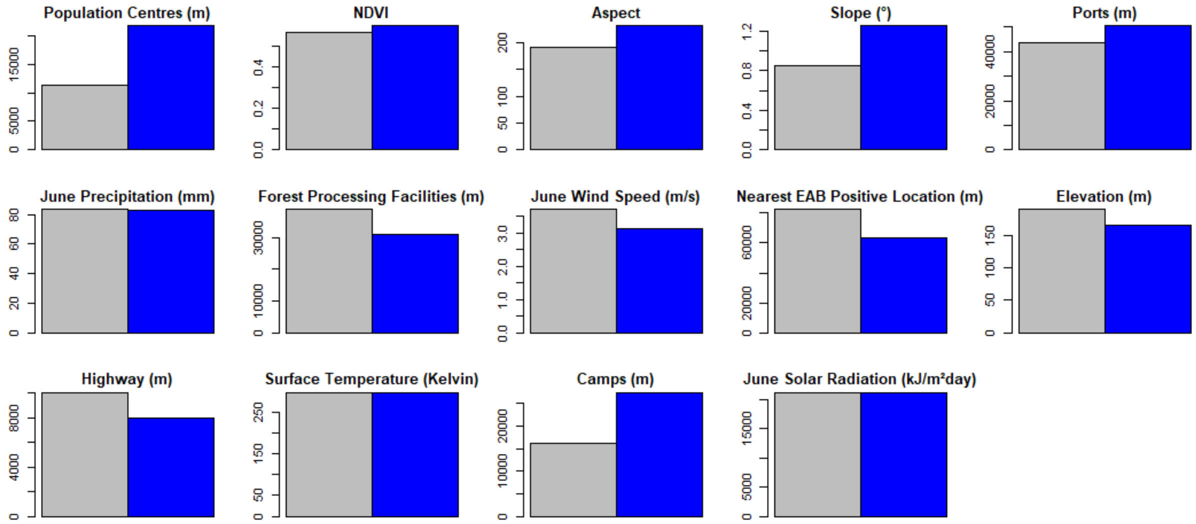


Figure 5.4. Bar plot outlining differences between the average values of the explanatory variables for EAB presence points from 2006-2012 (grey) vs. EAB presence points from 2013 (blue)

According to Figure 5.4, the average values of the climactic variables (i.e., June precipitation, surface temperature and solar radiation) for the 2006-2012 and 2013 EAB presence points were similar for both time periods suggesting that these variables did not affect the transferability of the models significantly. The average values of June wind speed, which was identified as the most important variable by all three models decreased slightly for the 2013 presence points. However, the mean values of the anthropogenic variables such as the distance from population centres, ports, forest processing facilities, highways and camps in 2013 differed substantially from the group of EAB presence points from 2006-2012. For instance, the EAB presence points were found significantly farther away from population centres and camps in 2013 than the points from 2006-2012. Although this finding reflects stricter regulations enforced by Canadian officials in these areas, the inconsistencies in the average values of the anthropogenic variables between the two time periods could explain the lower prediction accuracies of the models.

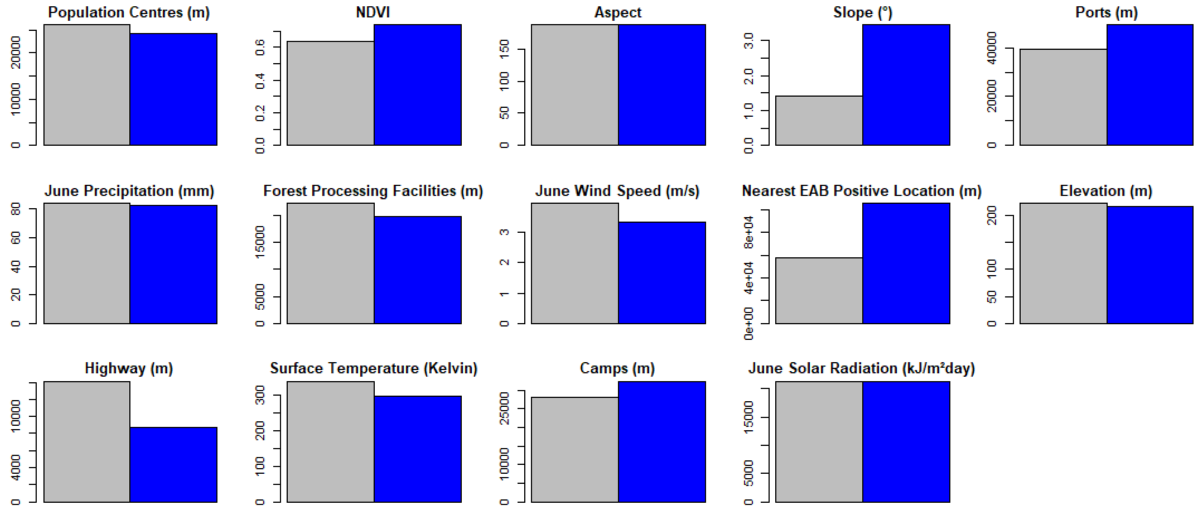


Figure 5.5. Bar plot outlining differences in the average values of the explanatory variables for EAB absence points from 2006-2012 (grey) vs. EAB abesence points (blue) from 2013

Similar to the EAB presence points, the average values of the climactic variables such as June precipitation and solar radiation in Figure 5.5 did not differ for the EAB absence points for both time periods (i.e., 2006-2012 and 2013). However, for June wind speed and surface temperature, the absence points from 2013 exhibited slight decreases. As for the anthropogenic variables, the three variables that differed the most between 2006-2012 and 2013 were distance from ports, highways and the nearest EAB positive location. The variable “distance from the nearest EAB positive location” showed some unexpected results as the EAB absence points appeared to be closer to the presence points during 2006-2012 compared to 2013 whereas this trend was the opposite for the EAB presence points. Nonetheless, the bar plots summarized the lack of consistency between the values of the predictor variables for the 2006-2012 and 2013 datasets which could have contributed to the lower prediction accuracies of the models.

