The Pennsylvania State University

The Graduate School

Department or Ecosystem Science and Management

HARNESSING THE POWER OF GEOSPATIAL DATA WITH RANDOM FOREST TO FORECAST GYPSY MOTH OUTBREAK

A Thesis in

Forest Resources

by

Zhiyue Xia

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

August 2018

The thesis of Zhiyue Xia was reviewed and approved* by the following:

Douglas Miller
Research Professor of Geography
Thesis Advisor

Laura Leites
Associate Research Professor of Quantitative Forest Ecology
Thesis Advisor

Shelby Fleischer
Professor of Entomology

Michael Messina
Head of Department of Ecosystem Science and Management

*Signatures are on file in the Graduate School

ABSTRACT

The gypsy moth (*Lymantria dispar*) is a non-native forest pest that was introduced to the USA in 1869. Since then it has spread continuously across most of the northeastern US. Larvae of this insect prefer feeding on oak species, although other species may also serve as host trees. During an outbreak, larvae defoliate forests across large regions and repeated defoliation can predispose the trees to attacks by secondary insect pests or fungal infections causing tree mortality.

Gypsy moth outbreaks are episodic and are difficult to predict. Development of forecasting models remains a challenge despite their potential usefulness in effectively mobilizing resources to deal with the outbreaks. Previous studies indicate that vegetation attributes measured through remote sensing, terrain, and climate characteristics influence the likelihood of gypsy moth outbreaks. In addition, temporal and spatial variables describing the cyclic and spatial patterns of the outbreaks could be very valuable in forecasting outbreaks.

In this thesis, a model is developed to forecast gypsy moth outbreaks using Pennsylvania as a case study. Systematic sampling was used to locate 5,042 sample pixels across forest areas of Pennsylvania and focus on defoliation episodes during the time period 2000-2016 to develop the model. For each pixel, a large suite of temporal and spatial predictor variables is derived from inventory data, climate, topography, and remote sensing measures of vegetation status, while the occurrence of defoliation is obtained from annual defoliation sketch maps. Machine learning modeling algorithm Random Forests was used in this study, which has a well-documented predictive ability and can deal with a large number of variables. The model performance is assessed by hindcasting defoliations in 1985, 1990 and 1995, and by cross validation leaving out one year of the fit dataset at a time. An accurate forecasting model is of critical importance for projecting the spatial extent of future defoliations and for forest management planning.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Thanks to my advisors, Drs. Douglas Miller and Laura Leites, who have guided me and given me advice on this thesis project for the past four semesters. Also, thanks to my committee members Dr. Shelby Fleischer and Dr. Andrew Liebhold for giving me valuable suggestions.

I appreciate the gypsy moth defoliation sketch maps provided by U.S. Forest Service Forest Health Protection and gypsy moth egg mass survey provided by the PA Bureau of Forestry, Division of Forest Health.

Chapter 1

Gypsy moth in North America

Impact of the gypsy moth in North America

The European gypsy moth, *Lymantria dispar*, is a non-native forest pest that was introduced to the United States in 1869 (Liebhold et al., 1995a). From 1869 to the early 1960s, defoliation by gypsy moth occurred in hardwoods and oak forests of New England and the Mid-Atlantic states. However, it has spread continuously across the majority of the northeastern United States and has attacked forest stand types that have never been subjected to previous outbreaks (Davidson et al., 1998).

Gypsy moth larvae prefer feeding on oak species, although other species also serve as host trees. Liebhold et al. (1995b) reported and summarized research literature to identify forest tree species preferred by gypsy moth (Liebhold et al., 1995b). Liebhold et al. (1997) characterized the distribution of these preferred hosts across the US; the most common gypsy moth hosts included white oak (*Quercus alba*), sweetgum (*Liquidambar styraciflua*), quaking aspen (*Populus tremuloides*), northern red oak (*Quercus rubra*) and black oak (*Quercus velutina*).

During an outbreak, larvae defoliate forests across large areas. The defoliation reduces carbohydrate production which increases the tree's susceptibility to other stressors (Twery, 1991). Similar to other defoliators, the effects of gypsy moth on trees primarily reduces growth, flowering, fruiting and increases mortality (Davidson et al., 1999). The reduction of diameter growth caused by gypsy moth defoliation is directly proportional to the extent of the defoliation (Twery, 1991; Baker, 1941). The stem volume growth of oak decreases, on average, by 20% in the year when the trees are attacked, compared to a year without defoliation (Twery, 1987). Tree

mortality is correlated with frequency, duration, and intensity of defoliation, which in turn are related to tree species susceptibility to gypsy moth (Twery, 1991). Frequency of defoliation measures how many times (years) a stand is defoliated in a given period. The duration of defoliation refers to the number of consecutive defoliation episodes in a given period. The probability of tree mortality increases with increasing length of defoliation (Davidson et al., 1999). Defoliation intensity is classified to three levels – light (percentage of defoliation is less than 30%), moderate (percentage of defoliation is from 30% to 60%), and heavy (percentage of defoliation is greater than 60%). Light defoliation causes tree physiologic damage, while moderate or heavy defoliation results in heavy damages, even tree mortality. In 1992, Tigner reported 23% oak mortality after the first heavy defoliation episode in Virginia oak-hickory forests. After the second defoliation episode in the same year, the mortality percent increased to 30%. After the third defoliation episode, oak mortality rose to 50% (Tigner, 1992).

Repeated defoliation affects not only tree growth; it may also predispose the trees to attacks by secondary insect pests or fungal infections causing tree mortality (Liehold et al., 1995a). The most common fungal organism that causes defoliation-related mortality is *Armillaria* species, a kind of root rot fungi (Davidson et al., 1999). *Agrilus bilineatus*, the two-lined chestnut borer, may also cause tree mortality by secondary infection following defoliation. During the defoliation, the tree's ability to resist infection decreases so that the fungal colonization is more likely to happen (Davidson et al., 1999).

Defoliation by gypsy moth not only reduces a tree's diameter and volume growth and contributes to tree mortality; it also affects ecosystem services such as wildlife habitat provisioning, water resources, and timber supply (Twery, 1991). Defoliation can affect wildlife species in both positive and negative ways. For example, defoliation of the tree species preferred by gypsy moth limits the habitat of wildlife that also favors those tree species, while the defoliation of overstory hardwoods may enhance the growth of understory herbs and shrubs,

which provide food for some wildlife species (Twery, 1991). The volume of water produced from watersheds may increase with defoliation since fewer leaves need soil moisture for photosynthesis (Twery, 1991). Other economic effects include timber value loss due to loss of wood quality and growth, as well as volume loss due to mortality (Twery, 1991).

Gypsy moth population dynamics

The gypsy moth life cycle can be divided into four stages – eggs, larvae, pupae, and adults. In the spring, larvae hatch from the overwintering eggs laid in the previous year. Larvae have 5-6 stages (instars). Larvae feed on foliage for about 8 weeks and then find a protected place to pupate. Adult females are not capable of flight. Adult males locate females using pheromones emitted by females. After mating, females lay eggs on tree trunks, branches or objects on the forest floor. In the following year, overwintering eggs hatch and young larvae disperse in the wind (U.S. Forest Service, 2018a).

In the region where the gypsy moth has been generally established, there are usually four population phases. The first phase is an endemic phase during which the population stays at low levels which are innocuous to trees. Next is the release phase during which the population increases rapidly. The third phase is the outbreak phase during which large areas are defoliated by the insect. Then the population comes to the last phase, the decline phase when populations quickly decline back to low levels (Elkinton and Liebhold, 1990).

Gypsy moth outbreaks are recurring events, with an irregular interval, so that outbreaks are difficult to predict. Several studies have demonstrated that gypsy moth in North American are cyclic (Elkinton and Liebhold, 1990; Liebhold et al., 2000; Johnson et al. 2005; Johnson et al. 2006). The most common explanation for the cyclic nature of forest insect populations is the

density-dependent mortality caused by parasitoids, predators and pathogens (Myers, 1988; Liebhold et al. 2000).

Gypsy moth parasitoids play a role in controlling gypsy moth population dynamics, but most studies indicate that they are not the major factor (Elkinton and Liebhold, 1990). The primary egg parasitoids in North American are *Ooencyrtus kuvanae* (Encyrtae) and *Anastatus disparis* (Eupelmidae). Egg parasitoids attack gypsy moth eggs and lower the gypsy moth population. Other parasitoids such as braconid, *Cotesia melanoscela* and the tachinid, *Blepharipa pratensis*, attack gypsy moth larvae and thus also play a limited role in regulating gypsy moth population (Liebhold el al., 2000).

Predators of gypsy moth include small mammals, birds, and invertebrate animals. Bess et al (1947) found that larval survival rate on trees inside an exclusive fence was significantly higher than the trees which were accessible to small mammals. Previous studies show some bird species feed on gypsy moth larvae, but larvae are generally not a major food source for bird species. Even though the mortality caused by vertebrate predators is much higher than invertebrate predators, invertebrate predators such as ground beetles (*Carabidae*) and ants (*Formicidae*) also attack the pupa and result in a decrease of the gypsy moth population (Liebhold et al., 2000).

The most important factor causing the collapse of gypsy moth populations is pathogen disease. The most common pathogens in North America are the nuclear polyhedrosis virus (NPV) and fungal infection *Entomophaga maimaiga*. The infection occurs when larvae eat foliage contaminated by the pathogens. In high density gypsy moth populations, NPV usually causes outbreaks to crash, however, NPV have little effect on gypsy moth populations in low population density. However, fungal infection can decline gypsy moth populations in both low and high population density (Elkinton and Liebhold, 1990; Andreadis et al. 1990; U.S. Forest Service, 2018b).

In addition to the parasitoids, predators, and pathogens that directly cause gypsy moth population fluctuations, the local environment also affects population dynamics. Habitat, weather and climate as well forest composition and forest susceptibility all play a role. The influence from these factors are complex and will be discussed in chapter 2.

Managing gypsy moth in United States

Several government programs have been organized to manage the gypsy moth spread in the past century. Usually, there are five types of management methods. Eradication and slowing the spread are used to prevent and postpone establishment of the gypsy moth in uninfested areas. For managing gypsy moth in the area where it has already established, the common methods are suppression, biological control and silviculture treatments (U.S. Forest Service, 2018c).

Eradication programs are used to eliminate gypsy moth populations from the region where it has not become established. The uninfested area comprises three-quarters of the land area of the contiguous United States. Egg masses can be transported from infested areas to the un-infested areas. Eradication programs survey for new gypsy moth populations by setting traps in a given area. Once the gypsy moth is detected, microbial pesticides are typically used to eliminate populations in order to prevent the introduction of gypsy moth species to this area (Tobin et al., 2011).

The gypsy moth Slow the Spread (STS) program is a national project to contain the gypsy moth which is conducted by state and federal government (Sharov et al., 2002). This program focuses on slowing the gypsy moth spread rather than eradicate or suppress the population. Gypsy moth female adults are not able to fly so natural dispersal speed is limited. Early instars are the most mobile life stage (Liebhold et al., 2000). The estimated range of expansion due to natural first instar dispersal is about 1 ¼ miles per year (Liebhold et al., 1992).

But humans accidentally move gypsy moth life stages (mostly egg masses) and this makes a greater contribution to spread. The STS program deploys thousands of traps to survey for spreading gypsy moth colonies. Once colonies are detected they are delimited and ultimately suppressed. The key treatment used to slow the spread is mating disruption. Releasing sex pheromones for several months disrupts normal mating. Adult females that are not mated lay unviable eggs (Sharov et al., 2002).

Eradication and slow the spread programs are only effective for preventing or postponing gypsy moth outbreak. In the area where gypsy moth has already established, other methods are used to decrease population density. Outbreak suppression uses chemical or biological pesticides to reduce the population density. It can be carried out by individual homeowners, local government or state and federal agencies. The most common chemical pesticides are diflubenzuron (Dimilin) and tebufenozide (Mimic). Two common biological pesticides are Btk (*Bacillus thuringiensis* var. *kurstaki*), derived from natural soil bacterium, and Gypchek, produced by NPV. Btk can sometimes kill non-target species such as butterflies. Gypchek infection is specific to gypsy moth, so it is safe to use in the area with other insects (State of Minnesota, Department of Natural Resources, 2018). Another potential method for managing gypsy moth is using silviculture knowledge to change forest composition and reduce susceptibility to gypsy moth (U.S. Forest Service, 2018c).

Gypsy moth modeling approaches

Many studies have focused on developing simulation models to describe gypsy moth population dynamics, explore the correlation between population and environmental factors, and forecast outbreaks.

Previous studies investigated links between local weather and climate with the gypsy moth defoliation. Williams and Liebhold (1995c) developed a discriminant function to classify the presence or absence of defoliation, using average historical climate variables and forest susceptibility. The model selected six variables through a stepwise procedure, including forest susceptibility, April maximum temperature, September minimum temperature, June precipitation, March minimum temperature and September maximum temperature. The model could correctly classify 49% of the presences of defoliation caused by gypsy moth, while 62.8% of the absences. Sharvo el al. (1999) developed a regression model to estimate the rate of spread with minimum January temperature. They used temperature data acquired from weather stations distributed over Michigan and found the spread rate was negatively related to the minimum January temperature.

Models exploring the gypsy moth population variation caused by habitat characteristics such as topographic position, forest composition, and presence of predators or pathogens have also been developed. Forest et al. (2013) developed a multiple regression model to estimate the defoliation intensity, the portion of area defoliated by gypsy moth, with environmental variables, including host abundance, topography, local phenological asynchrony and pesticide treatment. All the predictors were significant in the model which could explain 20-34% variance in defoliation intensity. Reilly et al. (2014) studied linking the probability of gypsy moth mortality caused by fungal infection to weather. They found a very strong positive relationship between mortality and moisture-related variables. Forest susceptibility was widely used as a predictor in gypsy moth models (Liebhold et al., 1993; Williams and Liebhold, 1995c; Liebhold et al., 2000; Forest et al., 2013). Forest susceptibility is quantified by the basal area of gypsy moth preferred host trees (Liebhold et al., 1995b). In addition to the variables derived from weather stations and field measurements, geospatial data also have been used as independent variables for simulating gypsy moth dynamics, such as defoliation by gypsy moth sketch maps and remotely sensed imagery (Townsend et al. 2004; De Berus et al. 2008; Spruce et al. 2011; Thayn, 2013).

Previous studies developed predictive models to forecast the outbreak using field measurements. The most common field measurements of gypsy moth population density used to predict defoliation are counts of overwintering egg masses. Most operational gypsy moth suppression programs use counts of egg masses in $1/40^{th}$ acre plots to predict defoliation and assess the need for aerial spraying (Liebhold et al. 1993). Some studies have explored prediction of defoliation from counts of male moths in sex pheromone traps as well as egg mass counts (Liebhold et al. 1994). Liebhold et al. (1993) developed stepwise linear regression models and nonlinear Weibull models to forecast percentage defoliation, using as predictors larval density, egg mass density, egg density (egg density was a production calculated from egg mass density and mean number of eggs per mass), mean egg mass length, basal area of host trees and other derived variables like $\log_{10}$ egg mass density, ratio of new-old egg mass and so forth. Results from the regression model indicated the ratio of new-old egg mass was a sufficient variable for predicting defoliation percent. Results for the Weibull model indicated that when the egg mass density was less than 125 egg masses per hectare, it was unlikely to result in defoliation. However, when egg mass density ranged from 250 to 2500 egg masses per hectare, the forecasting results were imprecise. Zhou and Liebhold (1995) developed a series of logistic regression models to forecast the probability of defoliation by gypsy moth from spatially interpolated egg mass counts. They chose different combinations of four variables as predictors, including egg mass density, male moth population density, the defoliation percent in the same cell in the previous year and the distance to the cell defoliated in the previous year. All coefficients were significant in their models. They simulated pest control decision-making by setting probability thresholds for treatment alternatives. For example, if the threshold was set to 0.7, when the probability of presence was greater than 0.7, pest control treatment would be carried out in that area. However, they indicated that the model results presented highly stochastic outcomes

that would result in erroneous decision-making. Considerable previous modeling efforts ultimately helped design more reliable pest management programs.

Previous studies on modeling gypsy moth dynamics indicate that forecasting gypsy moth outbreaks is an essential component of decision-making for managing gypsy moth outbreaks. A forecasting model with good predictive ability is crucial for making the decision whether to mobilize resources. In Chapter 2, I developed a model using a machine learning algorithm, Random Forests, to forecast gypsy moth outbreaks.

References

Andreadis, T. G., & Weseloh, R. M. (1990). Discovery of Entomophaga maimaiga in North American gypsy moth, Lymantria dispar. Proceedings of the National Academy of Sciences, *87*(7), 2461-2465.

Baker, W. L. (1941). Effect of defoliation by gypsy moth on certain forest trees. Journal of Forestry, 39(12), 1017-1022.

Bess, H. A., Spurr, S. H., Littlefiedl, E. W. (1947). Forest site conditions and the gypsy moth. Harvard Forest Bulletin. 22. 56 pp.

Davidson, C. B., Gottschalk, K. W., & Johnson, J. E. (1999). Tree mortality following defoliation by the European gypsy moth (*Lymantria dispar L.*) in the United States: a review. Forest Science, 45(1), 74-84.

De Beurs, K. M., & Townsend, P. A. (2008). Estimating the effect of defoliation by gypsy moth using MODIS. Remote Sensing of Environment, 112(10), 3983-3990.

Doane, C. C., & McManus, M. L. (1981). The gypsy moth: research toward integrated pest management (No. 1584). US Department of Agriculture.

Elkinton, J. S., & Liebhold, A. M. (1990). Population dynamics of gypsy moth in North America. Annual Review of Entomology, 35(1), 571-596.

Foster, J. R., Townsend, P. A., & Mladenoff, D. J. (2013). Spatial dynamics of a defoliation by gypsy moth outbreak and dependence on habitat characteristics. Landscape Ecology, 28(7), 1307-1320.

Johnson, D. M., Liebhold, A. M., Bjørnstad, O. N., & Mcmanus, M. L. (2005). Circumpolar variation in periodicity and synchrony among gypsy moth populations. Journal of Animal Ecology, 74(5), 882-892.

Johnson, D., M. Liebhold, A. M., & Bjørnstad, O. N. (2006). Geographical variation in the periodicity of gypsy moth outbreaks. Ecography, 29(3), 367-374.

