COLBY COLLEGE

HONORS THESIS

Species Distribution Model Development and Validation

Author:
Victoria CHISTOLINI

Supervisor:
Dr. Stephanie TAYLOR

A thesis submitted in fulfillment of the requirements for Honors in the degree of Computer Science

in the

Record Lab
Bigelow Laboratory for Ocean Sciences

May 10, 2018

COLBY COLLEGE

Abstract

Dr. Nick Record
Bigelow Laboratory for Ocean Sciences

Computer Science

Species Distribution Model Development and Validation

by Victoria CHISTOLINI

In the Northeastern United States, ticks bites and subsequent Lyme Disease infection are becoming a growing problem [1]. Data collected by the Maine Medical Center Research Institute from 1995 - 2013 from citizen reports of ticks, presents the unique opportunity to study patterns of human - tick encounters in the state of Maine [2]. The field of species distribution modeling seeks to develop mathematical and computational models to characterize the distribution of a species based on it environment, and often uses the Maximum Entropy model to accomplish this goal [3]. Tick activity is highly dependent on the time of year and the tick life cycle, thus it is reasonable to expect that a single model will not be able to capture the time-of-year specific drivers of activity.

Thus we seek to develop an ensemble of Maximum Entropy models that are tuned to forecasting tick encounter likelihood at different times of the year. Ensembles of models can effectively be created by altering the combination of predictor variables in the model [14]. We will use the Poisson point process model as a statistical framework for evaluating predictor combinations at different times of the year. Using the best performing parameter set at for each month of the year, we will build Maximum Entropy models on training data and evaluate model fit of training data and model performance on testing data using a modified AUC statistic [12].

Acknowledgements

’ I would like to thank Stephanie Taylor for her continued guidance, support and mentorship throughout my work on this project. Nick Record, thank you for welcoming me into your lab, introducing me to your research and helping me to learning more about your incredible field of computational - mathematical modeling. A big thanks to Ben Tupper for always welcoming my questions and helping guide me through the massive code base that he has built. Manny Gimond, thank you for introducing me to spatial point pattern analysis and for sharing your course resources and mentorship on the accompanying R programming.

Contents

Chapter 1

Introduction

1.1 Ticks in Maine

According to the Center for Disease Control (CDC) in 2015, 95% of reported Lyme Disease cases came from only 14 states; of these 14, 12 were on the east coast and all 6 of the New England States were represented. Since 1996 the annual number of confirmed cases of Lyme Disease has increased dramatically, with about 25,000 confirmed cases in 2016 [1]. Given the growing magnitude of the problem presented by ticks and Lyme Disease infection, there has been a deep interest to understand where Lyme Disease carrying ticks are located and how we can most effectively reduce human contact.

Beginning in 1995, the Maine Medical Center Research Institute (MMCRI) began a project to create detailed records of the locations of discovered infected ticks. Doctors were encouraged to have their patients bring in any ticks that they found on their bodies, in their homes or on their pets, for free testing to determine if the tick was carrying Lyme Disease [2]. Data about the locations of these tick sightings were recorded as part of the study until 2014.

1.2 What is Species Distribution Modeling?

Species Distribution Modeling (SDM) is a term used to categorize a whole class of models developed for the purpose of understanding the patterns and relationships of an observed species and its environment [3]. Often, the purpose of these models is to predict the range of a species based on where the species has been recorded during surveillance. Another application of SDM models is to predict whether or not a species could be found in a certain location based on the environmental variables of the location. It is with this latter focus that we will pursue Species Distribution models throughout this work. The focus of our research is to develop models to forecast the likelihood of a human-tick encounter in the State of Maine.

Chapter 2

Approach

2.1 Models used in Species Distribution Models

The database of tick encounters developed by MMCRI houses information from the sightings of over 4,000 sightings of female deer ticks in the state of Maine, however in its raw form, the data is missing a crucial metric: information about where ticks are not found.

Classical modeling techniques use a set of predictor variables to classify events under certain conditions as likely or unlikely to happen. A logistic regression model, for example could take predictor variables about patients' heart rates and temperatures to determine the likelihood that the patient has the flu. In order to make a good prediction about the patient, the model needs heart rates and temperatures from patients who are healthy and from those who are sick in order to be able to differentiate sick from healthy metrics. In machine learning terminology, models like logistic regression are part of a class of models called supervised learning algorithms, because in order to discern patterns these models need examples of each category to be classified.

Unfortunately, our tick dataset only provides information about locations where ticks are present and no information about the locations where ticks are absent, which motivates the need to use a different class of models. Unsupervised learning models are able to classify data that has not been labeled, by which group it belongs to, through using the patterns inherent in the data itself to distinguish and predict classification groups. However, in order to use unsupervised algorithms it is necessary that the data contain many examples from each category that you are trying to predict. Thus since our data contain only presence information and no absence data, we are unable build models for our data using unsupervised algorithms.

Clearly when selecting an appropriate model for a dataset, it is essential to be certain that the type and quality of the data fits what is necessary and expected by the model building procedure to produce valid results. If the model makes assumptions about the data, which do not hold, then the interpretability and usability of the model will be greatly impaired. In the literature on Species Distribution Modeling, the maximum entropy model is most commonly used, because it is able to work well with data containing only presence locations. Thus, we focus our research efforts on creating high performing models using the maximum entropy model.

2.2 The Maximum Entropy Model

To determine the appropriateness of any modeling approach we first start by summarizing what information we know in the beginning. We know a sample of locations (latitude and longitude coordinates), where ticks have been found, called presence points. Assuming the presence points are well collected and representative of the locations and environmental conditions where ticks are likely to be found by humans, then we can use this information to estimate and predict locations with high likelihood of encounter.

We have no information about unsuitable locations, where encounter risk is low, however, we can use our study region, the state of Maine as a constraint on our distribution. By taking a random sample of points from our study region, we create a representation of the diverse habitat conditions in our region, called *background points*. While we have a high degree of certainty that the *presence points* represent spaces where ticks should be found, we have no opinion at all about whether or not ticks should be found in the background locations. The theory of minimizing cross entropy (which can be proved analogous to maximizing entropy) says that given the information that we have, we can actually find a unique distribution that is optimized to the information that we know and does not penalize us for what we don't know, by selecting the distribution with minimal cross entropy [4].

Put more technically, what we begin with are two probability density functions, $p(x)$ the distribution of ticks from our presence and background points and $q(x)$, the probably density function of the tick encounter likelihood conditional on weather conditions and geographic constraints. The process of finding maximum entropy, then is accomplished by minimizing the cross entropy function:

$$\int q(x) log(\frac{q(x)}{p(x)}) dx \tag{2.1}$$

During the minimization process, the predictor variables undergo transformations that help to maximize the entropy of the solution; this is how the model is fit [5]. The covariates may be transformed into terms of the type : linear, quadratic, product (representing interaction), hinge, threshold or categorical variables in the final model. The effectiveness of the transformations are assessed through cross validation process, using L1 regulation to avoid overfitting [5].

