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

April 29, 2018

COLBY COLLEGE

Abstract

Dr. Nick Record
Bigelow Laboratory for Ocean Sciences

Computer Science

Species Distribution Model Development and Validation

by Victoria CHISTOLINI

The Thesis Abstract is written here . . .

Acknowledgements

Thank you to Stephanie and Nick and Ben.

Contents

Chapter 1

Introduction

1.1 What is Species Distribution Modeling

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 per year has increased by over 10,000 additional 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 were carrying Lyme Disease [2]. Data about the locations of these tick sightings were recorded as part of the study until 2014.

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 wether 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.

1.1.1 Data used in Species Distribution Models

The database of tick encounters developed by MMCRI houses an enormous amount of information about the distribution of locations where ticks have been found 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 your 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 algorithm as well.

Clearly when selecting an appropriate model for your data, it is essential to be certain that the type and quality of your data fits what is necessary and expected by the model building procedure to produce respectable results. If the model makes assumptions about your 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 the assumptions it makes about the data are well aligned to our purpose of modeling, thus we focus our research efforts on creating high preforming models using the maximum entropy model.

1.2   The Maximum Entropy Model

To determine the appropriateness of any modeling approach we first start by summarizing what information we know to begin with. 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 do 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 [10].

More technically put, what we begin with are two probability density functions, $p(x)$ the distribution of tick 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 minimize the cross entropy function:

$$\int q(x) log(\frac{q(x)}{p(x)}) dx \tag{1.1}$$

During the minimization process predictor variables undergo transformations that help to maximize the entropy of the solution; this is how the model is fit [6]. The co-variates 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 [6].

1.2.1 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 know to be able to survive, we also start 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. Perhaps the background points dataset consists of two environmental features, elevation and vegetation type look 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 equally 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 this is our best first guess at the distribution and is called the prior distribution. By preforming 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 reweigh the model. This time there is not as clear cut a way to reweigh 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 distribution that is most uniform the 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 [11].

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.

1.2.2   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. We have currently developed several different MaxEnt models using environmental covariates that best fit the expert knowledge of the parameters which affect tick survival. The core set of environmental covariates come from North American Mesoscale Forecast System (NAM), in the form of raster images of a certain 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, mean v-direction wind speed, 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:

$$q(x) = \frac{e^{\lambda * f(x)}}{Z} \tag{1.2}$$

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 [12]. 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.

Since we have created several MaxEnt models our goal is preform 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 a main goal of our work is to develop a rigorous understanding of how covariates influence model performance, when need a statistically based framework to assess significance of parameter effect. The MaxEnt model does not allow us to preform hypothesis tests on coefficient significance for the output expression, thus we must use other methodology from spacial-temporal analysis to develop inference about covariates.

1.3 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 [14]. We want to know what values our response and predictor variables can take on, what do their distributions look like? Is there missing data? 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 goals of the process is the answer questions in regard to the distribution and quality of the data, as well as how well data would be able to hold up the assumptions necessary for preforming future modeling efforts.

Aside from the broader questions mentioned above, our exploratory analysis seeks to also answer questions about the Spacial-Temporal aspects of our data. Spacial-Temporal analysis is the use of specialized statistical techniques that are designed to help identify patterns that derive from the spacial and temporal dimensions of the data [4]. In the sections below we address the questions we seek to answer through spacial-temporal analysis of the data, as well as the methodologies that we will use to accomplish these goals.

1.3.1   Spacial Analysis and Predictor Selection

A branch of spacial 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 distributions, that assumes points are generated according to some process, this process may be random, or it may be defined in terms of external factors [13].

One important kind of point process is that generated by a completely random mechanism, which we will call a point process with complete spacial 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 if we know a certain number of points are in one region we cannot predict the number of points in another region [13]. An inhomogeneous point process (IHP), is a point process such that intensity of points depends on spacial location [13]. We will want to develop a test that can determine if our tick data in a CSR pattern, 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, 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 [13].