Liebhold, A. M., Halverson, J., & Elmes, G. (1992). Quantitative analysis of the invasion of gypsy moth in North America. Journal of Biogeography, 19(5), 513-520.

Liebhold, A. M., Simons, E. E., Sior, A., & Unger, J. D. (1993). Forecasting defoliation caused by the gypsy moth from field measurements. Environmental Entomology, 22(1), 26-32.

Liebhold, A., Thorpe, K., Ghent, J., & Lyons, D. B. (1994). Gypsy moth egg mass sampling for decision-making: a user's guide. USDA-Forest Service, NA-TP-04-94.

Liebhold, A. M., MacDonald, W. L., Bergdahl, D., & Mastro, V. C. (1995a). Invasion by Exotic Forest Pests: A Threat to Forest Ecosystems. Forest Science Monographs 30. 49 p.

Liebhold, A. M., Gottschalk, K. W., Muzika, R. M., Montgomery, M. E., Young, R., O'Day, K., & Kelley, B. (1995b). Suitability of North American tree species to gypsy moth: a summary of

field and laboratory tests. Gen. Tech. Rep. NE-211. Radnor, PA: US Department of Agriculture, Forest Service, Northeastern Forest Experiment Station. 34 p.

Liebhold, A. M., Gottschalk, K. W., Mason, D. A., & Bush, R. R. (1997). Forest susceptibility to the gypsy moth. Journal of Forestry, 95(5), 20-24.

Liebhold A, Elkinton J, Williams D, Muzika R. 2000. What causes outbreaks of the gypsy moth in north america? Population Ecology 42(3):257-66.

Myers, J. H. (1988). Can a general hypothesis explain population cycles of forest Lepidoptera? Advances in Ecological Research 18, 179-242.

Reardon, R. C., Podgwaite, J. D., & Zerillo, R. T. (1996). Gypchek, the gypsy moth nucleopolyhedrosis virus product. USDA Forest Service, Northeastern Area, Forest Health Technology Enterprise Team.

Reilly, J. R., Hajek, A. E., Liebhold, A. M., & Plymale, R. (2014). Impact of *Entomophaga maimaiga* (Entomophthorales: Entomophthoraceae) on outbreak gypsy moth populations (Lepidoptera: Erebidae): the role of weather. Environmental Entomology, 43(3), 632-641.

Sharov, A. A., Pijanowski, B. C., Liebhold, A. M., & Gage, S. H. (1999). What affects the rate of gypsy moth (Lepidoptera: Lymantriidae) spread: winter temperature or forest susceptibility? Agricultural and Forest Entomology, 1(1), 37-45.

Spruce, J. P., Sader, S., Ryan, R. E., Smoot, J., Kuper, P., Ross, K., ... & Hargrove, W. (2011). Assessment of MODIS NDVI time series data products for detecting forest defoliation by gypsy moth outbreaks. Remote Sensing of Environment, 115(2), 427-437.

Thayn, J. B. (2013). Using a remotely sensed optimized Disturbance Index to detect insect defoliation in the Apostle Islands, Wisconsin, USA. Remote Sensing of Environment, 136, 210-217.

Tigner, T. (1992). Gypsy moth impact on Virginia's hardwood forests and forest industry. Virginia Department of Forestry, Charlottesville, Virginia, 36.

Tobin, P. C., Bai, B. B., Eggen, D. A., & Leonard, D. S. (2012). The ecology, geopolitics, and economics of managing *Lymantria dispar* in the United States. International Journal of Pest Management, 58(3), 195-210.

Townsend, P. A., Eshleman, K. N., & Welcker, C. (2004). Remote sensing of defoliation by gypsy moth to assess variations in stream nitrogen concentrations. Ecological Applications, 14(2), 504-516.

Twery, M. J. (1987). Changes in vertical distribution of xylem production in hardwoods defoliated by gypsy moth. Doctoral dissertation, Yale University.

Twery, M. J. (1991). Effects of defoliation by gypsy moth. IN: Gottschalk, Kurt W.; Twery, Mark J.; Smith, Shirley I., Eds. Proceedings, US Department of Agriculture interagency gypsy moth research review 1990; East Windsor, CT. Gen. Tech. Rep. NE-146. Radnor, PA: US Department of Agriculture, Forest Service, Northeastern Forest Experiment Station. p 27-39.

U.S. Forest Service. 2018a. Gypsy moth life cycle. www.nrs.fs.fed.us/disturbance/invasive_species/gm/biology_ecology/life-cycle/. Accessed 06/21/2018

U.S. Forest Service. 2018b. Gypsy Moth Nucleopolyhedrosis Virus. www.fs.fed.us/ne/morgantown/4557/gmoth/natenem/virus.html. Accessed 06/21/2018

U.S. Forest Service. 2018c. Gypsy Moth Management. www.fs.fed.us/ne/morgantown/4557/gmoth/manag/. Accessed 06/21/2018

Williams, D. W., & Liebhold, A. M. (1995a). Influence of weather on the synchrony of gypsy moth (Lepidoptera: Lymantriidae) outbreaks in New England. Environmental Entomology, 24(5), 987-995.

Williams, D. W., & Liebhold, A. M. (1995b). Detection of delayed density dependence: effects of autocorrelation in an exogenous factor. Ecology, 76(3), 1005-1008.

Williams, D. W., & Liebhold, A. M. (1995c). Forest defoliators and climatic change: potential changes in spatial distribution of outbreaks of western spruce budworm (Lepidoptera:

Tortricidae) and gypsy moth (Lepidoptera: Lymantriidae). Environmental Entomology, 24(1), 1-9.

Zhou, G., & Liebhold, A. M. (1995). Forecasting defoliation by gypsy moth with a geographical information system. Insect Science, 2(1), 83-94.

Chapter 2

Forecasting gypsy moth outbreaks by combining defoliation sketch maps, climate, terrain, and derived variables using Random Forests

Introduction

Gypsy moth outbreaks occur during the growing season from March to July and typically affect large forested areas (Liebhold et al., 1998). It is common for gypsy moth populations to stay at a low density for many years and suddenly erupt in a given year (Liebhold et al., 1998). Timing of outbreaks is generally cyclic events but with some irregularity (Liebhold et al., 2000) which makes gypsy moth outbreaks difficult to predict. According to historic defoliation time series, defoliated area has a significant positive correlation at time lag 1 and time lag 2 (Williams and Liebhold, 1995ab). The dispersal of gypsy moth is unlikely to influence large scale defoliation because the dispersal rate is low. Gypsy moth females in North America cannot fly. Late instars (the larvae at late larval stages) move between trees but their movement is not extensive (Elkinton and Liebhold, 1990). Even though gypsy moth dispersal rate is low, using a fine scale to research defoliation by gypsy moth is not a good choice, since the semi-variograms of population density vs. measurement distance demonstrates a nugget-effect. Nugget-effect refers to the phenomenon where the semi-variogram is discontinuous at the beginning, but then continuous. The nugget-effect is usually caused by nonuniform distribution of the sample objectives. Semi-variograms can provide some guidance for identifying a suitable sample size for population census. Gypsy moth male population density is not continuous within 3300 m sample distance. The egg mass density is not continuous within the 2300 m sample distance (Zhou and Liebhold, 1995). The discontinuity of the population density within a small distance indicates that

gypsy moth population cannot be analyzed at a fine spatial scale. Gypsy moth population dynamics presents an autocorrelated temporal pattern and a large scale spatial pattern.

Gypsy moth prefers oak species. However, other tree species can also be host trees. Liebhold et al (1995b, 1997a) reported and summarized previous studies to identify tree species preferred by gypsy moth and characterized the distribution of these preferred tree species across the United States.

The forest stands composed of more preferred host trees may be more susceptible to gypsy moth disturbance. Forest susceptibility affects speed of gypsy moth spread. Sharov et al. (1999) evaluated how forest susceptibility affected the gypsy moth spread. They quantified forest susceptibility by the percent of the land that has over 50% of tree basal area in gypsy moth preferred species and concluded that forest susceptibility was positively correlated with the speed of spread.

In addition to temporal information on defoliation and forest susceptibility to gypsy moth, terrain and climate are also important factors that influence defoliation by gypsy moth. Previous studies indicate that topography can reflect features of gypsy moth habitat. Elevation has a positive correlation with gypsy moth protandry, which refers to the phenomenon when the date of male gypsy moth maturity is earlier than female. In this way the reproductive asynchrony is disturbed, affecting gypsy moth spread and establishment (Walter at el, 2015).

Both weather and climate normals are correlated with the spatial distribution of outbreaks of gypsy moth. Williams and Liebhold (1995a) investigated how climate change projections over 100 years could affect defoliation by gypsy moth. They concluded that when temperature and precipitation increased, the area defoliated by gypsy moth increased. In contrast, the area defoliated by these insects decreased with increased spring temperature and decreased precipitation. They developed a discriminant function for defoliation by gypsy moth in Pennsylvania with six variables that were selected using a stepwise procedure with 28 climate

variables. The climate normal variables identified as important by this modeling approach were: April maximum temperature, September minimum temperature, June precipitation, March minimum temperature and September maximum temperature. All of these six climate variables are averages of historic 30 years data that can describe habitat climate characteristics in the growing season. In addition to climate normals, annual winter and summer temperatures are related to gypsy moth population. Cold winters with continuous low temperatures kill gypsy moth eggs (Sharov et al.,1999; Liebhold et al., 2000). The larvae grow to late instars (the larvae at late life stages) in late May or June. High early summer temperature restrains the spread of fungal pathogen *Entomophaga maimaiga* which is a natural disease that cause gypsy moth late instars died, so the mortality of late instars caused by fungal infection decreased (Reilly et al., 2014). Previous year summer temperature and current year winter temperature are perhaps potential predictors for predicting defoliation.

Geospatial data provides temporal and spatial information that can map the defoliation caused by gypsy moth through the years, so the historic trend of frequency and degree of gypsy moth disturbance can be modeled and used as predictor variables in forecasting models. Varied geospatial data types are available to the public. Harnessing the power of geospatial data to develop a predictive model is a promising way to forecast defoliation by gypsy moth.

Selecting a suitable modeling approach to forecast gypsy moth defoliation could increase the predictive accuracy. Previous studies used multiple linear regression to estimate the portion of defoliated area (Liebhold et al., 1993; Forster et al. 2013). But using the percentage of defoliation as a response variable is very challenging since qualifying defoliation intensity is complex. To overcome this issue, Zhou and Liebhold (1995) developed a classification model to predict defoliation or no defoliation rather than percent defoliation.

Random Forests is a machine learning algorithm with a good predictive ability (Breiman, 2001that trains multiple either classification or regression trees and outputs the class mode of the

trees when creating a classification, and the mean estimate of all trees when doing regression. The model is widely used in forestry and ecological studies, including forecasting future species distribution, estimating the likelihood of fire occurrence and predicting presence of invasive species (Cutler et al. 2007; Evans et al. 2011; Oliveira et al. 2012). Culter et al. (2007) compared Random Forests with linear models, using "presence or absence" ecology problems as case studies. They indicate that it is difficult to know which model will perform better for a specific problem. However, Random Forests models should perform better than linear models when the training dataset is big and the predictors are collinear (Culter et al., 2007). The potential predictors for forecasting defoliation by gypsy moth includes climate normals, weather, topography, forest composition, integrated moisture index, in which there might be strong interactions. Random Forests is a promising tool to handle these predictors.

In this chapter, I combine gypsy moth dynamics, climate, topography, integrated soil moisture, and forest inventory data, using a Geographic Information System (GIS) methodology and a Random Forests algorithm to develop a model that will forecast defoliation by gypsy moth using Pennsylvania as a case study. A model with good predictive ability will help project the spatial extent of the defoliation and aid in mobilizing mitigation resources efficiently.

Methodology

Study Area

The study area comprises all forestland in Pennsylvania. Forests were identified by using the forest canopy cover layer provided by Landscape Fire and Resource Management Planning Tools (LANDFIRE) program (Figure 2.1., www.landfire.gov) of the U.S. Department of Agriculture Forest Service and U.S. Department of the Interior. This layer maps the percent cover

of tree canopy in a 30 m by 30 m pixel. The original layer classifies the percent forest cover into ten classes from 0% to 95%, but it was reclassified into two classes: a) forest, if the percent cover of tree canopy was equal or larger than 15%; b) non-forests, if the percent cover is less than 15% (i.e. class 0% in the original) (Figure **2-1).** The sampling pixels were generated in the forest area.

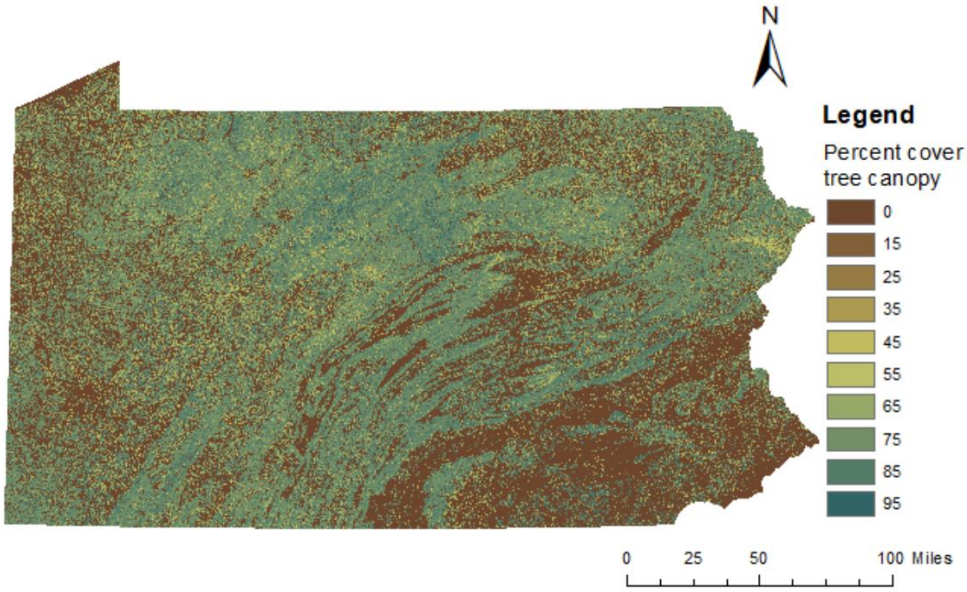

Figure **2-1**: LANDFIRE forest canopy cover raster layer in Pennsylvania

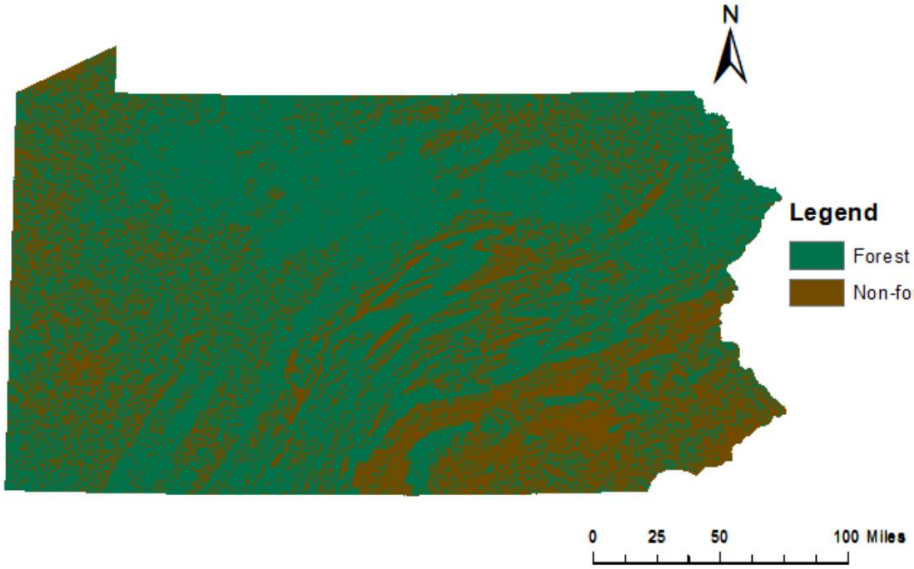

Figure **2-2**: Pennsylvania forest and non-forest raster layer

The climate in Pennsylvania is generally considered to be humid continental type, but varied physiography characterizes different climate and weather within the state. Summers are

generally warm and humid. Freezing temperature occurs in the winters. Average precipitation for the area is approximately 34-52 inches (The Pennsylvania State Climatologist, 2018). Topography factors, slope, aspect and elevation, influence the microclimate of these forests.

Sampling procedure

A systematic sampling procedure was used for selecting sample plots in Pennsylvania: a 4 km by 4 km grid was created and overlaid over the study area, with the center of each cell being the center of a potential sampling unit (sampling pixel hereafter). The sample pixels were set to a size of 1 km by 1 km. The potential sample pixels were filtered and only those with a forest cover percent larger than 50% were kept as a final sample pixel. A total of 5,042 sample pixels across Pennsylvania composed the sample (Figure **2-3).** For each of these sample pixels, data was obtained for each year in the period 2000-2016. The sample pixel-year combination was the modeling unit (observation). There was a total of 85,714 observations (5,042 sample pixels by 17 years). Table **2-1** presents a summary of number of sampling pixel-year combinations by presence/absence of defoliation from 2000 – 2016.