2.2.1 Maximum Entropy Example

When we began the modeling process, we had our dataset of tick observations, which highlighted the set of locations where ticks are known to be able to survive. We also started with another subset of point locations in the state of Maine, called the background points, which represent the landscape that we are trying to model. Suppose the background points dataset consists of two environmental features, elevation and vegetation type and looks like this:

x	Elevation	Vegetation Type	p(x)
1	H	1	0.1
2	H	3	0.1
3	H	2	0.1
4	M	2	0.1
5	L	4	0.1
6	L	3	0.1
7	M	5	0.1
8	M	4	0.1
9	L	3	0.1
10	M	3	0.1

We know nothing more than this about the background locations, so we distribute probability to the subset of locations uniformly. That is if we were to go back to each of these locations then the chances of us getting a tick is equal in each location; so if we have n observations in our set, then the probability of getting a tick at any of the locations would be $\frac{1}{n}$. This is a very naive estimate that each location is equiprobable, but this is our best first guess at the distribution and is called the prior distribution. By performing a spacial analysis we may be able to find patterns between the locations of our tick sightings in the observations dataset, for example that 90% of the locations were below a certain elevation. Based on this insight, which we call a constraint, we can recalculate the probabilities predicted for each location in our background point dataset.

x	Elevation	Vegetation Type	p(x)
1	H	1	1/30
2	H	3	1/30
3	H	2	1/30
4	M	2	9/70
5	L	4	9/70
6	L	3	9/70
7	M	5	9/70
8	M	4	9/70
9	L	3	9/70
10	M	3	9/70

Next we define another constraint on the model; we find that 82% of the tick sightings are happening at vegetation types 4 and 5. We can incorporate this new constraint and reweight the model. This time there is not as clear cut a way to reweight the probabilities. We know 18% of the probability should be represented in the set S_1= { 1, 2, 3, 4, 6, 9,10} and 80% should be represented in the set S_2 = {5,7,8 } but how should the probabilities be determined for each element in the set? As mentioned previously, based on the principle of maximum entropy, we want to select the most uniform distribution by minimizing the entropy function in equation 1.2. This equation is difficult to solve analytically, thus numerical methods will be used for the optimization procedure. This toy example is based off of work done in [6].

In building the maxent model distribution, we have used constraints in order to add information to our model. There are several methods for figuring out how to derive the constraint rules. One of these processes is similar to the construction of a

decision tree where at each new additional leaf we want to choose the feature that maximizes the information gain. We use the environmental features to develop the constraints which are the expected values of the features.

2.2.2 Current MaxEnt model

The algorithm that we will be using to create our maximum entropy model is called MaxEnt and is supported by the R programming language through the dismo package; the main algorithm is written in Java, by S.J. Phillips et al (2006) [7]. We have currently developed several different MaxEnt models using environmental covariates that best fit the expert knowledge of the parameters that affect tick survival. The core set of environmental covariates came from North American Mesoscale Forecast System (NAM), in the form of raster images of a 12 km resolution. Each pixel of the image represents a projected latitude-longitude coordinate, and the value stored at the pixel is the value of the covariate at that location. For example if the raster image represented elevation, then the value at each pixel would be the meters above sea level at that location.

The core parameter set includes: minimum air temperature, maximum air temperature, mean air temperature, mean percent vegetation cover, mean relative humidity, mean snow cover, mean snow depth, mean transpiration rate, mean u-direction wind speed (east-west winds), mean v-direction wind speed (north-south winds), mean wilt, sum of precipitation. Some models include additional parameters for previous-year's minimum winter temperatures.

Once a MaxEnt model is fit using the covariates mentioned above, the output model is of the form:

$$q(x) = \frac{e^{\lambda * f(x)}}{Z} \tag{2.2}$$

where λ is a set of weights on the features and Z is a scaling constant that makes sure the probability distribution $q(x)$ sums to 1 [7]. A new model is fit for each day of the year using all of the data in the database available for the particular day of the year. Since often there is not enough data for a single day of the year, data is taken from a window of time around the model day. The size for a suitable window depends on the time of year, with larger windows being needed during time of the year with fewer tick observations.

Our next goal is to perform exploratory analysis and look at potential sources of error and model refinement through the creation of new models with time-of-year specific parameter combinations to boost overall system performance. Since one main goal of our work is to develop a rigorous understanding of how covariates influence model performance, we need a statistically based framework to assess significance of parameter effect. The MaxEnt model does not allow us to perform hypothesis tests on coefficient significance for the output expression, thus we must use other methodology from spatial-temporal analysis to develop inference about covariates.

2.3 Spacial-Temporal Exploratory Analysis

Exploratory data analysis is a crucial step in the model development process. It should occur before any model is attempted. It is here that the statistician is able to get to know the data. The key component of exploratory analysis is data visualization [8]. Exploratory analysis tells us something about the patterns generated from the data. It is also useful in identifying problematic elements of the data such as missing values or the nature of the scales at which the measurements are made. For example, does the scale of the raw data make sense, or should we transform the data, perhaps using a $log(x)$ transform or discretizing data which only takes on certain integer values. There is no standard procedure for exploratory analysis, however the major goal of the process is to assess the data quality and to reveal data patterns. The same exploratory analysis tools are also used in examining the assumptions necessary for any parametric hypothesis tests.

Aside from the broader questions mentioned above, our exploratory analysis seeks to also answer questions about the Spatial-Temporal aspects of our data. Spatial-Temporal analysis is the use of specialized statistical techniques that are designed to help identify patterns that derive from the spatial and temporal dimensions of the data [9]. In the sections below we address the questions we seek to answer through spatial-temporal analysis of the data, as well as the methodologies that we will use to accomplish these goals.

2.3.1 Spatial Analysis and Predictor Selection

A branch of spatial analysis called *point pattern analysis* can help us understand how the tick observations are distributed spatially across different covariates and provide a statistical framework for hypothesis testing of these covariates. We will be using the *Poisson point process* model for our experiments. A point process is a way of mechanistically thinking about point patterns and the processes that generate these patterns. These processes may be random, or they may be defined in terms of external factors [10].

One basic point process model is that generated by a completely random mechanism, which we will call a point process with complete spatial randomness (CSR). CSR point processes have two special properties, (1) they are homogeneous, meaning that the intensity of points does not depend on the location within the experiment window, and (2) the points are independent of one another; meaning that the location of one point cannot influence the placement of a nearby point [10]. An inhomogeneous point process (IHP), is a point process such that the intensity of points depends on spatial location [10]. We will want to develop a test that can determine if our tick data follow a CSR process (the null hypothesis) or an IHP pattern (the alternative hypothesis).

In order to develop the statistical test mentioned above, we need to make some observations about point patterns. If we assume that the study space of the points is well defined and that an observation could have occurred anywhere in the study space, and that the set of observations is a complete enumeration of all observations, then it can be shown that the points follow a Poisson distribution. Informally,

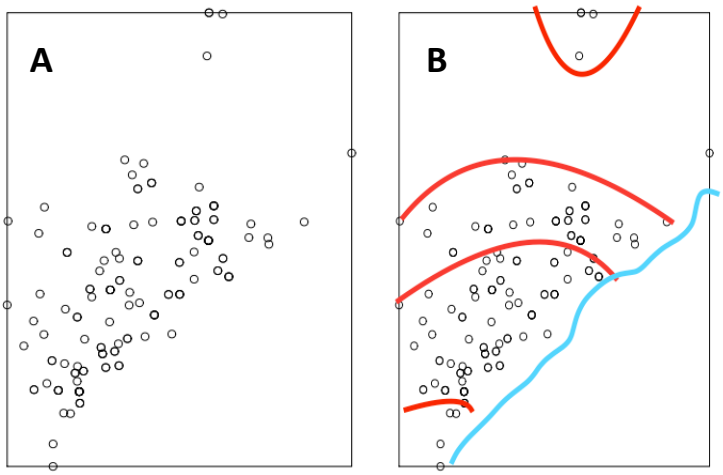

FIGURE 2.1: An example of Complete Spatial Randomness derived by controlling for covariates.

this can be seen by imagining the study space divided into infinitesimally small regions, such that most regions have no points, then by the law of the frequency of rare events, the points follow a Poisson distribution [10].

Since we suspect that environmental factors are the major mechanisms driving the human-tick observation pattern, then the human-tick observation pattern is an IHP type pattern. We want to test this hypothesis that human-tick observations follow an IHP, by controlling for the covariate process and then testing for a CSR process. Figure 1.1 provides a graphic explanation of this process: in A the IHP from day 150 and years 2006-2013 clearly is clustered in space. It is important to note that the right boundary, added in blue in B is the Atlantic Ocean and is not part of the study region for the purpose of calculations. In B the boundaries of a fictitious covariate are added, one could think of these boundaries as temperature gradient lines. The point patterns in each section appear to be more in line with a CSR process in that the points appear to be randomly distributed and independent of one another. Although this example has been fabricated it is a good illustration of what we seek to do. However, controlling for a single covariate may not be enough to achieve CSR, one might need to add several terms, including interaction and indicator terms as well.

