Gap Analysis of Five Orders in Great Smoky Mountain National Park:

A Quantification of Inventory Gaps

by
Micah Jasny
Stuart Pimm, Advisor

4/29/16

Masters project submitted in partial fulfillment of the

requirements for the Master of Environmental Management degree in Ecosysten Science & Conservation

the Nicholas School of the Environment of

Duke University

I. Executive Summary:

Global biodiversity is currently in a state of crisis with human alterations to the environment exacerbating extinctions so that extinction rates now exceed 1,000 times normal background rates (Lees & Pimm, 2015, Nicholas & Langdon, 2007). To better understand and protect global biodiversity, the All Taxa Biodiversity Inventory (ATBI) project was founded to determine the identity, distribution, and function of every species present within a specific study location (Sharkey, 2001). The most famous ATBI was established in 1998 at Great Smoky Mountain National Park and has since identified over 20,000 species with almost 1,000 species new to science (Discover Life in America, 2014; Nichols & Langdon, 2007; Parker & Bernard, 2006). To help the GSMNP ATBI project use its resources more efficiently, I conducted a taxonomic gap analysis for five orders to identify whether more species may potentially exist within the Park's boundaries that have yet to be added to species inventories. If inventory gaps were present, I then estimated the total species richness to determine which order had the largest taxonomic gap and should thus be the focus of future sampling efforts.

The Park had previously identified five orders that it believed may contain taxonomic gaps: crustaceans, diptera, hemiptera, hymenoptera, and acari. Given species presence locations for these orders, I generated species accumulation curves to determine if taxonomic gaps were present in the Park's inventories. The species richness modeling program EstimateS was then used to quantify total species richness for each of the focal orders within the Park (Colwell, 2013). Given the estimated total species richness and the number of species previously found by the Park for each order, I was able to quantify the taxonomic gaps in each order's inventory in terms of the total number of species and the percent of the order identified.

To determine where future sampling efforts should be focused to identify the remaining species, I used the species distribution program, Maxent to locate areas of high species richness for each order within the Park (Philips, Dudik, &Shapire, 2010). I compiled 15 environmental predictor layers at a 30m resolution which were uploaded into Maxent along with the presence points of all species that were present at 15 or more locations. Habitat suitability produced by the model was then thresholded and stacked for all species within each order to identify areas of high species overlap. For cases in which the data were spatially structured, bias files were constructed to remove the direct and indirect influences of spatial bias on the models.

From the species accumulation curves, it was determined that crustaceans, diptera, hemiptera, and acari all likely had species yet to be found within the Park, while the hymenoptera accumulation curve approached an asymptote at around 550 species. After conducting the species richness analysis, hemiptera was found have the largest gap with about 334 species potentially yet to be identified. However, acari has the largest gap in terms of the percent of the order identified by the Park (45.74%). This Park can now decide whether to view taxonomic gaps in terms of the potential number of species or percent identified.

After order-level species distributions were calculated, hemiptera and hymenoptera had the highest richness in the northern and central parts of the Park, diptera and acari were found primarily along streams, and crustaceans had the highest richness in the western parts of the Park. Distance to streams, soil type, vegetation, and slope all play critical roles in defining habitat suitability for these focal orders and all five focal orders can be sampled in the areas in close proximity to Park streams.

This analysis identified that past sampling efforts have primarily occurred along trails and roads within the Park so future sampling should be tailored to focus away from these

anthropogenic features to avoid under sampling. One of the benefits of this analysis is that its accuracy improves as more sampling is done. With more sample presence locations, both the species richness and species distribution models better reflect reality. Combining species richness and species distribution modeling to structure species inventory efforts will result in more efficient and effective use of resources which will allow the ATBI to better conserve and protect biodiversity within the Park.

II. Introduction:

Biodiversity is defined as the "variety and variability of organisms" and can be described as the richness of entities (genotypes, species, or even ecosystems), the evenness of the distribution of entities, and their interactions (Hooper et al., 2005; National Research Council, 1999). While there are many methods to estimate the number of species on Earth, the total is thought to range from 2 to 50 million different species (Scheffers et al., 2012). In 2014, over 1,730,725 species had been identified by taxonomists (World Conservation Union, 2014). Even in the most thoroughly studied taxonomic groups – birds, mammals, and amphibians – new species are constantly being discovered (Lees & Pimm, 2015).

Biodiversity plays a fundamental role in determining ecosystem structure and function as species composition, richness, and evenness impact and respond to ecosystem properties (Gamfeldt et al., 2008; Hooper et al., 2005). Different species can alter abiotic conditions which impact the energy and material fluxes that help to define an ecosystem (Chapin et al., 2000). The loss of even one species may impact multiple ecosystem functions which can lead to a destabilization of the ecosystem (Gamfeldt et al., 2008). With regards to humans, biodiversity provides goods and services that are necessary for sustaining human life and is linked to

numerous cultural, intellectual, aesthetic, and spiritual values (Chapin et al., 2000; Gamfeldt et al., 2008).

However, global biodiversity is presently in a state of crisis as the population trends of numerous species have continued to decline over the last four decades (Butchart et al., 2010; Hannah et al., 2002). The unprecedented rates of biodiversity change have been directly linked to human alterations of the environment including changes in land-use, atmospheric CO_2 concentrations, nitrogen deposition, acid rain, and the introduction of invasive species (Sala et al., 2000). Changes in biodiversity will lead to even greater changes in productivity, hydrology, carbon storage, and nutrient cycling which could result in a feedback loop of detrimental alterations to natural communities (Hooper et al., 2005).

The result of these changes has been a dramatic acceleration in extinction rates with current rates exceeding 1,000 times normal background levels (Lees & Pimm, 2015, Nicholas & Langdon, 2007). Humans have already caused the extinction of 5% to 20% of species in many groups and currently a third of all U.S. native flora and fauna are imperiled or of conservation concern (Chapin et al., 2000; Nicholas & Langdon, 2007). The rate of extinction has increased to the point that it is now predicted that species are going extinct before they can be discovered and many are identified only to die-out within the following few years (Lees & Pimm, 2015).

The ecologist and conservationist Daniel Janzen once said, “to truly insure its survival into perpetuity, we need to know what our biodiversity is” (Janzen, 2002). In response to the rapid loss of species and habitat around the world, a group of scientists began conducting organized inventory events aimed at further expanding our understanding of species composition and biodiversity (Nicholas & Langdon, 2007). One of these events is the All Taxa Biodiversity Inventory (ATBI) project, which is a continuous inventory project aimed at determining the

identity, distribution, and function of every species present within a specific study location (Sharkey, 2001).

The oldest and most well-known ATBI is currently being conducted at Great Smoky Mountains National Park (GSMNP) by Discover Life in America (DLIA) (Sharkey, 2001). This ATBI was formed on Earth Day, 1998 and has engaged the work of hundreds of scientists and thousands of volunteers (Nichols & Langdon, 2007). Since then, the ATBI has helped discover close to 20,000 species in within the park, including almost 1,000 species that were new to science (Discover Life in America, 2014; Parker & Bernard, 2006). The ATBI project is presently nearing the end of its second decade of work for GSMNP and while they have greatly expanded the Park's inventory, there is still more work to be done. The purpose of my analysis is to identify and quantify the remaining gaps in the Park's inventory to define the focus of future ATBI research.

In studying the gaps in the Park's inventories, I have conducted a taxonomic gap analysis. In past studies and conservation, gap analysis has referred to an assessment of whether the current range of a target species is being fully protected by conservation efforts (Gap Analysis Process, 2015; Jennings, 2000; Rodrigues et al., 2004; Scott et al., 1993). This approach is centered on the biogeography of species of conservation concern (Gap Analysis Process, 2015; Jennings, 2000; Rodrigues et al., 2004; Scott et al., 1993). When gap analysis is used in this study, it is referring to potential gaps in our current understanding of taxonomic species richness within the study area or a "taxonomic gap analysis".

For this gap analysis, I studied five orders suspected by the Park to contain gaps in the species inventories: crustaceans, diptera, hemiptera, hymenoptera, and acari.

The questions this analysis addresses are:

1.) Given five priority gap orders, do these orders have potentially more species to be found?
2.) If inventory gaps do exist, what is the estimated total species richness for these orders within the Park?
3.) Which order has the largest estimated taxonomic gap in terms of both the total number of species and the percent of the order yet to be found?
4.) Based on the species the Park has already identified, what areas within the Park have high species richness for the five orders and should be the focus of future sampling effort?

Currently, 42,000 species of crustaceans have been identified and described (Natural History Collections, 2007). Crustaceans are predominantly aquatic and are an essential basal group for many aquatic system food webs (Natural History Collections, 2007). Diptera include all "true flies" which have small, club-shaped hindwings and live in a variety of fresh water aquatic, semi-aquatic, and moist terrestrial habitats (Meyer, 2009). Around 95,800 species of diptera have been identified globally, most of which feed on dead organic material or parasitize other animals, though some are herbivorous (Meyer, 2009).

Hemiptera or "true bugs" are extremely diverse with between 50,000 and 60,000 hemiptera identified world-wide (Hemiptera, n.d.) Most hemiptera species are terrestrial and feed on plants, though some species are parasitic or live on the surfaces of water in aquatic habitats (Hemiptera, n.d.). Hymenoptera or "membrane-winged" insects contain a large number of insect taxa including bees and ants (Waggoner & Speer, 1997). Hymenoptera species are found in a variety of terrestrial habitats and play crucial ecological roles as primary pollinators for many plant species (Waggoner & Speer, 1997). The final focal order being analyzed for inventory gaps is acari. There are about 30,000 known acari species including tick and mite species, however it

is believed that many more species have yet to be discovered (Waggoner, 1997). Acari are parasitic organisms and occur in a great abundance of habitats (Waggoner, 1997).

This goal of this study is to help DLIA and the Park to use their resources more efficiently by providing insights into how inventory sampling has been conducted thus far, whether sampling design should be altered in the future, and determining which order(s) should be the focus of future sampling. While the GSMNP ATBI has made unprecedented strides in species identification and inventory work, it is essential that the ATBI project continue to grow and evolve, as it pioneers new efforts to conserve global diversity (Dunn, 1993).

III. Materials & Methods:

a. Study Area:

The focal study area for this analysis is the Great Smoky Mountains National Park (GSMNP). The Park was first established in 1934 and protects 800 square miles of the southern Appalachians, encompassing land in both Tennessee and North Carolina (White, et al., 2000). The Park's elevation ranges from 266m to 2,016m above sea level (U.S. National Park Service, 2016). The high country areas receive an average of 216cm of rain each year which help support the Park's five major forest types: cove hardwood forest, spruce-fir forest, northern hardwood forests, hemlock forest, and pine-and-oak forests (U.S. National Park Service, 2016).

The Smoky Mountains are also some of the most diverse natural areas in North America due to the wide variety of habitats created by high levels of precipitation, a large range in elevations, and complex geology (Ditmanson, 2008; White et al., 2000). At the beginning of the ATBI at GSMNP in 1998, the Park had identified a total of 9,800 species within Park boundaries. By 2001, the ATBI had helped to double that number with over 17,000 species identified in and

around GSMNP (Sharkey, 2001). However, this number is still far short of the 100,000 species biologists have hypothesized may exist within the Park boundaries (Sharkey 2001).

b. Data:

The data used for this analysis include 45,406 specimen presence locations for the five focal orders which include the specific project the specimens were collected for, the date the sample was collected, the site code, site description, spatial coordinates for the sampling location, the collector, the taxonomic identification for each specimen, and the taxonomist identifying the specimen (Great Smoky Mountains National Park Biodiversity Database, 2015). Any samples that lacked species level taxonomic identification for the specimen were removed from the dataset. In addition to the species presence points, 15 environmental predictors were compiled in ArcMap 10.3.1 to be used to generate species distribution models for the five orders (Table 1; Esri, 2015).

c. Species Accumulation Curves

The initial step of this analysis was to determine whether gaps actually existed within the inventories for the focal orders by generating species accumulation curves for each of the five orders. A species accumulation curve describes the identification of new species against some measure of sampling effort (usually either time spent sampling or number of sampling locations) (Llorente, 1993; Ugland, Gray & Ellingsen, 2003). For this study, the number of unique species identified was plotted as a function of the total number of sites sampled.

Using R programming software, a matrix was created that contained the unique site code for each site and the number of samples taken at each site (R Core Team, 2014). The column of unique site codes was then used to create a column of the cumulative number of sites sampled. A

list of the unique species was then generated for each site and a for-loop (section of code that iterates through different user defined inputs) was used to identify the number of unique species that were present at that site which had not been found at a previous site. Finally, the cumulative number of species was combined with the cumulative number of sites sampled and plotted as the species accumulation curve (see appendix for R code).

If the species accumulation curves approach an asymptote, the rate of finding new species drastically decreases, suggesting that most of the species within that order have been identified and that a large taxonomic gap may not be present in the Park's inventory for that order. If the accumulation curve does not approach an asymptote, it indicates that there may be more species to be identified by the Park. Because unique sampling sites were used as a measure of sampling effort, the accumulation curves describe whether more species will be discovered by sampling at additional locations (Ugland & Ellingsen, 2003).

d. Modeling Species Richness

Once it was determined whether or not species gaps exist within the five focal order inventories, the next step was to estimate the total number of potential species present in each of the five focal orders within the Park (the projected asymptote of the species accumulation curves). To identify the potential total number of species existing within each order within the park, the program EstimateS version 9 was used to generate species richness estimates (Colwell, 2013). Created in 1997, EstimateS has rapidly grown in number of users and applications and has been used in conservation biology, inventory science, landscape ecology, archaeology and many other fields (Colwell, 2013). Provided either incidence or abundance data for a number of sampling sites, the program can compute non-parametric asymptotic species richness estimates from a variety of richness estimators (Colwell, 2013).

EstimateS requires a .txt file containing a table that lists species presence for every species at every possible sampling location (Colwell, 2013). This was done in R by using the table function for species identity and sampling site code and exporting the resulting table as a .txt file (R Core Team, 2014). This file was then uploaded into the EstimateS program and the program run using 100 randomizations without replacement and 200 bootstraps (Colwell, 2013).

Because the data given to me by the Park was presence data and not species abundance data, I used three incidence-based species richness estimators (Chao2, Incidence-based Coverage Estimator (ICE), and second order Jackknives (Jack2)) to estimate species richness for the five focal orders (Chao, 1984; Chazdon et al., 1998; Gotelli & Colwell, 2011; Palmer, 1991). The means for the Chao2, ICE, and Jack2 richness estimators were than averaged together for an estimate of average species richness for each focal order within the Park. These values were then subtracted by the number of species the Park had previously identified for each order to identify the number of potential species yet to be identified. Finally, I divided the number of species identified for each order by the average species richness to get an estimate of the percent of each order currently identified.

c. Species Distribution Modeling:

After the inventory gaps were quantified, the data were then used to try and specify what areas present in the Park should be the focus of future sampling effort to locate these missing species. To do this, I used the species distribution modeling program, Maxent version 3.3.3k (Philips, Dudik, & Schapire, 2010). The Maxent program utilizes sample presence points and 10,000 pseudo-absence points (randomly generated points used to create an average estimate of background environmental conditions) to predict species distribution over the inputted study area (Phillips, Dudik, & Schapire, 2010). The model algorithm used by Maxent attempts to generate a

distribution that is close to random (average environmental conditions estimated using the pseudo-absences) while still fitting to the observed presences (Phillips, Dudik, & Schapire, 2010).