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 the hypothesis, if we control for the covariate mechanisms, does the IHP become a CSR. Figure 1.1 provides an 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 actually 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. One can see that each of the sections in between the gradient lines has a CSR pattern such that points are not spatially varying and the points are independent. Although this example has been fabricated it is a good illustration of what we seek to do, I make not be a simple as adding a single covariate 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:

$$\lambda(x) = e^{\alpha + \beta(x)} \tag{1.3}$$

We can conduct hypothesis tests on each of the β_i coefficients to discern the effectiveness of the parameters in the model.
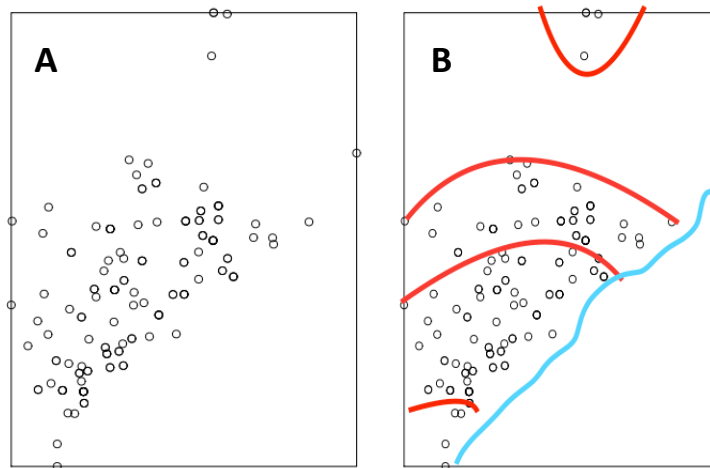
FIGURE 1.1: An example of Complete Spacial Randomness derived
by controlling for covariates.

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 points as it assess the patterns. Although this assumption might be too destructive 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 and so even though the places were no observations are seen are actually likely possible habitats, it is ok that they are penalized because these places with few to no observations have much lower populations and thus it is much less likely for people to encounter tick there in the first place.

1.3.2 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 off of 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.

Another temporal component of analysis concerns the amount of point data that the dataset has 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 in 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 preform an experiment to understand the impact of the amount of data used to build the model and the model's subsequent accuracy. We seek to understand if adding more data from the observations of surrounding months to the model when there are 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 [5]. At this point in our primary analysis we have not explicitly done any on how, if present at all the lack of stationarity of data is present in our database. 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.

1.4 Model Assessment

1.4.1 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 how well the model has done. 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.

$$
\begin{array}{c}
 \quad\quad\quad\quad Observed\ Presence \quad Observed\ Absence \\
\begin{array}{l}
Predicted\ Presence \\
Predicted\ Absence
\end{array}
\left(
\begin{array}{cc}
a & b \\
c & d
\end{array}
\right)
\end{array}
\tag{1.4}
$$

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

1.4.2 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 $\frac{b}{b+d}$ [7]. 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 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 [7].

Given the type of data we are using, presence only data, and the fact that we are generating a forecast, the traditional methods of model assessment fall short. Firstly, since the training dataset would be only a series of presence observations, so there would be 0 observations correctly classified for no encounter, cell d, which would make it impossible to calculate summary statistics, used to calculated the ROC curve. Given these two complicating factors, we will pursue other methods of assessment which employ modifications on the classical techniques to be functional for a presence only dataset.

Since we do not have absence data in our dataset and thus our training dataset is devoid of this essential information, then a more accurate substitute is using is using the proportion of area predicted present instead of 1-specificity [7]. With the simple modification under way we can proceed to interpret the AUC in a similar fashion as our interpretations of the traditionally defined AUC.

1.4.3 The F-AUC statistic for assessment of future predictive power

The F-AUC or forecast AUC, is another validation technique used to assess the strength a performance of models. I model is built with a given predictor set, using only a subset of the data in the database. For example years 2006-2010, then the model is tested using data from 2011-2013, using the 2006-2010 model to predict the outcomes for the 2011-2013 data. Using this metric we can get a more accurate estimate of the true predictive power of the model.

1.4.4 Limitations of AUC Interpretability

Another limiting factor that about AUC that we have previously mentioned is that it represents the classification ability of the model independent of a threshold. The threshold determines the minimum probability that will be considered a presence. In our application this threshold is important because we are creating a forecast and the interpretability of a forecast greatly depends on the ability to discriminate events from non-events, thus the there seems to be an implied necessity 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):

$$TSS = sensitivity + specificity - 1 = \frac{ad - bc}{(a + c)(b + d)} \qquad (1.5)$$

can be used used to quantify the strength of the classifier at a given threshold [9]. The TSS statistic has been shown to have a good behavior and is well correlated with the AUC statistic [9]. However, in order to calculate the TSS statistic we would need to have access to a complete confusion matrix which we do not have, thus in order to use the TSS, we would need to substitute our proxy metric calculated at a given threshold.