The fitted poison point process model is of the form:

$$\lambda(x) = e^{\alpha+\beta(x)} \tag{2.3}$$

We can conduct hypothesis tests on each of the β_i coefficients to discern the effectiveness of the parameters in the model.

We must be careful about some of the assumptions that we made, and how they relate to our interpretation of the model. The assumption of having the points be a complete enumeration is clearly not met. This causes the model to penalize all areas that do not have an observation where a tick incident occurs (due to the event not

be recorded). Although this assumption might be too restrictive if we were trying to create a mapping of all of the possible locations for ticks, this is not our goal, we are trying to map human-tick encounter risk. Thus we can argue in favor of penalizing the locations with few to no observations if they have lower populations and thus little chance of an encounter there in the first place

2.3.2 Temporal Analysis

Ideally we are trying to create a forecasting model for each day of the year. The need for models that are season or even month-day specific is based on the fact that the weather conditions can vary greatly throughout the year, as with the behavioral patterns of the humans who are going to encounter ticks. Therefore, when designing our experiments of predictor selection, it is important to repeat the experiments to capture the yearly variability. One way to do this is to run a predictor selection experiment for a day mid-month for each month in the year.

A temporal limitation to the analysis is the limited number of observations at different times of the year. Ideally we will be looking to make a model for each day of the year to use for forecast prediction. However, due to the fact that one is less likely to encounter a tick during the winter months, we have fewer observations for this time frame. With fewer observations to build the model, the forecast predictions are less accurate.

Thus, we must perform an experiment to understand the impact the limited amount of data has on the model's subsequent accuracy. We seek to understand if aggregating data from months with few observations will positively impact accuracy. We also seek to determine a threshold about the number of observations necessary to achieve a certain error tolerance and a bound on the window size based on day of the year.

Another area of interest is the stationarity of the data. There is evidence that the range of suitable tick habitats continues to expand as the effects of global climate change become more dramatic [11]. At this point in our primary analysis we have not explored stationarity in our dataset. We expect that there will be some correlation with the tick distribution, its predictors and time. In order to test this correlation, we can build regression models for each predictor at each location across time to quantify the degree and strength of non-stationarity in the data.

2.4 Model Assessment

2.4.1 Assessment of Performance

Assessment of classification algorithms is an essential part of determining their utility as valid predictive models. Classically testing the strength of a classification algorithm involves building the model with a subset of the dataset and keeping another subset for use in testing the model called the training dataset. Then the model is run with the training data and an assessment is made about the model performance. The construction of a confusion matrix helps to communicate how well the model did in

classifying the training data by tabulating the number of correctly classified results as well as false positives and false negatives.

$$\begin{array}{cc} & \begin{array}{cc} Observed\ Presence & Observed\ Absence \end{array} \\ \begin{array}{c} Predicted\ Presence \\ Predicted\ Absence \end{array} & \begin{pmatrix} a & b \\ c & d \end{pmatrix} \end{array} \tag{2.4}$$

The data in the confusion matrix can then be visualized in a plot called the ROC curve.

2.4.2 The AUC statistic for model assessment

The ROC curve plots the classifier's sensitivity versus 1-specificity, where sensitivity is easily calculated from the confusion matrix as $\frac{a}{a+c}$ and 1- specificity is $\frac{b}{b+d}$ [12]. Since the results of the MaxEnt model are a series of probabilities given environmental conditions of encountering a tick at a particular location, then in order to create a confusion matrix, we would need to decide on some arbitrary threshold at which we decide that a probability is high enough to be considered a presence of a tick.

Since the decision of a threshold is arbitrary, then an ROC is formed from finding the confusion matrix for each threshold, which will give a new sensitivity and 1-specificity value to plot for each matrix. The summary statistic used to characterize the ROC curve is the area under the ROC curve or AUC, which evaluates the strength of the classifier by the characteristics of the ROC curve. An AUC statistic of 0.5 represents a random classifier, and scores above 0.5 represent a better than random model [12].

Since we are using presence only data, the traditional methods of model assessment the ROC curve fall short. The training dataset is a series of presence observations, thus there would be 0 observations correctly classified for no encounter, cell d, which would make it impossible to calculate 1- specificity. Since we can't calculate 1-specificity, we can use a proxy for it using the proportion of area predicted present at each threshold [12]. With the simple modification, we can proceed to interpret the AUC in a similar fashion as our interpretations of the traditionally defined AUC.

2.4.3 Modified AUC Example

Using the method of area predicted present to calculate AUC, we are seeking to quantify model fit by how much area we need to include in order to capture all of the presence points. For example in figure 2.2 we see two mappings of the same set of presence points. The blue boarder line marks the boarder with the Atlantic Ocean. All points above this boarder line are in the study space, while the area below is part of the Atlantic Ocean.

In (A), there are two area bands present, the red band encircles points in the region that is predicted high likelihood of encounter at a given threshold t, the yellow band represents the study area that is predicted moderately high likelihood of encounter given t. In (B) there is a third green band that represents the area predicted low likelihood of encounter given t. Based on the area predicted present method, (A) would have a higher AUC than (B) because we only had to include area from highly likely

and moderately likely to capture almost all of the points, while in (B), a large chunk of presence points were only captured if we included a low likelihood chunk of area. In a perfectly fit model, we would capture all points by only including highly likely areas.

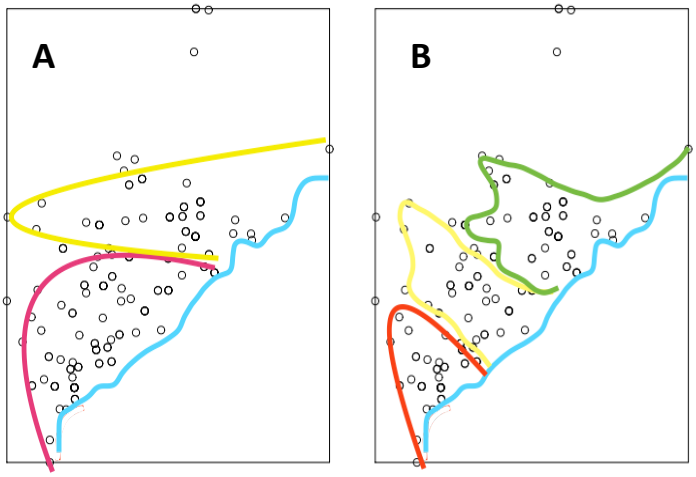

FIGURE 2.2: Example of modified AUC principle. Red bands represent high probability area, yellow bands represent medium probability area and green bands represent low likelihood probability area.

One caveat of the area predicted present method is that higher AUC values are easier to get when there are fewer observation points used to fit the model. In times of low tick activity, for example during the winter, the few observations that we do have are often highly clustered, thus it is easy to fit a model right to this single cluster of points. This results in an overfit model that has a high AUC score because all of the points will be captured in a high likelihood of encounter region. We can help identify the models with overfitting, and thus inflated AUC scores, but preforming forecast testing using the FAUC metric.

2.4.4 FAUC: a forecast testing metric

When we build our model using the training dataset, we are interested in how well the fitted model fits the data. We use AUC as the primary metric to assess the model fit under the training data. AUC only provides information on how well the model represents the data used to tune it. The forecast AUC (FAUC), is a validation technique used to assess the strength and performance of models in a forecasting context.

First we build a MaxEnt model using only a subset of the data in the database, a training dataset, then performance on predicting the rest of the data, the testing dataset can be measured by the FAUC. For example years 2006-2010 of data of day 150 are used to build the model, then the model is tested using data from 2011-2013 of day 150, using the 2006-2010 model to predict the outcomes for the 2011-2013 data. Partitioning our data into testing and training sets allows us to more accurately estimate of the true predictive power of the model.