CHAPTER 6: CONCLUSIONS AND REMARKS

In conclusion, this research established a framework to model the spread of the EAB in Southern Ontario by designing methods to improve the quality of the species and predictor data in order to achieve high model accuracies. Since the EAB outbreak in Ontario, early detection of the beetle in uninhabited areas is paramount to slowing its spread. That said, in order to successfully build a predictive model to identify potential at-risk areas, high quality historic species data is required. However, there were a few discrepancies with the collection of the sampled data in this research that posed challenges for the modelling process which should be discussed in detail. For instance, during the surveying process, a majority of the data acquisition was performed through visual survey during which the trees were briefly examined at the ground or canopy level. As a result, the EAB would be virtually undetectable at low densities if the trees did not exhibit visual signs of infestation. Although visual surveying was the preferred method over other labour and cost-intensive sampling methods such as branch sampling, the possibility of mislabelling a tree as “absence” introduced significant omission errors in the detection of the EAB, thereby reducing the quality of the data.

Secondly, although the sampled data provided by the CFIA provided an excellent temporal coverage of the EAB from the years 2002-2013 in Ontario, this meant the data exhibited high spatial heterogeneity which was evident in the descriptive statistics. For instance, the descriptive statistics in Tables 3.2 and 3.3 indicated the anthropogenic variables had high standard deviations for the sampled points. A solution to address the issue of varying spatial scales on the distribution of a species is to create models with hierarchical structures where the predictors are separated into sub-models (Mackey & Lindenmayer, 2001). However, the efficiency of such models has not been tested thoroughly to be used as a reliable method to deal with different scales (Elith & Leathwick, 2009). In order to overcome the scale issue pertaining to the predictor variables, a solution proposed in this research was to eliminate all points that fell within a distance that is equivalent to the scale of the coarsest predictor variable. This way, no two sampled EAB presence and absence points shared the same cell of a predictor variable. With this in mind, a future research objective is to implement a generalized linear mixed model (GLMM) that can include random variation within the two classes (i.e., EAB presence and absence) even when there is inconsistency between the scales of the predictor variables.

Aside from addressing the scale issue, the quality of the predictor data was also assessed such as testing for multicollinearity using a correlation matrix and VIF values. Although the effect of multicollinearity in generalized linear models is clear, such that it inflates the regression coefficients of the predictor variables, its effect on machine learning models lacks a proper understanding. In this research, the effects of excluding variables that exhibited high VIF values was assessed using Random Forest where a full model containing all the variables was compared to reduced models excluding the multicollinear variables and a model containing variables retained by the RFE method. Overall, it was concluded that the exclusion of variables that exhibited high VIF values from the models resulted in greater classification accuracies than the RFE method which eliminated too many variables.

A prominent issue with the species data was high levels of clustering as indicated by the nearest neighbour index which was problematic from a modelling perspective. As a result, the effects of clustering were tackled by filtering the species clusters to a single point based on a distance threshold. Using a distance threshold of 100 meters for the EAB presence points and 150 metres for the EAB absence points, the classification accuracies achieved on the validation dataset by logistic regression, Random Forest, and RGLM were 93%, 95%, and 95%, respectively. Another method of pre-processing the EAB dataset to maximize model performance was determining the most appropriate ratio of EAB presence to absence points. The optimal ratio was determined to be 1:1 whereas using greater number of absence points to presence points deteriorated the models' performances.

Although all the models performed well on the validation dataset, the performances of logistic regression and Random Forest were significantly lower for the prediction dataset: logistic regression (70%) and Random Forest (52%). On the other hand, RGLM outperformed both models by achieving 84% for the prediction accuracy. It is a well known historical issue with species distribution models that are tested for their transferability, especially for extrapolation purposes (Elith & Leathwick, 2009) to exhibit a lower performance due to an assumption of equilibrium conditions of the predictor variables in the predicting region. For this reason, variability among the predictor variables due to natural phenomenon such as climate change in the predicting region cannot be reflected by the species points used to train the model. Differences among the predictor variables between the EAB presence and absence points in the sampling region (2006-2012) vs. the prediction region (2013) was assessed using bar plots for the two spatio-temporal periods to substantiate any differences. The findings concluded that various predictor variables, particularly the anthropogenic variables exhibited large variations between the two sets of points which sheds some light on the poor prediction capabilities of the models as a result of conflicting values of the predictor variables. Above all, even with spatio-temporal variations in the prediction dataset, RGLM performed the best due to its sophisticated variable selection methods over Random Forest and logistic regression such as administering two rounds of variable ranking prior to selection into the models. A drawback of RGLM over Random Forest is the longer computation time.

Aside from the classification accuracies of the three modeling methods, their variable importance rankings were consistent. According to the results, logistic regression, Random Forest, and RGLM identified June wind speed as the most significant variable. The other variables had significantly lower rankings which raised the question of data separation. When a contingency table was created for June wind speed, it was determined that the variable contributed to a quasi-separation of the two classes which could explain its superiority over other variables. From a climactic perspective, it suggests that the EAB do not require high wind speeds to travel long distances and that other transport mechanisms, most likely anthropogenic vectors are responsible. Among the anthropogenic variables, population centres and forest processing facilities were ranked amongst the top variables by the machine learning models suggesting that the influence of humans on the spread of the EAB is profound.

A variable that surprisingly was not identified as one of the most important variables as suggested by literature (Huset, 2013; Prasad et al., 2010) was the distance to the nearest EAB positive location. This variable was overwhelmingly the most important variable in the logistic regression and Maxent models used by Huset (2013). However, although its negative effect on EAB risk was consistent with Huset (2013), such that as the distance from an EAB positive location increased, the EAB risk decreased, it was among the lower ranked variables in the models. This relationship is also authenticated by the descriptive statistics where the average distance from a presence point to a nearby presence points was shorter than from the absence points which suggests that the EAB thrive in areas that were previously infested.