Before the distribution models could be built, the species and environmental data needed to be prepared. All of the species sample locations were plotted in ArcGIS 10.3.1 (Esri, 2015). The specified datum for the spatial positions of the samples, which defines the coordinate samples of the sample locations within the park, was divided between NAD83 Zone 17N and NAD27 UTM Zone 17N with some specimens having no specified coordinate system. For the points without a coordinate system, I identified the sample collector and determined if a coordinate system was listed for other data they had collected could be used for the data missing the coordinate system. If no coordinate system was listed for any of the samples collected, a visual analysis was done projecting the points using both NAD83 and NAD27 and a coordinate system selected. The NAD83 data was then re-projected into NAD27 and combined with the rest of the NAD27 data.

The 15 environmental layers were also projected using NAD27. Because some of the environmental layers did not cover the full extent of the Park, a GIS mask was generated based on where all of the environmental layers overlapped. This was done by creating a copy of each environmental layer where cells containing data were converted to a value of 1. I added all 15 environmental layers and exported the areas where they overlapped (cell value = 15) as the study area mask. This mask was then applied to the original environmental layers to standardize the spatial extents of the environmental predictor variable layers. When the GIS mask was built using the overlap of all of the environmental layers, it did not cover the entire extent of the Park. For this reason, any specimen points existing outside of the study area were removed from the

dataset since habitat could only be modeled for areas where data points co-occurred with areas of measured environmental data.

In order to construct species distribution models with reasonable accuracy, a minimum number of presence points needs was established as a cutoff for distribution model building. It has been a general rule of thumb that 30 or more presence points are required to build an accurate species distribution model (Wisz, Mary Suzanne, et al., 2008). However, recent studies have found that certain distribution modeling methods, including Maxent, can still accurately predict the extent of species distribution models for species with fewer than 30 presence points (Papes & Gaubert, 2007; Proosdij et al., 2015). Specifically, these studies found that when using Maxent, accurate species distribution models could be created using as few as 15 presence points (Papes & Gaubert, 2007; Proosdij et al., 2015). For this reason, only species with 15 or more presence points were used to build the species distribution models for the five orders.

Once species with 15 or more points were selected for each order, .csv files were built for each order containing the x and y coordinates for each presence point along with the name of the species identified at that location. To locate species-rich areas, I decided that I was going to threshold the continuous predictions of species habitat suitability for each species to produce binary maps of species habitat and non-habitat within the study area. Species-rich areas within the Park would then be discerned by identifying which areas had the greatest number of overlapping species habitat for that order.

To build the species distribution models, I created a GIS tool that runs a Python script. The initial part of the script is a modified script of J. Donoghue (2013) which runs the Maxent program based on the parameters set by the user (see appendix for python script). I augmented the script so that the user simply had to enter the locations of the species .csv file, the file holding

the environmental ascii layers and the output file, and the model would run the program and distinguish continuous environmental predictor variables from categorical environmental predictor variables. Once Maxent was run, the GIS tool identifies the location of the continuous predictions of habitat suitability for each species and converts it from ascii back into a raster format.

Included in the Maxent results are a variety of common threshold values and corresponding omission rates used to threshold the continuous Maxent outputs. For this analysis, I decided to use the balance training omission, predicted area and logistic threshold value which balances the training omission rate, the cumulative threshold, and the fractional predicted area (Phillips, Dudik & Schapire, 2010). The GIS tool opens the Maxent results file created by the program, identifies this threshold value for each modeled species and then applies that threshold to that specific species' continuous habitat suitability prediction to create a binary map of suitable habitat. Finally, my GIS tool uses cell statistics to stack the binary species distribution rasters for all of the species within one order and counts the number of species with habitat in each cell to identify species-rich areas (Esri, 2015).

From visual inspection of species presence points displayed over the distance to trails and distance to road layers, it seems that most of the inventory sampling effort occurred primarily along these man-made or anthropogenic features. This makes sense since many parts of the Park are largely inaccessible except by traveling along either roads or trails. This sampling design, however, results in an artificial inflation of the impact of roads and trails on species habitat suitability. If this spatial bias is not addressed, species habitat would occur predominantly along trails and roads within the Park.

To decrease bias inherent in the data, I generated bias files for the Maxent program, which have been shown in recent studies to increase the accuracy of models based on spatially biased data (Kramer-Schadt et al. 2013, Phillips et al. 2009, Syfert et al. 2013). The Maxent program default assumes that the presence points used to build the model were systematically sampled throughout the study area, which is not usually the case (Kramer-Schadt et al. 2013). The bias file addresses the flaws in this assumption by allowing the user to input a grid that biases the Maxent pseudo-absence points in the same way the data is biased (Kramer-Schadt et al. 2013, Phillips et al. 2009).

By analyzing which environmental variables were the most important for delineating suitable habitat for each species I was able to discern if distance to trails, distance to roads, or a combination of the two was the biggest source of spatial bias for each order. In the case that both trails and roads biased the data, the two layers were combined and the distance from both trails and roads generated using the Euclidean distance tool in ArcMap (Esri, 2015). The distance of each presence point to the bias-generating layer was then extracted and the distribution of the distances of all of the points within the order analyzed. Because multiple species were identified at the same sampling points in all five orders, it was decided that order-level bias files were sufficient to remove the bias instead of building bias files for each individual species.

Once the distribution of distances was analyzed, a distance which would encompass a large majority of the presence points was selected as a cutoff distance. The distances within and beyond the cutoff value was then reclassified to represent proportional sampling effort within the study area. When a bias file is uploaded into the Maxent program, it is used to bias the program's selection of background pseudo-absence points (Phillips, Dudik & Schapire, 2010). This

decreases the spatial bias "noise" within the data, allowing the ecological "signal" to be more prominent in the results.

Once the bias files were created for all five of the focal orders, new species distribution models were created. The GIS tool I wrote for Maxent did not include bias file entry, so I ran the Maxent program by hand and thresholded the continuous predictions of habitat suitability to binary maps using a subset of the Python code. The resulting distribution maps and variable importance for defining habitat were analyzed and compared to the original, spatially biased results.

The reason that a bias file is required to address the spatial bias instead of simply removing the spatially biased predictor layers is because of the possibility of correlation with other environmental layers. For example, because a large amount of sampling was conducted along trails and roads, there may be bias in elevation or habitat type if most trails and roads occur at a specific elevation or in a particular habitat type. By including a bias file when constructing species distribution models, it addresses the direct bias as well as the indirect bias created by correlation with the spatially biased layers.

Because this analysis is focused on the environmental factors that affect the priority orders, the species distribution models were conducted a third time using the bias files generated from roads and trails, but removing roads and trails as possible explanatory variables for predicting species habitat. To determine which environmental factors were the most important for delineating habitat for each order, the top three predictor variables that defined habitat suitability for each species were selected from the Maxent output and a plot of the number of times one of the environmental predictor variable was found to be important was plotted and compared to the others.

Since this is an All Taxa Biodiversity, and the remaining inventory gaps are going to have to be filled at some point in the future, I reclassified and combined the distributions for the five focal orders to identify locations within the Park where at least one species from all five of the focal orders could potentially be found. Each of the order-level distribution maps were reclassified so either no species where found (0) or at least one species within the order could potentially exist (1). The reclassified layers were then added together to produce a map depicting where the focal orders overlapped which may be useful for future sampling efforts.

e. Analysis of Sampling Design:

To study whether the past trends of sampling predominantly near roads and near trails was sufficient for capturing species composition for the five focal orders within the Park, a comparative analysis of species composition both near and far from roads and trails was conducted using the mean sample distance for each order. Species accumulation curves were then created using the data within and beyond the median sampling distance for trails and roads and the total number and number of unique species (species only identified either within or beyond the median distance) were identified and compared. If the species accumulation curves within the median distance approached an asymptote it would mean that the Park has identified most of the species in that area and needs to focus future sampling efforts further away from trails. However, if the species accumulation curves do not reach an asymptote and it is discovered that no new species are being identified further away from trails and roads, it means that the Park has yet to identify all of the species near trails and roads and should continue sampling as it has done in the past.

For this analysis, the datasets that had been loaded into Arcmap had the distance from trails and distance from roads extracted to each sample point and exported to a new .csv file. The

median distances for trails and roads were then calculated and the datasets split into two groups containing points within and beyond the median distance for trails and roads. Species accumulation curves were then built for each distance data subset and compared using the code that had been applied earlier in the analysis. The data within and beyond the median distance for trails and for roads were then compared in terms of the total number of species and number of unique species found within each distance regime using R (R Core Team, 2014).

IV. Results:

From the species accumulation curves (Figures 1-5), there is a clear difference between the orders in terms of the number of species in each order identified by the Park. Crustaceans had the fewest identified species with only 11 species followed by acari with 40 species. Hymenoptera, hemiptera, and diptera had high species richness with 528, 618, and 677 identified species. The species accumulation curves for diptera, hemiptera, and hymenoptera follow a sigmoidal pattern while the curves for crustaceans and acari follow a more linear trend (Figures 1-5). Out of the five accumulation curves, only hymenoptera appears to be approaching an asymptote at approximately 550 hymenoptera species (Figure 4).

With the species accumulation curves generated, the total species richness for the five focal orders within the Park were then estimated by means of EstimateS (Table 2). A total of 14 crustacean species, 962 diptera species, 952 hemiptera species, 721 hymenoptera, and 87 acari species were potentially present within Park boundaries, as determined from the average of the means for ICE, Chao2, and Jack2 richness estimators (Table 2). In terms of which order had the greatest number of potential species yet to be found, hemiptera had the largest mean gap with about 334 potential unidentified species (Table 2). Hemiptera was followed by diptera with about

285 unfound species, hymenoptera with 193 unfound species, acari with 40 unfound species, and crustaceans with only 3 unfound species (Table 2).

However, when the number of species identified by the Park was divided by the mean estimated richness to identify percent of species previously discovered by the Park, a different picture emerged. According to this analysis, the Park has identified roughly 80.90% of crustaceans, 73.19% of hymenoptera, 70.40% of diptera, 64.92% of hemiptera, and 45.74% of acari (Table 2). This means that acari have the biggest gap of the five focal orders in terms of the percent of the order previously identified.

Once the gaps in the Park's inventories had been quantified, the next step was to spatially model species richness to identify areas for future sampling. The number of presence points for identified species ranged from 1 (a large proportion of identified species had only one recorded presence) to 88 presence points. Removing all species with less than 15 presence points removed a large fraction of the data (Figure 6). Only 2 of the original 11 crustacean species had 15 or more presence points, and of the 677 diptera species identified within the Park, only 23 species met this modeling requirements (Figure 6). Most of the hemiptera, hymenoptera, and acari species were also relatively rare with only 10 hemiptera species, 26 hymenoptera species, and 3 acari species having 15 or more presence points (Figure 6).

Despite this decrease in the number of species, the species distribution models for the five orders were generated and the amount of overlap identified (Figure 7). From the initial overlap richness maps produced by the models, all of the orders except for diptera have a dendritic pattern of habitat distribution that follow the trails and roads throughout the Park which is likely caused by spatial bias present in the data (Figure 7). When analyzing variable importance for these four orders, distance to trails and roads explained a large proportion of what

defined habitat suitability for most species. Diptera also had a dendritic pattern for habitat, but most habitat seemed to be located in close proximity to streams instead of near roads and trails (Figure 7).

To remove the spatial bias in the data and the influence that trails and roads were having on habitat delineation, the distribution models were re-run using the bias files (Figure 8). When analyzing variable importance for the modelled species, distance to trails and roads were no longer the most important variables in delineating habitat. Despite removing the main influence distance to trails and roads, these anthropogenic layers still had some influence on the distribution models as they were still used as possible explanatory predictor variables.

To create a solely ecological image of the factors influencing habitat suitability for the species within the five focal orders, the models were generated a third time using the bias files, but removing the distance to trails and roads layers as environmental predictor variables. This removes the anthropogenic layers as explanatory variables, but still reduces bias caused by correlations between the remaining environmental predictor variables and the anthropogenic layers. The resulting order-level species distributions further reduced the dendritic pattern in habitat within the Park (Figure 9). The resultant models are built using only the ecological variables important to the species within the five focal orders.

The top three variables defining habitat suitability for each modeled species within each focal order was taken and plotted to identify which variables are important for identifying habitat for the order (Figures 10-14). For crustaceans, vegetation type was important for both of the modeled species (figure 10). Distance to streams, disturbance, elevation, and understory vegetation type were also found to be important variables in defining habitat for one of the two modeled crustacean species (Figure 10).

For diptera, distance to streams was found to be the most important variable for delineating habitat for all 23 modeled species (Figure 11). This was followed by slope, which was important for 21 out of 23 diptera species (Figure 11). Additionally, vegetation type, solar radiation, organic soil content, soil type, and understory vegetation type were also found to be important generating diptera habitat (Figure 11). For hemiptera, soil type was the most important variable for 8 of the 10 modeled species (Figure 12). Slope and disturbance were also found to be important for delineating species habitat for 8 out of the 10 hemiptera species (Figure 12).

For the 25 modeled hymenoptera species, soil type was important for 20 species, followed by vegetation and disturbance which were important for 18 and 17 of the modeled species, respectively (Figure 13). Both elevation and understory vegetation type were important for 9 of the 25 modeled hymenoptera species (figure 13). Slope, distance to streams, and hillshade were also found to be important for delineating the habitat of a few of the hymenoptera species (Figure 13). Finally, distance to streams was also found to be significant for all three acari species followed by slope which was important for 2 out of the 3 modeled species (Figure 14). Disturbance, soil type, elevation and organic soil content were also found to be important for modeling individual acari species (Figure 14).

Once the distribution and variable importance were produced, the next step was to combine the distribution layers for the five focal orders to see where they overlapped (Figure 15). Once the layers were reclassified and stacked, it appears that the greatest areas of overlap have a dendritic pattern and decrease towards the center of the Park (Figure 15). The dendritic pattern of habitat is similar to that found in the habitat distributions for diptera and acari and indicates that all five focal orders can be found in close proximity to streams (Figure 15).

To determine if alterations need to be made in ATBI sampling design, the datasets were divided on the basis of the median distance and compared in terms of species accumulation and composition (Figures 16 – 25, Tables 3 – 12). Crustaceans had a median distance of 205m for roads and 513m for trails. While neither the distance to roads nor distance to trails species accumulation curves reached an asymptote, the species accumulation curves beyond the median appeared to rise at a greater slope than within the median distances (Figures 16 & 17). Comparing species composition for crustaceans, a greater total of species and unique species were identified beyond the median distance for roads (Table 3). This trend was reversed for the distance to trails comparison for crustaceans which had a total of 11 species within the median trail distance, 4 of which were not found beyond the median distance (Table 4). Crustaceans had a total of 8 species beyond the median trail distance with only 1 species found solely away from trails (Table 4).

Diptera samples had a median distance of 150m for roads and 108m for trails. From the accumulation curves built from the diptera data, it seems that the curves built from the data within the median distances for both trails and roads are closer to approaching asymptotes than the curves built using the data beyond the median distances (Figures 18 &19). However, a greater number of diptera species were found within the median distances for both trails and roads with almost twice as many unique species present within the median distance than beyond it (Tables 5 & 6).

Hemiptera had median distances of 202m for roads and 108m for trails and had similar trends in accumulation curves both within and beyond the median distances for trails and roads (Figures 20 & 21). From the accumulation curves it looks like hemiptera are closer to approaching an asymptote beyond the median distance for roads, but hemiptera within the

median distance for trails are visibly closer to approaching an asymptote than the curve for beyond the median distance (Figures 20 & 21). While hemiptera had close to even numbers of total species found within and beyond the median distances for trails and roads, within the median distances always had slightly higher totals (Tables 7 & 8). A greater number of species for hemiptera were found that only occurred within the median distances for both trails and roads and not beyond the median distances (Tables 7 & 8).