2.4.5 Other metrics of forecast performance

AUC represents the classification ability of the model independent of a threshold, however interpretation of a forecast depends on eventually selecting a threshold. The threshold determines the minimum probability that will be considered a presence. Since the utility of our forecast greatly depends on the ability to discriminate events from non-events, it is useful to define the threshold as a single number when calculating a summary statistic about the skill of the forecast. The statistic True Skill Statistic (TSS):

$$TSS = sensitivity + specificity - 1 = \frac{ad - bc}{(a+c)(b+d)} \tag{2.5}$$

can be used used to quantify the strength of the classifier at a given threshold [13]. The TSS statistic has been shown to have a good behavior and is well correlated with the AUC statistic [13]. However, in order to calculate the TSS statistic we would need to have access to a complete confusion matrix which we do not have. Therefore, in order to use the TSS, we would need to substitute our proxy metric calculated at a given threshold.

2.5 Ensemble Models

Since we are trying to develop models to forecast human-tick encounter for the entire year, it is reasonable to assume that it is unlikely that a single model will be able provide the best predictions for every day of the year. However, it is reasonable to believe that a collection of models with different parameter combinations could be specialized to perform better at different times of the year. Evidence has shown that collections of models, called ensemble models, provide more robust forecast models, where each of the individual models in the ensemble provides independent and novel information to contribute to a collective consensus [14].

An ensemble of models is defined as creating duplicated models with altered initial conditions, boundary conditions, types of model, and parameter combinations [14]. When we create a model there are many sources of uncertainty, we do not know the true mechanisms that drive human-tick encounters. As a result, we must take educated guesses at which predictors are useful to include. We began by intuitively selecting core parameters and then use the statistical framework of the Poisson point process to guide predictor selection for different times of the year. We also performed exploratory experiments on the window size parameter at different times of the year, perturbing it based on season and subsequent observation density.

Despite having created many new models, looking specifically at conditions at certain times of the year to fit the highly variable observation patterns, it is possible that our ensemble is underdispersive. An underdispersive ensemble means that the ensemble is not as variable as would be expected given the diversity of its members [15].

Variability is desirable in an ensemble because, it ensures robustness and representation of future potential variability [14] Underdispersion occurs because it is very difficult to capture all of the types of uncertainty that exist. Therefore, is a trade off

between exhaustively searching the multi-dimensional space of model uncertainty versus spending effort in creating an ensemble of fewer, perhaps less variable, but individually more skillful members. There is evidence that focusing energy on improving the quality of individual models produces higher quality results. [14]

2.5.1 Future Work with Ensembles

Currently we have done the first step of creating an ensemble by selecting a set of skillful models that have strengths at different times of the year. Currently we do not have enough information to distinguish which models perform better than others conditional on the time of the year. Future work includes more rigorous testing of the ensemble members on future data, and storing model performance over the long term. With this information we would be able to discern which models are more skillful and properly calibrate the ensemble as described in [15].

Chapter 3

Exploratory Analysis and Model Development

We will begin a close analysis of the Maine Tick database following the outline of the methodology in the Approach section. We begin by conducting an exploratory analysis on the tick data, covariates and window size model parameter. We want to understand how the tick data and covariates are distributed and how this may change throughout the year. We want to understand the implications of the window size parameter in the model. Specifically, we want to know how the model fit (AUC) is related to different values of window size.

Then we will use the statistical framework of the Poisson point process model to evaluate how the predictive effectiveness of different covariates changes throughout the year by studying 13 days of the year (the 15th day of each month). The Poisson point process models will guide our predictor selection conditional on time of the year. We will then use the time of year specific covariate combinations to build maximum entropy (MaxEnt) models using training data (data from 2006-2010) with each of these combinations, in an effort to build a forecast ensemble.

We will evaluate the model fit against the training data using AUC and then evaluate model performance on testing data (data from 2011-2013) using the FAUC. The FAUC will guide a preliminary analysis of the ensemble, but further work will be required in developing rigorous tests for ensemble performance.

3.1 Looking at distributions

We start by looking at how the raw data of our observations set is distributed. We are looking to create a testing and training dataset that are distributed similarly. Figure 3.1 shows two plots of the number of observations recorded on a given day of the year for the time period 2006-2010 (A) and 2011-2013 (B). Although in (A), we see that there are overall more points per day, the yearly shape of observation records is strikingly similar between these two sets. The first and third quarter of the year have few observations per day, while the second and forth quarters of the year have increasing activity, which spikes mid quarter and then declines for the second half of the quarter.

Given the yearly activity cycle of the tick, we need to be cognizant of the number of points being used to create each model. Clearly, there will be a much smaller pool of observations during the first and third quarters. Many days may have zero observations even as we are aggregating data from a span of almost 10 years. On days

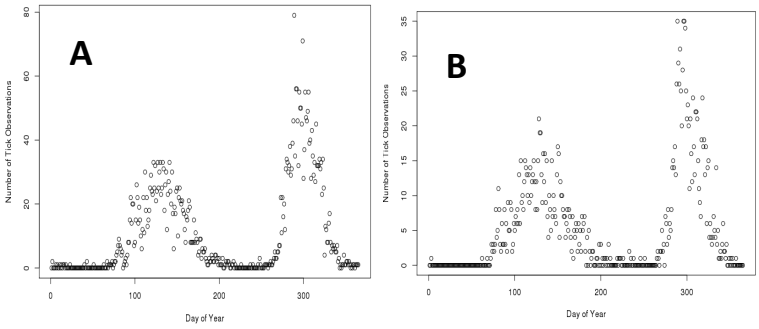

FIGURE 3.1: Distribution of tick observations during the year for (A) years 2006-2010 and (B) year 2011-2013.

with so few data points it becomes impossible to create a model. One way around this obstacle of low observations counts is to pull in observations from a window of time around the forecast date. We will call this new parameter the window size of the model.

3.2 Window size implications

The window size parameter has a lot of uncertainty around it. It is unclear what values this parameter should take on at different times of the year. Further we do not know the impact of increasing the window size parameter on the accuracy and precision of the model. One hypothesis is that it is necessary to keep the window size parameter large enough so that the model has exposure to enough data to create well-informed predictions. Another hypothesis is that if the window size is increased by too much, then the accuracy of the predictions will be weakened due to the presence of data irrelevant to the current stage of tick activity.

In order to assess the impact of window size on forecast skill, we run a MaxEnt model with a subset of the 13 core parameters as predictors against 7 candidate window sizes: $\pm$ 2, 3, 7, 15, 20, 30, 40 days, for 13 days of the year (approximately the 15th day of each month is tested); days 15, 46, 76, 105, 135, 146, 166, 196, 227, 258, 288, 319, 349 of the year.

The general trend of figure 3.2 independent of window size is the inverse of figure 3.1. Highest AUC's are seen during the first and third quarters of the year, while steep crashes in AUC scores are experienced during the second and forth quarters. It is important to note that as a rule, the fewer points we have in our model, the higher the AUC will be because there is a lower chance of being wrong as described in section 2.4.3 and demonstrated in figure 2.2. Thus figure 3.2 shows that models during low activity are better fit to the training data, than those models developed

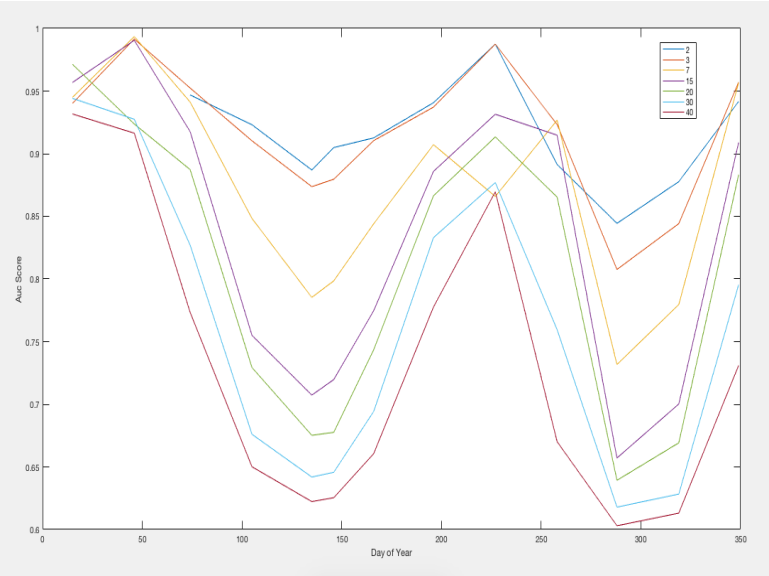

FIGURE 3.2: AUC score by day of year for the 2006-2013 time period of observations. Each line represents a different window size from the 7 candidate sizes.

during high activity periods, however it is unclear if high AUC is correlated with good long term forecast performance or if it is an indicator of an overfit model.