The last objective of this research was to create an automated risk map tool which streamlined the visualization of risk probabilities outputted by the machine learning models. The advantage of the tool is that it can be used on any spatial scale to calculate the risk of a species. That said, the disadvantage is that it is quite computationally intensive as it requires the collection of an adequate amount of sample points pertaining to the desired resolution of the risk map (i.e., the higher the resolution of the risk map, the more sample points are required). Following the acquisition of the sample points, the rest of the steps such as obtaining the probabilities for each sampled point does not require much computer time. Nonetheless, the creation of the risk map tool using the ArcGIS-R bridge for the Random Forest and RGLM models allowed the opportunity to compare all the models in the same GIS platform. An extension of this tool is to develop a mobile and desktop application on the early detection of the EAB in Canada. The application is currently being developed by Esri Canada in conjunction with our research team at the Earth Observation Laboratory of York University. The scope of the application is to allow users to visualize previously EAB-infested trees, enter locations of newly EAB-infested trees and visualize the risk of the area based on previously detected EAB positive locations and various explanatory layers. The application holds a lot of promise but the true method of eradication of the EAB consists of a multi-integrated plan consisting of stricter regulations on all pathways of exposure.

Ultimately, a comprehensive analysis was conducted on the preparation of the EAB dataset in Southern Ontario in order to model its dispersal patterns using various distribution models. From the classification results, it is abundantly clear that strictly machine learning methods cannot be used to predict the emergence of the EAB but rather a combination of statistical modelling with some aspects of randomness such as a random generalized linear model should be used. Using the methods and results from this research as a stepping stone, a similar procedure can be used by conservation authorities to visualize the dispersal of the EAB in any region. For the most part, complete eradication of the EAB is not in the foreseeable future, but its spread can certainly be impeded with the right vision and tools.

BIBLIOGRAPHY

Akinwande, M. O., Dikko, H. G., & Samson, A. (2015). Variance Inflation Factor: As a Condition for the Inclusion of Suppressor Variable(s) in Regression Analysis. *Open Journal of Statistics*, *5*(December), 754–767.

Appleton, E., Kimoto, T., Holmes, J., & Turgeon, J. J. (2017). Surveillance Guidelines for Emerald Ash Borer. *Canadian Food Inspection Agency*, 1–12.

Archer, K. J., & Kimes, R. V. (2008). Empirical characterization of random forest variable importance measures. *Computational Statistics and Data Analysis*, *52*(4), 2249–2260. https://doi.org/10.1016/j.csda.2007.08.015

Austin, P. C., & Tu, J. V. (2004). Automated variable selection methods for logistic regression produced unstable models for predicting acute myocardial infarction mortality. *Journal of Clinical Epidemiology*, *57*(11), 1138–1146. https://doi.org/10.1016/j.jclinepi.2004.04.003

Barbet-Massin, M., Jiguet, F., Albert, C. H., & Thuiller, W. (2012). Selecting pseudo-absences for species distribution models: How, where and how many? *Methods in Ecology and Evolution*, *3*(2), 327–338. https://doi.org/10.1111/j.2041-210X.2011.00172.x

Bebber, D. (1999). Spatial autocorrelation. *Trends in Ecology & Evolution*, *14*(5), 196.

BenDor, T. K., Metcalf, S. S., Fontenot, L. E., Sangunett, B., & Hannon, B. (2006). Modeling the spread of the Emerald Ash Borer. *Ecological Modelling*, *197*(1–2), 221–236. https://doi.org/10.1016/j.ecolmodel.2006.03.003

Bento, A. P., Gaulton, A., Hersey, A., Bellis, L. J., Chambers, J., Davies, M., … Hatko, S. (2013). Classification and Regression by randomForest. *Nucleic Acids Research*, *5*(1), 983–999. https://doi.org/10.1023/A:1010933404324

Boria, R. A., Olson, L. E., Goodman, S. M., & Anderson, R. P. (2014). Spatial filtering to reduce sampling bias can improve the performance of ecological niche models. *Ecological Modelling*, *275*(March), 73–77. https://doi.org/10.1016/j.ecolmodel.2013.12.012

Braunisch, V., Coppes, J., Arlettaz, R., Suchant, R., Schmid, H., & Bollmann, K. (2013). Selecting from correlated climate variables: A major source of uncertainty for predicting species distributions under climate change. *Ecography*, *36*(9), 971–983. https://doi.org/10.1111/j.1600-0587.2013.00138.x

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140. https://doi.org/10.1007/BF00058655

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brown, J. L. (2014). SDMtoolbox: a python-based GIS toolkit for landscape genetic, biogeographic and species distribution model analyses. *Methods in Ecology and Evolution*, *5*(7), 694–700. https://doi.org/10.1111/2041-210X.12200

Bursac, Z., Gauss, C. H., Williams, D. K., & Hosmer, D. W. (2008). Purposeful selection of variables in logistic regression. *Source Code for Biology and Medicine*, *3*. https://doi.org/10.1186/1751-0473-3-17

Courville, T., & Thompson, B. (2015). Use of Structure Coefficients in Published Multiple

Regression Articles: B Is Not Enough, (April 2001). https://doi.org/10.1177/0013164401612006

Cushman, S. A., & Huettmann, F. (2010). *Spatial complexity, informatics, and wildlife conservation*. *Spatial Complexity, Informatics, and Wildlife Conservation*. https://doi.org/10.1007/978-4-431-87771-4

Darst, B. F., Malecki, K. C., & Engelman, C. D. (2018). Using recursive feature elimination in random forest to account for correlated variables in high dimensional data. *BMC Genetics*, *19*. https://doi.org/10.1186/s12863-018-0633-8

de Groot, P., Biggs, W. D., Lyons, D. B., Scarr, T., Czerwinski, E., Evans, H. J., … Marchant, K. (2006). A Visual Guide to Detecting Emerald Ash Borer Damage. *Canadian Forest Service*, 1–20. Retrieved from http://cfs.nrcan.gc.ca/pubwarehouse/pdfs/26856.pdf

Dix, M. E., Buford, M., Slavicek, J., Solomon, A. M., & Conard, S. G. (2010). Invasive Species and Disturbances: Current and future roles of Forest Service research and development. *A Dynamic Invasive Species Research Vision: Opportunities and Priorities 2009-2029*, 91–102.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., … Lautenbach, S. (2013). Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, *36*(1), 027–046. https://doi.org/10.1111/j.1600-0587.2012.07348.x

Drake, J. M., Randin, C., & Guisan, A. (2006). Modelling ecological niches with support vector machines. *Journal of Applied Ecology*, *43*(3), 424–432. Retrieved from http://www.blackwell-synergy.com/doi/abs/10.1111/j.1365-2664.2006.01141.x

Elith, J., & Leathwick, J. R. (2009). Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. *Annual Review of Ecology Evolution and Systematics*, *40*, 677–697. https://doi.org/10.1146/annurev.ecolsys.110308.120159

Emerald Ash Borer. (2012). *Invasive Species Centre*.