Hymenoptera was the only order where all accumulation curves approached an asymptote with median distances of 390m for roads and 108m for trails (Figures 22 & 23). For both roads and trails, the accumulation curves approached an asymptote within the median road distance than beyond it (Figure 22 & 23). Hymenoptera also had greater numbers of total and unique species found within the median distances than beyond them (Tables 9 & 10).

Acari, however, had slightly different trends. With a median sample distance of 43m from roads, the accumulation curve for beyond the median distance for roads increased at a much steeper slope within the median road distance (Figure 24). When comparing the species composition within and beyond the median distance for roads, greater numbers of acari species and greater numbers of unique species are found beyond the median road distance (Table 11). For trails, acari samples had a median distance of 234m and while the accumulation curves seemed very similar, greater numbers of total and unique species were found within the median trail distance than beyond it (Figure 25, Table 12).

V. Discussion:

The species accumulation curves revealed a distinct difference in the richness of these orders within the Park. From 500-700 species of diptera, hemiptera, and hymenoptera had

previously been discovered. However, only 11 crustaceans and 40 acari species had been identified within the Park. This brings up the question of whether the Park inventory reflects the actual species richness or the relative sampling effort. Crustaceans and acari each had fewer than 100 sites sampled while diptera, hemiptera, and hymenoptera all had upwards of 250 sampling sites. However, since only the hymenoptera species accumulation curve approaches an asymptote, it is possible there are still a large number of crustacean and acari species yet to be found within the Park.

The species richness estimation component defined species gaps in two different ways: the average potential number of species yet to be found and the percent of the order already identified by the Park. While hemiptera had the greatest number of potential species yet to be discovered, the Park has identified less than half of the acari species present within its boundaries. Additionally, acari had the least number of sampling sites visited with almost half as many as crustaceans (the next lowest sampled order).

Another interesting result from the species richness estimation was that hymenoptera did not have the smallest taxonomic gap in either the number of species or the percent of the order identified by the Park. Based on the species accumulation curve for hymenoptera, I expected it to have the smallest remaining inventory gap, but hymenoptera had the third largest number of the number of species potentially yet to be identified and third largest in terms of the percent of the order identified by the Park. The species accumulation curves do not account for differences in habitats sampled, so it is possible that more hymenoptera species may be present in under sampled habitats.

When conducting the total species richness estimation analysis, a trend in the Chao2, ICE, and Jack2 estimators was apparent. For four out of the five focal taxa, Chao2 produced the

most conservative estimate of species richness followed by ICE. Finally, the mean for the Jack2 estimator was usually larger than the other two estimators. The one instance where this trend did not occur was with acari, where the trend actually reversed with Jack2 having the lowest estimated richness and Chao2 having the greatest estimated richness. By averaging the three estimators together for an average total estimate of species richness, I reduced the influence each estimator's bias had on the final results.

From the different species distribution model runs, it seems that the addition of the bias files reduced the amount of influence that the distance to trails and distance to roads had on the models. For crustaceans, both species were found throughout the Park, but had the highest amount of overlapping habitat in the western parts of the Park. Diptera retained a dendritic pattern of species habitat throughout the park which followed the streams throughout the park (as supported by the table of variable importance for diptera). Hemiptera were found throughout the park with greater species overlap occurring in the northern parts of the Park. This trend was also shared by hymenoptera, though hymenoptera seemed to have more habitat and greater species overlap than hemiptera. Finally, acari also had a dendritic pattern of habitat that followed streams, but less pronounced than diptera. Acari habitat seemed concentrated around the edges of the Park rather than the center of the Park.

Hymenoptera had two species that did not have overlapping habitat (only 25 out of the 26 species overlapped). This suggests that there is niche differentiation between some of the hymenoptera species. Clear definitions of niche partitioning for the modeled species may be hard to address in this analysis since many of the modeled species depend on micro-habitats and the finest resolution available for the environmental variables was 30m. Nevertheless, since this

analysis is focused on the general trends in habitat selection shared by the order, I believe the data is sufficient to draw inferences on order-level habitat preferences.

The analysis of variable importance for defining species habitat for the five focal orders identified vegetation type, soil type, slope, and distance to streams as being important factors in determining habitat. For crustaceans, vegetation type was found to be the most important variable defining habitat. Soil type was the most important environmental variable for hemiptera and hymenoptera species habitat. Finally, distance to streams was the main factor delineating diptera and acari species habitat. By identifying the important environmental variables for each order and identifying the similarities shared by species within a focal order, it will help better define future sample site selection.

One assumption of this analysis is that the modeled species serve as good indicators of what defines habitat for the rest of the order. Constructing models for species with less than 15 points would have increased the number of species defining the habitat preferences, but it would also have drastically reduced the accuracy of the models and introduce error into the final results. While the spatial distributions of order species may not fully depict the habitat preferences of the order, it does serve as a way to focus future sampling efforts. By using the maps generated in this analysis as a guide to future sampling efforts, it will allow for the testing and tuning of the species distribution maps which will make them better predictors of order-level habitat preferences.

To determine if sampling along trails and roads was sufficient to capture Park species composition, or if future sampling should altered to capture a greater range of distances from anthropogenic features, the data for the five orders were split along median distances and compared. From the species accumulation curves for within and beyond the median distances for

trails and roads, only the accumulation curves for hymenoptera for both within and beyond the median distance for trails and roads reached an asymptote. Because none of the other accumulation curves for the other focal orders approach asymptotes, it indicates that more species can be found both near and beyond the median sample distances.

When comparing the species composition within and beyond the median sample distances for trails and roads, most of the previously sampled species occurred within the median sample distances. The only two exceptions were for crustaceans and acari with both having more total species found and unique species found beyond the median sample distances. However, both of these species also had the lowest number of sampling sites so this trend may change with additional sampling.

Based on these results, it is clear that additional sampling needs to be conducted both near and far from trails and roads. However, because most of the previous sampling has occurred along trails and roads, I suggest that future sampling be conducted further away from trails and roads when possible. This will reduce the amount of spatial bias currently present in the Park's inventory and allow for more accurate comparisons of species composition.

It is important to note that this is an ecological analysis and does not account for other constraints that may produce or exacerbate inventory gaps. These non-ecological factors include finding taxonomists skilled enough to actually identify collected specimens for example. Another factor is that some orders and species require more effort and equipment to identify specimens down to species-level. For example, some of the focal orders, such as crustaceans, are larger and more easily identifiable. Acari on the other hand are much smaller, and have fewer differentiating features to distinguish different species.

It is my hope that this analysis serves as the foundation for new methods for conducting inventory analyses. By combining statistical estimation to identify and quantify inventory gaps with overlapping species distribution models as a proxy for species richness, this analysis creates a two-pronged approach to species inventory efforts. This approach will be useful for different reasons at different periods in future inventory efforts. At the onset of future inventory projects, this analysis could be done to identify and prioritize taxonomic groups within the study area. Throughout the lifespan of the inventory, this analysis can be used to measure the project's progress in identifying taxonomic groups and allow the project to make faster decisions on which taxonomic groups to prioritize. If used periodically throughout an inventory, the rate of change in the percentage of species identified within a focal group can be measured which will allow the project to make predictions of when the inventory for that group will be concluded.

While this analysis includes several assumptions about the data including the spatial scale used for this analysis, the environmental predictor variables used, and the number of species used to make predictions about the rest of the order, these assumptions can easily be remedied. The assumptions of spatial scale and the environmental factors used in this analysis can be fixed by generating more environmental predictor layers at finer spatial resolutions. This will help capture microhabitat conditions and niche partitioning throughout the study area which will result in more accurate models. Finally, as more sampling is conducted, more species will be above the 15-point modeling cutoff and be used to model order-level distribution and describe important environmental predictors.

As populations decline and extinction rates continue to grow, it is essential that we strive to identify and understand biodiversity in an effort to better conserve and protect it. As E. O. Wilson said in *The Diversity of Life*, "I will argue that every scrap of biological diversity is

priceless, to be learned and cherished, and never to be surrendered without a struggle" (Wilson, 1999). It is my hope that this analysis will aid in the fight for biodiversity and will be used to better conserve and protect the amazing variety of life that currently surrounds us.

VI. Figures:

Table 1: List of environmental predictors

Environmental Predictor	Data Type	Source
Elevation	Continuous	*The National Map*: New viewer, services, and data download: U.S. Geological Survey
General vegetation	Categorical	*The National Map*: New viewer, services, and data download: U.S. Geological Survey
Understory vegetation	Categorical	Madden D. Understory Vegetation at Great Smoky Mountains National Park, Tennessee and North Carolina. The Center for Remote Sensing and Mapping Science, University of Georgia. Geospatial Dataset-1047499.
Soil type	Categorical	U.S. Department of Agriculture, Natural Resources Conservation Service:. 2006. Digital General Soil Map of the Great Smoky Mountains National Park. Geospatial Dataset-2197976.
Disturbance history	Categorical	National Park Service, Great Smoky Mountains National Park. 2007. Vegetation Disturbance History at Great Smoky Mountains National Park, Tennessee and North Carolina. National Park Service, Great Smoky Mountains National Park, Resource Management and Science. Geospatial Dataset-1045868
Distance to streams	Continuous	National Park Service: Thomas Colson. 2015. Great Smoky Mountains National Park Streams and Rivers. Geospatial Dataset-2202817.
Distance to trails	Continuous	National Park Service: Thomas Colson. 2013. Trails. Geospatial Dataset-2202813.
Distance to roads	Continuous	National Park Service: Thomas Colson. 2015. Road Centerlines, Great Smoky Mountains National Park. Geospatial Dataset-2219243
Ph	Continuous	National Park Service: Thomas Colson. 2011. Soil pH. Geospatial Dataset-2198022.
Organic soil matter	Continuous	National Park Service: Thomas Colson. 2011. Organic Matter Content of Soil. Geospatial Dataset-2198011.
Solar Radiation	Continuous	National Park Service: Thomas Colson. 2014. 30-m Potential Solar Radiation (2014). Raster Dataset-2208716
Slope	Continuous	Derived from elevation layer in GIS
Aspect	Continuous	Derived from elevation layer in GIS
Hillshade	Continuous	Derived from elevation layer in GIS
Topographic position index (TPI)	Continuous	Derived from elevation layer in GIS