The question of window size is particularly relevant during the low activity of the first and third quarters of the year. We want to increase the window size to get more data and more information about trends in tick encounters during the time of year, but we don't want the window size to be too big and provide irrelevant information. For the first 25% of the year it seems that the accuracy is about equal for all window sizes, and thus it is likely useful to include a larger window size so that the model has access to more information at fit time. By around day 70 though, this is not entirely true, windows that are too large such as the sizes 30 and 40 produce much weaker AUC scores.

The window size of 15, however, continue to perform well high, so it would be advisable to use window sizes up to 15 during this time of the year. Even maintaining a window size as large as 7 into the beginning and at the end of second quarter seems justified based on continued high performance of the 7 day window during these transitional times. For the third quarter a window of size 15 seems to have the best balance of performance and inclusion of data. For the second and fourth quarters, windows sizes or 1 - 2 days are sufficient for high performance because there is plenty of data amassed even for small windows during this time. By the end of the fourth quarter, however, as winter begins to take hold, window sizes of 15 or 20 are best.

Another metric that we can use to assess the window size parameter is by the number of points contained within a given window size. Figure 3.3 shows a decreasing trend of AUC values produced by having too large of a window size. Up to around 200 points the models seem to produce high quality results with AUC scores hovering around 0.9. However, once we go past 500 points, AUC scores start to tank, indicating that window sizes generating such high point values are probably too big

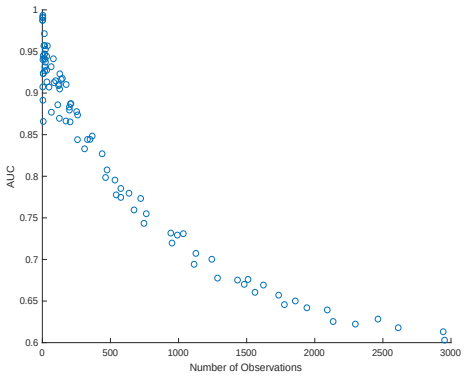

FIGURE 3.3: AUC score by number of observations for the 2006-2013 time period of observations. Each dot comes from one of the 13 days of the year mentioned above, with one of the 7 window sizes.

to provide valuable information.

3.3 Understanding predictor variables

Before we try to fit models to our observation set, it is necessary to understand what the distributions of our predictor variables look like. This will allow us to identify is there might be some kinds of transformations to the covariates that might be beneficial to our model. In Figure 3.4, we have created histograms of all of the predictor variables for day 150 of 2007. Day 150 is around the height of the first peak in tick activity. We can see that variables like mean air temperature and wind have distributions that appear continuous in nature. However, transpiration rate, wilt and vegetation cover appear to have discrete distributions of values.

The vegetation cover variable is categorical, with 20 different categories, based on the International Geosphere-Biosphere Programme (IGBP) land cover classification system. According to figure 3.4, the most popular vegetation type for tick observations is 11, permanent wetlands, as well as 12-15 which represent cropland and mix vegetation types of forest, shrub and grassland. A more through examination about how vegetation cover of where ticks are found changes throughout the year, reveled another important vegetation type in the second and fourth quarters of the year to be 5, mixed forests. Since a categorical variable with 20 categories does not seem reasonable since so few categories are actually represented, we create a binary

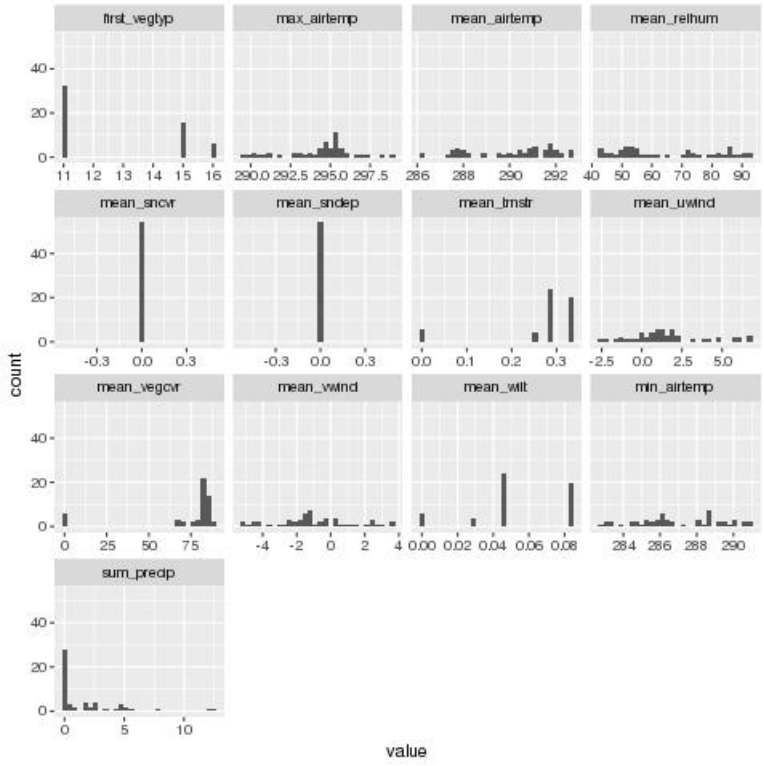

FIGURE 3.4: Histograms of the distributions of core predictor variables for day 150 of 2007.

indicator variable, called v4 that represents if a tick is found in vegetation type 11-15 or not.

3.4 Poisson point process models for predictor selection and evaluation

Using the insights that we gained from the exploratory analysis of our dataset we begin with the first step of the ensemble creation process. In section 1.5, we outlined the process for creating an ensemble which involves perturbing initial conditions (in our case window size and study space boundaries), using different model types or creating novel parameter combinations. We will focus on creating novel parameter combinations that are specialized to certain times of the year and using the statistical frame work of the poisson point process model to assess the significance of each of the parameters in the new models.

We build poisson point process models for 13 days of the year (approximately the 15th day of each month is tested); days 15, 46, 76, 105, 135, 146, 166, 196, 227, 258, 288, 319, and 349 of the year. We derive the window size parameter by trying to minimize the window size while maintaining a good coverage of observation data over the study space. If increasing window size does not seem to add more coverage and 50 or more points are included, then we do not expand the window size. These judgements are made qualitatively by looking at point spread across the map of Maine. Sometimes, however, we increase the window size as much as up to $\pm$ 20 days, a size which evidence from the previous exploratory window experiments reveals is about as large as possible without sacrificing performance over the long

term, and we still only have a few points. Despite having little information, we build a model with what we have.

```
predictor_selection(days)

    for day in days:

        data = extract_data(day)

        correlated_predictors = test_correlation(data)
        models = build_uncorrelated_models(data,correlated_predictors)

        for model in models:
            while model.z-scores.min() > -1.96 &&  model.z-score.min() < 1.96:
                model.predictor.remove(z-scores.min())
```

FIGURE 3.5: Pseudo code for the predictor selection algorithm.

3.4.1 Predictor reduction algorithm

For each of the 13 days of the year mentioned above, download the tick observation data from the time span of 2006-2013, with the window size chosen by hand using the method described above. Data about the window sizes selected for each of the experiments run is shown in table 3.1. We also extract data for each of the core 13 covariates from the raster images at the pixels where each observation in the day and window range being studied occurred.