Emerald Ash Borer Management Plan update. (2013). *City of Peterborough*, 8. Retrieved from https://www.markham.ca/wps/wcm/connect/markhampublic/fd25855e-ead9-415e-a232-2325f884f608/emerald_ash_borer_management_plan_update_report.pdf?MOD=AJPERES&CACHEID=fd25855e-ead9-415e-a232-2325f884f608

Evans, J. D. (1996). Straightforward Statistics for the Behavioral Sciences. *Pearson's Correlation*, 122.

F. Dormann, C., M. McPherson, J., B. Araújo, M., Bivand, R., Bolliger, J., Carl, G., … Wilson, R. (2007). Methods to account for spatial autocorrelation in the analysis of species distributional data: A review. *Ecography*, *30*(5), 609–628. https://doi.org/10.1111/j.2007.0906-7590.05171.x

Fairmaire, A., & Parsons, G. L. (2008). Emerald Ash Borer. *America*, (November).

Firth, D. (1993). Bias Reduction of Maximum Likelihood. *Biometrika*. https://doi.org/10.1093/biomet/80.1.27

Fisher, J. B., Trulio, L. A., Biging, G. S., & Chromczak, D. (2007). An analysis of spatial clustering and implications for wildlife management: A burrowing owl example. *Environmental Management*, *39*(3), 403–411. https://doi.org/10.1007/s00267-006-0019-y

Freeman, E. A., & Moisen, G. G. (2008). A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa. *Ecological Modelling*, *217*(1–2), 48–58. https://doi.org/10.1016/j.ecolmodel.2008.05.015

Gaetz, N., & Hildebrand, T. (2012). Recommended Approach for the Management of Emerald Ash Borer. *TRCA Forest Health Working Group*, (July).

Gregorutti, B., Michel, B., & Saint-Pierre, P. (2017). Correlation and variable importance in random forests. *Statistics and Computing*, *27*(3), 659–678. https://doi.org/10.1007/s11222-016-9646-1

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, *46*(1–3), 389–422. https://doi.org/10.1023/A:1012487302797

Hanou, I. (2011). 2010 Town of Oakville Hyperspectral EAB Analysis. *AMEC Earth & Environmental, Inc.*, 1–20.

Haque, M., Rahman, A., Hagare, D., & Chowdhury, R. (2018). A Comparative Assessment of Variable Selection Methods in Urban Water Demand Forecasting. *Water*, *10*(4), 419. https://doi.org/10.3390/w10040419

Hegyi, G., & Laczi, M. (2015). Using Full Models, Stepwise Regression and Model Selection in Ecological Data Sets: Monte Carlo Simulations. *Annales Zoologici Fennici*, *52*, 257–279. https://doi.org/https://doi.org/10.5735/086.052.0502

Hijmans, R. J., & Elith, J. (2013). Species distribution modeling with R Introduction. *October*, 71. https://doi.org/10.1016/S0550-3213(02)00216-X

Hosmer, D. W., & Lemeshow, S. (2000). Applied logistic regression. *Wiley Series in Probability and Statistics*, 373. https://doi.org/10.1198/tech.2002.s650

Hulley, G., Hook, S. (2017). MOD21A2 MODIS/Terra Land Surface Temperature/3-Band Emissivity 8-Day L3 Global 1km SIN Grid V006 [Data set]. NASA EOSDIS Land Processes DAAC. doi: 10.5067/MODIS/MOD21A2.006

Hur, J. H., Ihm, S. Y., & Park, Y. H. (2017). A variable impacts measurement in random forest for mobile cloud computing. *Wireless Communications and Mobile Computing*, *2017*. https://doi.org/10.1155/2017/6817627

Huset, R. (2013). A GIS-based Analysis of the Environmental Predictors of Dispersal of the Emerald Ash Borer in New York, 120. Retrieved from http://surface.syr.edu/geo_thesis/3/

Iwundu, M. P., & Efezino, O. P. (2015). On the Adequacy of Variable Selection Techniques on Model Building - ProQuest. *Asian Journal of Mathematics and Statistics*, *8*(1), 19–34. https://doi.org/10.3923/ajms.2015.19.34

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). Statistical Learning. In *An Introduction to Statistical Learning with Applications in R* (pp. 15–57). https://doi.org/10.1016/j.peva.2007.06.006

Kgosiesele, E. (2010). Predictive Distribution Modelling of Timon lepida in Spain.

Knight, K. S., Brown, J. P., & Long, R. P. (2013). Factors affecting the survival of ash (Fraxinus spp.) trees infested by emerald ash borer (Agrilus planipennis). *Biological Invasions*, *15*(2), 371–383. https://doi.org/10.1007/s10530-012-0292-z

Koedel, C., & Betts, J. R. (2010). Value-added to what? How a ceiling in the testing instrument influences value-added estimation. *Education Finance and Policy*, *5*, 54–81.

Kramer-Schadt, S., Niedballa, J., Pilgrim, J. D., Schröder, B., Lindenborn, J., Reinfelder, V., …

Wilting, A. (2013). The importance of correcting for sampling bias in MaxEnt species distribution models. *Diversity and Distributions*, *19*(11), 1366–1379. https://doi.org/10.1111/ddi.12096

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., … Hunt, T. (2018). Package “caret,” 215. Retrieved from https://cran.r-project.org/web/packages/caret/caret.pdf

Landsat 4-7 Surface Reflectance (LEDAPS) Product Guide. (2018). *U.S. Geological Survey (USGS)*, (March), 1–38.