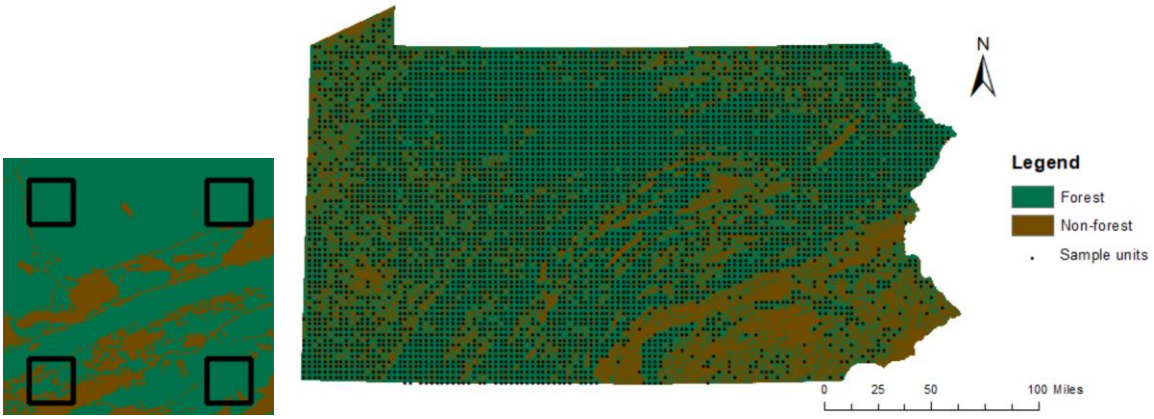

Figure **2-3**: Distribution of sample pixels in Pennsylvania

Table **2-1**: Summary of sampling pixels from 2000 – 2016 with and without recorded defoliation by gypsy moth

Year	Number of defoliated pixels	Number of no-defoliated pixels
2000	422	4620
2001	62	4980
2002	41	5001
2003	7	5035
2004	4	5038
2005	98	4944
2006	345	4697
2007	376	4666
2008	437	4605
2009	239	4803
2010	44	4998
2011	17	5025
2012	3	5039
2013	203	4839
2014	103	4939
2015	296	4746
2016	87	4955

Data sources and derived variables

Six types of data were identified to develop the predictive model of defoliation by gypsy moth. The first one was the LANDFIRE forest canopy cover layer. This layer was used to identify forest area and determine the percent of forest in the sample pixels. The second type was the sketch maps of defoliation by gypsy moth for the period 1983-2016. The sketch maps provided the response variable and additional landscape metrics to be considered as predictor variables such as distance to the nearest defoliation patch in the previous year. The layer basal area of gypsy moth host tree species described the forest susceptibility. Terrain data were used to derive topographic metrics, including elevation, slope, aspect, topographic position index and topographic roughness index. Climate data characterizes climate in the previous year and the forecasted year. The integrated moisture index (IMI) describes long term soil moisture conditions and ranges from wet to dry. In addition to climate and terrain, IMI maps provides additional information on site quality (Iverson et al., 1997; Iverson and Prasad, 2003).

Sketch map of defoliation by gypsy moth

Pennsylvania defoliations by gypsy moth were mapped, using annual aerial surveys, by the Pennsylvania Department of Conservation and Natural Resources, Bureau of Forestry (DCNR-BoF). The US Department of Agriculture Forest Service (USFS) Forest Health Protection combined all states sketch maps and provided maps for the entire country. These maps (sketch maps hereafter) are available as shapefiles from the 1980's to present and they provide spatial and temporal observations of the defoliation (e.g. Figure **2-4**).

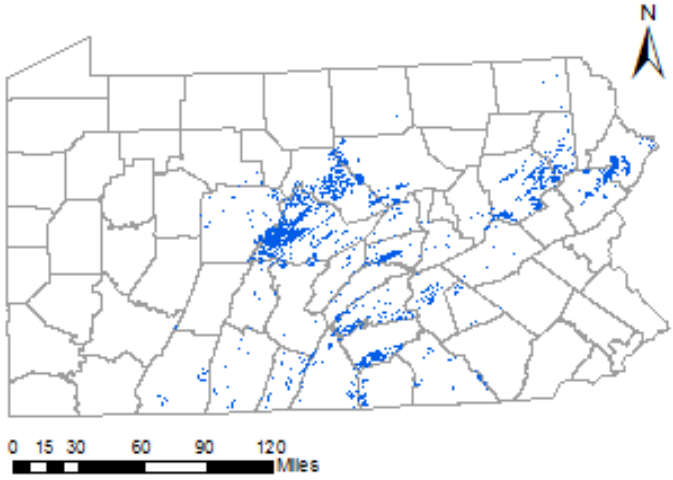

Figure **2-4**: Defoliation by gypsy moth in year 2008

Four gypsy moth dynamics variables were derived from the sketch maps. The first one is the response variable: presence or absence of defoliation by gypsy moth in the sample pixel in year t. The three others are predictor variables that attempt to capture the cyclical temporal nature of the defoliation. For this, the percent of defoliation in an area neighboring the sample pixel was calculated for years t-1, t-2, and t-3. The neighboring area was delineated using a buffer of 4 km radius centered in the sample pixel.

LANDFIRE forest canopy cover

The LANDFIRE program (www.landfire.gov) provides varied national scale geo-spatial datasets (e.g. disturbance, vegetation, fuel etc.) to the public. The forest canopy cover layer provides estimates of the percent forest cover by vertically projecting the tree canopy into a horizontal representation of the ground's surface and estimating the percent of this projection of the total area (LANDFIRE, 2018). The spatial resolution of the LANDFIRE layer is 30 m by 30

m. This layer was used to select the study area by classifying forest and non-forested areas, and to calculate the percent of forest in the sample pixels.

Basal area of gypsy moth host trees

The basal area (ft^2 per acre) of gypsy moth host trees layer describes the density of gypsy moth hosts and this has previously been found to be correlated with forest susceptibility to defoliation by gypsy moth (Liebhold et al. 1997a). The layer used in this study was provided by the USFS, Northeastern Research Station. Forest susceptibility was assessed by combining the individual trees basal area. The list of gypsy moth host trees that was used to combine individual host species was summarized by Liebhold et al (1995b). The individual trees basal area maps were interpolated from 2009 USFS Forest Inventory and Analysis (FIA) data. The 20 most common host species (Appendix A list species names) include 13 oak species (white oak, northern red oak, black oak, chestnut oak, post oak, water oak, southern red oak, scarlet oak, laurel oak, willow oak, California red oak, Canyon live oak and bur oak), two aspen species (quaking aspen and bigtooth aspen), and five other species (sweetgum, paper birch, eastern larch, tanoak, eastern hophornbeam). Basal area of host trees is generally highly correlated with historic average defoliation across the Conterminous United States (Liebhold et al., 1997b).

Terrain data

A LiDAR-derived Digital Elevation Model (DEM) was available for the entire state from Pennsylvania Spatial Data Access (PASDA, available at www.pasda.psu.edu). This DEM was constructed by the Pennsylvania Map (PAMAP) program in 2006-2008. The horizonal ground resolution of the DEM is 3.2 ft.

Because the high spatial resolution of elevation points may have noise, the original DEM was resampled from 1 m by 1 m to a lower resolution of 10 m by 10 m. From this layer, Five topography variables were calculated: elevation, slope, aspect, topographic position index (TPI) and topographic roughness index (TRI). TPI is also called “difference from mean elevation”. TPI measures the relative position of the center by calculating the difference between elevation of center pixel and mean elevation of 8 adjacent pixels (De Reu et al., 2013). TRI measures the roughness degree of the landform by calculating the mean absolute difference of center pixel and other adjacent pixels.

To combine the high-resolution topography data with other data sources of lower spatial resolution, the spatial resolution of the topographic data was lowered by averaging elevation, slope, aspect, topographic position index, and topographic roughness index for each of the sample pixels.

Climate data

Temporal climate data describing weather conditions for the period 1999-2016 were obtained from the TopoWx dataset (Oyler et al., 2015; available at www.scrimhub.org/resources/topowx) and the PRISM climate group (Daly et al. 1994; available at www.prism.oregonstate.edu). January temperature was used to represent winter temperature. Low winter temperatures can kill gypsy moth eggs and reduce gypsy moth growth rate (Sharov et al, 1999; Liehold et al, 2000). June temperature and the previous year annual precipitation were also extracted and used as predictors in the Random Forests model because these weather factors can affect the gypsy moth reproductive process and are related to gypsy moth populations (Reilly et al, 2014).

Climate normals from Rehfeldt et al. (2006) for the period based on 1961-1990 were obtained and used to characterize the sample pixel long term climate.

Integrated moisture index

The Integrated moisture index (IMI) was obtained from USFS Northeastern Research Station and describes the long-term soil moisture condition (Iverson et al., 2003). IMI is created from four landscape features (topography, slope, cumulative flow of water and water capability of soil) and has an approximate spatial resolution of 10 m by 10 m. Combined with other forest inventory data, IMI is useful for predicting forest productivity and species composition (Peters et al., 2010). Forest productivity and composition affect forest susceptibility to disturbance. IMI was therefore included as a potential predictor variable in the gypsy moth forecasting model.

The data sources have different spatial references and spatial resolutions. For example, the original spatial reference of PRISM climate data is GCS North American, with spatial resolution 4 km by 4 km. While the LiDAR topography data is 1 m by 1 m high spatial resolution data in State Plane Pennsylvania spatial reference. In order to work with these different data sources, all layers were reprojected to a Pennsylvania Albers projection.

Table **2-2** lists all predictors used in the model, also including the link between the predictors and defoliation.

Table **2-2**: Summary of variables input to Random Forests model

Variables	Reflecting	Relationship	References
Forest composition			
Basal area of host trees	Forest susceptibility	Forest containing a larger proportion of preferred species are at a higher risk	Liebhold et al., 1997. Sharov et al., 1999.
Forest cover percent	Landscape composition		
Autocorrelative variables			
Estimated defoliated area in the entire state	Gypsy moth population cycle	Time series of historic defoliation presents autocorrelation	Williams and Liebhold, 1995ab
Percent of defoliated area in the neighboring area for year t-1			
Percent of defoliated area in the neighboring area for year t-2			
Percent of defoliated area in the neighboring area for year t-3			
Number of years since last defoliated in the neighboring area			
Geographic location			
Longitude	Geographic location		
Latitude			
Terrain			
Elevation	Topography characteristics	High elevation decreases population growth rates	Forster et al., 2013; Walter et al., 2015
Slope		Topography is related to humidity, temperature and presence of host trees	
Aspect			
Topographic Position Index			
Topographic Ruggedness Index			
Soil Moisture			
Integrated Moisture Index	Long term soil moisture conditions	Related to forest tree species composition	(Iverson et al., 1997)

Table **2-2**: Summary of variables input to Random Forests model (continued)

Variables	Reflecting	Relationship	References
Weather			
Minimum and maximum summer temperature for year t-1	Environment temperature during adults and new eggs stage	High temperature decreases mortality of gypsy moth caused by fungal infection	Reilly et al., 2014
Minimum and maximum January temperature for year t	Winter temperature	Cold winter can kill gypsy moth eggs	Liebhold et al., 2000 Sharov et al., 1999
Annual precipitation for year t-1	Drought-stressed trees, Fungal infection	Positively related with gypsy moth mortality due fungal infection, stressed trees more susceptible	Reilly et al., 2014 Williams et al., 1995
Climate normals			
All variables available from Rehfeldt, 2006	Long-term survival environment for gypsy moth and related to host tree presence	Climate normals describe the habitat environment. They are related with gypsy moth population.	Williams et al., 1995

Estimate of total area of defoliation by gypsy moth using an autoregressive model

Gypsy moth populations have a cyclical behavior with years of low density populations that build to dense levels through several years. Using area defoliated in previous years as a proxy for population sizes may have predictive power. Therefore, I used an autoregressive model, which is useful to account for cyclical behavior (Moran, 1953), to estimate the total area of defoliation in Pennsylvania in the forecasted year using the total area of defoliation in the previous two years. Williams and Liebhold, (1995ab) show that a 2nd order autoregressive process is sufficient to capture the cyclical nature of the gypsy moth outbreaks.

The 2nd order autoregressive model was developed using historic data on defoliated area recorded from 1986-2016 (available from the USDA Forest Service "Gypsy Moth Digest" https://www.fs.usda.gov/naspf/programs/forest-health-protection/gypsy-moth-digest). The original units of these data are acres but were converted to hectares for this study.
The following is the autoregressive model equation:

$$log\_defol(t) = 4.4255 + 0.7048 * log\_defol(t-1) - 0.2314 * log\_defol(t-2)$$

where *log_defol(t)* is the log10 total area (ha) defoliated in Pennsylvania in year t; *log_defol(t-1)* is the log10 total area (ha) defoliated in Pennsylvania in year t-1; *log_defol(t-2)* is the log10 total area (ha) defoliated in Pennsylvania in year t-2.

Figure **2-5** compares the predicted (from the previous two years data using the above model) total defoliated area with actual defoliated area in Pennsylvania from 1985-2016. The y-axis is displayed on a logarithmic scale. Since the recorded defoliated area for 2010 and 2012 are 0 ha, they are not displayed in the figure. The estimated defoliated area for 2014 was extremely high since defoliation occurred in 2013 in 135,860 ha which was positively correlated with the estimated defoliated area for 2014, while there was no area in Pennsylvania attacked by gypsy moth for 2013, which was negatively correlated with the estimated defoliated area for 2014. This indicates that the autoregressive model is imprecise for predicting a year before which the large-scale defoliation occurred in year t-1 and no defoliation occurred in year t-2.

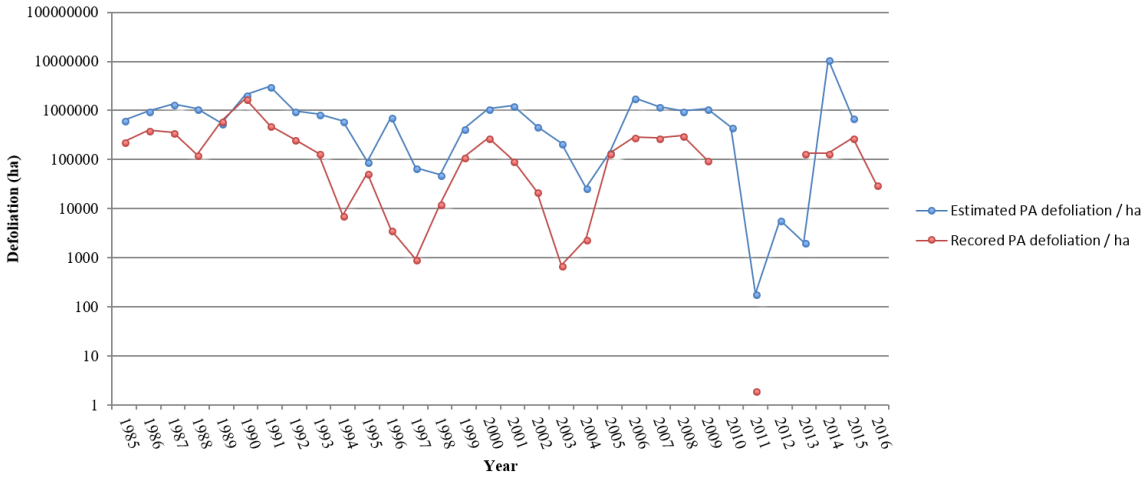

Figure **2-5**: Recorded Pennsylvania defoliated area and estimated defoliated area by gypsy moth from 1985-2016

The estimated total defoliated area in the entire state for year t was added to the model to forecast defoliation for a sample pixel in year t. For all sample pixels in year t, the value of estimated total defoliated area is the same.

Modeling approach

The Random Forests algorithm (Breiman, 2001) was used to develop the model. Random Forests is a machine learning method for classification and regression. The model trains multiple decision or regression trees and outputs the mode class when doing classification or mean prediction when doing regression (Ho, 1995). To forecast the defoliation by gypsy moth at year t, the input variables (Table **2-2**) included climate normal variables, temporal weather for year t-1 and t, and terrain variables, IMI variable, basal area of host tree, and gypsy moth dynamic variables, while the output is presence or absence of the defoliation by gypsy moth in year t.

A flowchart of the methodology is presented in Figure **2-6**.

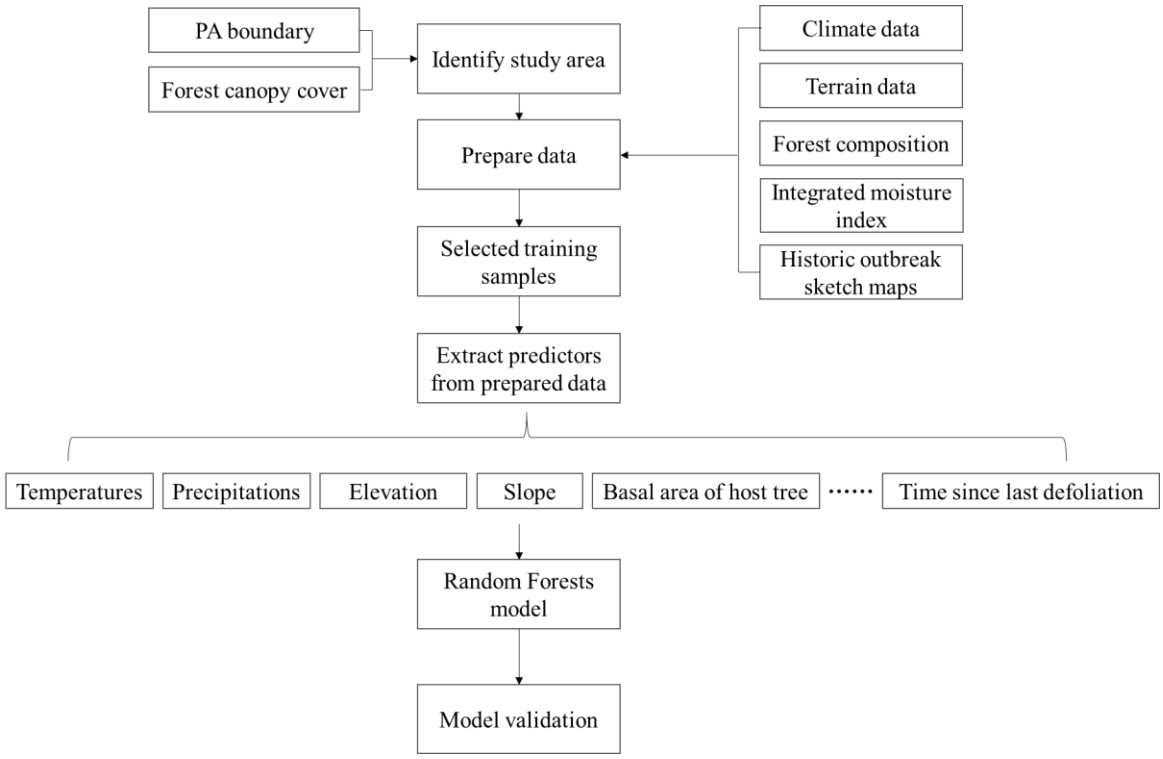

Figure **2-6**: Flowchart of the methodology

Model validation

The predictive ability of the model was assessed in two ways: 1) by hindcasting using independent sample of pixels and years (validation dataset hereafter), 2) by using k-fold cross-validation.