We begin parameter selection by generating a correlation matrix of all of the 13 covariates. Trying to create a poisson point process model with covariates that are highly correlated impairs the performance of the model. Once we have information about the highly correlated predictors we are able to create subsets of the parameters that do not contain any of the highly correlated predictors. Then we build a poisson point process model using the command ppm() from the spatstat package in r. Once we have a model built, we prune unnecessary predictors using a greedy, stepwise algorithm, which removes the predictor with the lowest z-score (an indication of its statistical significance) until all predictors are significant at the 5% level.

3.4.2 Results from predictor selection experiments

Figure 3.6 shows the summarized results of the predictor selection experiments, while table 2.1 shows detailed descriptions of the number of observations and number of points used for each experiment. In figure 3.6, highlights many striking patterns throughout the year. We see that during the winter and early spring, the predictors mean air temperature and v-direction wind appear consistently up to day 150 in the year.

Based on figure 3.1, day 150 occurs approximately at the first peak in tick activity, which coincides with the onset of late spring / early summer. The predictors of highest importance during this time period are the relative humidity, the temperature range (minimum air temperature and maximum air temperature), the type of

vegetation and the amount of precipitation. Late summer and fall show a renewed importance of mean air temperature, u-direction wind and vegetation type.

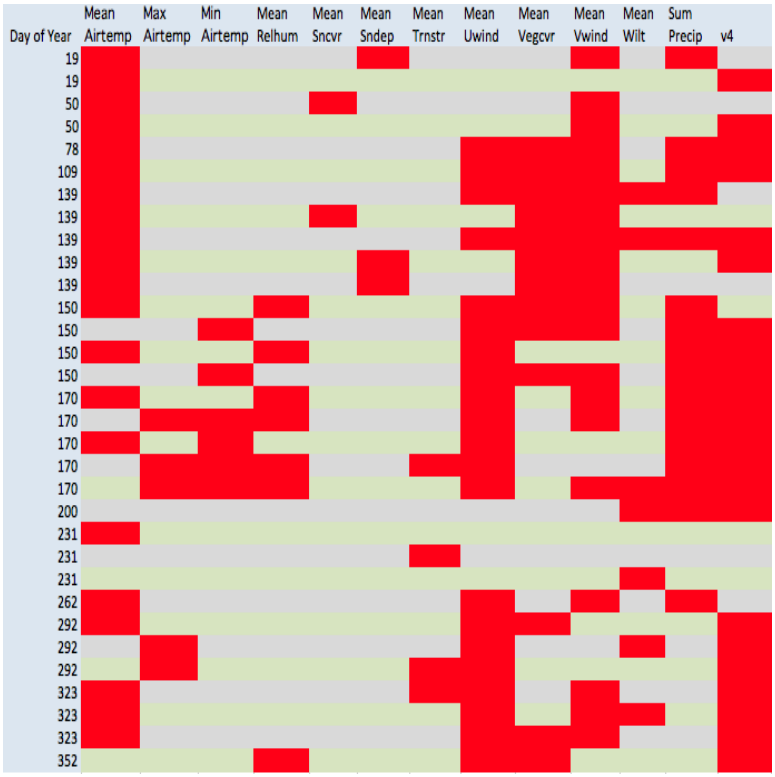

FIGURE 3.6: Histograms of the distributions of core predictor variables for day 150 of 2007.

TABLE 3.1: Parameter combinations and meta data of ensemble members

Selected Models			
Day of Year	Window Size	Number of Observations	Predictors
19	20	12	meanAirtemp, vwind, sndepth, sumPrecip
19	20	12	meanAirtemp, v4
50	20	11	meanAirtemp, vwind, sncvr
50	20	11	meanAirtemp, vwind, v4
78	7	117	meanAirtemp, meanVegcvr, vwind, sumPrecip, sncvr, v4
109	3	118	meanAirtemp, uwind, meanVegcvr, vwind, minAirtemp, sumPrecip, v4
139	3	233	meanAirtemp, uwind, meanVegcvr, vwind, wilt, sumPrecip
139	3	233	meanAirtemp, meanVegcvr, vwind, sncvr
139	3	233	meanAirtemp, uwind, meanVegcvr, vwind, wilt, sumPrecip, v4
139	3	233	meanAirtemp, meanVegcvr, vwind, sndepth, v4
139	3	233	meanAirtemp, meanVegcvr, vwind, sndepth
150	7	445	meanAirtemp, meanHumidity, meanVegcvr, uwind, vwind, sumPrecip

Selected Models Continued			
Day of Year	Window Size	Number of Observations	Predictors
150	7	445	minAirtemp, meanVegcvr, uwind, vwind, sumPrecip, v4
150	7	445	meanAirtemp, meanHumidity, uwind, sumPrecip, v4
150	7	445	minAirtemp, meanVegcvr, uwind, vwind, sumPrecip, v4
170	7	262	meanAirtemp, meanHumidity, uwind, vwind, sumPrecip, v4
170	7	262	minAirtemp, meanHumidity, uwind, sumPrecip, v4
170	7	262	minAirtemp, maxAirtemp, meanHumidity, uwind, vwind, sumPrecip, v4
170	7	262	minAirtemp, maxAirtemp, meanHumidity, uwind, trnstr, sumPrecip, v4
170	7	262	minAirtemp, maxAirtemp, meanHumidity, uwind, trnstr, sumPrecip, v4
170	7	262	minAirtemp, maxAirtemp, meanHumidity, uwind, vwind, wilt, sumPrecip, v4
200	10	61	wilt, sumPrecip, v4
231	13	13	meanAirtemp
231	13	13	trnstr
231	13	13	wilt

Selected Models Continued			
Day of Year	Window Size	Number of Observations	Predictors
262	10	62	meanAirtemp, uwind, vwind, sumPrecip
292	3	553	meanAirtemp, uwind, meanVegcvr, v4
292	3	553	maxAirtemp, uwind, wilt, v4
292	3	553	maxAirtemp, uwind, trnstr, v4)
323	3	250	meanAirtemp, trnstr, uwind, vwind, v4)
323	3	250	meanAirtemp, wilt, uwind, vwind, v4)
323	3	250	meanAirtemp, meanVegcvr, uwind, vwind, v4)
353	10	34	meanHumidity, uwind, meanVegcvr, v4)

3.5 Creating MaxEnt models

Having determined combinations of predictor variables that are specialized for different times of the year, we can now build maximum entropy models with these new predictor combinations. We create a new maxent model for each of the 13 studied days of the year, using the window sizes listed in table 3.1, for each of the 32 predictor sets, resulting in 416 maxent models. We calculate the AUC statistic of each model to indicate how well the models fit the data that they were trained on.

Looking back at figure 3.1 we see the two plots of the number of tick observations on given days of the year. We see that the year trace in figure 3.1 (A) from 2006-2010 and 2011-2013 in (B) have the same trajectory and thus represent a good split of the data. We use the time span from 2006-2010 to build the 416 maxent models and will use the time span 2011-2013 in section 3.6 to evaluate model performance on testing data that it has never seen before.

3.5.1 MaxEnt model fit results

One way to easily assess model performance is to evaluate how well the model fits the data that it was trained on. We can think of this step in model evaluation like interpreting a correlation coefficient. The correlation coefficient tells us how well a linear model fits the trends in the data, however it does not indicate anything about

how well the model will preform in the task of accurately classifying future data. The AUC statistic provides the same function for MaxEnt models, acting as an indication of goodness of fit.

