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

December 28, 2017

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 State Government 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 tick sightings, tick gender and disease status, 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 will 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.

1.1.1 Data used in Species Distribution Models

This database of tick encounters developed by the Maine state government houses an enormous amount of information about the distribution of likely locations for ticks sightings in the state of Maine, however in its raw form it is missing a crucial metric: information about where ticks are not likely to be found. This lack of essential data is the crux of what makes species distribution models difficult to develop.

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 rate and temperature to determine the response of 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 those who are sick in order to be able to discriminate sick from healthy metrics. Unfortunately, our tick dataset does not provide information about the absence of ticks in locations, which motivates the need to use different models. The primary goal of this thesis is to develop a model to best predict human-tick encounters based on our data. We will be using methods such as the maximum entropy model MaxEnt, and randomly generated ensembles of models, which have been shown to be effective, with data such as ours [cite].

1.2 Predictor Selection and Exploratory Analysis

Our dataset is special in several ways: one way which is mentioned above, is that contains a subset of human-tick encounter observations from the state of Maine over a 15 year period, but no information about the subset of human-no tick encounters, a crucial piece of information necessary in classification algorithms. Another complicating factor of this dataset is that it contains data concerning spacial locations across time; spacial and temporal dimensions require explicit analysis and dictate the use of distinct statistical techniques. Spacial analysis is the use of specialized statistical techniques that are designed to help identify patterns that derive from the spacial dimensions of the data [4].

1.2.1 Spacial Analysis

In species distribution modeling there are two components to the data set that will be used to make predictions on the location of ticks in the environment. We have described the first part of the dataset, the tick observations, a set of latitude, longitude pairs that uniquely define a point in space where a tick has been sighted by a person, as well as the date of observation. The second part of the dataset are the predictors, which are a set of environmental variables that will be used to define the specie's environmental niche [3]. Environmental data, collected by the weather service can be downloaded for the location and date of the sightings and will be stored in memory as a raster object, which is a grid of pixels used to represent spacial data[4].

Selection of appropriate predictors to use for model building can be very difficult because it takes a good deal of understanding about the way in which the species interacts with its environment to understand what types of environmental parameters influence a its ability to live in an environment. In our case, many additional factors influence the ability of a tick to survive and disperse throughout its environment such as availability of host animals, whose distribution may also be unknown. Further, once we have an idea of what parameters are appropriate to use in to model, based on the literature of known tick habitat conditions, we must be careful to select only variables that are not correlated with one another to include in model building, otherwise contribution estimates for the predictors are confounded.

Based on our insights, before attempting to build a predictive model we must preform an exploratory analysis of the relationships between our predictors and our observations. Creating pairwise regression models of each predictor against one another, help us to identify and quantify correlation between covariates. Once we

have identified which predictors are significantly correlated with one another, we must select a subset of parameters which optimally characterize tick behavior while reducing the possibility of correlation. For example, if we hypothesize that the two predictors vegetation cover and elevation are important parameters in characterizing a tick's suitable habitat, but find that the type of vegetation in an ecosystem is almost completely determined by the ecosystem's elevation, we cannot include both parameters in our model, but which should we include?

A branch of spacial analysis called point pattern analysis can help us understand how the tick observation points are distributed differently across different covariates and may help us to further understand the utility of each parameter. There are several types of point pattern analysis algorithms, however, we will focus on two in our analysis: quadrant density and Poisson Point process. The premise of both of these techniques is the same: to understand how the distribution of the observation points relate to the distribution of the covariate.

In the quadrant density method as the researcher you must look at the distribution of the covariate and break it into discrete categories, in our example of elevation, we might separate the predictor into a high, medium and low elevation group for example and then compute density of the region, defined as the number of points over region area [4]. Since the categorization of the variable is determined at the discretion of the researcher, very different conclusions can be made based on the categorization strategy, which can lead to unreflective conclusions, however the Poison Point process will allow us to use a less biased approach based on creating a logistic regression model of points based on the predictor:

$$\lambda(i) = e^{\alpha+\beta(i)} \tag{1.1}$$

This logistic model allows the researcher to understand the density of point observations at any arbitrary input location, and thus understand more deeply the relationship that the point observations have with a particular covariate.

Also, other examples have included the use of PCA analysis as a preprocessing step to reduce the features used in modeling building to independent highly effective features, with reduced dimensionality [8].

1.2.2 Temporal Analysis

Time also plays a key role in our model. 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 analysis on the tick distribution and time. 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. We may display the resulting correlation coefficients for the a particular covariate at a particular location on a heat map to get a picture of how for a given predictor it correlated throughout time in space.

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.

1.2.3 Communicating Uncertainty

There is immanent uncertainly in the models that we build to predict human-tick encounters, and clearly communicating this expected margin of error is important for those who seek to use the models. One source of error for our models comes from the data that we use as predictor variables. Our models inherit the uncertainty of the tools used to measure these environmental variables. There is generally well understood data on the accuracy and expected error for satellite weather data.