To obtain the validation dataset, three years were selected: 1980, 1990, 1995. For each of those years, 200 sample pixels were randomly located within the forest area with the restriction of the sample pixel having at least 50% of canopy cover (Figure **2-7**). For the k-fold cross-validation, year was used as the fold, this is, the model was trained using 16 years (total 80,672 observations, 16 years by 5042 pixels) of the 17 available and validated with the 17^{th} year (total 5042 observations). This was done by removing one available year at a time.

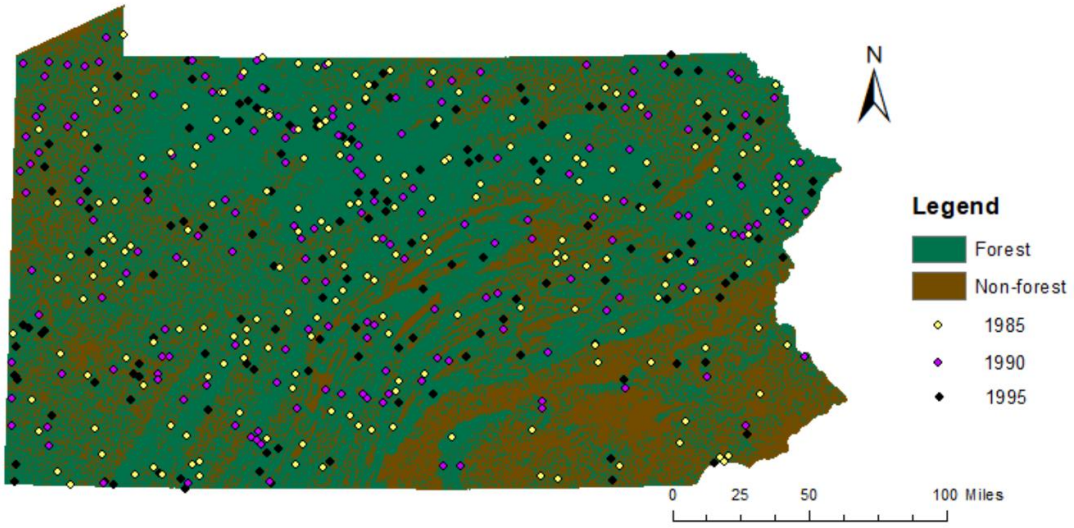

Figure **2-7**: Distribution of sample pixels for validation

Results

The forest composition in the gypsy moth defoliated and non-defoliated areas are different. The most common gypsy moth hosts in the region are oak species, though some less abundant species, such as birch and aspen, are also hosts. Comparing the basal area of host trees in the pixels that defoliation and no defoliation by gypsy moth, the average basal area of the host tree in the defoliated pixels is higher than that in the non-defoliated pixels (Figure **2-8**a). In the same way, the forest percent in the defoliated pixels is higher than no-defoliated pixels (Figure **2-8**b).

There is also a statistically significant difference between the climate of defoliated areas and no-defoliated areas. For example, the difference in mean minimum January temperature for year t in defoliated area and non-defoliated area is significant different (t-test, p-value < 2.2*10e-16). The mean minimum January temperature at year t in the defoliated pixels is -7.6 °C, while the mean in the no-defoliation pixels is -8.5 °C (Figure **2-8**c).

In addition to forest composition and climate, defoliation by gypsy moth history and topography in the defoliated pixels and non-defoliated pixels are also compared. According to Figure **2-8**e, in most defoliated pixels, the percent of defoliation in neighboring areas (circle with a radius of 4 km) in the previous year is 0%. However, most defoliated pixels are in areas that have been attacked by gypsy moth in the previous year. In another words, the forest stands seem more likely to have defoliation present if defoliation was present in the neighboring area in the previous year.

A small difference of no biological importance was detected in topography features of defoliated area and not defoliated pixels. Mean elevation in the defoliated area 1,401 ft and 1,421 ft in the non-defoliated area. (Figure **2-8**f).

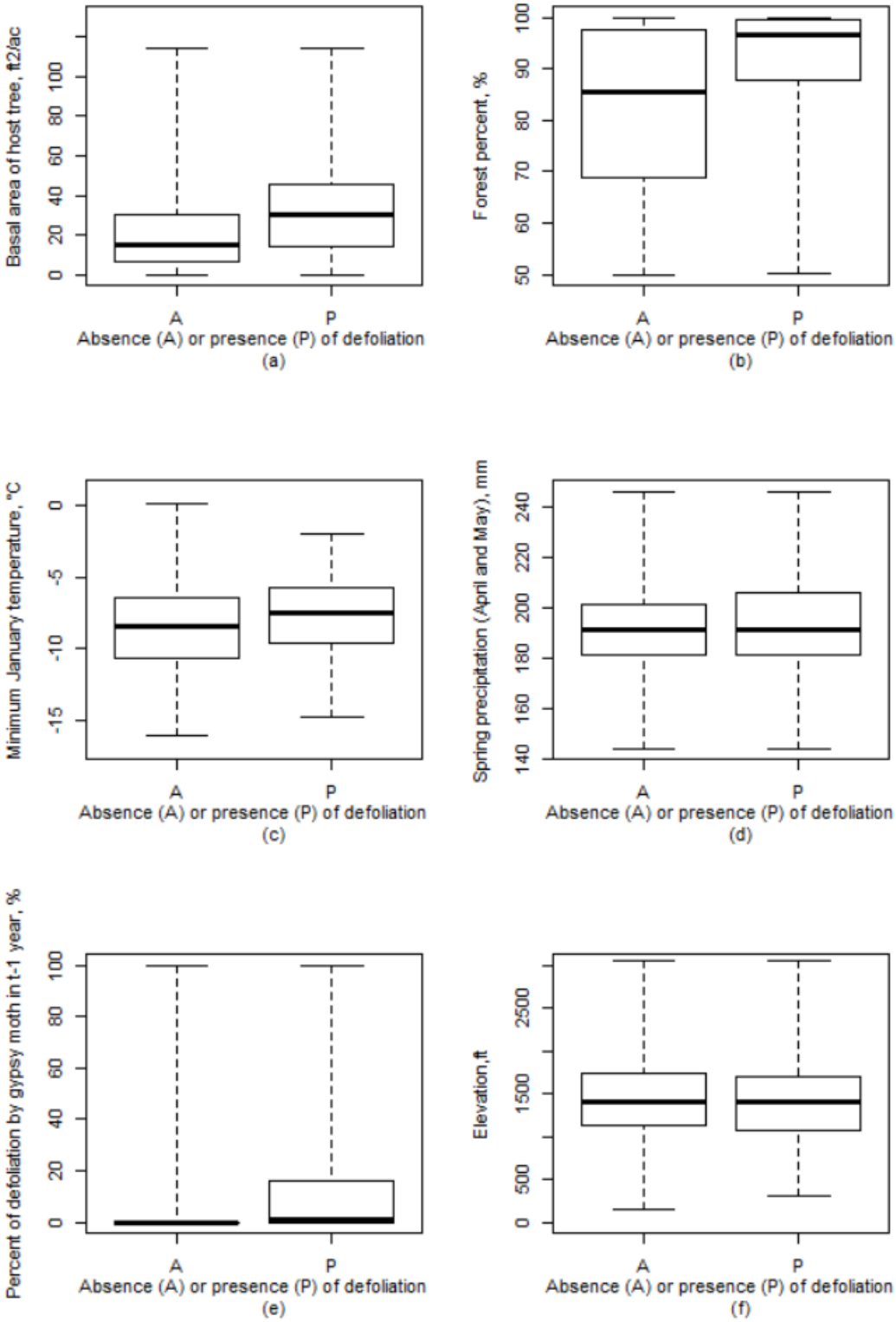

Figure **2-8**: Boxplots of selected variables in areas where the defoliation is absent (A) or present (P)

The Random Forests model classified the pixels into defoliation and non-defoliation. To explore the importance of different types of variables, two models were fitted: a reduced model using only six predictor variables reflecting gypsy moth dynamics (model 1 hereafter), and a full model utilizing all predictor variables (model 2 hereafter).

Predictor variables included in model 1 are: basal area of gypsy moth host trees, estimated total area of defoliation in the forecasted year, number of years since last defoliation, and defoliation percent across the neighborhood in the previous three years (t-1, t-2, and t-3). In the dataset, the number of pixels in the defoliation class is very small (2,784 versus 82,930 in the no-defoliation class, Table 2-1). Therefore, the classes to be predicted by Random Forests are unbalanced. To balance them, balanced Random Forests was used (Kuhn and Johnson, 2013). This approach allows each tree to be developed using a user-determined sample size for each class. The sample sizes to be used for the development of the individual trees were set to be approximately 60% of the defoliation class (1,600 observations per class). The out of bag estimate of the overall error rate for the model was 18.9%. The confusion matrix (Table **2-3**) shows that the classification error for predicting non-defoliation (absence) is 18.9%, while the classification error for predicting defoliation (presence) is 20.2%. The most important predictor variable in model 1 is basal area of host tree for which the mean decrease in accuracy was 47.8% (Figure **2-9**). Some variables were important for predicting defoliation but not important for predicting no-defoliation. For example, estimated defoliated area was the most important variable for predicting defoliation, but it was not important for predicting no-defoliation. All six gypsy moth dynamics predictor variables have a mean decrease in accuracy for predicting defoliation larger than 10%, indicating that they are all important for predicting defoliation.

Table **2-3**: Model 1 confusion matrix

Observed \ Predicted	A	P	Classification error rates
A	67255	15675	18.90%
P	562	2222	20.19%

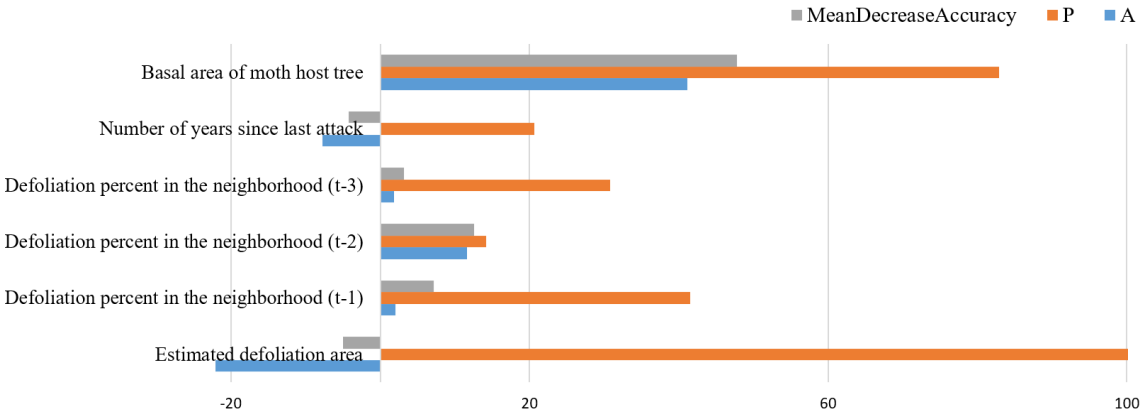

Figure **2-9**: Model 1 variable importance

Partial dependence plots for variables important for predicting defoliation are presented in Figure **2-10**. The partial dependence plot for basal area of host tree fluctuates when the basal area is small, but the probability of predicting the pixel to defoliation class is higher when the basal area of host tree is larger than 40 square feet per acre (Figure **2-10**a). The partial dependence plot for the estimated total defoliated area in the entire state indicates that low percent is associated with absence of defoliation predictions, while larger estimates are associated with presence of defoliation predictions. Values above two million hectares are unreliable as seeing by the scarcity of data in that range (rug under the figure indicates where the data lies, Figure **2-10**b). The higher the defoliation percent in the area neighboring the sample pixel in years t-1, t-2, t-3, the higher probability of classifying the pixel as defoliation in year t (Figure **2-10**cde). The shorter the time since last attack in the area neighboring the sample pixel, the higher probability of the pixel being classified as defoliation (Figure **2-10**f).

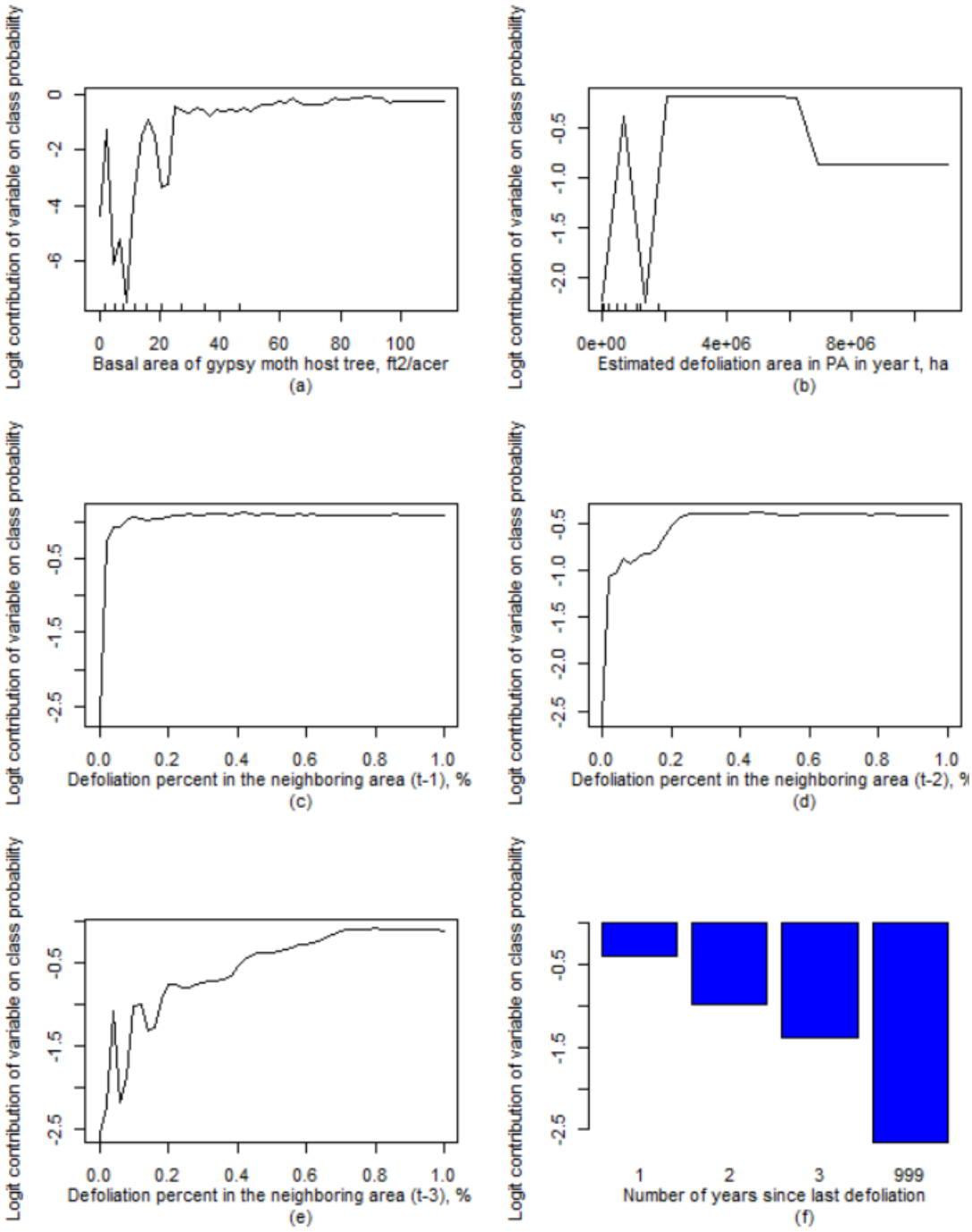

Figure **2-10** Model 1 partial dependence plots for variables predicting defoliation. Selected variables: a) basal area of host trees; b) estimated defoliated area in the forecasted year; c) defoliation percent in the neighboring area in year t-1; d) defoliation percent in the neighboring area in year t-2; e) defoliation percent in the neighboring area in year t-2; f) number of years since last defoliation in the neighboring area

Model 2 used gypsy moth dynamics variables, forest percent, geographic location, terrain, climate and IMI as predictor variables. There were 40 predictor variables in total. Random Forests parameters were set to 500 trees, the number of variables tried at each split to 13 (one third of the total number of variables), and the sample size of each class for each tree was set to 1600 to balance the classes (60% of the number of defoliation observations). After fitting the model, variables with very low mean decrease in accuracy percent (<5%), were removed and the model fit again without them. The classification error rate for model 2 is 10.2%. The confusion matrix (Table **2-4**) shows that the classification error for predicting absence of defoliation is 10.2%, improving 8.7% from model 1. The error rate for predicting presence of defoliation is 9.8%, which is 10.4% lower than that of model 1.

Table **2-4**: Model 2 confusion matrix

Observed \ Predicted	A	P	Classification error rate
A	74434	8496	10.24%
P	273	2511	9.81%

Figures **2-11** and **2-12** show the ten most important variables for predicting defoliation and no-defoliation in model 2. Figure **2-13** shows the ten most important variables for classification.

The most important variable for predicting defoliation is estimated defoliated area (Figure **2-11**). The geographic location of the sample pixels, longitude and latitude, rank second and sixth, respectively, for predicting defoliation. All temporal climate variables (forecasted year minimum and maximum January temperature, previous year minimum and maximum June temperature and previous year annual precipitation) are in the top ten important variables for predicting defoliation. The forest percent in the sample pixel ranks fourth for predicting

defoliation. Defoliation percent in the neighborhood in t-1 year is more important than in t-2 and t-3 year for predicting defoliation.

Comparing the ten most important variables, temporal variables are important for predicting the defoliation by gypsy moth, while the variable related to local environment and normal climate are important for predicting non-defoliation.