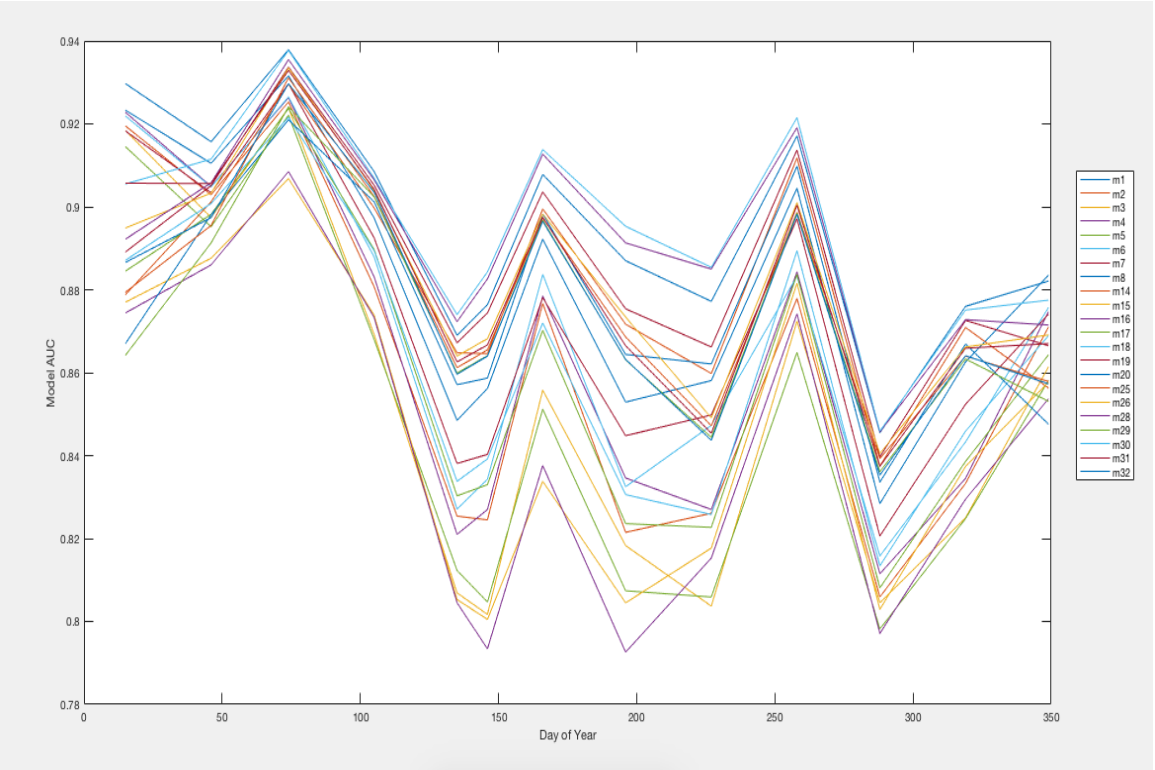

FIGURE 3.7: AUC results from running the 32 different covariate combinations.

Figure 3.7 shows the AUC for 32 predictor combinations MaxEnt models fitted at each of the 13 time points in the year. In the winter months, there is a large degree of variability between the models, which tightens during the early spring. As we approach mid-way through the second quarter mode fit tanks and there is enlarged variability again. In the third quarter, the models regain some of their quality performance around mid summer, which remains stable although highly varied until fall what the variability decreases and peaks quickly, followed by a quick crash and slow regain of fit into the new year.

3.6 Evaluating MaxEnt models

Now that we have built and assessed the goodness of fit of all of our 416 MaxEnt models, we want to utilize the testing dataset that we saved from years 2011-2013 to evaluate the model's performance on data that is has not yet seen. Figures 3.8 and 3.9 show the F-AUC scores of each of the models at the 13 time points of the experiment. The models in figure 3.8 so poor performance up through the first half of the first quarter at around day 60. From a peak at day 60, the models begin a gradual decline in performance up through the beginning of the fourth quarter. During this time period the variability between models is rather low. During the fourth quarter we see increased performance and greater between model variability.

Although it intuitively make seem that it is good for the ensemble to have low variability, since it may seem like the models are providing a consistent answer, this is not the case, it is ok to have variability between models and in fact this is a desired quality of an ensembles. When ensemble have too little variability between models it is a sign of ucnderdispersion and indicates that the ensemble needs calibration [15]. So it is a good sign that we see variability in our ensemble and not one model dominating the other models, a contrast to what we saw in figure 3.7.

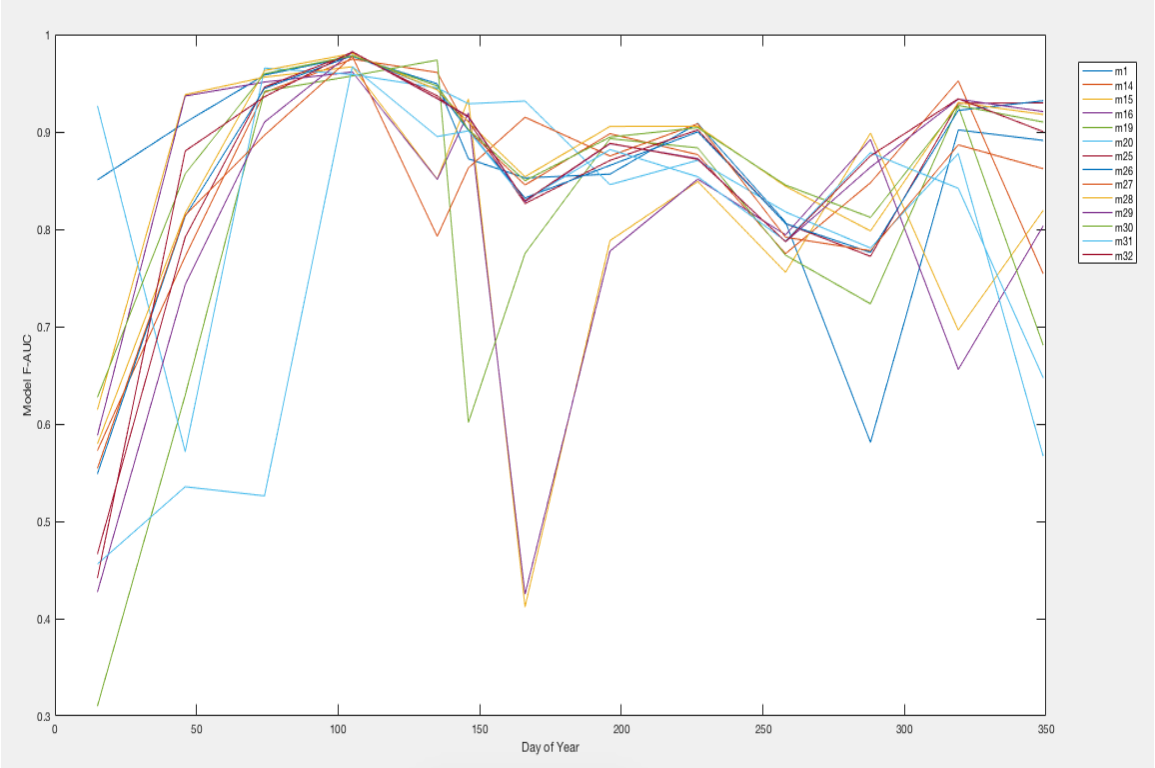

FIGURE 3.8: F-AUC results from a subset of the 32 predictor combinations that have high performance year round.

Figure 3.9 shows the second group of models from the 32 original sets. This group is characterized by two steep crashes in performance in the third and fourth quarters of the year. At the beginning of the year, these models do much better than those in figure 3.8, along with their high performance there is a healthy variability and no single model is dominating the field. High volatility is seen in the third and beginning of the fourth quarter, and it is unclear what is causing this, since the steep descents do not seem to line up with periods of highest activity unlike in figure 3.7. The models recover performance by the end of the fourth quarter.

If we look at the patterns of the covariate combinations that make up the models in figure 3.8 and figure 3.9, we can see some interesting patterns. In Figure 3.10, we can see that the models in figure 3.8, who have a fairly consistent trend year long, have a high emphasis on the vegetation cover and vegetation type parameters as well as both u and v direction wind and mean air temperature.