Lee, Y. K., Lee, E. R., & Park, B. U. (2012). Principal Component Analysis in very High-Dimensional spaces, *22*, 933–956.

Lobo, J. M., Jiménez-valverde, A., & Real, R. (2008). AUC: A misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*. https://doi.org/10.1111/j.1466-8238.2007.00358.x

Louppe, G. (2014). Understanding Random Forests. *Cornell University Library*, (July), 1–225. https://doi.org/10.13140/2.1.1570.5928

Louppe, G., Wehenkel, L., Sutera, A., & Geurts, P. (2013). Understanding variable importances in forests of randomized trees. *Advances in Neural Information Processing Systems 26*, 431–439. https://doi.org/NIPS2013_4928

Mac Nally, R. (2000). Regression and model-building in conservation biology, biogeography and ecology: The distinction between - and reconciliation of - “predictive” and “explanatory” models. *Biodiversity and Conservation*, *9*(5), 655–671. https://doi.org/10.1023/A:1008985925162

Mackey, B. G., & Lindenmayer, D. B. (2001). Towards a Hierarchical Framework for Modelling the Spatial Distribution of Animals. *Biogeography*, *28*(9), 1147–1166.

Magnuson, J. J., Crowder, L. B., & Medvick, P. A. (1979). Temperature as an ecological resource. *Integrative and Comparative Biology*, *19*(1), 331–343. https://doi.org/10.1093/icb/19.1.331

Marchant, K. R. (2012). Emerald Ash Borer Management Plan. *City of Missisauga*, 1–74.

Matsuki, K., Kuperman, V., & Van Dyke, J. A. (2016). The Random Forests statistical technique: An examination of its value for the study of reading. *Scientific Studies of Reading*. https://doi.org/10.1080/10888438.2015.1107073

McCullagh, P., & Nelder, J. A. (1989). *Generalized Linear Models, Second Edition*. *Generalized Linear Models Second Edition*. https://doi.org/10.1007/978-1-4899-3242-6

McCullough, D. G., Poland, T. M., & Cappaert, D. L. (2009). Attraction of the emerald ash borer to ash trees stressed by girdling, herbicide treatment, or wounding. *Canadian Journal of Forest Research*, *39*(7), 1331–1345. https://doi.org/10.1139/X09-057

Midi, H., Sarkar, S. K., & Rana, S. (2010). Collinearity diagnostics of binary logistic regression model. *Journal of Interdisciplinary Mathematics*, *13*(3), 253–267. https://doi.org/10.1080/09720502.2010.10700699

Mousavi, S. Z., Kavian, A., Soleimani, K., Mousavi, S. R., & Shirzadi, A. (2011). GIS-based spatial prediction of landslide susceptibility using logistic regression model. *Geomatics, Natural Hazards and Risk*, *2*(1), 33–50. https://doi.org/10.1080/19475705.2010.532975

NOAA Biogeography Branch. (2013). Sampling Design Tool for ArcGIS. Retrieved from http://www2.coastalscience.noaa.gov/publications/detail.aspx?resource=mgAPjEXoVKE

OwI+9IXOjBThd9nQ55Og19XX2UqnBEdY=

Northern Research Station. (2016). Modeling Spread of Emerald Ash Borer. Retrieved from https://www.nrs.fs.fed.us/disturbance/invasive_species/eab/risk_detection_spread/modeling_spread/

Ohlmacher, G. C., & Davis, J. C. (2003). Using multiple logistic regression and GIS technology to predict landslide hazard in northeast Kansas, USA. *Engineering Geology*, *69*(3–4), 331–343. https://doi.org/10.1016/S0013-7952(03)00069-3

P. Vatcheva, K., & Lee, M. (2016). Multicollinearity in Regression Analyses Conducted in Epidemiologic Studies. *Epidemiology: Open Access*, *06*(02). https://doi.org/10.4172/2161-1165.1000227

Peng, C. J., Lee, K. L., & Ingersoll, G. M. (2002). An Introduction to Logistic Regression Analysis and Reporting. *The Journal of Educational Research*, *96*(1), 3–14. https://doi.org/10.1080/00220670209598786

Phillips, S. J., & Dudík, M. (2008). Modeling of species distributions with Maxent: New extensions and a comprehensive evaluation. *Ecography*, *31*(2), 161–175. https://doi.org/10.1111/j.0906-7590.2008.5203.x

Pontius, J., Martin, M., Plourde, L., & Hallett, R. (2008). Ash decline assessment in emerald ash borer-infested regions: A test of tree-level, hyperspectral technologies. *Remote Sensing of Environment*, *112*(5), 2665–2676. https://doi.org/10.1016/j.rse.2007.12.011

Prajwala, T. R. (2015). A Comparative Study on Decision Tree and Random Forest Using R Tool. *Ijarcce*, *4*(1), 196–199. https://doi.org/10.17148/IJARCCE.2015.4142

Prasad, A. M., Iverson, L. R., Peters, M. P., Bossenbroek, J. M., Matthews, S. N., Sydnor, T. D., & Schwartz, M. W. (2010). Modeling the invasive emerald ash borer risk of spread using a spatially explicit cellular model. *Landscape Ecology*, *25*(3), 353–369. https://doi.org/10.1007/s10980-009-9434-9

Prinzie, A., & Van den Poel, D. (2008). Random Forests for multiclass classification: Random MultiNomial Logit. *Expert Systems with Applications*, *34*(3), 1721–1732. https://doi.org/10.1016/j.eswa.2007.01.029

Provincial Digital Elevation Model Technical Specifications, v3.0. (2013). *Ontario Ministry of Natural Resources*, 1–23.

Robnik-Šikonja, M. (2004). Improving Random Forests, 359–370. https://doi.org/10.1007/978-3-540-30115-8_34

Rodriguez, G. (2007). Logit Models for Binary Data. *Princeton University*. Retrieved from http://data.princeton.edu/wws509/notes/c3.pdf

Royo, A. A., & Knight, K. S. (2012). White ash (Fraxinus americana) decline and mortality: The role of site nutrition and stress history. *Forest Ecology and Management*, *286*, 8–15. https://doi.org/10.1016/j.foreco.2012.08.049

Sainani, K. L. (2013). Multivariate Regression: The Pitfalls of Automated Variable Selection. *PM and R*, *5*(9), 791–794. https://doi.org/10.1016/j.pmrj.2013.07.007

Sarkar, S. K., Midi, H., & Rana, S. (2010). Model Selection in Logistic Regression and Performance of its Predictive Ability. *Computational Statistics*, *4*(12), 5813–5822.