1.5   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 models with different parameter combinations could be specialized to preform 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 [15].

An ensemble of models is defined as creating duplicated models with altered initial conditions, boundary conditions, types of model, and parameter combinations [15]. We we create a model there are many sources of uncertainty, we do not know what the true mechanisms that drive human-tick encounters, so we must take educated guesses at which predictors are useful to include. We began by intuitively selecting parameters and used the statistical frame work of the poisson point process to guide predictor selection for different times of the year. We also preformed 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 patterns of observations conditional on time of year, it is possible that our ensemble, as many ensembles are, is underdispersive meaning that the ensemble is not as variable as would be expected given the diversity of its members [16].

Variability is desirable in an ensemble because, it ensures robustness and representation of future potential variability [15] Underdispersion occurs because it is very difficult to capture all of the types of uncertainty that exist and thus there 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. [15]

1.5.1   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 preform better than others conditional on the time of the year. Future work includes more rigorous testing of the ensemble members on future collected data and storing model performance over the long term. With this information we would be able to discern which models are more skill full an properly calibrate the ensemble as described in [16].

Chapter 2

Exploratory Analysis

2.1 Temporal Stationarity Experiments

The initial experiment in exploratory analysis was to determine how each of the environmental parameters that we are potentially going to be using in the MaxEnt models are related to one another. In order to do this we carry out a correlation analysis to understand how each pairwise set of parameters varies with one another.

We broke the experiment into four seasons to see if there might be different highly correlated covariates depending on the season. Then we looked at the entire year a more general picture. Below is a section of code used to run the experiment. The dataTools library contains all of the main functions used throughout the code snippet. Here we define

```
library(dataTools)

window = c(-3,3)
params = c("first_vegtyp", "max_airtemp", "mean_airtemp", "mean_relhum",
           "mean_sncvr", "mean_sndep", "mean_trnstr", "mean_uwind",
           "mean_vegcvr", "mean_vwind", "mean_wilt", "min_airtemp",
           "sum_precip")


fall_days = c(273, 280, 287, 294, 301, 308, 315, 322, 329, 336)
data = create_multi_week_set(fall_days, window, params)
corr_matrx(data,length(params),"fall.csv")

data_matrx = read.csv("fall.csv")
process_corr_matrix(data_matrx,params)
```

2.2 Poisson Point Process Models for Predictor Selection and Evaluation

2.2.1 Building and Interpreting Point Process Models

My first look at the Poisson Point Process Models for explaining the distribution of points across the state focused on Day 150 over the years of our experiment 2006-2013. I tried two different window sizes: +/- 3 days and +/- 7 days. Both window periods have lead me to similar conclusions, however the 7 day window had twice

as many points so I will discuss my findings for this window.

Figure 2.2 shows the points observations for the tick sightings on this day of the year for the experiment time period plotted over the state of Maine. We see that there is a good geographic distribution along the highly populated areas of the state.

We have found three different models that show all highly significant predictors for this day in time:

$$\lambda(i) = e^{-432.54 + 1.37*airtemp^*(i) + 0.49*meanHumidity(i) + 0.038*meanVegcvr(i) - 5.10*uwind(i) + 2.69*vwind(i) + 1.0*sumPrecip(i)}$$

$$(2.1)$$

We say *airtemp**because you can include here mean, max or min airtemp and come to similar results.

$$\lambda(i) = e^{-483.93 + 1.54*meanAirtemp(i) + 0.58*meanHumidity(i) + 4.46*trnstr(i) - 5*uwind(i) + 1.35*vwind(i) + 0.81*sumPrecip(i)}$$

$$(2.2)$$

2.3   Predictor Selection Process

TABLE 2.1: Long table caption.