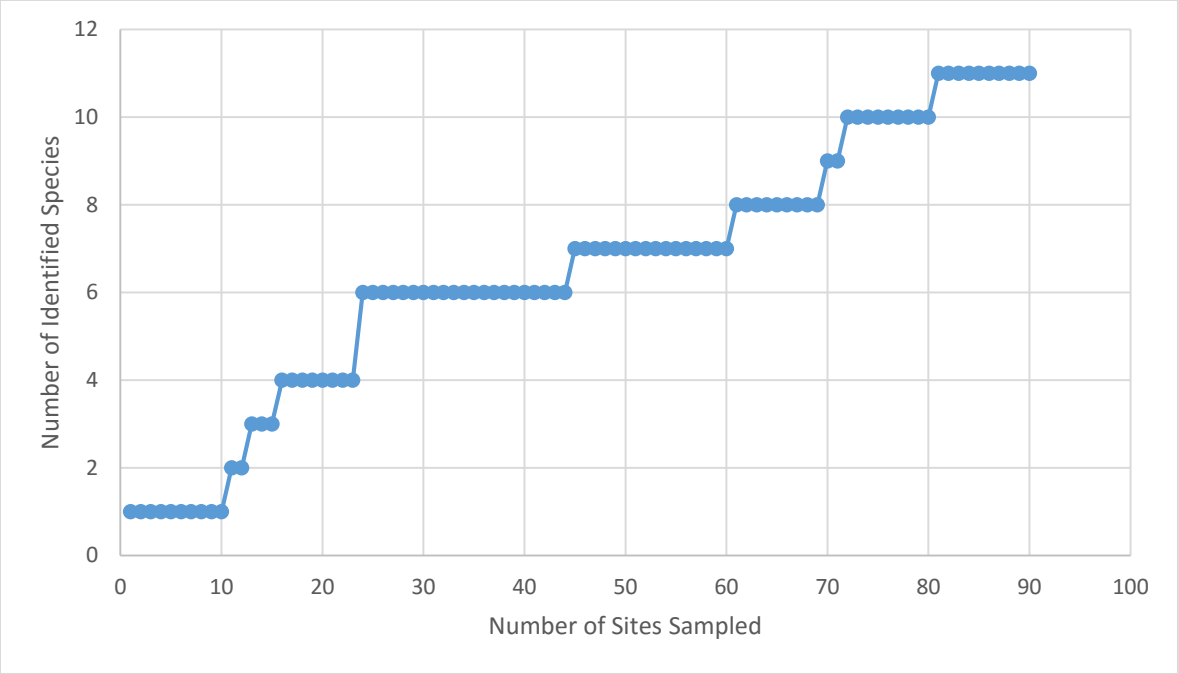


Figure 1: Species Accumulation Curve for Crustaceans

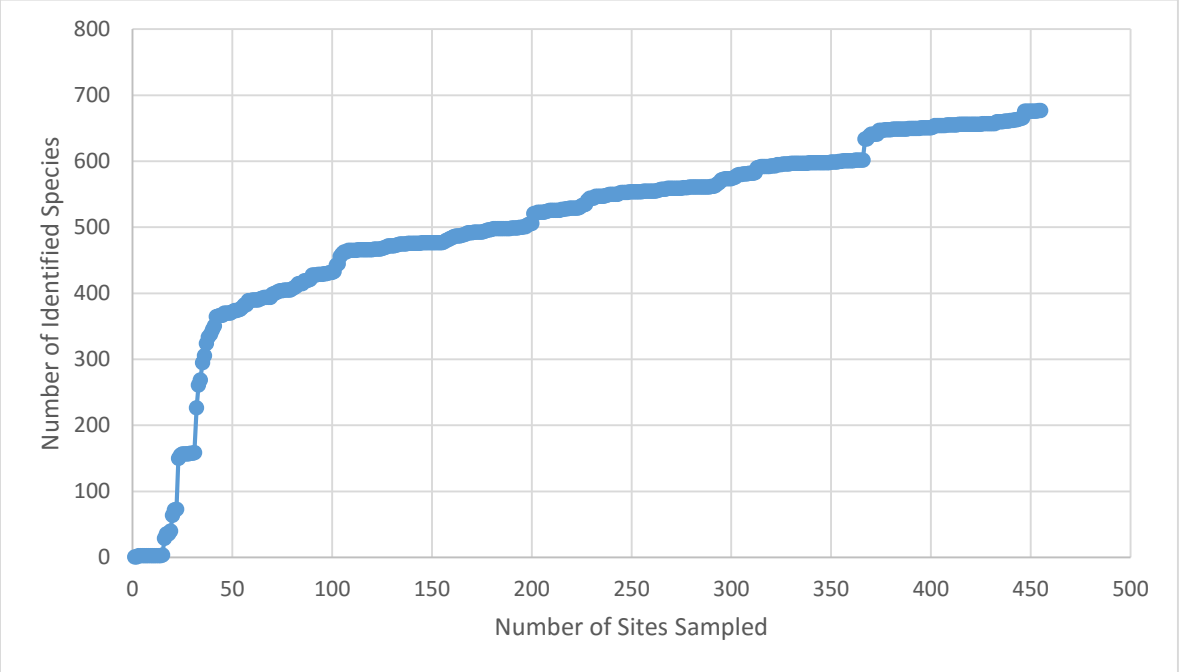


Figure 2: Species Accumulation Curve for Diptera

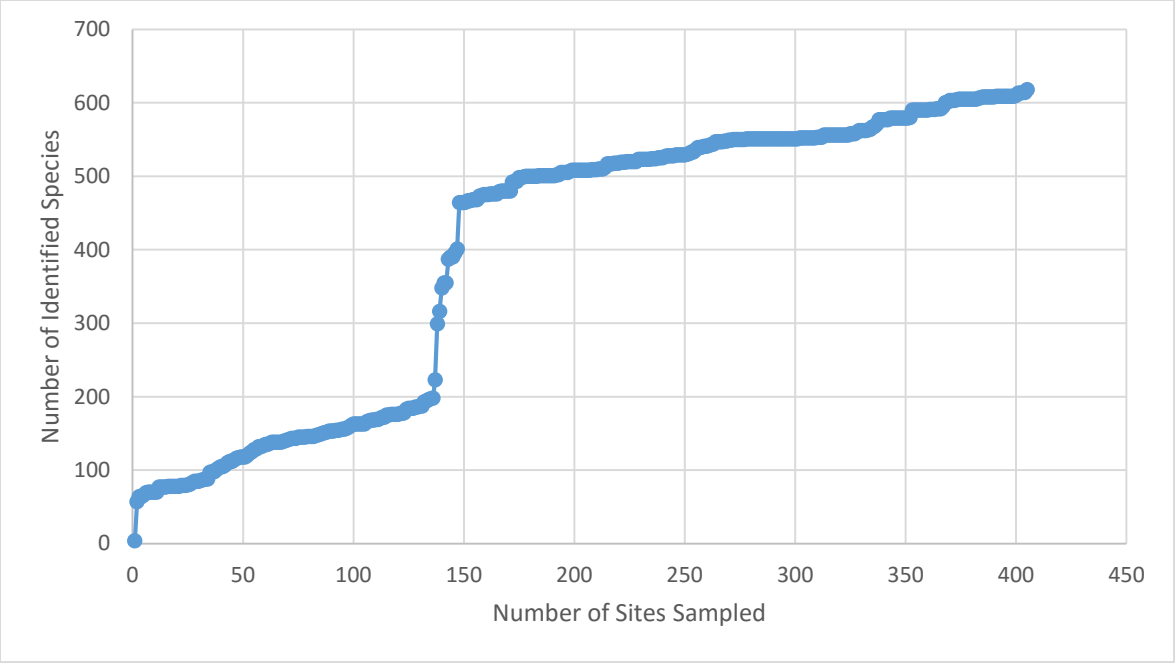


Figure 3: Species Accumulation Curve for Hemiptera

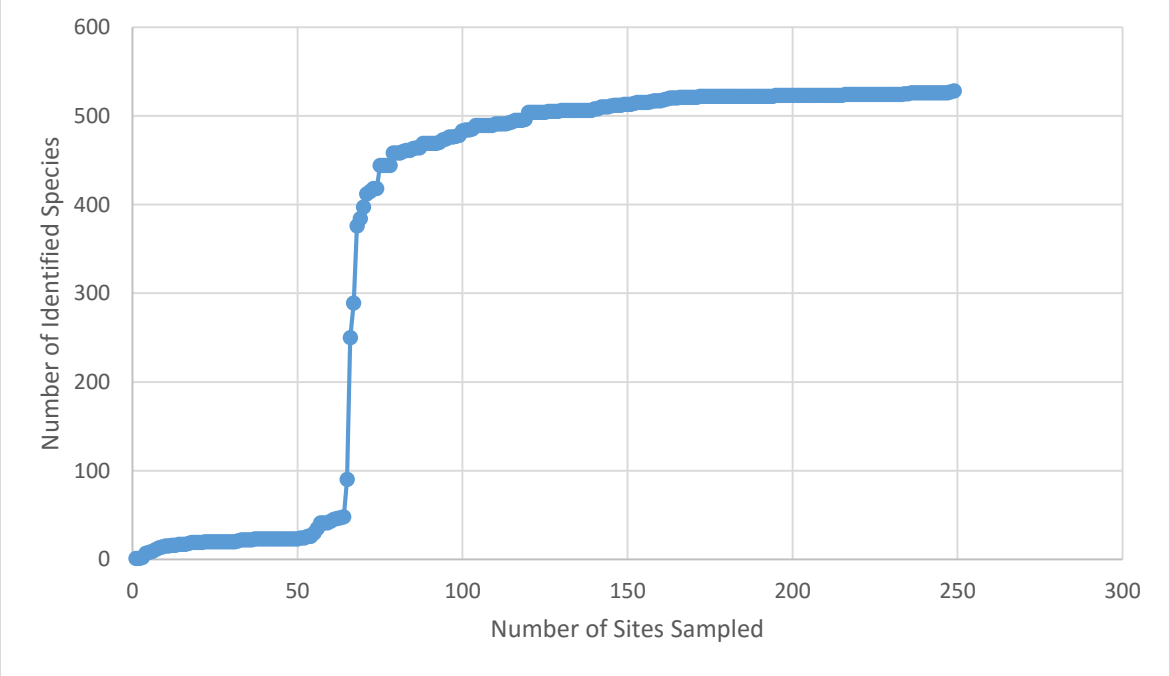


Figure 4: Species Accumulation Curve for Hymenoptera

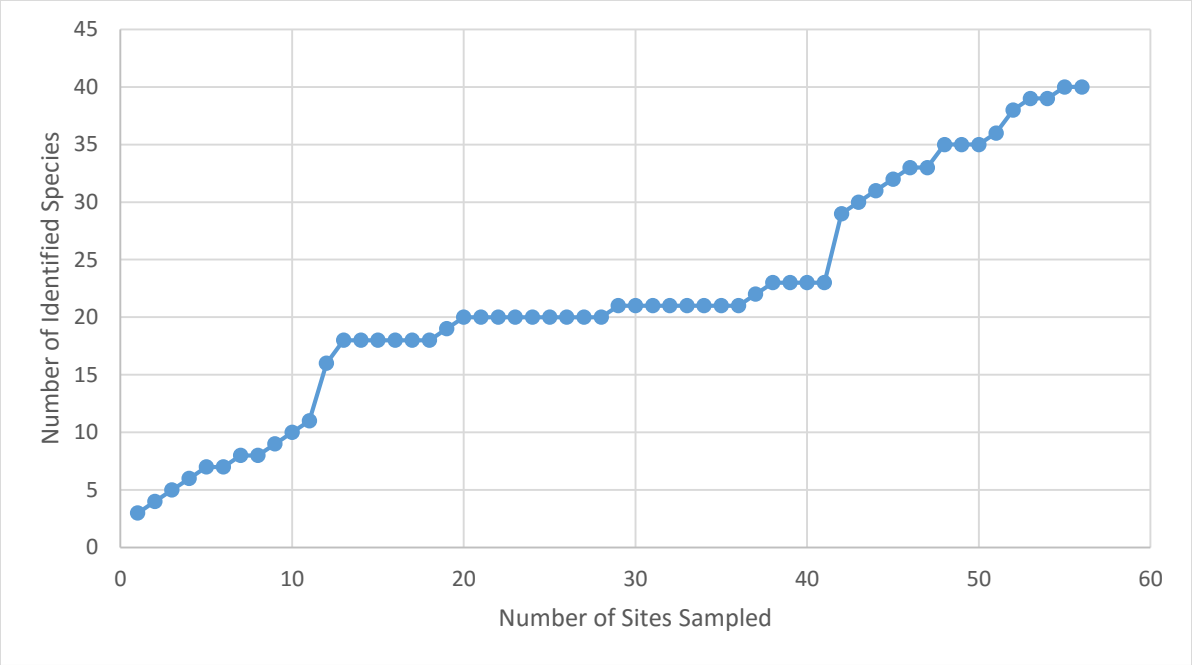


Figure 5: Species Accumulation Curve for Acari

Table 2: Species Richness Estimation Results

Order	Sites Sampled	Species Found	Chao2	ICE	Jack2	Mean Richness	Chao2 Gap	ICE Gap	Jack2 Gap	Mean Gap	Percent Found
Crustaceans	90	11	11.99	13.83	14.97	13.60	0.99	2.83	3.97	2.60	80.90%
Diptera	455	677	910.09	934.61	1040.22	961.64	233.09	257.61	363.22	284.64	70.40%
Hemiptera	405	618	908.1	927.82	1019.93	951.95	290.1	309.82	401.93	333.95	64.92%
Hymenoptera	249	528	681.82	700.21	782.07	721.37	153.82	172.21	254.07	193.37	73.19%
Acari	56	40	96.72	85.76	79.89	87.46	56.72	45.76	39.89	47.46	45.74%