However, we see in figure 3.11, that the models that have more yearly volatility emphasize parameters such as the range of air temperature (minimum and maximum

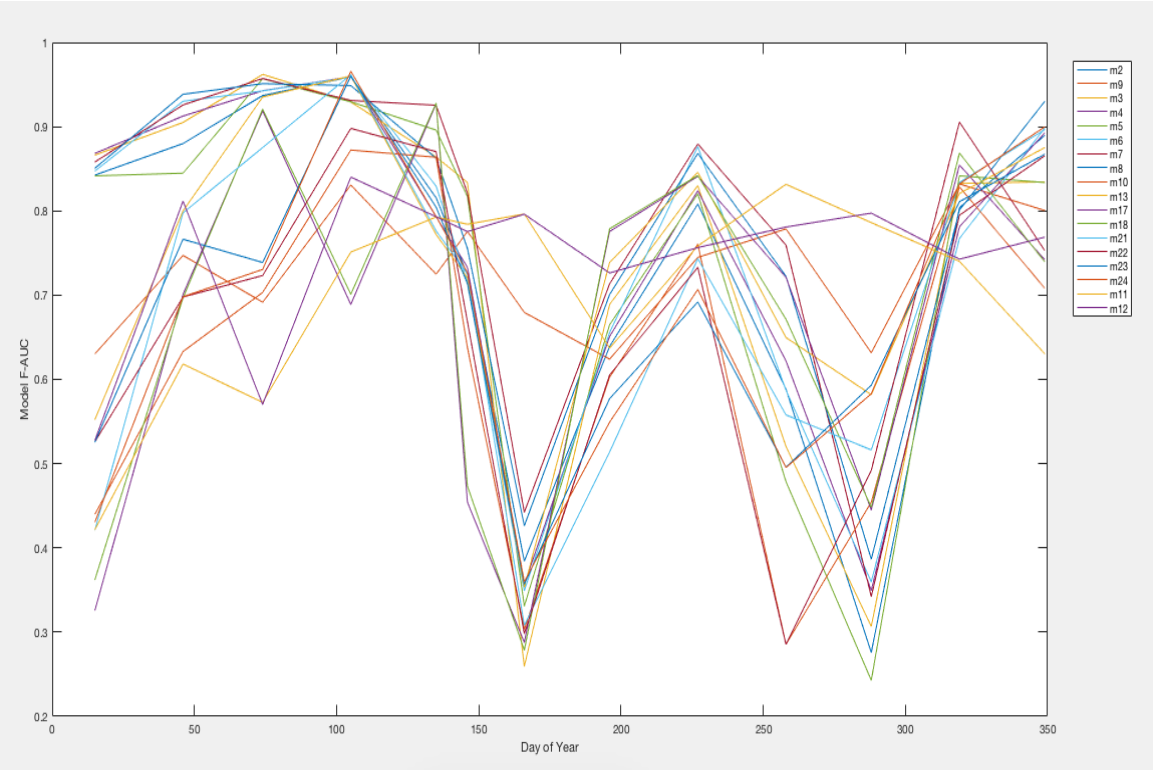

FIGURE 3.9: F-AUC results from a subset of the 32 predictor combinations that have two steep dips in performance throughout the round.

air temperature) as well as sum of precipitation, u-direction wind, but most surprisingly is the almost complete absence of the mean vegetation cover predictor.

3.7 Future directions

Now that we have identified a set of candidate models and have evidence that these models provide a decent amount of variability and quality performance, it is necessary to collect more data about their performance over the log term. By keeping track of how well the models do on future data, we will then be able to build up a better understanding of long term performance of the models and different points in the year so that we can come up with a posterior weighted averaging of the models based on the knowledge that we have gained. This will help calibrate the ensemble and create a further robust and effective system.

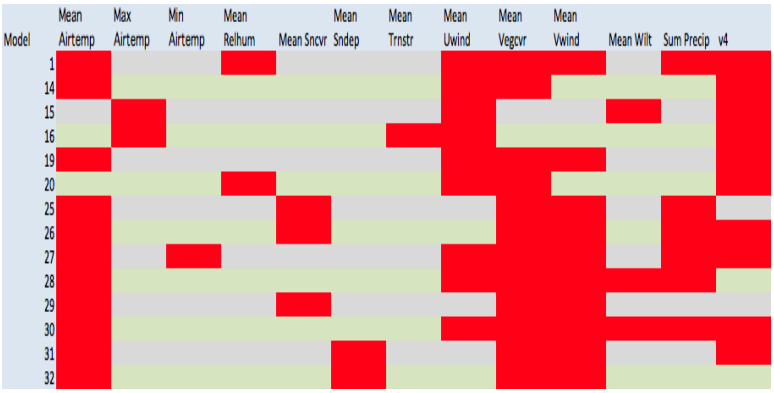

FIGURE 3.10: Summary of predictor combinations from figure 2.8.

FIGURE 3.11: Summary of predictor combinations from figure 2.9.

Chapter 4

Sources

[1] "Lyme Disease." Centers for Disease Control and Prevention,
13 Nov. 2017, www.cdc.gov/lyme/stats/index.html.

[2] Record, Nick

[3] Elith, Jane Leathwick, John R, "Species Distribution Models: Ecological Explanation and Prediction Across Space and Time"
https://doi.org/10.1146/annurev.ecolsys.110308.120159

[4] Shore, J., and R. Johnson. "Axiomatic Derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross-Entropy." IEEE Transactions on Information Theory, vol. 26, no. 1, 1980, pp. 26-37., doi:10.1109/tit.1980.1056144.

[5] Elith, Jane, et al. "A Statistical Explanation of MaxEnt for Ecologists." Diversity and Distributions, vol. 17, no. 1, 2010, pp. 43?57., doi:10.1111/j.1472-4642.2010.00725.x.

[6] Berger, Adam. "A Brief Maxent Tutorial." A Brief Maxent Tutorial,
Carnegie Mellon University, 5 June 1995,
https://www.cs.cmu.edu/afs/cs/user/aberger/www/html/tutorial/tutorial.html

[7] Phillips, Steven J., et al. "Maximum Entropy Modeling of Species Geographic Distributions." Ecological Modelling, vol. 190, no. 3-4, 2006, pp. 231-259.,
doi:10.1016/j.ecolmodel.2005.03.026.

[8] Gimond, Manuel. "What Is Exploratory Data Analysis." Mgimond.github.io,
Colby College, 2018, mgimond.github.io/ES218/Week01.html.

[9] Gimond, Manuel. "Intro to GIS and Spatial Analysis." Mgimond.github.io,
2018, mgimond.github.io/Spatial/index.html.

[10] Spatial Point Patterns Methodology and Applications with R Adrian Baddeley-Ege Rubak-Rolf Turner - Crc Press - 2016

[11] J. S. Gray, H. Dautel, A. Estrada-Pe, O. Kahl, and E. Lindgren, "Effects of Climate Change on Ticks and Tick-Borne Diseases in Europe," Interdisciplinary Perspectives on Infectious Diseases, vol. 2009, Article ID 593232, 12 pages, 2009. doi:10.1155/2009/593232

[12] Peterson, A. Townsend, et al. "Rethinking Receiver Operating Characteristic Analysis Applications in Ecological Niche Modeling."

Ecological Modelling, vol. 213, no. 1, 2008, pp. 63-72., doi:10.1016/j.ecolmodel.2007.11.008.

[13] Allouche, Omri, et al. "Assessing the Accuracy of Species Distribution Models: Prevalence, Kappa and the True Skill Statistic (TSS)." Journal of Applied Ecology, vol. 43, no. 6, Dec. 2006, pp. 1223-1232., doi:10.1111/j.1365-2664.2006.01214.x.

[14] Araujo, Miguel, and Mark New. "Ensemble Forecasting of Species Distributions."
TRENDS in Ecology and Evolution, Elsevier,
29 Sept. 2006, www.sciencedirect.com/science/article/pii/S016953470600303X.

[15] Raftery, Adrian E., et al. "Using Bayesian Model Averaging to Calibrate Forecast Ensembles." 2003,
doi:10.21236/ada459828.