Another source of uncertainty and error however, comes from the point observations of the human-tick encounters. Since there is likely sampling bias in the data collection methods due to human activity not being randomly distributed throughout the state of Maine, some areas of the state or more represented and over-sampled, while other areas, particularly in the Northern regions may be left un-sampled. Thus, if we have weather data accurate to the county level in Maine, then we may report our forecast predictions at this level of aggregation, however there will be unequal uncertainty in each of the counties which must be quantified and reported.

1.2.4 Data Quality and Selection Bias

We need to address selection bias, spacial-auto correlation, correlation with roads, sampling effort is correlated with population density.

1.3 The MaxEnt Model

Since we have previously described that we cannot use traditional models to build a tick-human encounter predictive model, due to the fact that our data contains only information about the presence locations of observations and no information on their absence locations, we will begin by discussing the MaxEnt model and how methods of its model building process allow it to work well with presence only data. The basic premise of the MaxEnt model is that the model has two probability densities as input: the point observations of tick locations and the environmental predictor variables [6].

The algorithm at its core needs to solve the problem of minimizing the entropy between the the two probability densities. It is through this process that the predictor variables undergo transformations that help to maximize the entropy of the solution during the model fitting process [6]. The covariates can become terms of the following type in the final model: linear, quadratic, product (representing interaction),

hinge, threshold or categorical, and the terms are deemed effective through cross validation and an L1 regulation process [6].

To better understand the MaxEnt model, we first need to get a better idea of what the maximum entropy method is. Maximum entropy and minimum cross entropy are methods which are based in Bayesian statistics and information theory. If we begin by thinking in the discrete setting, we understand that there is a set of ticks and they are distributed with a certain probability function based on the environmental parameters with a given probability. In a way, we already have an idea about where tick are distributed in space based on our observations and so we know some information about their distribution, but we are missing information. We do not have the complete picture of all of the suitable locations and we have no information about unsuitable locations.

Since we do have some information about the distribution however, we can use a cross entropy approach [10]. Minimizing the cross entropy of a system is analogous to finding the relative entropy which is analogous to maximizing entropy [10]. At this point, we have some information about the distribution of ticks and we also have constraints on our distribution, the region that we are mapping to, which is represented by background points. The theory of minimizing cross entropy says that given the information that we do have we can actually find a unique distribution that 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. Basically what we need to do is minimize the cross entropy function:

$$\int q(x) log(\frac{q(x)}{p(x)}) dx \tag{1.2}$$

where $p(x)$ is the probability density function of the distribution of ticks and $q(x)$ is the probably density function of the tick encounter likelihood based on weather conditions. We have an estimate of $q(x)$ because of our observations, but we don't know anything about q(x) other than some constraints about geographic location and weather conditions.

When we began the modeling process, we had our dataset of tick observations, which highlight 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 with 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.

The algorithm that we will be using to create our maximum entropy model is called MaxEnt and is supported by the R programming language, although the main program is written in Java, by S.J. Phillips et al. The MaxEnt model is of the form:

$$q(x) = \frac{e^{\lambda * f(x)}}{Z} \tag{1.3}$$

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].

The L1 regularization method for feature selection to avoid overfitting []...

1.3.1 Current MaxEnt model

-> predictors used, model structure

1.3.2 Additional Features of MaxEnt

-> population density,
host density,
weightings versus adding predictors

1.4 Ensemble Models

...

1.4.1 Ensembles with MaxEnt

* model physics
* initial conditions
* training data

1.5 Model Assessment

1.5.1 Assessment of a single model

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.

$$\begin{array}{cc} & \begin{array}{cc} Observed\ Presence & Observed\ Absence \end{array} \\ \begin{array}{c} Predicted\ Presence \\ Predicted\ Absence \end{array} & \begin{pmatrix} a & b \\ c & d \end{pmatrix} \end{array} \tag{1.4}$$

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

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 their 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):

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

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 this powerful statistic, we will need to determine if there is a modified version that we can use.

A few other metrics to consider are forecast AUC, AIC, and the procedure of cross validation which we will discuss later on ....

1.5.2 Assessment of an ensemble of models

* summarizing ensembles -> a bayesian approach

Appendix A

Code for Experiments

A.1 Experiment 1

The code for this experiment has been written in R

Appendix B

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-Peña, 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] A statistical explanation of MaxEnt for ecologists Jane Elith, Steven J. Phillips, Trevor Hastie, Miroslav Dudi, Yung En Chee and Colin J. Yates

[7] A. Townsend Peterson, Monica Papes, Jorge Sobero, "Rethinking receiver operating characteristic analysis applications in ecological niche modeling"

[8] Peterson, A.T., Papes, M., Eaton, M., 2007. Transferability and model evaluation in ecological niche modeling: a comparison of GARP and Maxent. Ecography 30, 550–560.

[9] OMRI ALLOUCHE, ASAF TSOAR and RONEN KADMON. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS)

[10] Axiomatic Derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross-Entropy JOHN E. SHORE, MEMBER, IEEE, AND RODNEY w. JOHNSON

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

[12] S.J. Phillips et al. Maximum entropy modeling of species geographic distributions