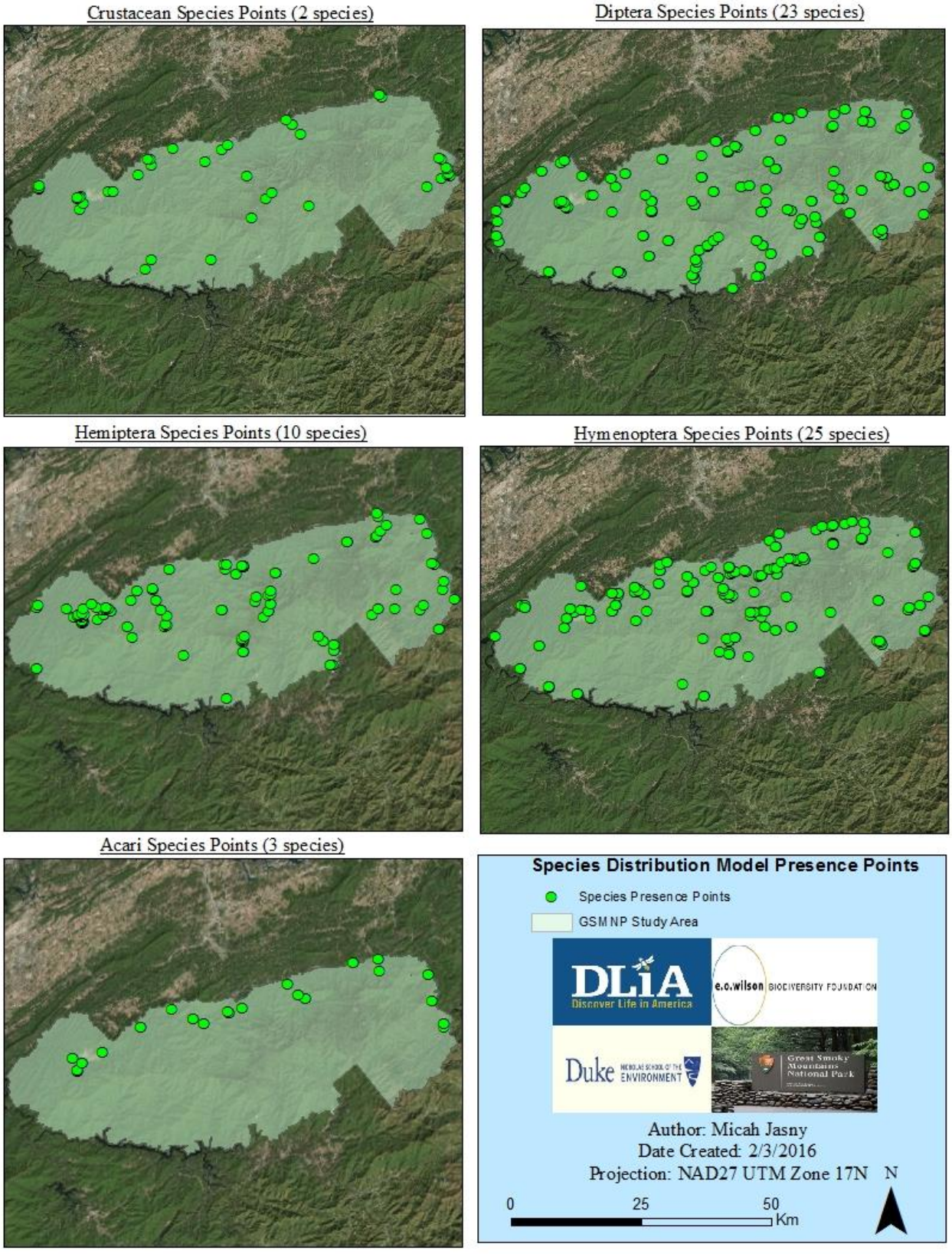


Figure 6: Presence Point Locations Used to Build Species Distribution Models

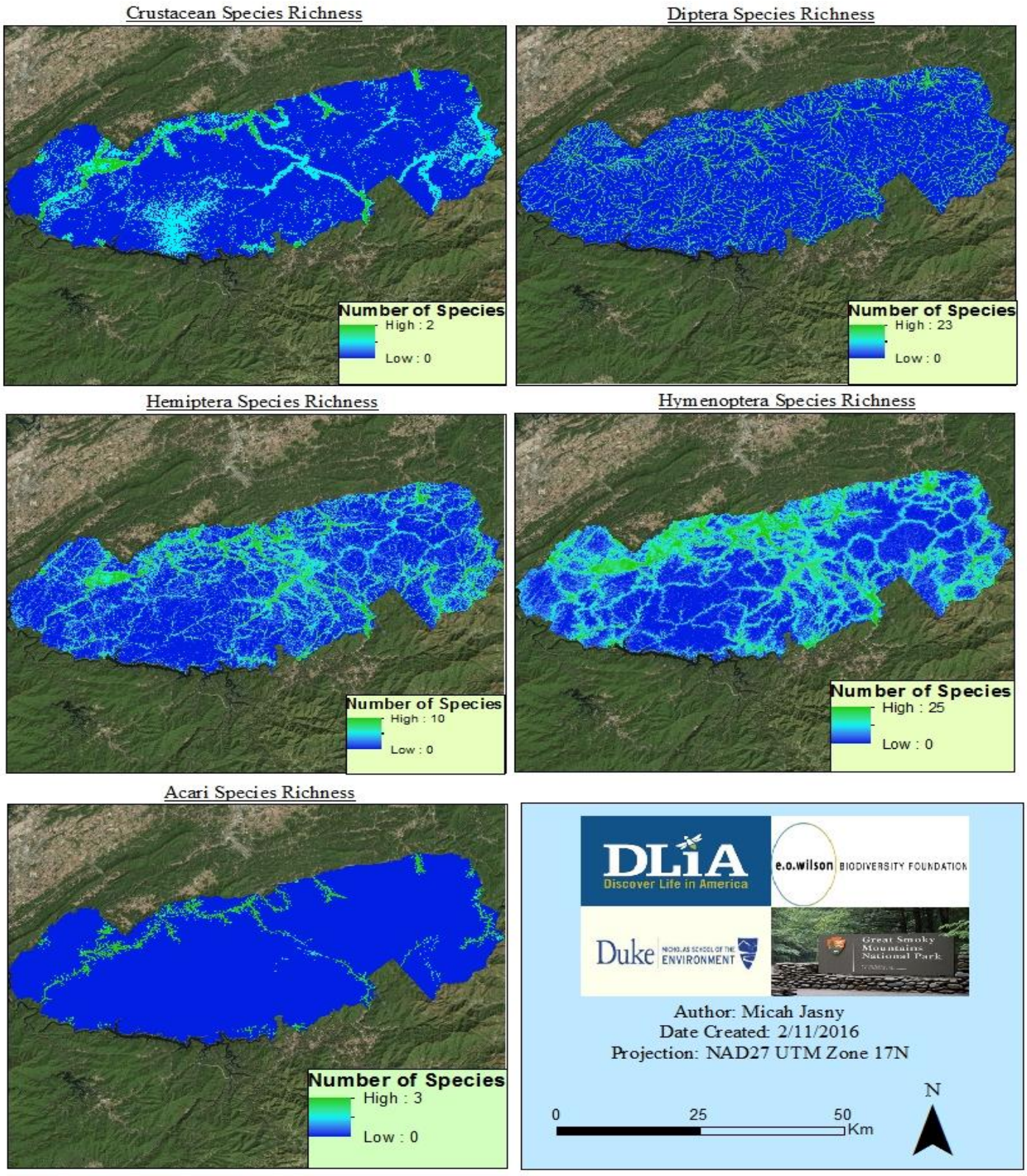


Figure 7: Initial Species Distribution Models for Five Focal Order

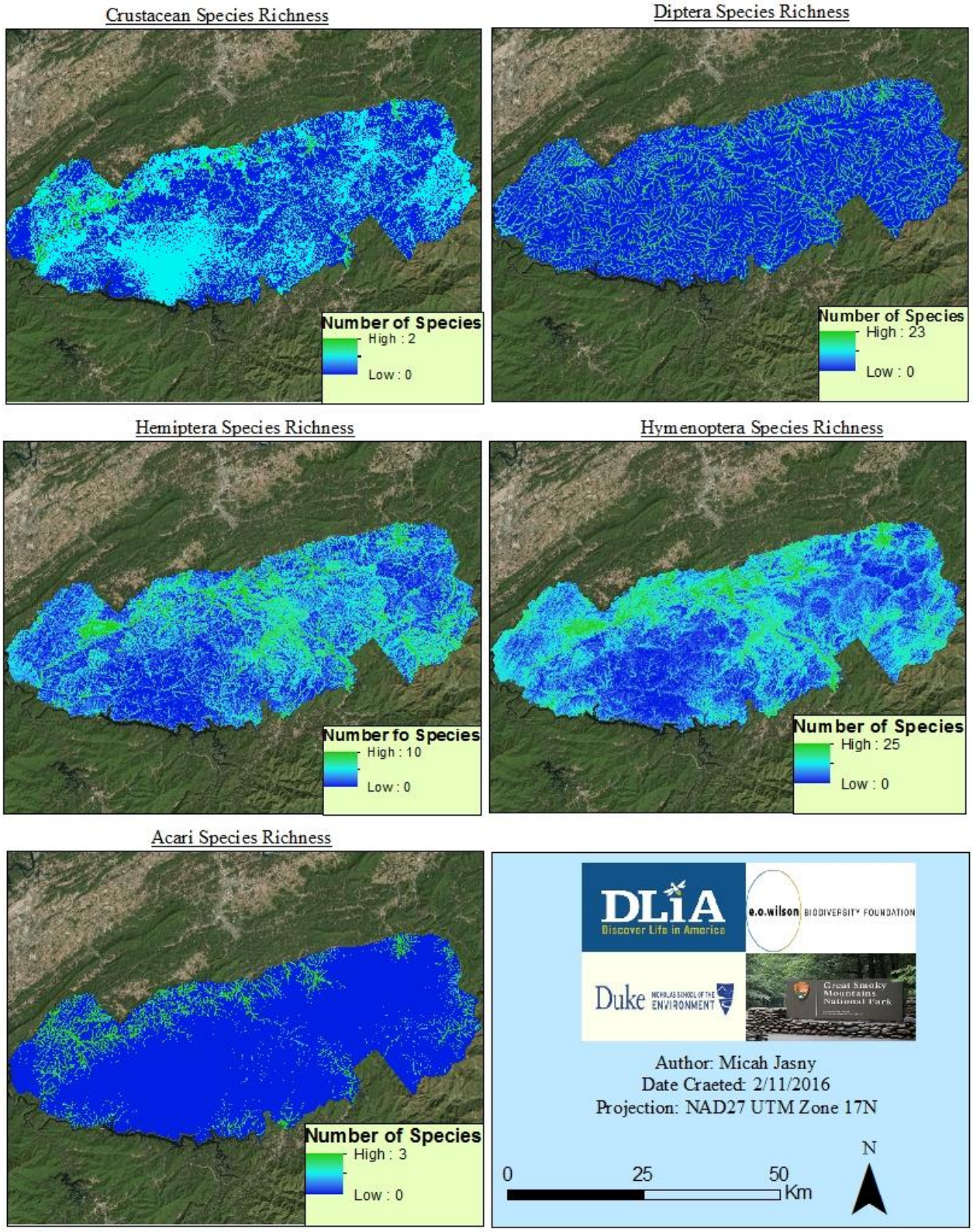


Figure 8: Species Distribution Models Using Bias Files for Five Focal Orders

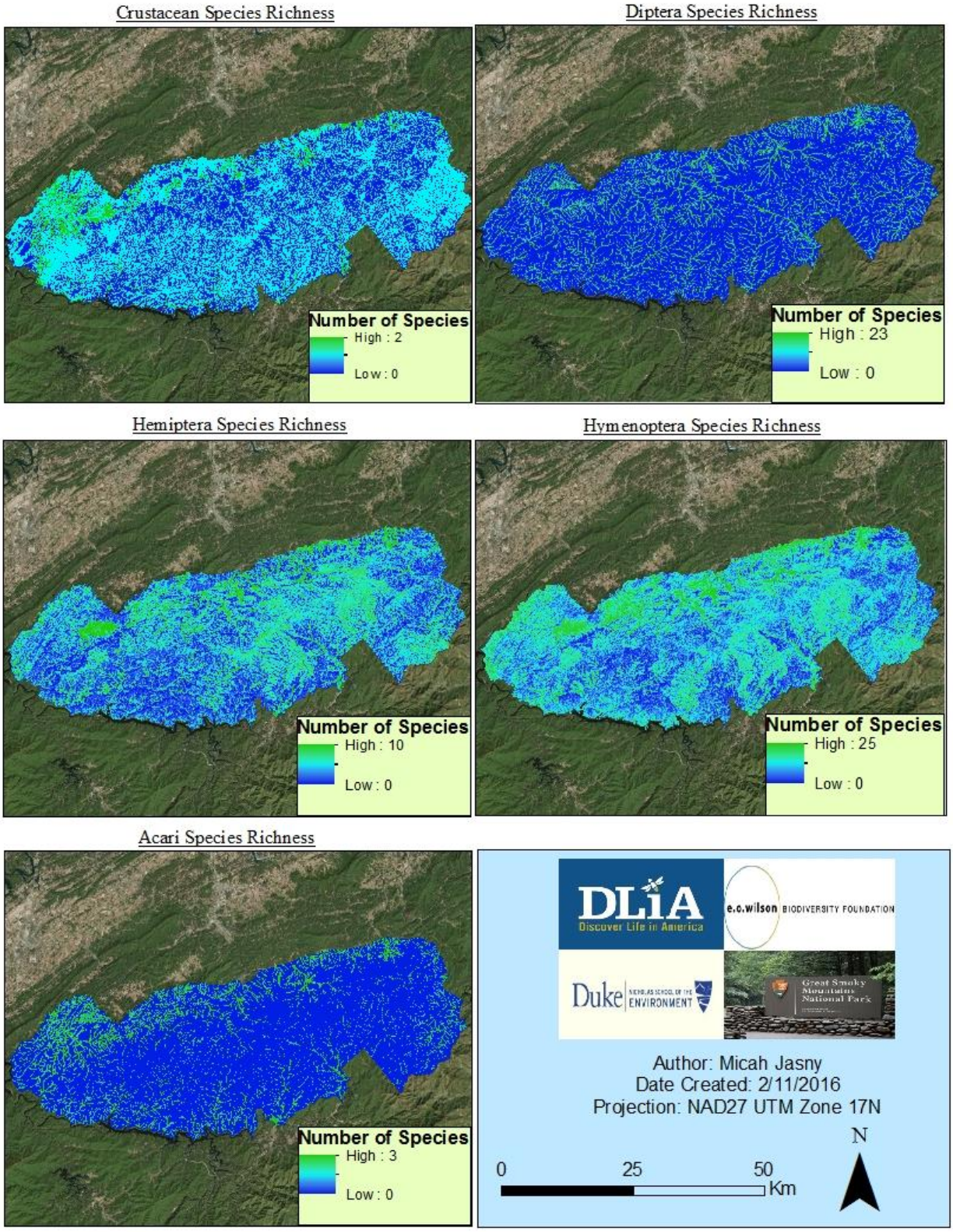


Figure 9: Species Distribution Models Using Bias Files without Anthropogenic Predictor variables

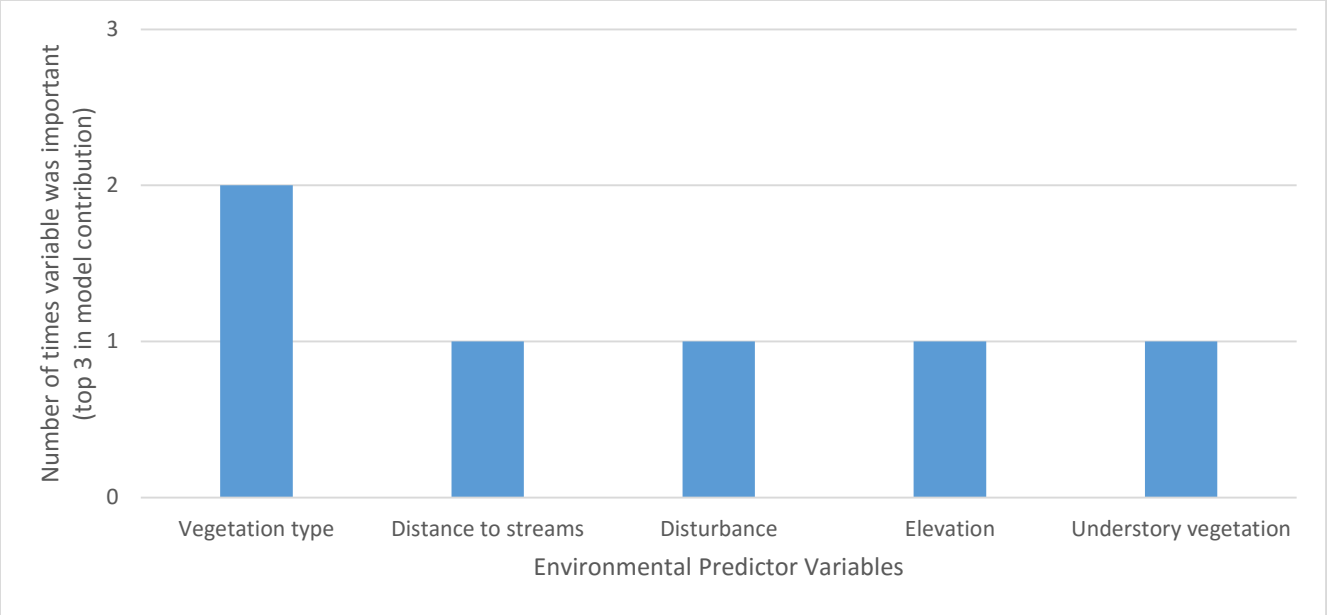


Figure 10: Environmental Predictor Variables Important for Crustacean Habitat Suitability

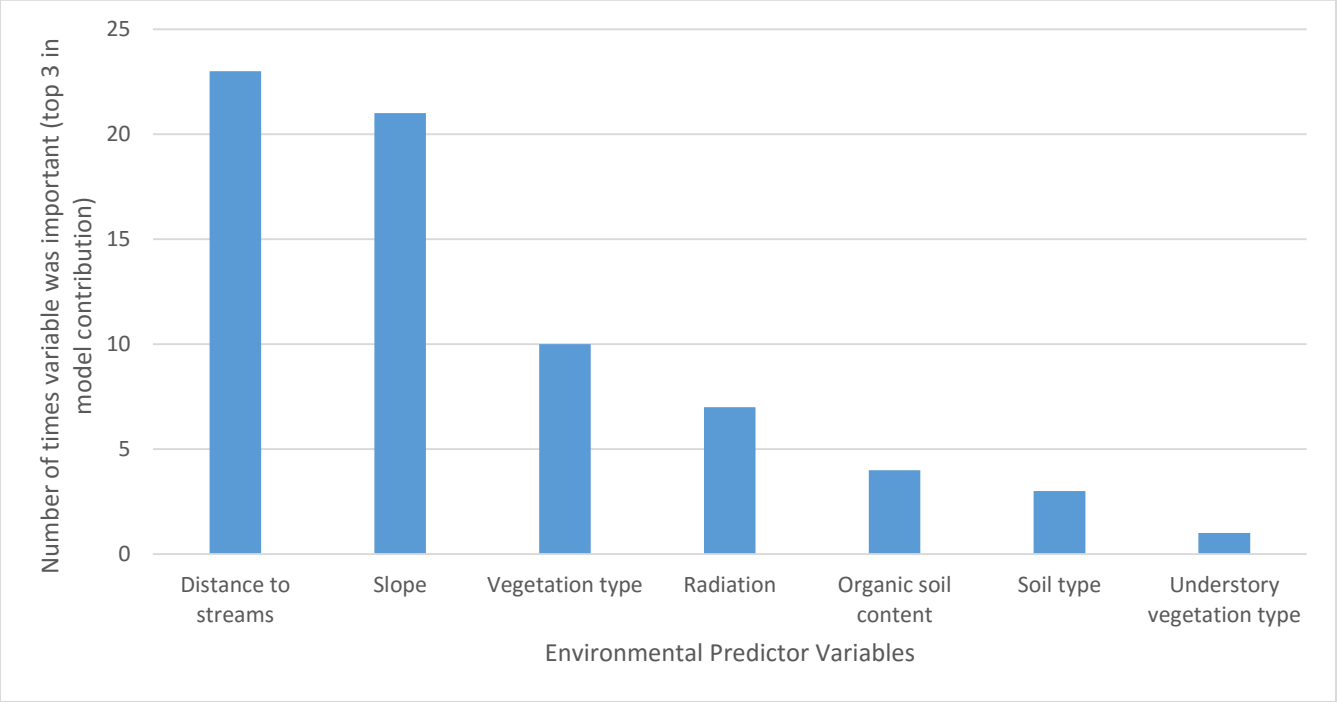


Figure 11: Environmental Predictor Variables Important for Diptera Habitat Suitability