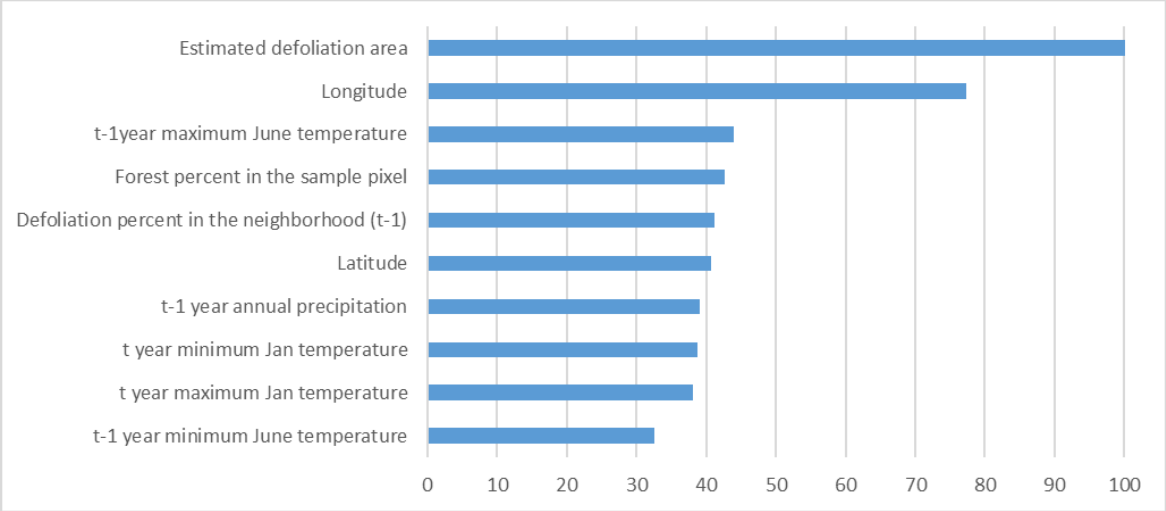

Figure **2-11**: Model 2 ten most important variables for predicting defoliation

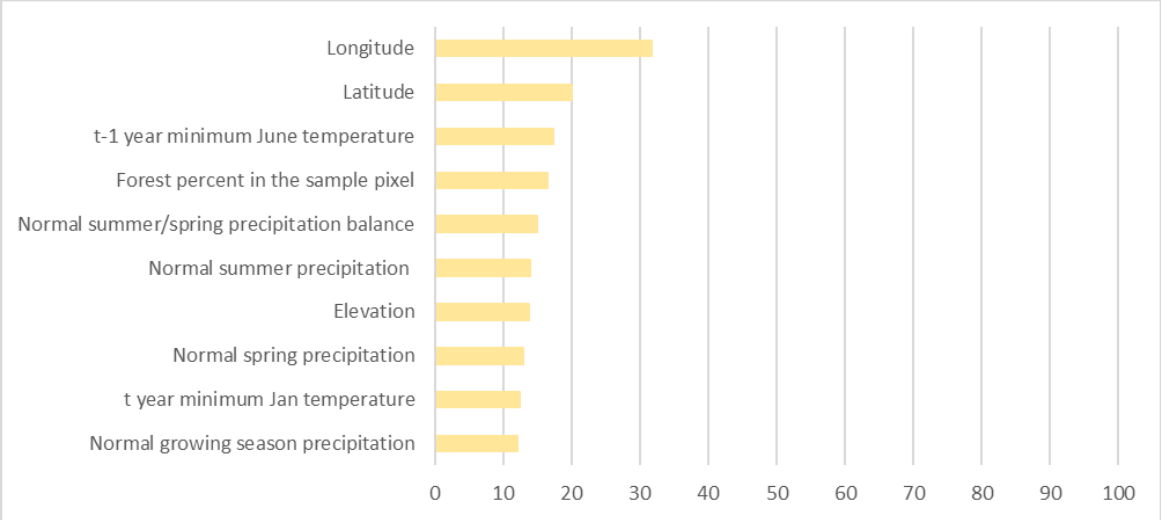

Figure **2-12**: Model 2 ten most important variables for predicting no-defoliation

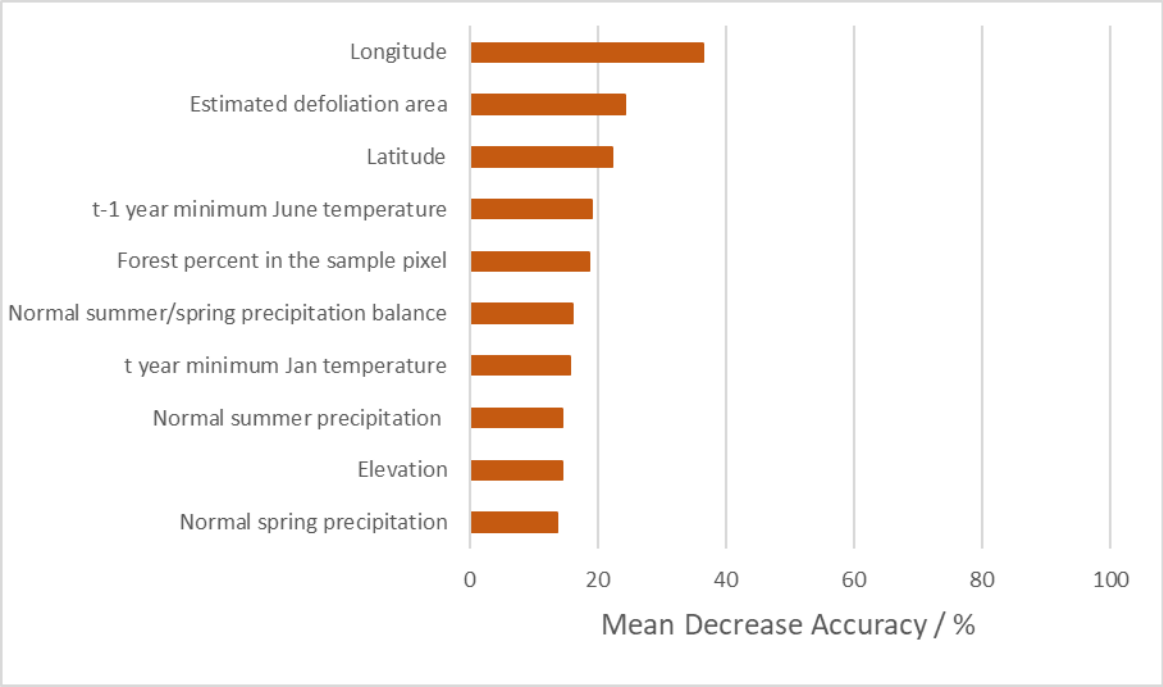

Figure **2-13**: Model 2 ten most important variables for overall classification

Geographic location is important for predicting defoliation by gypsy moth. The eastern region in Pennsylvania has higher probability to be attacked than the western region (Figure **2-14**a). This is likely associated with the distribution of different forest types in the state. The partial dependence plot for latitude shows that the pixels in the location where latitude is about 35 N and 41.5 N have higher probability to be classified to presence class (Figure **2-14**b).

The partial dependence plots for variables reflecting gypsy moth dynamics are not different from those for model 1 (Figure **2-14**cd). The partial dependence plot on defoliation percent in the neighborhood in year t shows the same trend. The same is the case for the partial dependence on estimated defoliated area (values above two million are unreliable because of lack of that in that range).

Temporal climate and normal climate variables are also important for predicting defoliation. When the maximum summer temperature is higher than 26 °C, the pixel is more likely to be classified as defoliation (Figure **2-14**e). The probability of predicting defoliation decreases when normal spring (April and May) precipitation reaches 180 mm and increases at 200 mm (Figure **2-14**f).

Topographic variables are not as important as other variables, but they still provide some improvements to classification in the model. For example, the likelihood of classification as presence of defoliation decreases when the elevation increases from 0 to 1300 ft, then increases after 1300 ft (Figure **2-14** g). The likelihood of classification as defoliation is low when TPI is 0 (Figure **2-14**h).

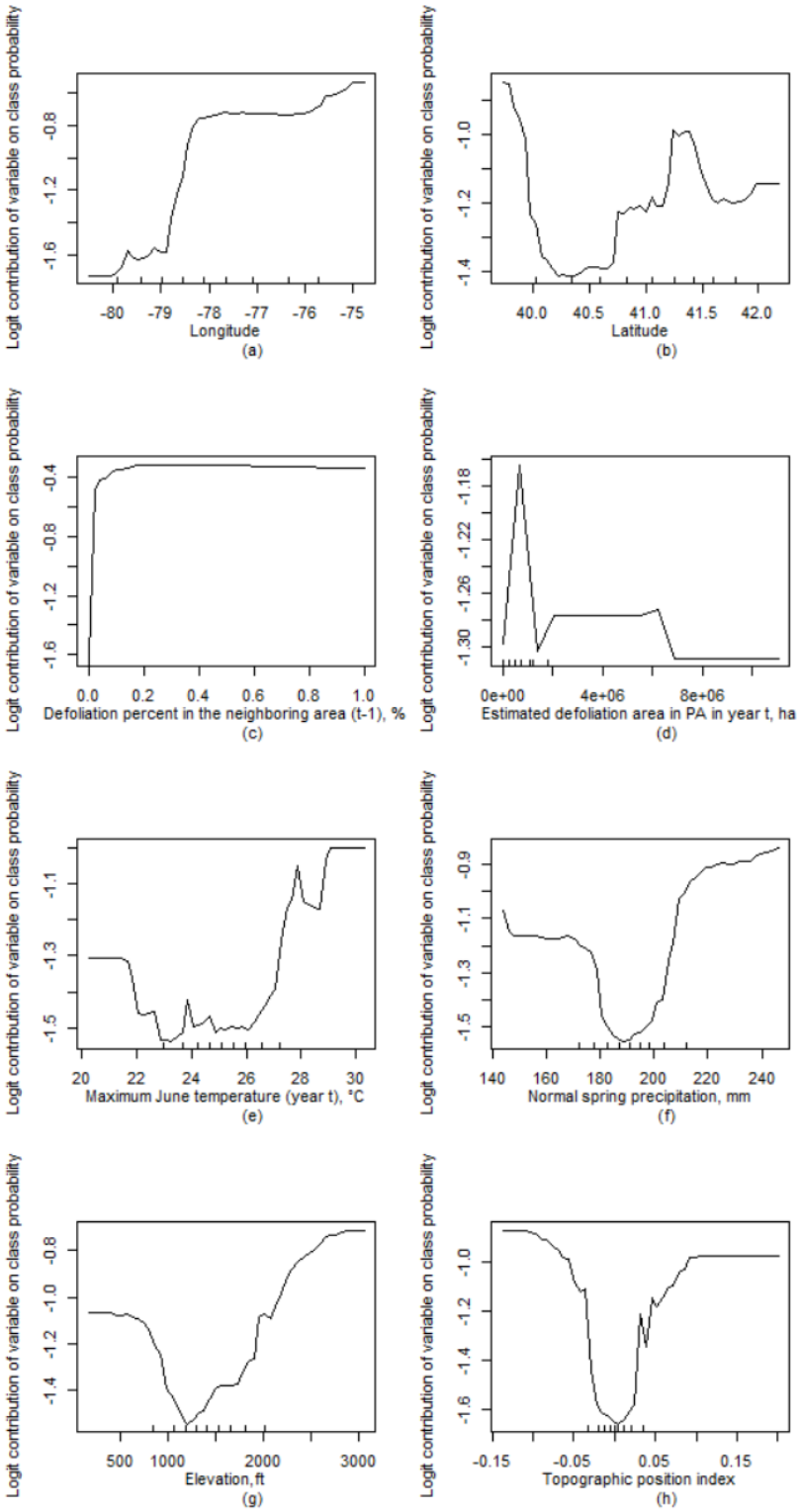

Figure **2-14**: Model 2 partial dependence plots on selected variables a) longitude; b) latitude; c) defoliation percent in the neighboring area in year t-1; d) estimated defoliated area in PA in year t; e) maximum June temperature; f) normal spring precipitation; g) elevation; h) topographic position index

To validate the full model, two approaches were used. In the first one, 200 pixels in each of the years 1985, 1990 and 1995 were randomly selected as a validation dataset. The results are presented in Table **2-5**. In 1985, the classification error for the absence of defoliation class was 15.8%, while the classification error for presence of defoliation was 37.5%. In 1990, large areas in Pennsylvania were attacked by gypsy moth. The classification errors for presence and absence of defoliation are 19.5% and 39.05% respectively. In 1995, very few areas were attacked by gypsy moth. No pixels in the sample had a recorded defoliation by gypsy moth. The classification error for two classes was 0.

Table **2-5**: Confusion matrices for validation years 1985, 1990 and 1995

(a) 1985

Predicted / Observed	A	P	Classification error rate
A	162	30	15.63%
P	3	5	37.50%

(b) 1990

Predicted / Observed	A	P	Classification error rate
A	95	23	19.49%
P	32	50	39.03%

(c) 1995

Predicted / Observed	A	P	Classification error rate
A	199	1	0%
P	0	0	0%

The hindcasting validation was supplemented with a k-fold cross-validation using year as the fold. For this, the model was trained using 16 years of the 17 available and validated with the 17^{th} year. This was done by removing one year at a time. The results are presented in Table **2-6**.

The model had good predictive ability for 8 of the 17 years, with error rates for predicting no-defoliation ranging between 8.1% to 22.1%, and error rates for predicting defoliation ranging between1.6% to 41.6%.

For years 2002, 2004, 2005, 2011, 2012, the error rates are high because of the small number of defoliated sample pixels, a few of them misclassified would result in a high percent of the total.

The validation results for year 2013 are surprisingly poor. This may be explained by the poor prediction of the autoregressive model of the estimated defoliated area for that particular year. The observed defoliated area is 67 times the estimated defoliated area in 2013. This is because the observed defoliated area for year 2012 and 2011 is 0 ha and 2 ha. The defoliation in 2013 was an eruptive event and the 2nd order autoregressive model underestimated the defoliated area for 2014.

Even though the estimated defoliated area for 2014 also had a large residual (Figure **2-5**), the validation results were better than those for 2013. The defoliated area for 2013 was underestimated, while the defoliated area for 2014 is overestimated. Underestimation of the total defoliated area decreases the likelihood of the pixels being classified as presence pixels. Another reason is that the spatial location of the defoliated area in 2013 was different from other years. The defoliated areas for the years included in the model development occur mostly in eastern PA. However, the defoliated area in 2013 did not occur in east. This points to the importance of using as many years as possible to develop the model and the need to use independent years for validation.

For year 2006 and year 2015, the error for predicting defoliation is about 50%. No clear explanation was found for this poor predictive performance.

Table **2-6**: Validation classification error rates for in each year.

Year	Number of observed absence samples	Number of observed presence samples	Error rate % (absence)	Error rate % (presence)
2000	4620	422	11.49	38.86
2001	4980	62	25.90	1.61
2002	5001	41	0.62	75.61
2003	5035	7	16.52	14.29
2004	5038	4	18.06	75.00
2005	4944	98	0.00	100.00
2006	4697	345	15.27	53.33
2007	4666	376	14.27	28.72
2008	4605	437	22.10	10.30
2009	4803	239	22.11	25.10
2010	4998	44	15.51	0.00
2011	5025	17	3.20	100.00
2012	5039	3	4.50	100.00
2013	4839	203	14.49	99.51
2014	4939	103	8.12	41.75
2015	4746	296	2.70	53.04
2016	4955	87	22.12	5.75

Discussion

The Random Forests model handled multiple correlated variables and found the important variables for predicting defoliation by gypsy moth. In this study, Random Forests performed well in classifying presence and absence of defoliation by gypsy moth. The overall error rate of model 2 is 9.81%. The error rate for predicting presence of defoliation is 9.81%; for predicting absence is 10.24%. Comparing with other classification models, Random Forests has higher accuracy. Zhou and Liebhold (1995) developed a logistic regression model to classify presence and absence. It was very difficult to set a probability threshold to discriminate presence and absence in the logistic model that would result in low classification error levels for both

presence and absence. Previous studies using linear regressions to estimate the percentage of defoliation (Liebhold et al., 1993; Williams and Liebhold, 1995c; Forster et al. 2013) did not explained a large portion of the response variance, mostly because quantifying defoliation intensity is challenging. Comparing with multiple linear regression models, logistic regression models, and non-linear discriminant functions, Random Forests was able to handle a large dataset in which there are interactions between predictors and resulted in a model with low error rates for predicting defoliation by gypsy moth.

Previous studies indicate defoliation by gypsy moth is an autocorrelated process. The defoliation series for different states identified it as a 1st order autoregressive or 2nd order autoregressive process (Williams and Liebhold, 1995ab). In this study, total defoliated area in Pennsylvania in the forecasted year was estimated by a 2nd order autoregressive model. The estimated defoliated area in the entire state was identified as the most important variable for predicting defoliation. When the estimated defoliated area increases, the probability of a sample pixel being classified as defoliation increases. As the significance of time lags are not the same in different states, defoliation in the three previous years was included in the model. Williams and Liebhold (1995a) used autoregression to model the gypsy moth defoliated area. They found that the gypsy moth defoliated area is positively autocorrelated with lag 1 defoliation and lag 2 defoliation. In the Random Forests classification model, the percent defoliation in the neighboring area (a circle with a 4 km radius) in year t-1, t-2, t-3 ranked 5th, 26th, and 36th respectively in total 40 variables (the variable importance for all variables is presented in Appendix B). The model shows that the partial dependence plot for these three variables has the same trend: as the area increases the probability of a sample pixel being classified as defoliation in the following year increases as well (Figure **2-15**).