Settur, B., Rajan, K. S., & Ramachandra, T. V. (2013). Land Surface Temperature Responses to Land Use Land Cover Dynamics. *Geoinfor Geostat An Overview*, *1*(4). https://doi.org/10.4172/2327-4581.1000112

Shtatland, E. S., Cain, E., & Barton, M. B. (2001). The perils of stepwise logistic regression and how to escape them using information criteria and the output delivery system. *Proceedings from the 26th Annual SAS Users Group International Conference*, 222–226. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.122.6652&amp;rep=rep1&amp;type=pdf

Song, L., Langfelder, P., & Horvath, S. (2013). Random generalized linear model: a highly accurate and interpretable ensemble predictor. *BMC Bioinformatics*, *14*, 5. https://doi.org/10.1186/1471-2105-14-5

Statistics Canada. (2011). *Geography Catalogue, Census Year 2011*. *Statistics*.

Stohlgren, T. J., Ma, P., Kumar, S., Rocca, M., Morisette, J. T., Jarnevich, C. S., & Benson, N. (2010). Ensemble habitat mapping of invasive plant species. *Risk Analysis*, *30*(2), 224–235. https://doi.org/10.1111/j.1539-6924.2009.01343.x

Strobl, C., Boulesteix, A. L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, *9*, 1–11. https://doi.org/10.1186/1471-2105-9-307

Syfert, M. M., Smith, M. J., & Coomes, D. A. (2013). The Effects of Sampling Bias and Model Complexity on the Predictive Performance of MaxEnt Species Distribution Models. *PLoS ONE*, *8*(2). https://doi.org/10.1371/journal.pone.0055158

Toloşi, L., & Lengauer, T. (2011). Classification with correlated features: Unreliability of feature ranking and solutions. *Bioinformatics*, *27*(14), 1986–1994. https://doi.org/10.1093/bioinformatics/btr300

User Guide for ORN Segment with Address. (2016). *Ministry of Natural Resources and Forestry*, 60. Retrieved from https://www.javacoeapp.lrc.gov.on.ca/geonetwork/srv/en/main.home?uuid=c7c7202d-942d-47dc-bb15-259eb71f2551

Václavík, T., Kupfer, J. A., & Meentemeyer, R. K. (2012). Accounting for multi-scale spatial autocorrelation improves performance of invasive species distribution modelling (iSDM). *Journal of Biogeography*, *39*(1), 42–55. https://doi.org/10.1111/j.1365-2699.2011.02589.x

Veloz, S. D. (2009). Spatially autocorrelated sampling falsely inflates measures of accuracy for presence-only niche models. *Journal of Biogeography*, *36*(12), 2290–2299. https://doi.org/10.1111/j.1365-2699.2009.02174.x

Vlassova, L., Perez-Cabello, F., Nieto, H., Martín, P., Riaño, D., & de la Riva, J. (2014). Assessment of Methods for Land Surface Temperature Retrieval from Landsat-5 TM Images Applicable to Multiscale Tree-Grass Ecosystem Modeling, 4345–4368. https://doi.org/10.3390/rs6054345

Wang, D., Zhang, W., & Bakhai, A. (2004). Comparison of Bayesian model averaging and stepwise methods for model selection in logistic regression. *Statistics in Medicine*, *23*(22), 3451–3467. https://doi.org/10.1002/sim.1930

Wenger, S. J., & Olden, J. D. (2012). Assessing transferability of ecological models: An underappreciated aspect of statistical validation. *Methods in Ecology and Evolution*, *3*(2), 260–267. https://doi.org/10.1111/j.2041-210X.2011.00170.x

Werkowska, W., Márquez, A. L., Real, R., & Acevedo, P. (2017). A practical overview of transferability in species distribution modeling. *Environmental Reviews*, *25*(1), 127–133. https://doi.org/10.1139/er-2016-0045

Whittingham, M. J., Stephens, P. A., Bradbury, R. B., & Freckleton, R. P. (2006). Why do we still use stepwise modelling in ecology and behaviour? *Journal of Animal Ecology*, *75*(5), 1182–1189. https://doi.org/10.1111/j.1365-2656.2006.01141.x

Wille, L. T. (2004). The Challenges of Clustering High Dimensional Data. In New Directions in Statistical Physics: Econophysics, Bioinformatics, and Pattern Recognition. Berlin, Heidelberg: Springer Berlin Heidelberg

Williams, R., & Dame, N. (2018). Analyzing Rare Events with Logistic Regression, 1–5.

Wolfinger, R., & O'connell, M. (1993). Generalized linear mixed models a pseudo-likelihood approach. *Journal of Statistical Computation and Simulation*, *48*(3–4), 233–243. https://doi.org/10.1080/00949659308811554