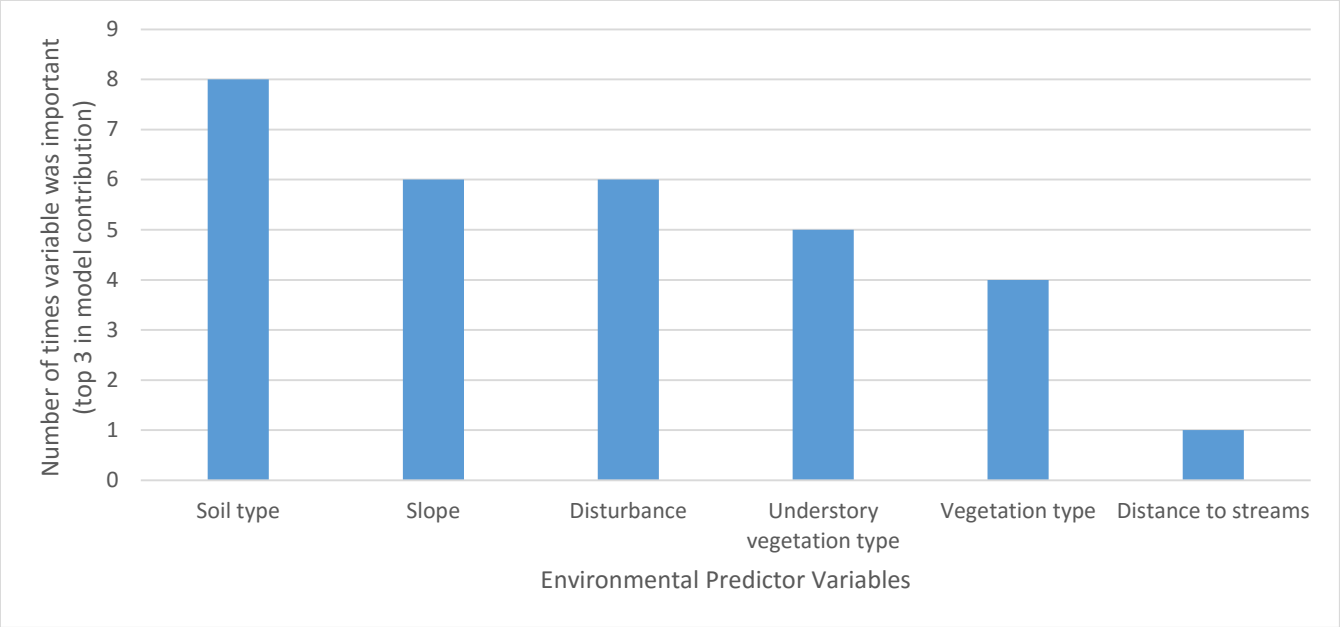


Figure 12: Environmental Predictor Variables Important for Hemiptera Habitat Suitability

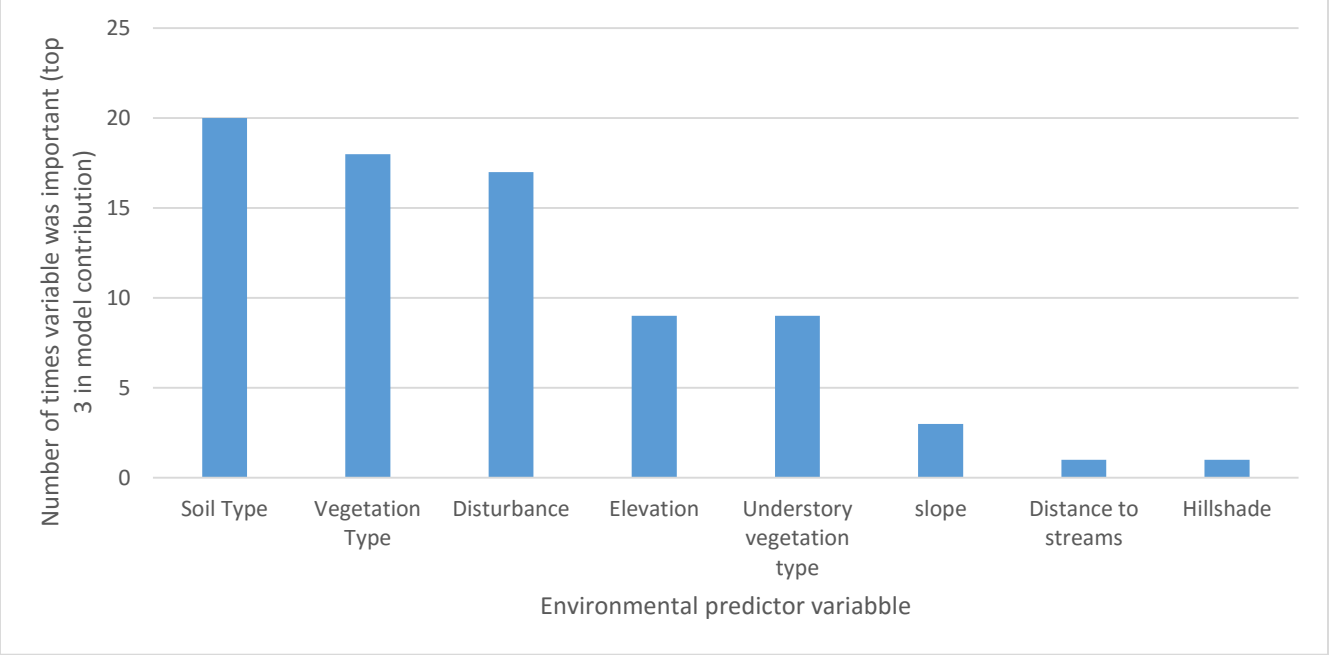


Figure 13: Environmental Predictor Variables Important for Hymenoptera Habitat Suitability

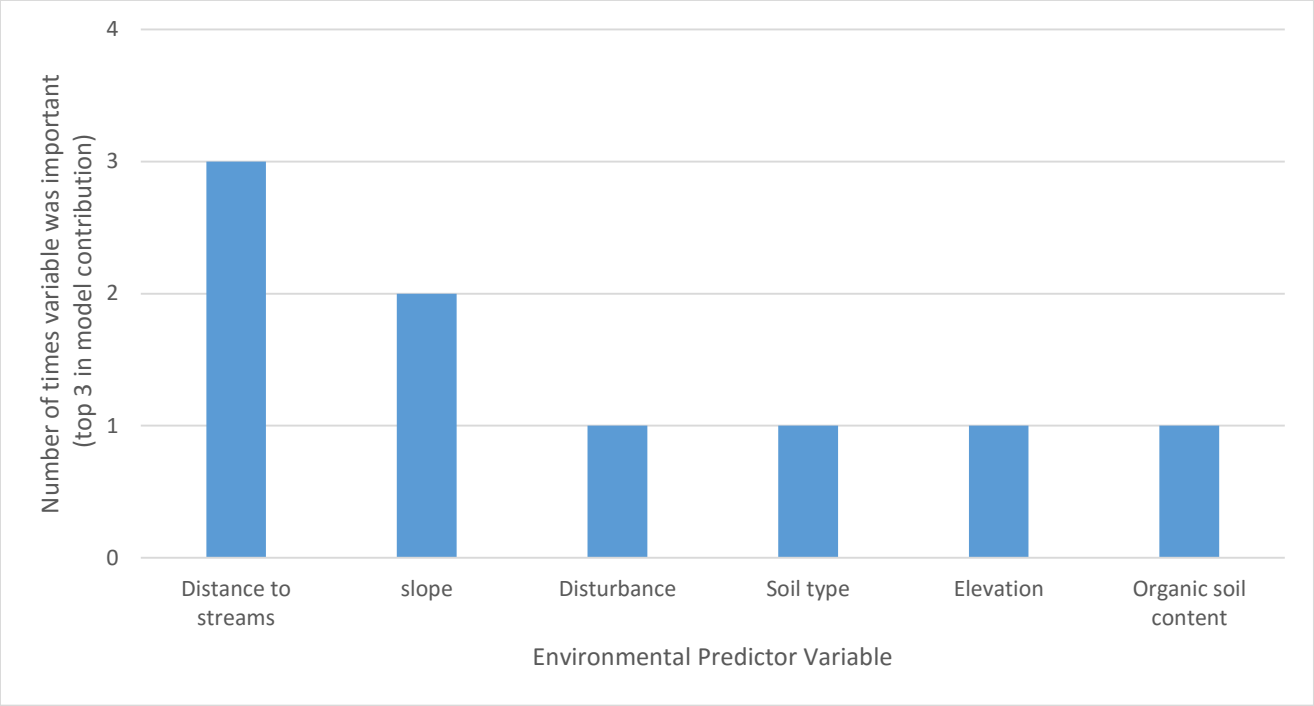


Figure 14: Environmental Predictor Variables Important for Acari Habitat Suitability

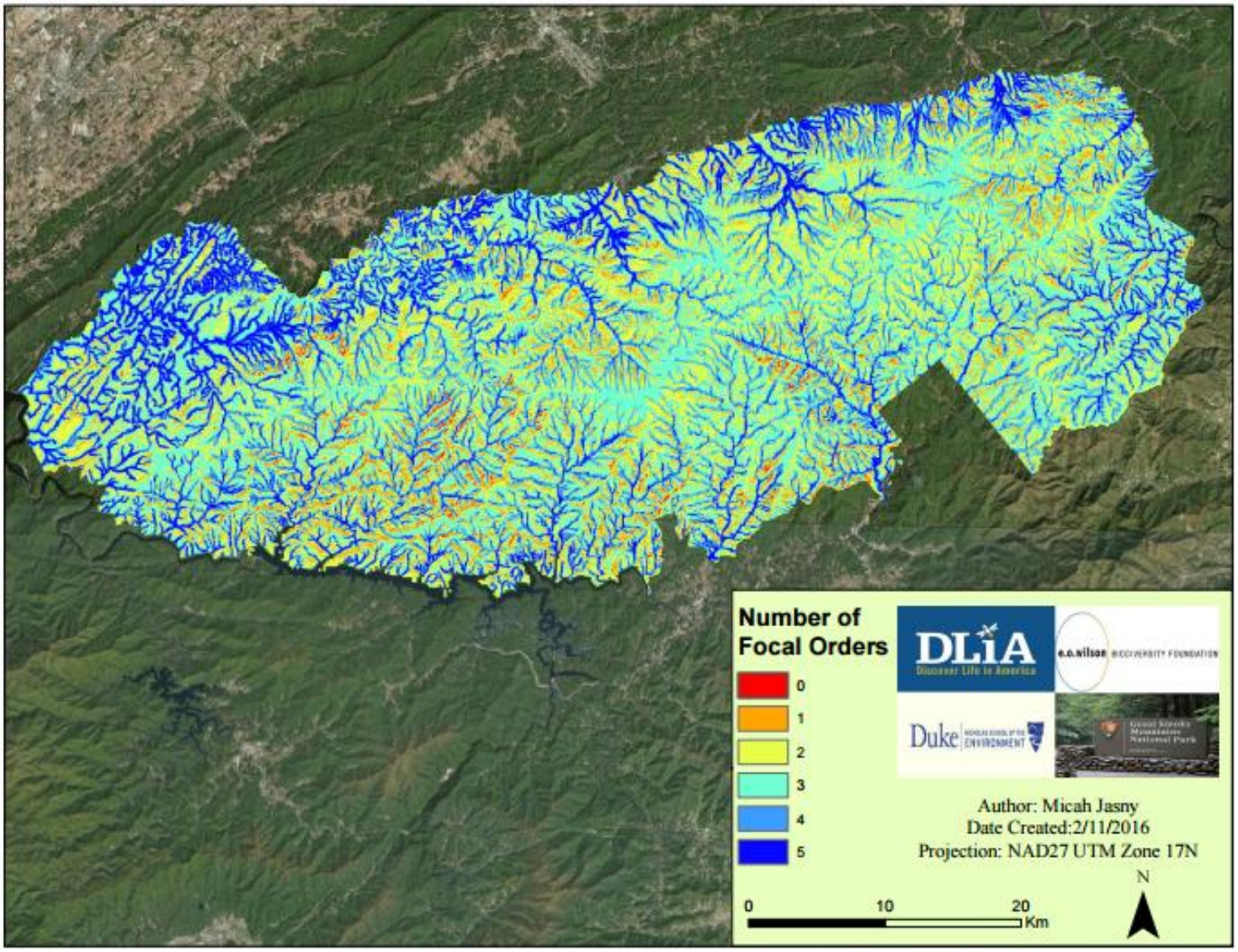
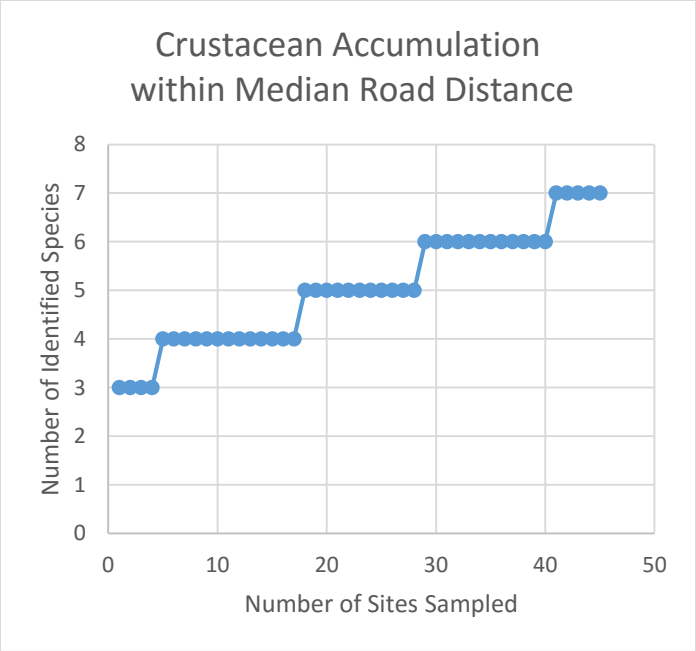
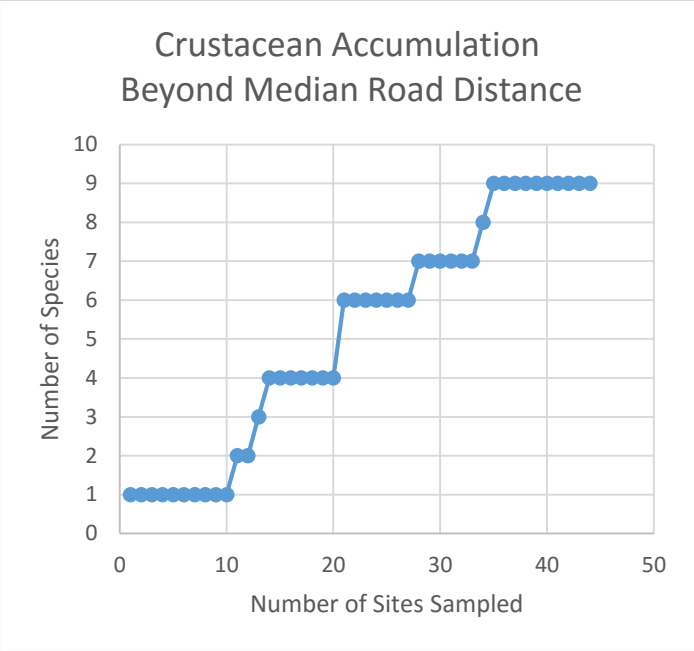
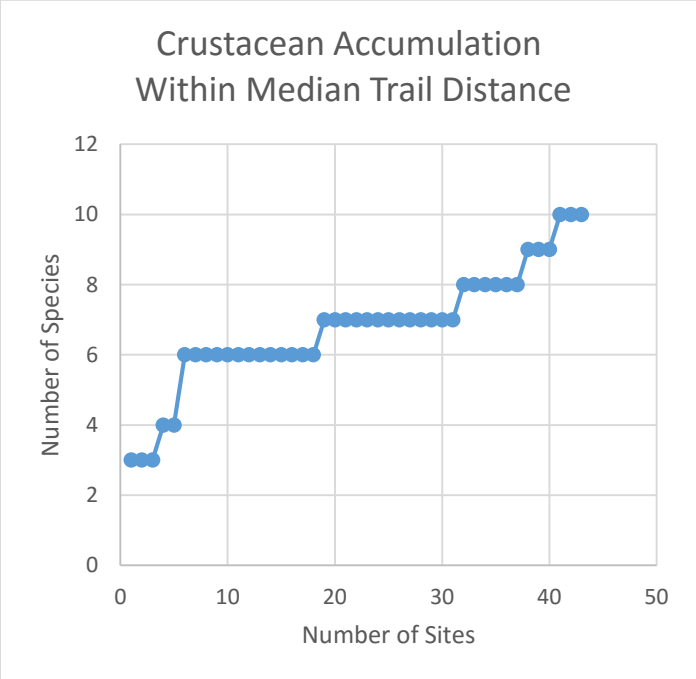