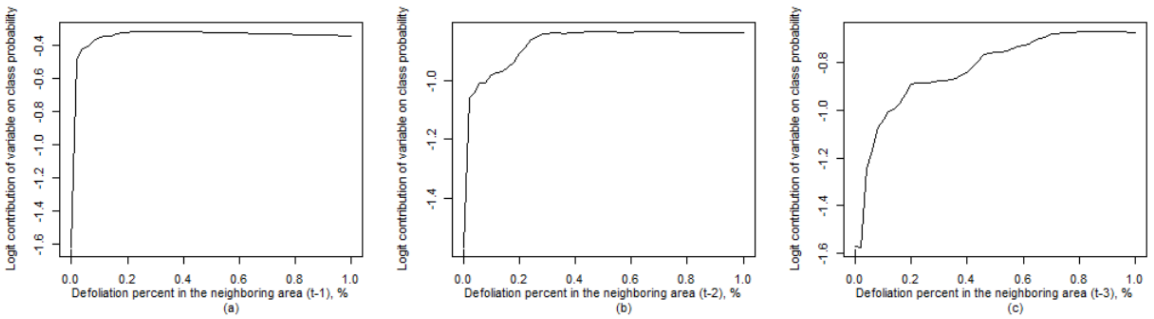

Figure **2-15**: Partial dependence plots for defoliation percent in the neighboring area in year t-1, t-2, t-3 for predicting defoliation

Forest composition is related to defoliation by gypsy moth. Previous studies provide the ranking of foliage preference by gypsy moth based on field survey and lab feeding trails (Liebhold et al., 1995b). Liebhold et al. (1993) used only one predictor, basal area of the host tree, to estimate percentage defoliation. They found a significant positive correlation between the predictor and response, with a slope of 0.003. In model 1, the basal area of the host tree is the most important variable for classifying defoliation and no-defoliation. The mean decrease in accuracy of host basal area is 47.7%, which means the model classification error rate increased 47.7% when the values of the host basal area variable are shuffled. However, the basal area of the host tree is not the most important variable only in terms of predicting defoliation. In model 1, the most important variable for predicting defoliation is estimated defoliated area.

In addition to host tree basal area, forest percent was included as a predictor in model 2. It is surprising that forest percent ranked 5th and host basal area ranked 16th in the list of variable importance for predicting defoliation. Host tree basal area represents the forest susceptibility to gypsy moth. Forest percent represents the landscape fragmentation. Intuitively, host tree basal area should be more important than forest percent. However, this is not verified in the model results. This might be caused by the spatial resolution of the data sources. Forest percent was calculated using LANDFIRE data which is generated from 30 m by 30 m Landsat images. The host tree basal area layer was a 1 km by 1 km raster interpolated from 2009 Forest Inventory

Analysis (FIA) data. Even though all layers were resampled and averaged to the same resolution of 1 km by 1km, the variables that were calculated from the high-resolution data may have more accurate values.

Climate and weather variables are important variables in model 2. The three most important weather variables are maximum June temperature in t-1, precipitation in t-1, and minimum January temperature in t (Figure **2-16**). Low winter temperature kills gypsy moth eggs and is negatively related with defoliation by gypsy moth (Sharov et al., 1999). High summer temperature decreases the mortality of gypsy moth caused by fungal infection so this is positively related with gypsy moth outbreak (Reilly et al., 2014). In model 2, results suggest the same association: high previous year maximum June temperature and high current year minimum January temperature increases the probability of a pixel being classified as defoliation (Figure **2-16**ac). The probability of a pixel being classified as defoliation decreases when previous year precipitation increases, but when the precipitation exceeds 1300 mm per year, the probability increases (Figure **2-16**b). In model 2, climate normal variables are not in the ten most important variables for predicting defoliation, but some of them are the ten most important variables for predicting no-defoliation. Summer/spring precipitation balance: (Jul+Aug)/(Apr+May), summer precipitation (Jul+Aug), spring precipitation (Apr+May) and growing season precipitation (April to September) are the four most important climate normal variables for predicting no-defoliation. IMI was also incorporated to the model but it was not important for predicting defoliation so it was removed from the final model.

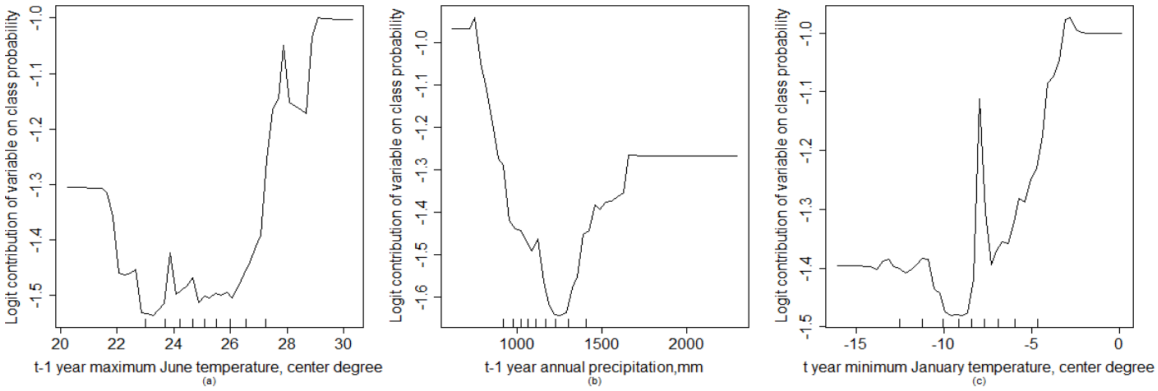

Figure **2-16**: Partial dependence plots for temporal climate variables

Geographic location is important for predicting defoliation by gypsy moth (Figure **2-11**). Pixels in the eastern region of the state are more likely to be classified as presence of defoliation than pixels in the western region (Figure **2-14**a). Elevation is the most important terrain variable in the model. Gypsy moth population growth rate declined with increased elevation due to reproductive asynchrony in Virginia and West Virginia where elevation ranges from 95m to > 1900m (Walter et al., 2015). Other terrain variables are not very important in the model. The variable importance of the topographic roughness index, slope, topographic position index, and aspect ranked 33th, 37th, 38th, 39th out of 40 variables.

The validation of model 2 shows that for years 2000, 2001, 2003, 2007, 2008, 2009, 2014 and 2016, the error rate for predicting presence of defoliation ranges from 1.6% to 41.6%, while predicting absence ranges from 8.1% to 22.1%. For the years when few or even no area in Pennsylvania was defoliated by gypsy moth, the error rates are high because of the small number of presence sample pixels. A few of them misclassified caused a high error rate percent. Comparing the validation results for 2013, in which the total defoliated area was underestimated, with those for 2014, in which the total defoliated area was overestimated, the error rate for predicting presence in 2013 was much higher than 2014, which indicates that underestimation of the defoliated area would decrease the model predictive ability for presence.

Conclusions

In this chapter, I used spatial data including terrain, climate, sketch maps, forest composition, and IMI, together with the Radom Forest algorithm to predict the defoliation by gypsy moth. The model handled a large number of variables and found the most important predictors for predicting defoliation by gypsy moth. The three most important variables for predicting defoliation are estimated defoliated area, geographic location and previous year summer temperature. The Random Forests model works effectively for processing a big training dataset and have high accuracy for classification defoliation and no-defoliation. The model validation shows the model has good predictive ability in some years. However, the model is not accurate in others. This highlights the importance of training the model with as many years as possible and of using independent years for validation.

References

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random forests for classification in ecology. Ecology, 88(11), 2783-2792.

De Reu, J., Bourgeois, J., Bats, M., Zwertvaegher, A., Gelorini, V., De Smedt, P., ... & Van Meirvenne, M. (2013). Application of the topographic position index to heterogeneous landscapes. Geomorphology, 186, 39-49.

Daly, C., Neilson, R. P., & Phillips, D. L. (1994). A statistical-topographic model for mapping climatological precipitation over mountainous terrain. Journal of Applied Meteorology, 33(2), 140-158.

Elkinton, J. S., & Liebhold, A. M. (1990). Population dynamics of gypsy moth in North America. Annual Review of Entomology, 35(1), 571-596

Evans, J. S., Murphy, M. A., Holden, Z. A., & Cushman, S. A. (2011). Modeling species distribution and change using random forest. In Predictive species and habitat modeling in landscape ecology (pp. 139-159). Springer, New York, NY.

Foster, J. R., Townsend, P. A., & Mladenoff, D. J. (2013). Spatial dynamics of a defoliation by gypsy moth outbreak and dependence on habitat characteristics. Landscape Ecology, 28(7), 1307-1320.

Ho, T. K. (1995, August). Random decision forests. In Document analysis and recognition, 1995., proceedings of the third international conference on (Vol. 1, pp. 278-282). IEEE.

Iverson, L. R., Dale, M. E., Scott, C. T., & Prasad, A. (1997). A GIS-derived integrated moisture index to predict forest composition and productivity of Ohio forests (USA). Landscape Ecology, 12(5), 331-348.

Iverson, L. R., & Prasad, A. M. (2003). A GIS-derived integrated moisture index. In: Sutherland, Elaine K.; Hutchinson, Todd F., eds. Characteristics of mixed oak forest ecosystems in southern Ohio prior to the reintroduction of fire. Gen. Tech. Rep. NE-299. Newtown Square, PA: US Department of Agriculture, Forest Service, Northeastern Research Station. 29-41.

Kuhn, M., & Johnson, K. (2013). Applied predictive modeling (Vol. 26). New York: Springer.

LANDFIRE. (2018). Forest canopy cover. www.landfire.gov. Accessed 06/12/2018

Liebhold, A. M., Simons, E. E., Sior, A., & Unger, J. D. (1993). Forecasting defoliation caused by the gypsy moth from field measurements. Environmental Entomology, 22(1), 26-32.

Liebhold, A. M., MacDonald, W. L., Bergdahl, D., & Mastro, V. C. (1995a). Invasion by exotic forest pests: a threat to forest ecosystems. Forest Science, 41(30).

Liebhold, A. M., Gottschalk, K. W., Muzika, R. M., Montgomery, M. E., Young, R., O'Day, K., & Kelley, B. (1995b). Suitability of North American tree species to gypsy moth: a summary of field and laboratory tests. Gen. Tech. Rep. NE-211. Radnor, PA: US Department of Agriculture, Forest Service, Northeastern Forest Experiment Station. 34 p.

Liebhold, A. M., Gottschalk, K. W., Mason, D. A., & Bush, R. R. (1997a). Forest susceptibility to the gypsy moth. Journal of Forestry, 95(5), 20-24.

Liebhold, A.M., Gottschalk, K.W., Mason, D.A., & Bush, R.R. (1997b). Evaluation of Forest Susceptibility to the Gypsy Moth across the Conterminous United States. Journal of Forestry 95: 20-24

Liebhold, A., Luzader, E., Reardon, R., Roberts, A., Ravlin, W. F., Sharov, A., & Zhou, G. (1998). Forecasting gypsy moth (Lepidoptera: Lymantriidae) defoliation with a geographical information system. Journal of Economic Entomology, 91(2), 464-472.

Liebhold, A., Elkinton, J., Williams, D., & Muzika, R. M. (2000). What causes outbreaks of the gypsy moth in North America? Population Ecology, 42(3), 257-266.

Moran, P. A. P. (1953). The statistical analysis of the Canadian Lynx cycle. Australian Journal of Zoology, 1(3), 291-298.

Oliveira, S., Oehler, F., San-Miguel-Ayanz, J., Camia, A., & Pereira, J. M. (2012). Modeling spatial patterns of fire occurrence in Mediterranean Europe using Multiple Regression and Random Forest. Forest Ecology and Management, 275, 117-129.

Oyler, J. W., Ballantyne, A., Jencso, K., Sweet, M., & Running, S. W. (2015). Creating a topoclimatic daily air temperature dataset for the conterminous United States using homogenized station data and remotely sensed land skin temperature. International Journal of Climatology, 35(9), 2258-2279.

Peters, M., Iverson, L. R., & Prasad, A. M. (2010). Using an intergrated moisture index to assess forest composition and productivity. Chapter 11. In: Eredics, Peter, ed. Mapping Forsetry. Redlands, CA: ESRI Press: 41-44.

Rehfeldt, G. E. (2006). A spline model of climate for the western United States (p. 21). Fort Collins, CO: US Department of Agriculture, Forest Service, Rocky Mountain Research Station. Available at charcoal.cnre.vt.edu/climate.

Reilly, J. R., Hajek, A. E., Liebhold, A. M., & Plymale, R. (2014). Impact of Entomophaga maimaiga (Entomophthorales: Entomophthoraceae) on outbreak gypsy moth populations (Lepidoptera: Erebidae): the role of weather. Environmental entomology, 43(3), 632-641.

Sharov, A. A., Pijanowski, B. C., Liebhold, A. M., & Gage, S. H. (1999). What affects the rate of gypsy moth (Lepidoptera: Lymantriidae) spread: winter temperature or forest susceptibility? Agricultural and Forest Entomology, 1(1), 37-45.

The Pennsylvania State Climatologist. (2018). State wide data. www.climate.psu.edu/data/state/. Accessed 06/12/2018

Walter, J. A., Meixler, M. S., Mueller, T., Fagan, W. F., Tobin, P. C., & Haynes, K. J. (2015). How topography induces reproductive asynchrony and alters gypsy moth invasion dynamics. Journal of Animal Ecology, 84(1), 188-198.

Williams, D. W., & Liebhold, A. M. (1995a). Influence of weather on the synchrony of gypsy moth (Lepidoptera: Lymantriidae) outbreaks in New England. Environmental Entomology, 24(5), 987-995.

Williams, D. W., & Liebhold, A. M. (1995b). Detection of delayed density dependence: effects of autocorrelation in an exogenous factor. Ecology, 76(3), 1005-1008.

Williams, D. W., & Liebhold, A. M. (1995c). Forest defoliators and climatic change: potential changes in spatial distribution of outbreaks of western spruce budworm (Lepidoptera: Tortricidae) and gypsy moth (Lepidoptera: Lymantriidae). Environmental Entomology, 24(1), 1-9.

Zhou, G., & Liebhold, A. M. (1995). Forecasting defoliation by gypsy moth with a geographical information system. Insect Science, 2(1), 83-94.

Chapter 3

Improving the forecasting model results using remote sensing and egg mass data

Introduction

Gypsy moth outbreaks are notoriously difficult to predict (Liebhold and Elkinton, 1989). The population of gypsy moth may stay at a low level for years and abruptly erupt on a large scale (Liebhold et al., 1995). Remote sensing data have been used to monitor the defoliation caused by gypsy moth imagery (Townsend et al. 2004; De Beurs et al. 2008; Spruce et al. 2011; Thayn, 2013). Egg mass field surveys are the standard method for predicting defoliation in individual stands in gypsy moth management programs. Remotely sensed data and egg mass density data are potential predictors for modeling and forecasting the frequency and degree of defoliation by gypsy moth.

MODIS (Moderate Resolution Imaging Spectro-Radiometer) time-series data, provided since 2000, has been widely used to analyze insect damage (Eklundh et al., 2009; Jepsen et al., 2009; Spruce et al., 2011). De Berus et al. (2008) used MODIS data to calculate several vegetation indices before and during the gypsy moth outbreak. These vegetation indices were used as variables to develop linear regression models to predict biomass loss estimates derived from Landsat. Spruce et al. (2011) assessed MODIS NDVI time series data products for detecting forest defoliation by gypsy moth outbreaks. They concluded that MODIS-based products provide spatial information on defoliation severity of gypsy moth defoliated area, which can be compared with information from sketch maps and help to estimate annual defoliation.

Remotely sensed data, then, can be used to detect defoliation locations and severity. Defoliation in the previous year may affect gypsy moth survival the following year. Liebhold et al. (1993) studied the relationship between some pre-season measurements and defoliation by gypsy moth in the subsequent year. They measured field variables such as egg density, larval density and egg mass density to forecast defoliation severity. Previous year defoliation may be related to the forest health status and disturbance resistance. During gypsy moth outbreaks, defoliation commonly persists for two or more years and then declines after defoliation reaches its peak (Liebhold and Elkinton, 1989). Using remote sensing variables that reflect previous year's defoliation may have predictive ability.

The two most common measurements for sampling gypsy moth population are counts of male moths in pheromone traps and counts of over-wintering egg mass populations (Liebhold et al., 1994). Egg masses are censused by sampling in plots with fixed-radius, or a variable-radius (Liebhold et al., 1994). Egg mass density is calculated from egg mass data, expressing the numbers of egg masses per unit land area (Liebhold et al., 1994). Semi-variograms derived from egg mass samples typically exhibit nugget-effect discontinuous densities at short distances. It is usually caused by nonuniform distribution of the censused objectives. This indicates that egg mass density cannot be analyzed at a fine scale (Zhou and Liebhold, 1994). Interpolating egg mass density to the raster with a spatial resolution of 1 km by 1 km is appropriate and can add value to other variables used for predicting defoliation in Chapter 2.

Egg mass density is positively correlated with the subsequent percent defoliation (Liebhold et al., 1994). Defoliation by gypsy moth typically occurs in forest stands where gypsy moth egg mass densities exceed 6,000 per hectare (Zhou and Liebhold, 1995). Gypsy moth egg mass exist in the field from late summer through spring of the following year, and this provides sufficient time for censusing egg mass population in order to assess the need for pest control treatment (Liebhold et al, 1994). Several previous studies used egg mass data to forecast

defoliation by gypsy moth (Leibhold et al., 1993; Zhou and Liebhold, 1994; Gansner et al., 1995). Zhou and Liebhold (1995) developed logistic regression models to forecast defoliation by gypsy moth using egg mass density, the presence or absence of defoliation in the previous year, distance to the nearest defoliated area in the previous year, and male moth density. Their model results indicated that all coefficients of these four variables were significant.

The purpose of this chapter is to explore additional data sources that may show promise for predicting defoliation by gypsy moth. Remote sensing and gypsy moth egg mass survey data are provided for recent years. To evaluate the importance of remote sensing variables and egg mass survey data for prediction, I use a subset of years within the 2000-2016 period for which egg mass data was available.

Methodology

Data sources and derived variables

Remotely sensed imagery derived variables and gypsy moth egg mass density were used as predictors. MODIS-based products and egg mass survey data are identified as data sources.

Remote sensing variables

MODIS-based remote sensing phenology metrics (phenology.cr.usgs.gov), provided by the U.S. Geological Survey Land Remote Sensing Program from 2000 to present were used. The essence of remote sensing phenology is tracking the annual trajectory of normalized difference vegetation index (NDVI) change and extracting phenological information. NDVI trajectory has

been used to simulate plant community photosynthesis functions that describe vegetation seasonal dynamics (Reed et al., 2009).

Variables included in the model were: previous year time-integrated NDVI (TIN), end of season time (EOST), and end of season NDVI (EOSN). The methods used to generate the metrics include smoothing the NDVI trajectory and tracking the NDVI change (Figure **3-1**).