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)

A

Correlation_Coef	Params
0.9931825	mean_airtemp ~ min_airtemp
0.9670209	max_airtemp ~ min_airtemp
0.8898020	mean_trnstr ~ mean_wilt
0.7620282	max_airtemp ~ mean_vegcvr
0.7494500	mean_airtemp ~ mean_vegcvr
0.7367836	mean_vegcvr ~ min_airtemp
0.6233336	mean_sncvr ~ mean_sndep
0.5207012	mean_trnstr ~ mean_vegcvr
0.4164788	mean_vegcvr ~ mean_wilt
0.4158549	mean_relhum ~ sum_precip

B

Correlation_Coef	Params
0.9835252	mean_airtemp ~ min_airtemp
0.9225929	max_airtemp ~ min_airtemp
0.8718058	mean_sncvr ~ mean_sndep
0.8650734	mean_trnstr ~ mean_wilt
0.8425126	mean_trnstr ~ mean_vegcvr
0.6824610	mean_vegcvr ~ mean_wilt
0.4330516	mean_airtemp ~ mean_vegcvr
0.4275942	max_airtemp ~ mean_vegcvr
0.4120246	mean_vegcvr ~ min_airtemp

C

Correlation_Coef	Params
0.9760519	mean_airtemp ~ min_airtemp
0.8892744	mean_trnstr ~ mean_wilt
0.8836858	max_airtemp ~ min_airtemp
0.6941048	mean_sncvr ~ mean_sndep
0.6445540	mean_sncvr ~ mean_trnstr
0.6340704	mean_relhum ~ sum_precip
0.5949007	max_airtemp ~ mean_vwind
0.5920286	mean_sncvr ~ mean_wilt
0.5221874	mean_relhum ~ min_airtemp
0.5121170	min_airtemp ~ sum_precip
0.5108828	mean_airtemp ~ mean_vwind
0.4763725	mean_airtemp ~ mean_relhum
0.4563481	mean_vwind ~ min_airtemp
0.4557932	mean_airtemp ~ sum_precip

D

Correlation_Coef	Params
0.9914439	mean_airtemp ~ min_airtemp
0.9482529	max_airtemp ~ min_airtemp
0.8684307	mean_trnstr ~ mean_wilt
0.8034460	mean_vegcvr ~ min_airtemp
0.7789160	mean_airtemp ~ mean_vegcvr
0.7072997	max_airtemp ~ mean_vegcvr
0.6891395	mean_sncvr ~ mean_sndep
0.4807520	mean_trnstr ~ mean_vegcvr
0.4720017	mean_relhum ~ sum_precip
0.4414010	max_airtemp ~ mean_vwind
0.4279856	mean_vegcvr ~ mean_wilt
0.4262981	mean_vwind ~ min_airtemp
0.4238627	mean_airtemp ~ mean_vwind

E

Correlation_Coef	Params
0.9943225	mean_airtemp ~ min_airtemp
0.9704051	max_airtemp ~ min_airtemp
0.8769499	mean_trnstr ~ mean_wilt
0.8148860	mean_airtemp ~ mean_vegcvr
0.8098032	max_airtemp ~ mean_vegcvr
0.8075747	mean_vegcvr ~ min_airtemp
0.6161498	mean_sncvr ~ mean_sndep
0.5336131	mean_trnstr ~ mean_vegcvr
0.4719720	mean_vegcvr ~ mean_wilt
0.4664747	mean_relhum ~ min_airtemp
0.4231692	mean_airtemp ~ mean_relhum
0.4116809	mean_relhum ~ sum_precip

FIGURE 2.1: Correlation Analysis between parameters for four seasons, fall (A), summer (B), winter (C), Spring (D) and all year (E).

Chapter 3

Sources

[1] CDC https://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] Manuel Gimond, "Introduction to GIS", Colby College

[5] 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

[6] 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.

[7] 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.

[8] Peterson, A. Townsend, et al. "Transferability and Model Evaluation in Ecological Niche Modeling: a Comparison of GARP and Maxent." Ecography, vol. 30, no. 4, 2007, pp. 550?560., doi:10.1111/j.2007.0906-7590.05102.x.

[9] 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.

[10] 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.

[11] Berger, Adam.
http://www.cs.cmu.edu/afs/cs/user/aberger/
www/html/tutorial/node2.htmlSECTION00011000000000000000

[12] 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.

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

[14] https://mgimond.github.io/ES218/Week01.html

[15] Thuiller, Wilfried, et al. "BIOMOD - a Platform for Ensemble Forecasting of Species Distributions." Ecography, vol. 32, no. 3, 2009, pp. 369?373., doi:10.1111/j.1600-0587.2008.05742.x.

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