Figure 15: Overlap in Distributions of Five Focal Orders





Figures 16: Crustacean Species Accumulation Within and Beyond Median 205m from Roads



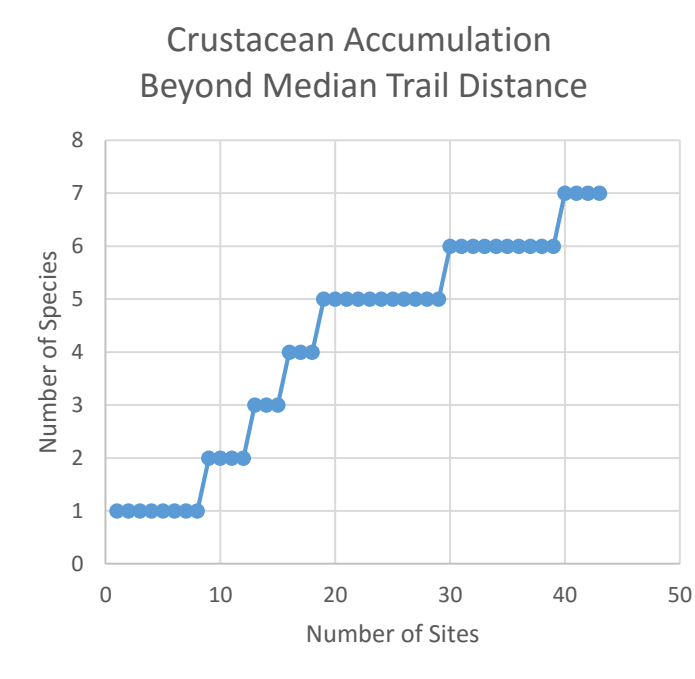


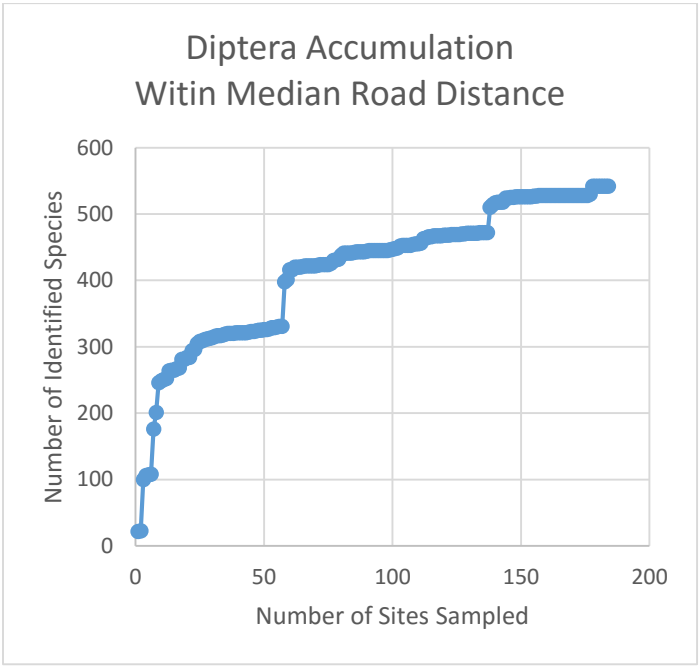
Figure 17: Crustacean Species Accumulation Within and Beyond 513m from Trails

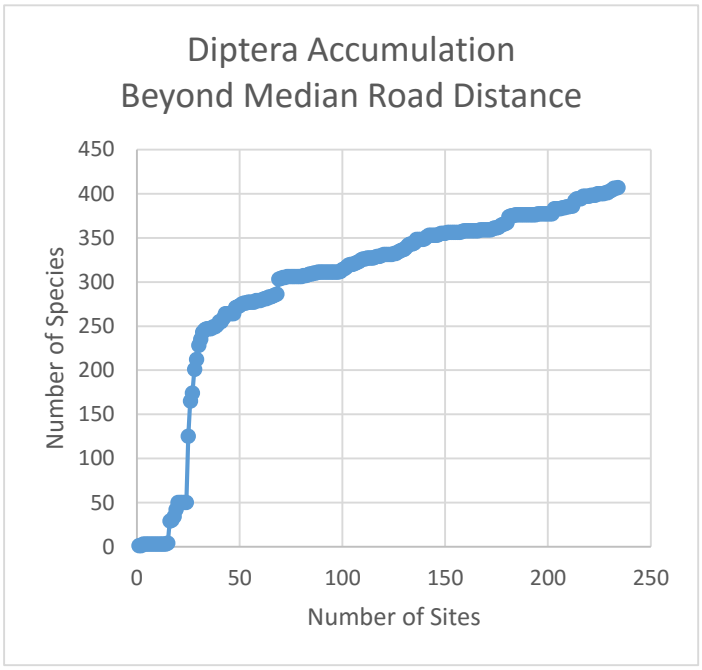
Table 3: Crustacean Road Distance Species Composition Comparison

Median Distance: 205m	Total number of species	Number of Unique species
Within median distance	7	2
Beyond median distance	10	5

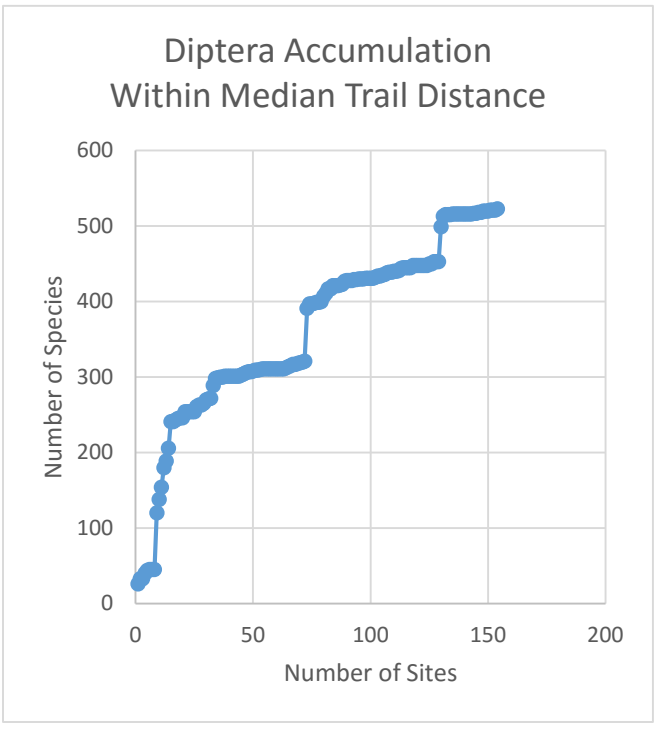
Table 4: Crustacean Trail Distance Species Composition Comparison

Median Distance: 513m	Total number of species	Number of Unique species
Within median distance	11	4
Beyond median distance	8	1





Figures 18: Diptera Species Accumulation Within and Beyond Median 150m from Roads



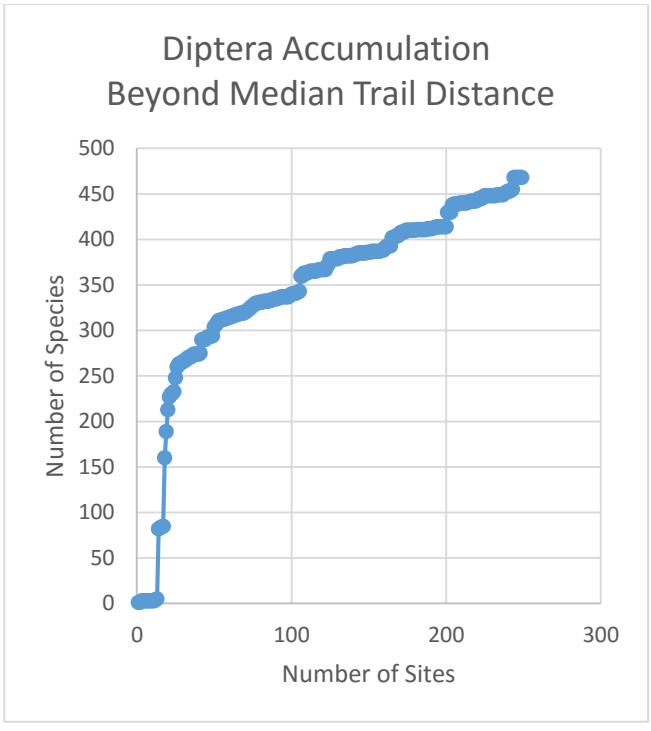


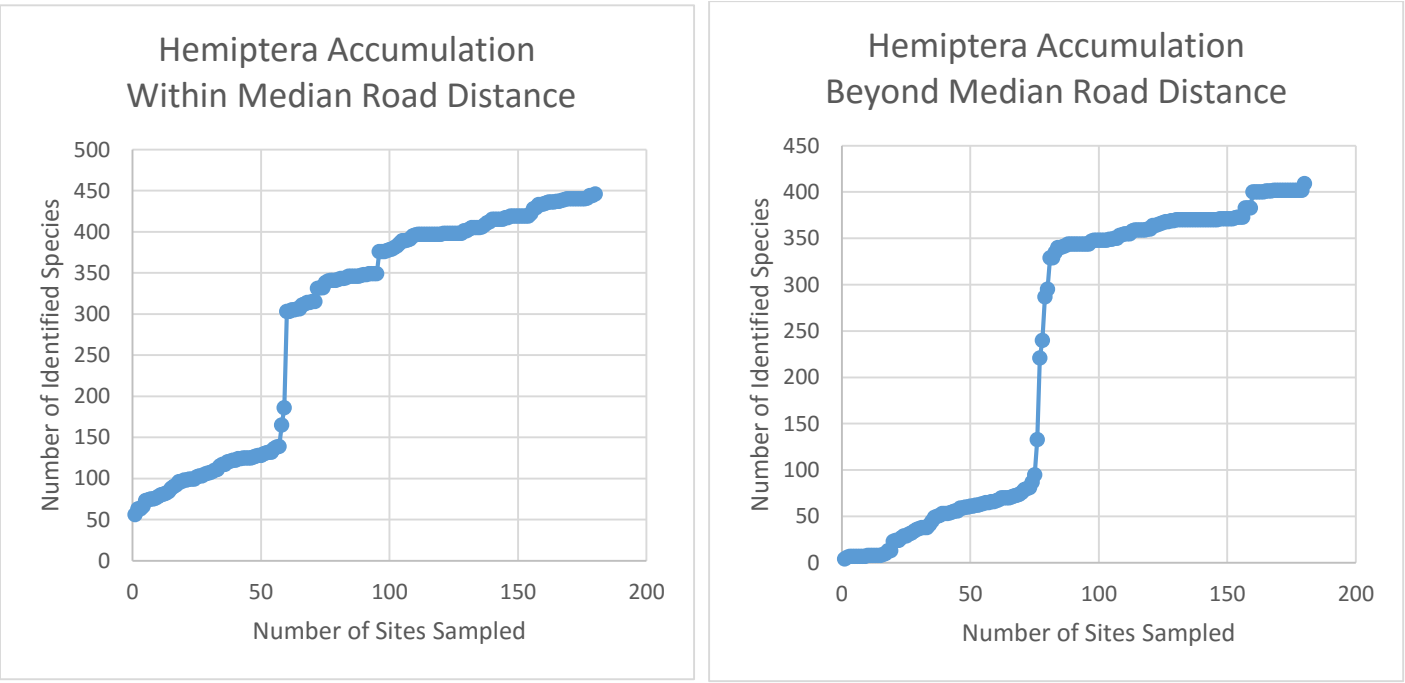
Figure 19: Diptera Species Accumulation Within and Beyond 108m from Trails

Table 5: Diptera Road Distance Species Composition Comparison

Median Distance: 150m	Total number of species	Number of Unique species
Within median distance	596	323
Beyond median distance	408	136

Table 6: Diptera Trail Distance Species Composition Comparison

Median Distance: 108m	Total number of species	Number of Unique species
Within median distance	572	253
Beyond median distance	470	151



Figures 20: Hemiptera Species Accumulation Within and Beyond Median 202m from Roads

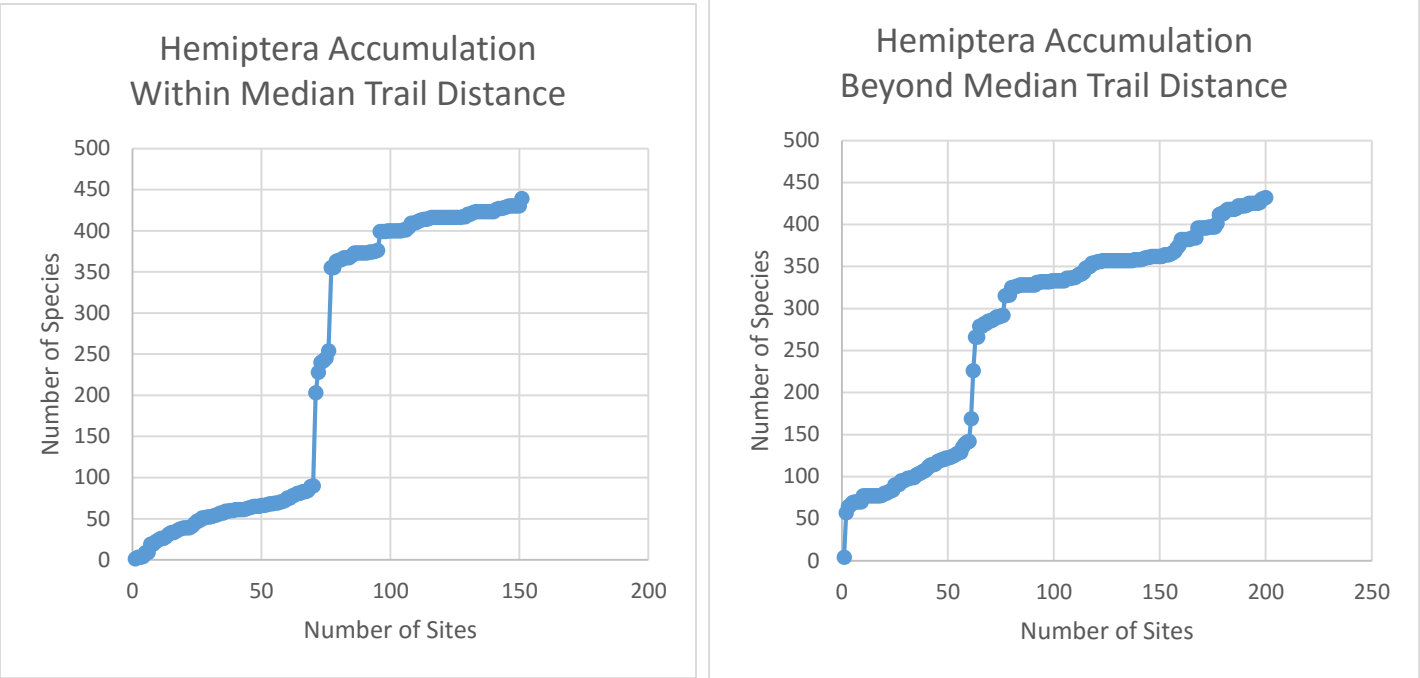


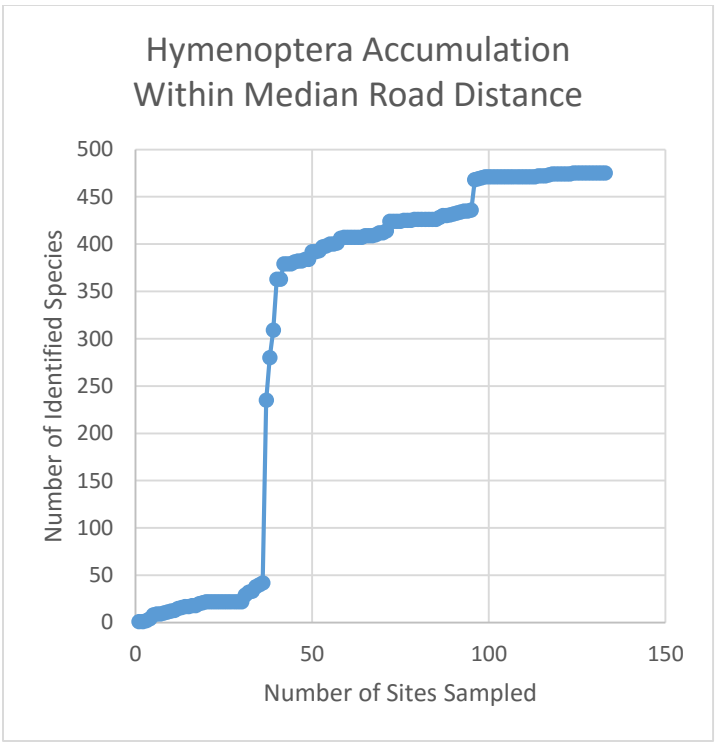
Figure 21: Hemiptera Species Accumulation Within and Beyond 108m from trails