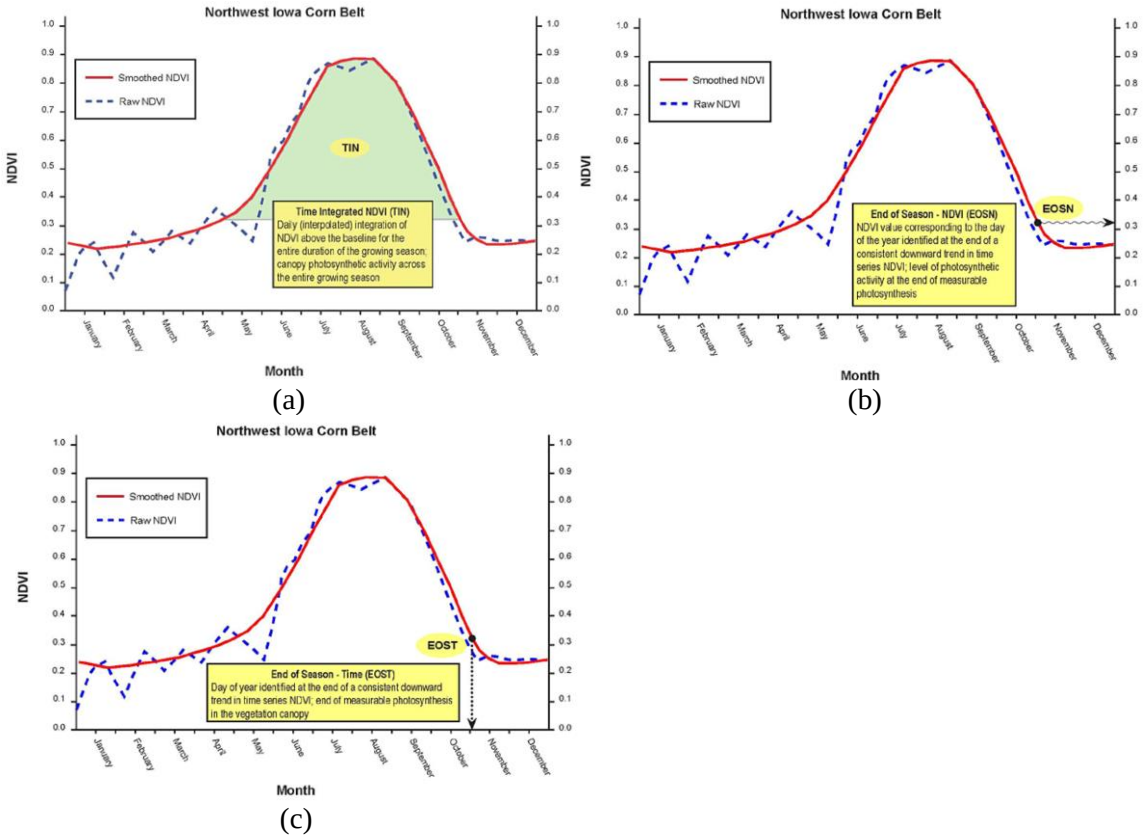

Figure **3-1**: Methods of tracking NDVI trajectory to calculate a) TIN, b) EOSN, c) EOST metrics (pictures accessed from phenology.cr.usgs.gov/methods_metrics.php)

Egg mass survey data

Gypsy moth mass survey data was provided by the DCNR-BoF, Division of Forest Health. Gypsy moth egg mass censuses were conducted from late summer to the following year spring. The completed egg mass survey data shapefile for Pennsylvania was organized for year

2010, 2013, 2014, 2015 and 2016. Ordinary kriging was used to interpolate an egg mass density surface (Liebhold et al., 1991, Figure **3-2**).

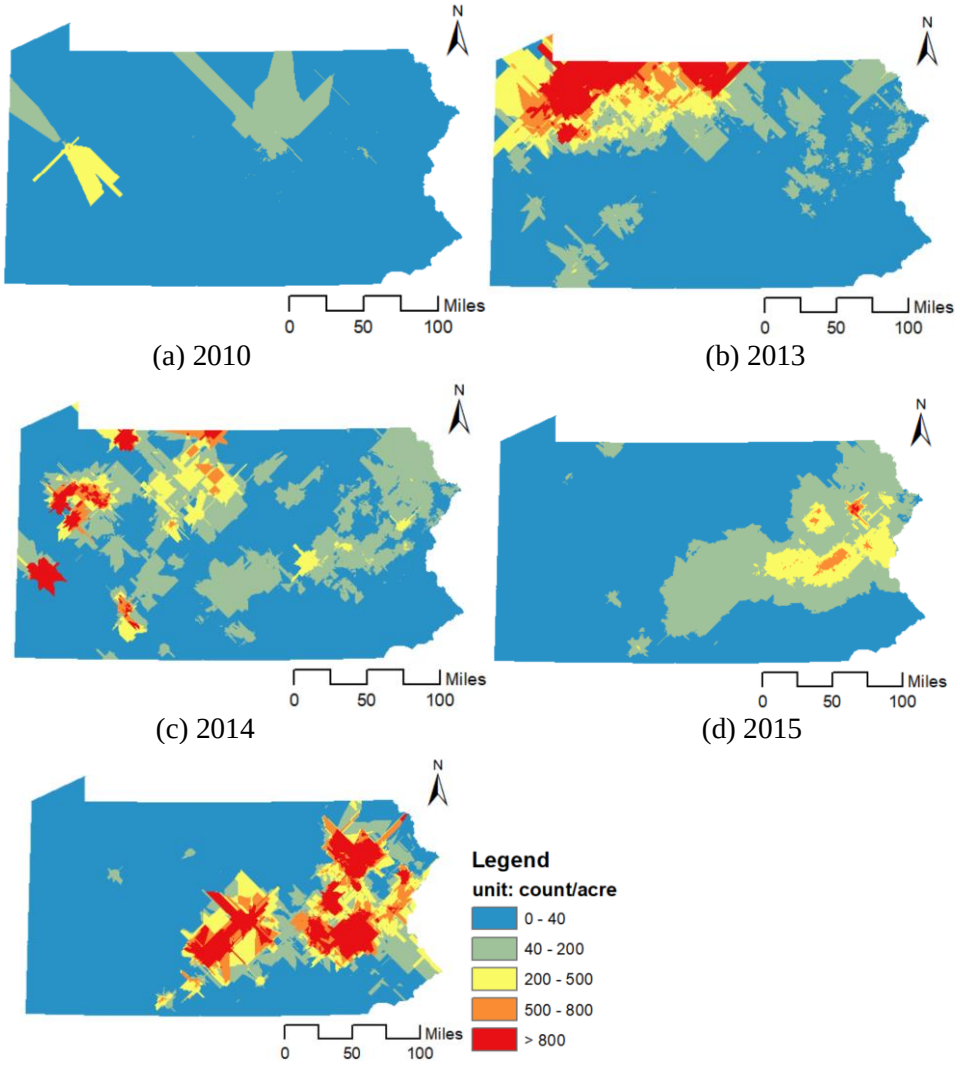

Figure **3-2**: Interpolated egg mass density metrics in Pennsylvania for five years

Model validation

To validate the model, an approach different form that of Chapter 2 was used. In Chapter 2, the forecasting model was validated by predicting defoliation in an independent year. However, since the egg mass data were only available for five years, excluding a year for validation would

greatly reduce the training dataset. Therefore, 200 pixels from the total of 5,042 sample pixels were selected for each year to validate the model (1,000 sample pixels total) and fit the model with the rest of the pixels.

Results

The pre-spring egg mass density in the defoliated area is higher than in the no-defoliated area (Figure **3-2**a). The mean density in the no-defoliated area is 98.95 egg mass per acre, while the mean density in the defoliated area is 584.42 egg mass per acre. This difference is statistically significant (t-test, p-value < 2.2*10e-16).

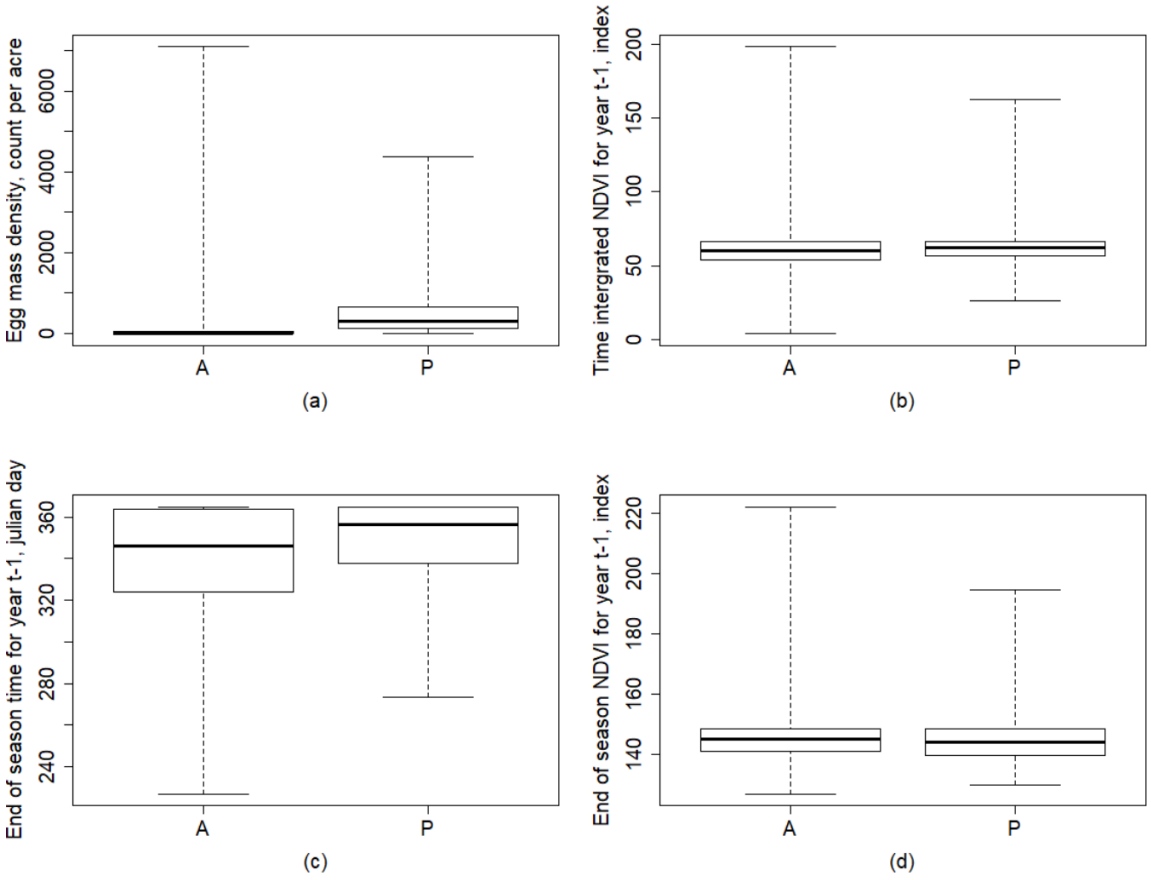

Figure **3-3**: Box plots of egg mass density and remote sensing phenological variables in areas where the defoliation is absent (A) or present (P)

The remote sensing phenology in the defoliated area and no-defoliated area are slightly different (Figure **3-2**bcd). The t-test shows the differences are significant (alpha = 0.001) but not biologically meaningful. The mean time-integrated NDVI for the defoliated sample pixels is 62.2, slightly higher than that in the no-defoliation sample pixels 60.6. The previous year growing season in the defoliated area ends on average one week later than the no-defoliated area. The end of season NDVI for the previous year for the defoliated sample pixels is lower than for the no-defoliation sample pixels.

The fit dataset comprises 24,210 instances (4,842 pixels x 5 years). In this dataset, the sizes of the defoliation class and no-defoliation class are unbalanced. The number of defoliated sample pixels is 700, while the number of no-defoliated sample pixels is 23,510. The classes were balanced by setting the sample size in Random Forests to (400, 400), about 60% of the total defoliated sample pixels. The model thus fits each classification tree with 400 randomly selected defoliation instances and 400 randomly selected no-defoliation instances.

Model 2 included the same 38 variables used in Chapter 2 but the years included were reduced to those for which egg mass data was available. Model 3 included the 38 variables and remote sensing variables (previous year TIN, EOSN, EOST). Model 4 included 38 variables and egg mass density. Model 5 used 38 variables, remote sensing variables, and egg mass density. Table 3-1 compares the performance of the four models.

For Model 2, the out of bag estimate of the overall classification error rate is 9.2%. The confusion matrix (Table **3-1**a) shows the classification error rate for predicting no-defoliation (absence) is 9.3%, while the classification error rate for predicting defoliation (presence) is 7.9%. The classification error rate for model 3 is 9.3%, which indicates that the remote sensing predictors did not improve the predictive ability of the model (Table **3-1**b). The classification error rate for model 4 is 9.0%, (Table **3-1**c). Adding egg mass density as a predictor improves the

accuracy of the fitting model. The classification error rate for model 5 is 9.0%. Model 5 has the lowest error rate, 6.1%, for predicting defoliation out of the four models.

Table **3-1**: Confusion matrices for models 2, 3, 4 and 5

(a) Model 2, without egg mass density and remote sensing variables

Observed \ Predicted	A	P	Classification error rate
A	21330	2180	9.27%
P	55	645	7.85%

(b) Model 3, with remote sensing variables

Observed \ Predicted	A	P	Classification error rate
A	21323	2187	9.30%
P	56	644	8.00%

(c) Model 4, with egg mass density

Observed \ Predicted	A	P	Classification error rate
A	21385	2125	9.04%
P	46	654	6.57%

(d) Model 5, with egg mass density and remote sensing variables

Observed \ Predicted	A	P	Classification error rate
A	21371	2139	9.10%
P	43	657	6.14%

The ten most important variables of model 5 for predicting defoliation, no-defoliation, and overall classification are shown in Figures **3-4**, **3-5**, and **3-6**. Egg mass density is the most important variable for predicting defoliation. The mean decrease in accuracy of egg mass density for predicting defoliation is 79.3% (Figure **3-4**). The remote sensing variables are not in the most ten important variables (Figure **3-5**). EOST, EOSN and TIN rank 21th, 32th, 38th out of 42 variables for predicting defoliation. However, TIN ranks 8th for predicting no-defoliation. The

egg mass density and remote sensing variables are not among the ten most important variables for the overall classification (Figure **3-6**).

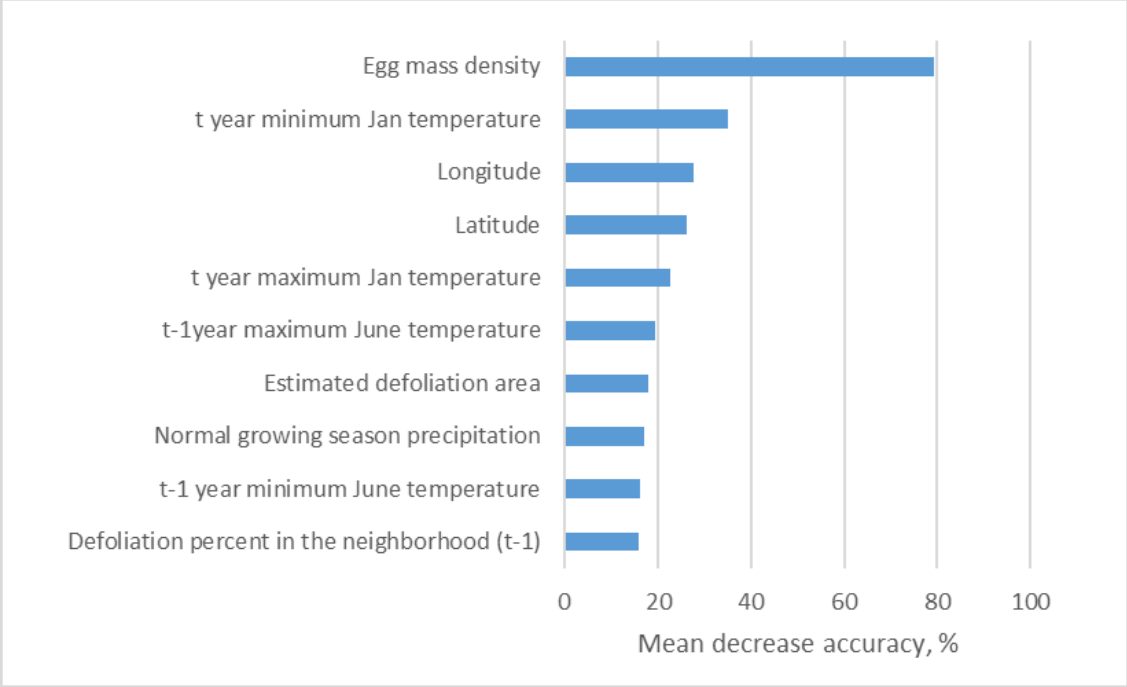

Figure **3-4**: Model 5 ten most important variables for predicting defoliation

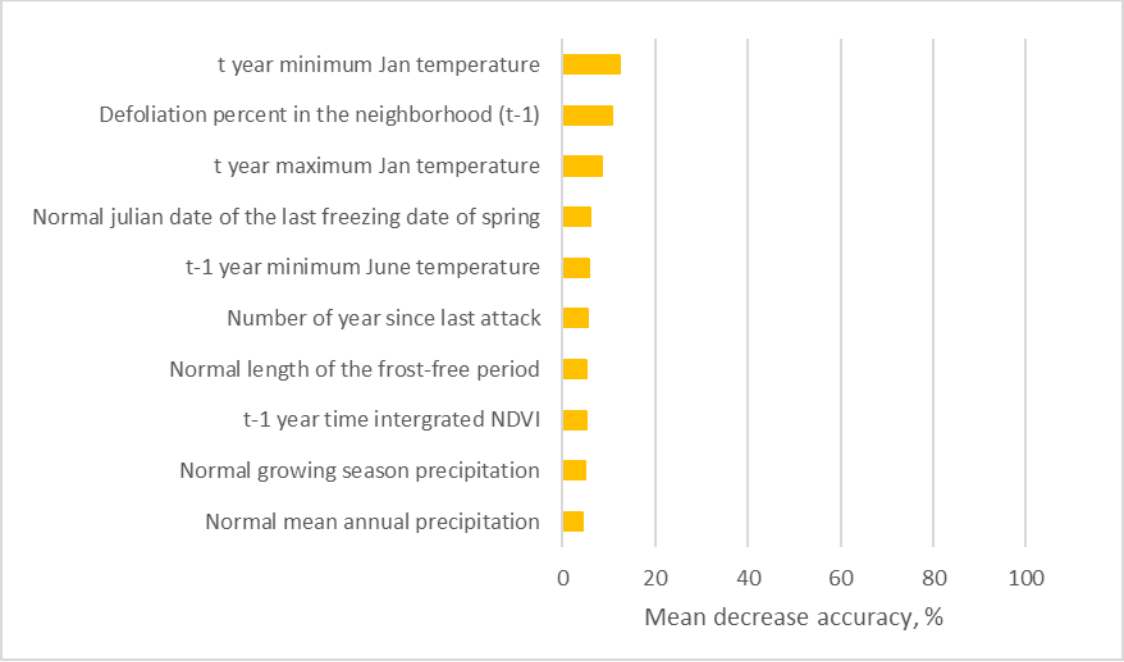

Figure **3-5**: Model 5 ten most important variables for predicting non-defoliation