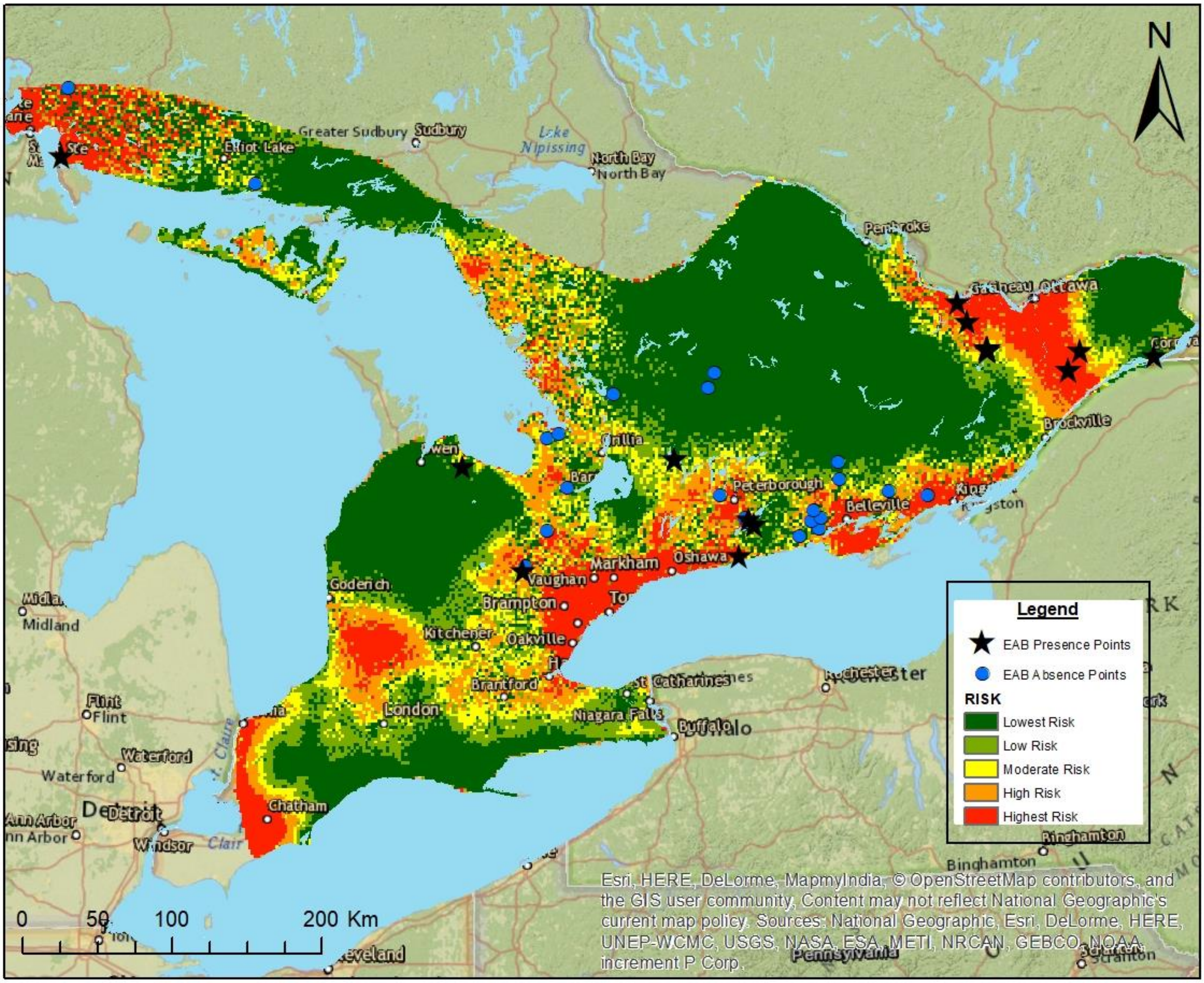
Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez and Markus Müller (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics,* 12, p. 77. <doi:10.1186/1471-2105-12-77> http://www.biomedcentral.com/1471-2105/12/77/

XLSTAT. (2017). *Addinsoft*, 1–1446.

Yu, C. H. (2000). An Overview of Remedial Tools for Collinearity in SAS.