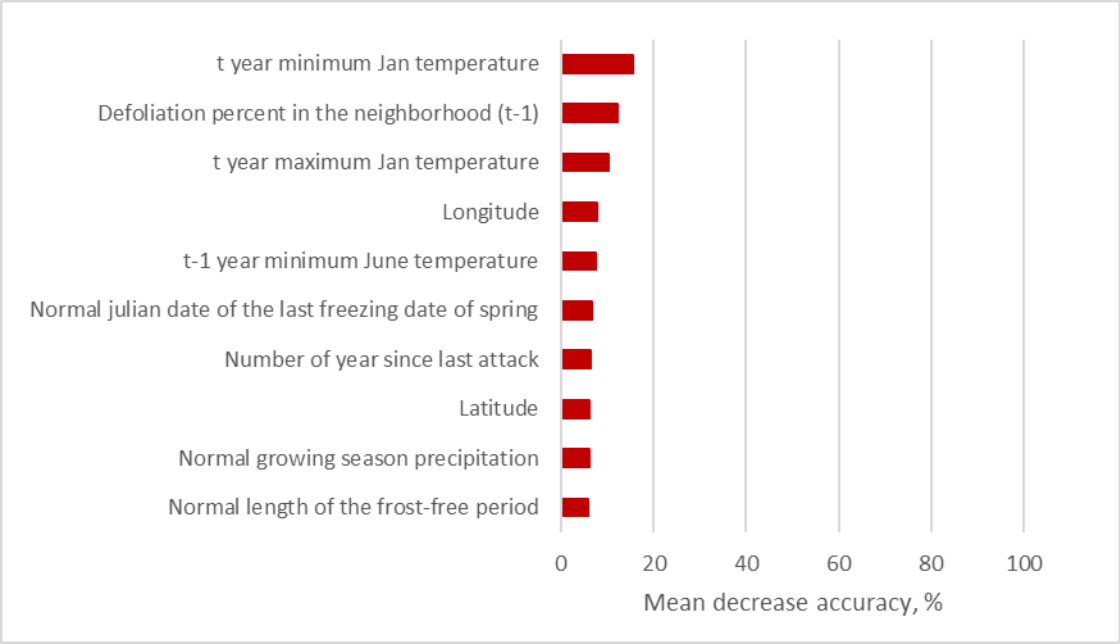

Figure **3-6**: Model 5 ten most important variables for overall classification

Table **3-2** shows the validation results for all models. The egg mass density variable slightly improved the model ability to predict no defoliation (Table **3-2**cd). Remote sensing variables did not change the model's prediction ability. Even though the classification error rates for predicting defoliation in models 4 and 5 are slightly higher than for model 2, it is only the result of the small number of observations (one single misclassified observation has a large weight in the percent calculation).

Table **3-2**: Confusion matrices for validation of models 2, 3, 4, and 5

(a) Model 2 validation, without egg mass density and remote sensing variables

Observed \ Predicted	A	P	Classification error rate
A	867	100	10.34%
P	1	32	3.03%

(b) Model 3 validation, with remote sensing variables

Observed \ Predicted	A	P	Classification error rate
A	867	100	10.34%
P	1	32	3.03%

(c) Model 4 validation, with egg mass density

Observed \ Predicted	A	P	Classification error rate
A	876	91	9.41%
P	2	31	6.06%

(d) Model 5 validation, with egg mass density and remote sensing variables

Observed \ Predicted	A	P	Classification error rate
A	875	92	9.51%
P	2	31	6.06%

Discussion

Several previous studies (De Beurs et al., 2008; Eklunndh et al., 2009; Spruce et al., 2011) used MODIS data to map, estimate, and monitor defoliation caused by gypsy moth. De Beurs et al. (2008) found that the map of defoliation by gypsy moth generated from a defoliation index calculated from MODIS products closely matched the sketch map. To forecast the defoliation by gypsy moth, previous year MODIS phenological variables were used as predictors and the link between the previous year remote sensing phenology and the defoliation was

explored. In this study, none of the remotely sensed variables made a substantial improvement in the model predictive accuracy.

Egg mass density is positively related with the percentage of defoliated area in the following year (Liebhold et al., 1993; Liebhold et al., 1994; Zhou and Liebhold, 1995). In this study, the partial dependence plot for egg mass density for predicting defoliation indicates that the probability of a pixel being classified as defoliation increases with increased egg mass density. The probability of a pixel being classified as defoliation arrives at a stable high level when the egg mass density is higher than about 500 egg mass per acre. The egg mass density is an important measurement for censusing the gypsy moth egg population. It is highly correlated with the larva population and mature moth population. Even though egg mass density is the most important variable for predicting defoliation, it is not important for predicting no defoliation.

In Chapter 2, including more variables in model 2 improved classification error rates compared to model 1. In model 2, the overall error rate decreases 8.5% with respect to model 1; while the error rate of predicting presence decreases 10.4%; and that for predicting absence improves by 8.7%. Comparing model 2 and model 4 in Chapter 3, the overall accuracy improves only 0.9% because of adding egg mass density variables. The remote sensing variables did not improve the model predictive ability.

The classification error rate for the models in this chapter and their validation are higher than those for the models in Chapter 2. In the latter, an independent year was used to validate the models. In Chapter 3, however, independent sample pixels within years included in the model development were used. This explains the similarity of classification error rates for the fit and validation datasets; the datasets are not independent. The remote sensing and egg mass variables improve the model prediction ability for no-defoliation, but the validation results show no change in the predictive ability. The egg mass density data are only provided for five years in which

Pennsylvania had not been defoliated on a large scale. I expect an improved predictive ability with a larger training dataset generated over more years.

Conclusions

In this chapter, remote sensing and egg mass variables were added as predictors, together with terrain, climate, sketch maps and host tree basal area data to build a Random Forests model to forecast defoliation by gypsy moth. The variable importance of egg mass density and remote sensing variables for predicting defoliation were compared with other variables. The egg mass density was the most important variable for predicting defoliation. All Random Forests models fitted had low classification error rates. The model validation results indicate the models have a relatively good predictive ability. The training dataset size was constrained by the availability of remote sensing and egg mass data. I would expect remote sensing and egg mass variables would improve the predictive ability of the models if longer time series were available.

References

De Beurs, K. M., & Townsend, P. A. (2008). Estimating the effect of defoliation by gypsy moth using MODIS. Remote Sensing of Environment, 112(10), 3983-3990.

Eklundh, L., Johansson, T., & Solberg, S. (2009). Mapping insect defoliation in Scots pine with MODIS time-series data. Remote Sensing of Environment, 113(7), 1566-1573.

Elkinton, J. S., & Liebhold, A. M. (1990). Population dynamics of gypsy moth in North America. Annual Review of Entomology, 35(1), 571-596.

Gansner, D. A., Herrick, O. W., & Ticehurst, M. (1985). A method for predicting defoliation by gypsy moth from egg mass counts. Northern Journal of Applied Forestry, 2(3), 78-79.

Jepsen, J. U., Hagen, S. B., Høgda, K. A., Ims, R. A., Karlsen, S. R., Tømmervik, H., & Yoccoz, N. G. (2009). Monitoring the spatio-temporal dynamics of geometrid moth outbreaks in birch forest using MODIS-NDVI data. Remote Sensing of Environment, 113(9), 1939-1947.

Liebhold, A. M., & Elkinton, J. S. (1989). Characterizing spatial patterns of gypsy moth regional defoliation. Forest Science, 35(2), 557-568.

Liebhold, A. M., Zhang, X., Hohn, M. E., Elkinton, J. S., Ticehurst, M., Benzon, G. L., & Campbell, R. W. (1991). Geostatistical analysis of gypsy moth (Lepidoptera: Lymantriidae) egg mass populations. Environmental Entomology, 20(5), 1407-1417.

Liebhold, A. M., Simons, E. E., Sior, A., & Unger, J. D. (1993). Forecasting defoliation caused by the gypsy moth from field measurements. Environmental Entomology, 22(1), 26-32.

Liebhold, A., Thorpe, K., Ghent, J., & Lyons, D. B. (1994). Gypsy moth egg mass sampling for decision-making: a user's guide. USDA-Forest Service, NA-TP-04-94.

Liebhold, A. M., MacDonald, W. L., Bergdahl, D., & Mastro, V. C. (1995). Invasion by exotic forest pests: a threat to forest ecosystems. Forest Science, 41(30).

Reed, B. C., Schwartz, M. D., & Xiao, X. (2009). Remote sensing phenology. In Phenology of ecosystem processes (pp. 231-246). Springer, New York, NY.

Spruce, J. P., Sader, S., Ryan, R. E., Smoot, J., Kuper, P., Ross, K., ... & Hargrove, W. (2011). Assessment of MODIS NDVI time series data products for detecting forest defoliation by gypsy moth outbreaks. Remote Sensing of Environment, 115(2), 427-437.

Thayn, J. B. (2013). Using a remotely sensed optimized Disturbance Index to detect insect defoliation in the Apostle Islands, Wisconsin, USA. Remote Sensing of Environment, 136, 210-217.

Townsend, P. A., Eshleman, K. N., & Welcker, C. (2004). Remote sensing of defoliation by gypsy moth to assess variations in stream nitrogen concentrations. Ecological Applications, 14(2), 504-516.

Zhou, G., & Liebhold, A. M. (1995). Forecasting defoliation by gypsy moth with a geographical information system. Insect Science, 2(1), 83-94.

Appendix A

List of 20 common host species by gypsy moth

(Liebhold, et al. 1997ab)

Common names	**Species**
White oak	*Quercus alba*
Sweetgum	*Liquidambar styraciflua*
Quaking aspen	*Populus tremuloides*
Northern red oak	*Quercus rubra*
Black oak	*Quercus velutina*
Chestnut oak	*Quercus prinus*
Post oak	*Quercus stellata*
Water oak	*Quercus nigra*
Paper birch	*Betula papyrifera*
Southern red oak	*Quercus falcata*
Scarlet oak	*Quercus coccinea*
Western larch	*Larix occidentalis*
Laurel oak	*Quercus laurifolia*
Bigtooth aspen	*Populus grandidentata*
Tanoak	*Notholithocarpus densiflorus*
Willow oak	*Quercus phellos*
California red oak	*Quercus kelloggi*

Appendix A (continued)

Common names	**Species**
Eastern hophornbeam	*Ostrya virginiana*
Canyon live oak	*Quercus chrysolepis*
Bur oak	*Quercus macrocarpa*

Appendix B

Variable importance for model 2

Variable	A	P	Mean Decrease Accuracy	Mean Decrease Gini
Forest Percent	16.48	42.72	18.74	67.56
Longitude	31.69	77.37	36.37	149.96
Latitude	19.96	40.67	22.18	72.25
Estimated defoliated area	5.95	111.81	24.19	177.97
Defoliation percent in the neighborhood (t-1)	0.62	41.21	7.56	209.97
Defoliation percent in the neighborhood (t-2)	6.11	8.98	6.64	11.85
Defoliation percent in the neighborhood (t-3)	5.70	16.33	6.67	11.67
Number of year since last defoliation	-0.11	20.69	5.36	157.90
Basal area of host trees	10.90	17.65	11.62	37.18
Elevation	13.76	20.75	14.51	24.31
Slope	8.44	8.66	9.10	17.36
Aspect	5.58	2.05	5.81	20.44
Topographic position index	5.98	7.06	6.41	21.89
Topographic ruggedness index	8.09	10.73	8.72	17.24
Integrated moisture index	8.95	-1.05	8.94	20.29
t year maximum January temperature	9.85	38.13	13.33	54.68
t year minimum January temperature	12.43	38.74	15.59	57.91
t-1year maximum June temperature	9.76	43.90	12.61	44.98
t-1year minimum June temperature	17.32	32.55	19.13	58.85
t-1 year annual precipitation	4.97	39.02	7.99	45.88

Appendix B (continued)

Variable	**A**	**P**	**Mean Decrease Accuracy**	**Mean Decrease Gini**
Mean annual temperature*	6.56	9.90	6.84	6.58
Mean annual precipitation*	9.96	22.48	10.93	19.84
Growing season precipitation, April to September*	12.12	24.32	13.16	20.62
Mean temperature in the coldest month*	6.53	10.50	6.80	8.19
Mean minimum temperature in the coldest month*	8.65	16.52	9.19	15.57
Mean temperature in the warmest month*	7.09	11.60	7.30	8.00
Mean maximum temperature in the warmest month*	10.10	15.94	10.53	12.55
Julian date of the last freezing date of spring*	8.78	18.84	9.34	16.62
Julian date of the first freezing date of autumn*	10.04	16.01	10.56	14.37
Length of the frost-free period*	11.96	18.09	12.42	18.90
Degree-days >5 degrees C*	9.13	12.95	9.42	11.29
Degree-days >5 degrees C accumulating within the frost-free period*	10.93	15.44	11.25	14.51
Julian date the sum of degree-days >5 degrees C reaches 100*	6.24	11.34	6.50	6.45
Degree-days <0 degrees C*	9.13	16.58	9.55	13.52
Degree-days <0 degrees C*	11.65	21.20	12.05	22.87
Summer precipitation balance: (jul+aug+sep)/(apr+may+jun)*	8.01	19.25	8.60	14.59
Summer/Spring precipitation balance: (jul+aug)/(apr+may) *	15.01	24.60	16.08	34.94

* are climate normals variables

Appendix B (continued)

Variable	**A**	**P**	**Mean Decrease Accuracy**	**Mean Decrease Gini**
Spring precipitation (apr+may)*	12.85	22.96	13.68	20.86
Summer precipitation (jul+aug)*	13.96	21.09	14.56	22.39
Winter precipitation (nov+dec+jan+feb)*	10.65	19.70	11.37	16.97

* are climate normals variables

Appendix C

Variable importance for model 5

Variable	A	P	Mean Decrease Accuracy	Mean Decrease Gini
Forest Percent	4.31	15.94	5.41	7.08
Longitude	3.86	27.59	7.72	27.10
Latitude	3.69	26.29	6.12	19.40
Estimated defoliated area	2.64	17.84	5.73	13.93
Defoliation percent in the neighborhood (t-1)	10.79	16.00	12.19	16.42
Defoliation percent in the neighborhood (t-2)	-3.99	11.25	-2.48	4.95
Defoliation percent in the neighborhood (t-3)	-5.05	12.43	-2.65	4.36
Number of year since last defoliation	5.53	11.63	6.48	6.85
Basal area of host trees	0.81	10.34	1.63	6.22
Elevation	1.38	11.88	2.47	4.06
Slope	2.83	10.05	3.73	4.50
Topographic position index	2.90	2.56	3.13	4.03
Topographic ruggedness index	3.18	9.54	3.89	4.23
t year maximum January temperature	8.46	22.53	10.12	11.45
t year minimum January temperature	12.43	35.17	15.52	23.09
t-1year maximum June temperature	1.06	19.29	3.13	6.64
t-1year minimum June temperature	5.66	16.28	7.55	13.29
t-1 year annual precipitation	3.12	15.40	3.96	5.41
t-1 year time integrated NDVI	5.14	8.15	5.57	4.40
t-1 year end of season time	3.65	11.85	4.45	4.57

Appendix C (continued)

Variable	**A**	**P**	**Mean Decrease Accuracy**	**Mean Decrease Gini**
t-1 year end of season NDVI	3.78	9.32	4.50	4.96
Mean annual temperature*	1.51	7.25	2.06	1.54
Mean annual precipitation*	4.33	12.34	4.94	6.31
Growing season precipitation, April to September*	4.94	17.01	6.02	9.76
Mean temperature in the coldest month*	2.26	9.22	2.72	2.22
Mean minimum temperature in the coldest month*	3.37	9.03	4.01	2.77
Mean temperature in the warmest month*	3.57	6.87	3.86	1.72
Mean maximum temperature in the warmest month*	1.83	9.82	2.35	2.43
Julian date of the last freezing date of spring*	5.86	10.15	6.56	4.61
Julian date of the first freezing date of autumn*	3.84	8.47	4.37	2.73
Length of the frost-free period*	5.20	10.37	5.75	4.23
Degree-days >5 degrees C*	1.01	8.92	1.72	2.23
Degree-days >5 degrees C accumulating within the frost-free period*	3.65	9.26	4.20	3.20
Julian date the sum of degree-days >5 degrees C reaches 100*	0.85	8.15	1.41	1.48

* are climate normals variables

Appendix C (continued)

Variable	**A**	**P**	**Mean Decrease Accuracy**	**Mean Decrease Gini**
Degree-days <0 degrees C*	2.78	10.67	3.44	3.01
Degree-days <0 degrees C*	2.98	10.79	3.67	3.98
Summer precipitation balance: (jul+aug+sep)/(apr+may+jun)*	3.85	14.03	4.54	3.61
Summer/Spring precipitation balance: (jul+aug)/(apr+may) *	0.96	14.74	2.49	7.01
Spring precipitation (apr+may)*	3.55	12.73	4.75	5.19
Summer precipitation (jul+aug)*	3.65	12.41	4.15	4.54
Winter precipitation (nov+dec+jan+feb)*	4.11	12.54	4.65	4.74
Egg mass density	-5.16	79.25	5.33	125.76

* are climate normals variables