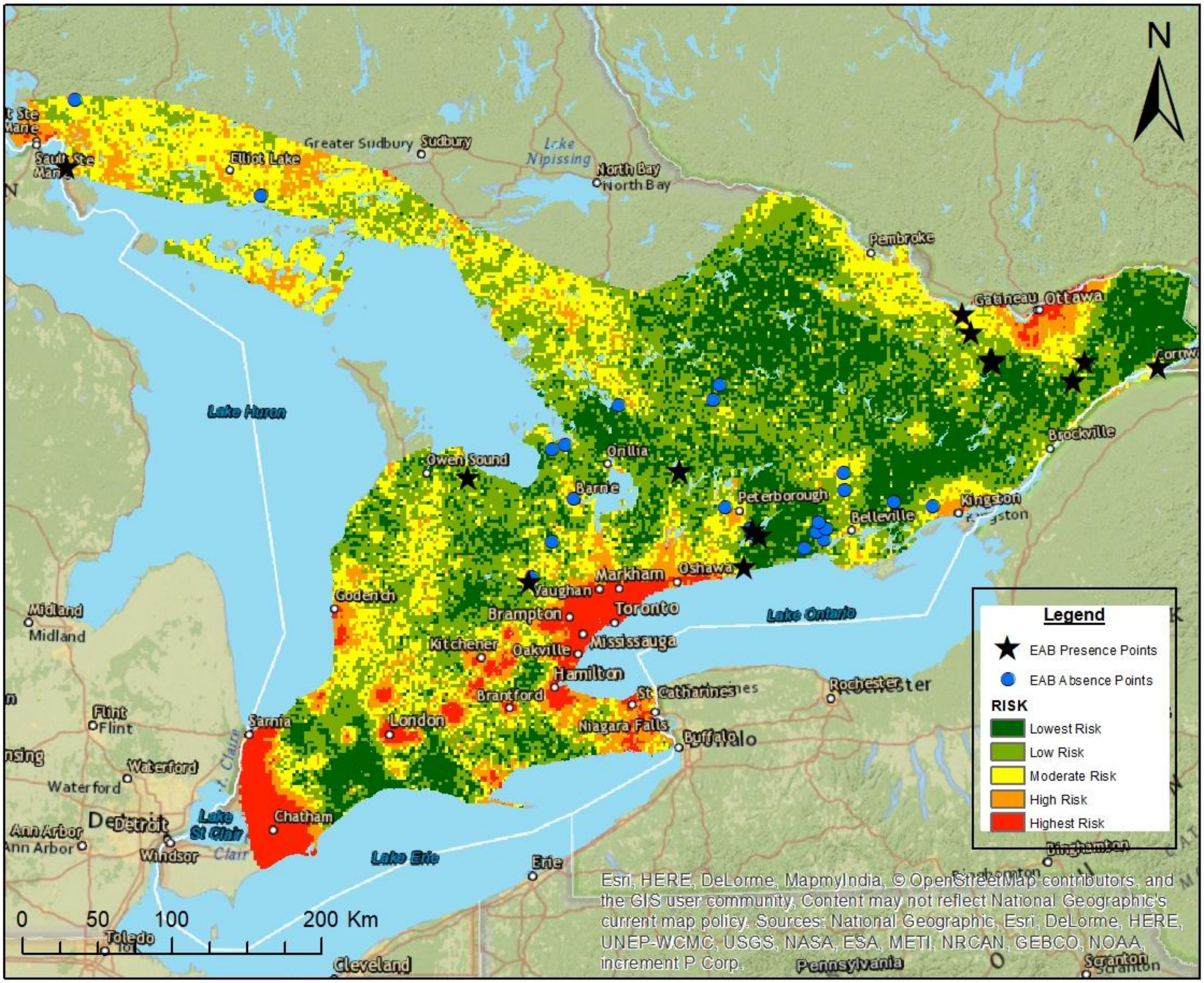
APPENDIX

Logistic Regression EAB Risk Map (2013)



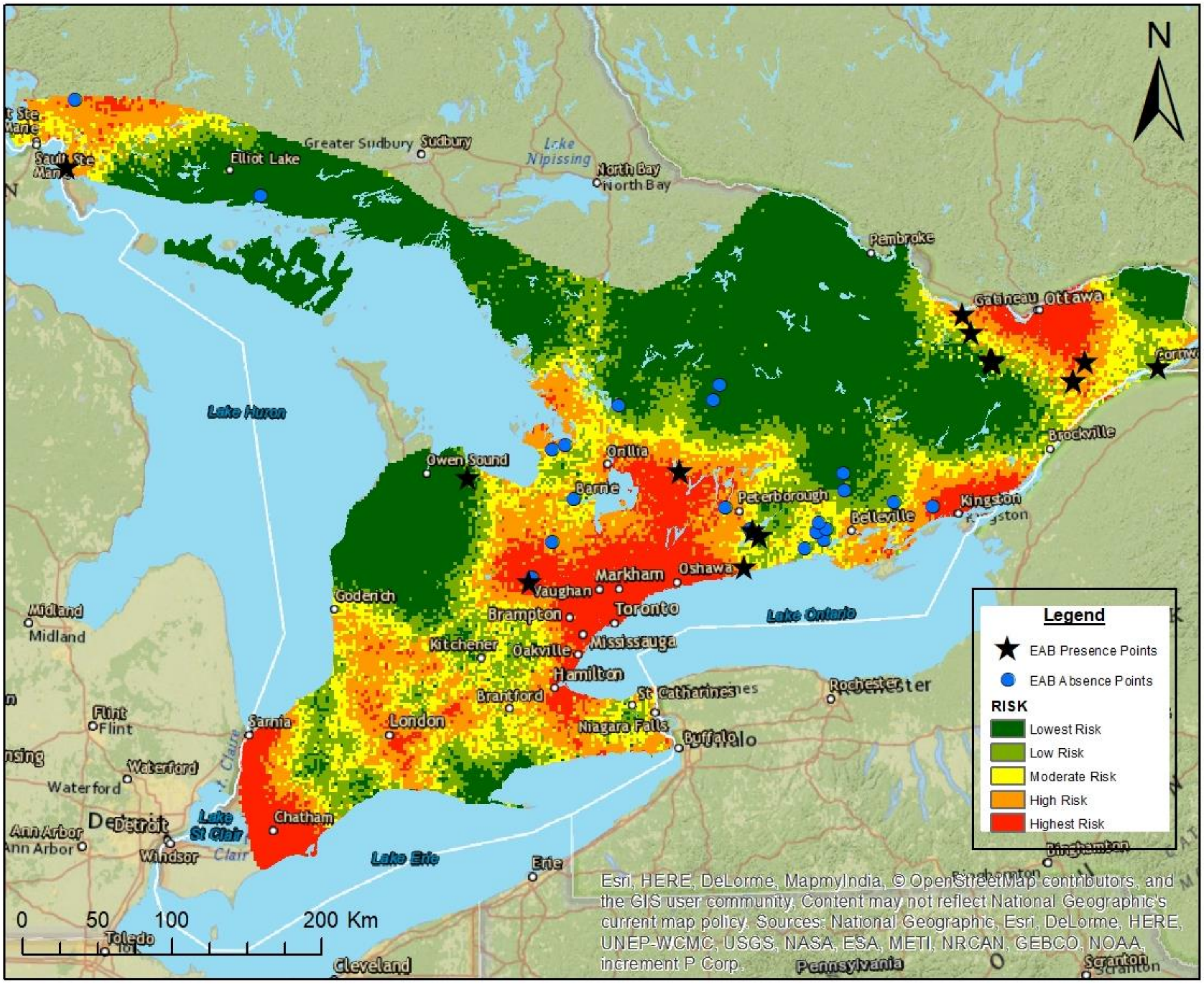
Esri, HERE, DeLorme, MapmyIndia, © OpenStreetMap contributors, and the GIS user community, Content may not reflect National Geographic's current map policy. Sources: National Geographic, Esri, DeLorme, HERE, UNEP-WCMC, USGS, NASA, ESA, METI, NRCAN, GEBCO, NOAA, increment P Corp.

Random Forest EAB Risk Map (2013)



Esri, HERE, DeLorme, MapmyIndia, © OpenStreetMap contributors, and the GIS user community. Content may not reflect National Geographic's current map policy. Sources: National Geographic, Esri, DeLorme, HERE, UNEP-WCMC, USGS, NASA, ESA, METI, NRCAN, GEBCO, NOAA, increment P Corp.

Random Generalized Linear Model (RGLM) EAB Risk Map (2013)



Esri, HERE, DeLorme, MapmyIndia, © OpenStreetMap contributors, and the GIS user community, Content may not reflect National Geographic's current map policy. Sources: National Geographic, Esri, DeLorme, HERE, UNEP-WCMC, USGS, NASA, ESA, METI, NRCAN, GEBCO, NOAA, increment P Corp.