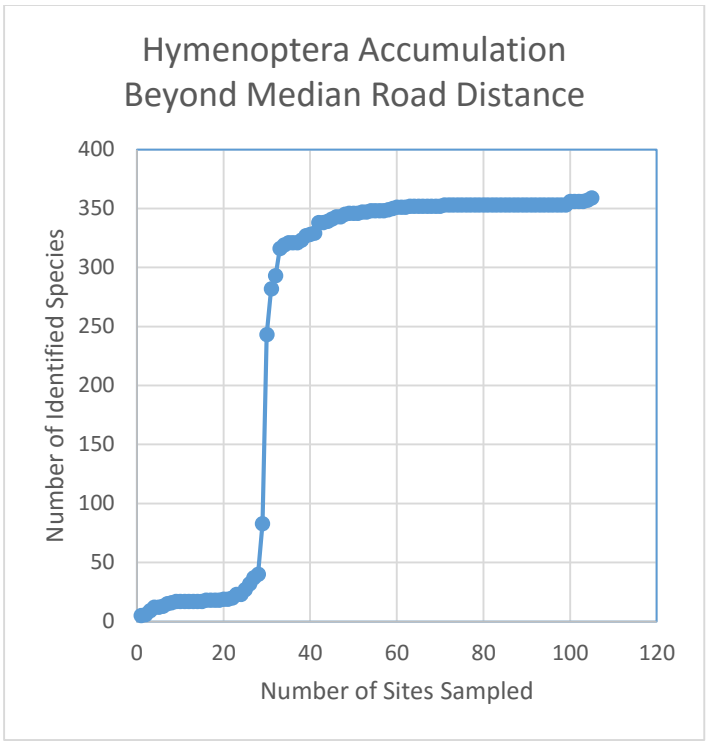
Table 7: Hemiptera Road Distance Species Composition Comparison

Median Distance: 202m	Total number of species	Number of Unique species
Within median distance	469	219
Beyond median distance	416	166

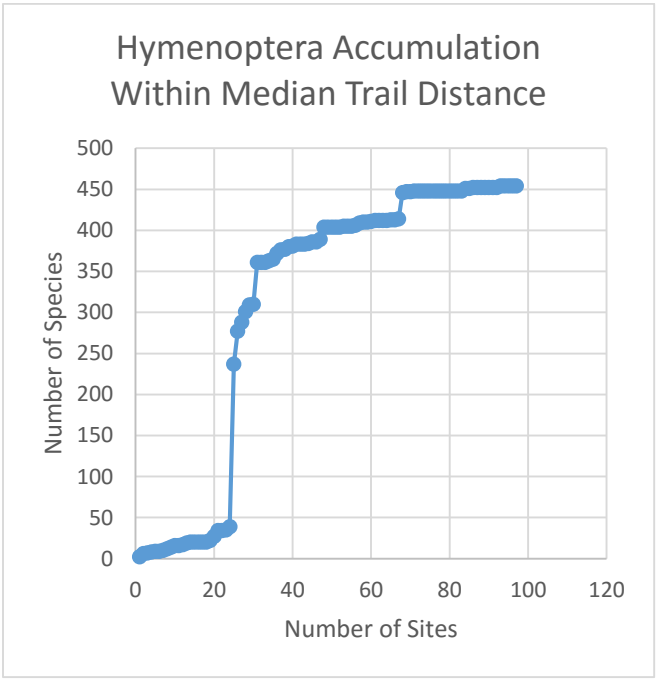
Table 8: Hemiptera Trail Distance Species Composition Comparison

Median Distance: 108m	Total number of species	Number of Unique species
Within median distance	469	201
Beyond median distance	433	165





Figures 22: Hymenoptera Species Accumulation Within and Beyond 390m from Roads



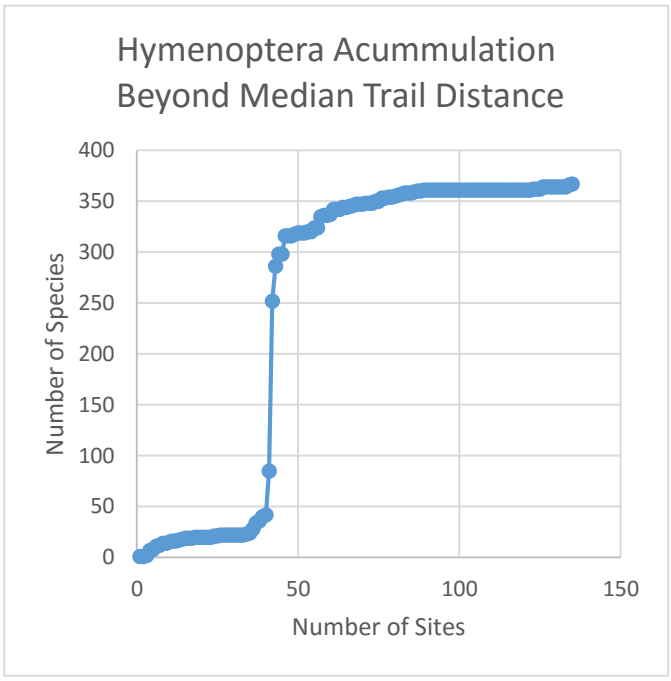


Figure 23: Hymenoptera Species Accumulation Within and Beyond 108m from Trails

Table 9: Hymenoptera Road Distance Species Composition Comparison

Median Distance: 390m	Total number of species	Number of Unique species
Within median distance	475	199
Beyond median distance	360	84

Table 10: Hymenoptera Trail Distance Species Composition Comparison

Median Distance: 108m	Total number of species	Number of Unique species
Within median distance	456	192
Beyond median distance	368	104

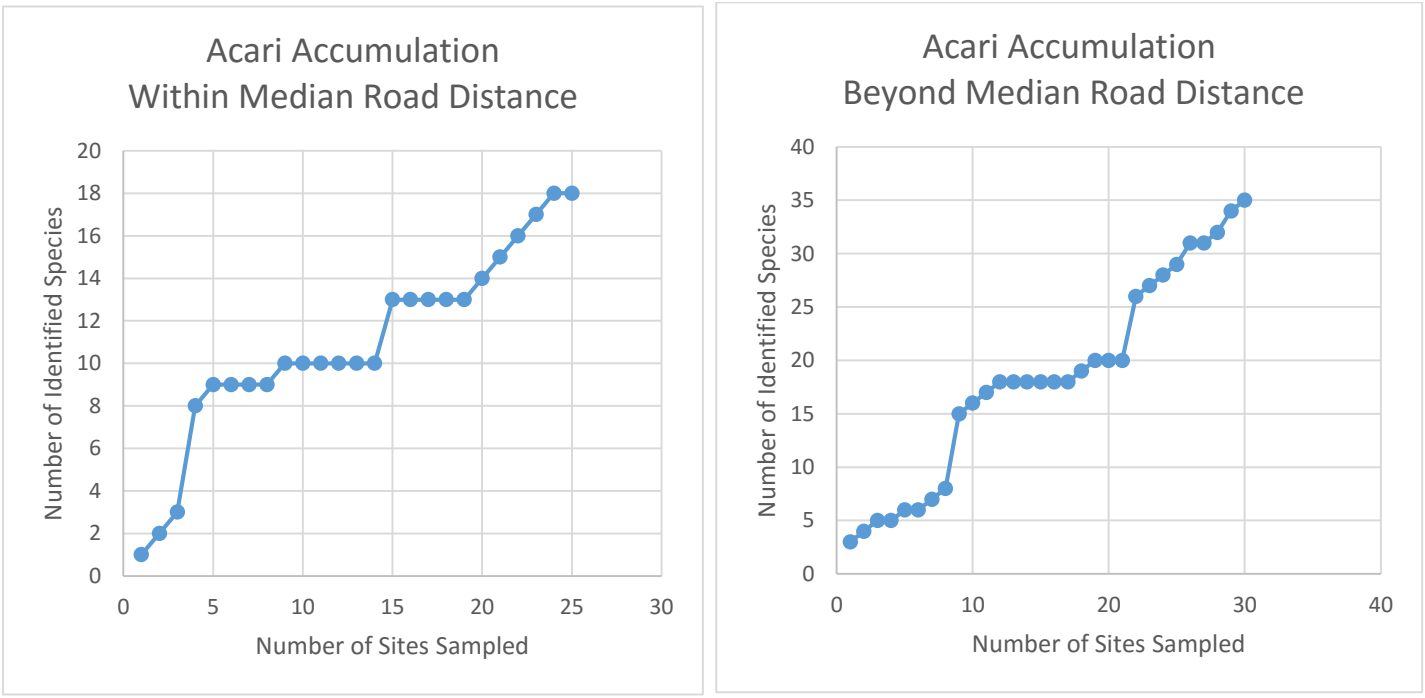


Figure 24: Acari Species Accumulation Within and Beyond 43m from Roads

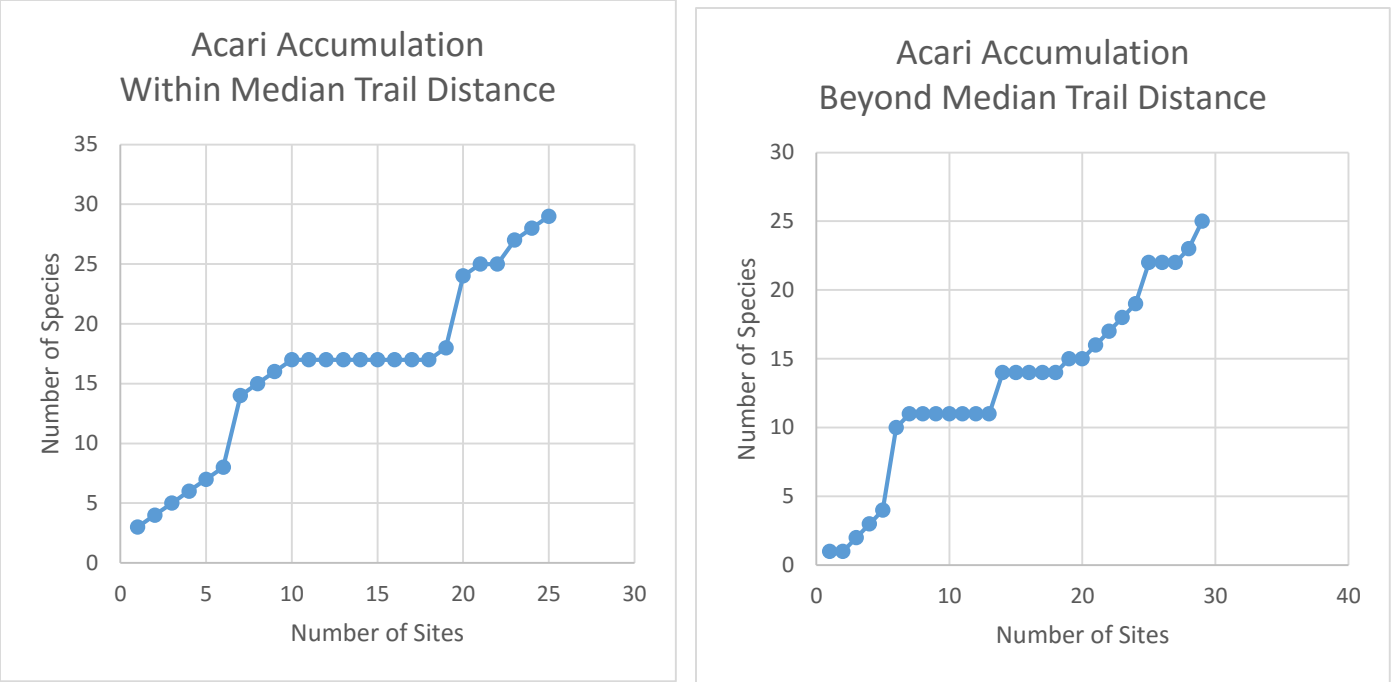


Figure 25: Acari Species Accumulation Within and Beyond Median 234m from Trails

Table 11: Acari Road Distance Species Composition Comparison

Median Distance: 43m	Total number of species	Number of Unique species
Within median distance	19	5
Beyond median distance	36	22

Table 12: Acari Trail Distance Species Composition Comparison

Median Distance: 234 m	Total number of species	Number of Unique species
Within median distance	31	15
Beyond median distance	26	10

VII. References:

Butchart, Stuart HM, et al. "Global biodiversity: indicators of recent declines." Science 328.5982 (2010): 1164-1168.

Chao, Anne. "Nonparametric estimation of the number of classes in a population." Scandinavian Journal of statistics (1984): 265-270.

Chapin III, F. Stuart, et al. "Consequences of changing biodiversity." Nature 405.6783 (2000): 234-242.

Colwell, Robert. "EstimateS: Biodiversity Estimation." Viceroy. University of Connecticut, 2013. Web.

Discover Life in America. (2014). Smokies Species Tally. Retrieved from http://www.dlia.org/smokies-species-tally#tallynote

Ditmanson, Dale. "Biodiversity at Great Smoky Mountains National Park." Office of Congressional and Legislative Affairs. U.S. Department of the Interior, 21 July 2008. Web.

Donoghue, J. (2013, February 19). A Maxent Script Tool for ArcGIS. Retrieved from https://geoapplications.wordpress.com/2013/02/19/a-maxent-script-tool-for-arcgis/

Elith, J., Phillips, S. J., Hastie, T., Dudík, M., Chee, Y. E., & Yates, C. J. (2011). A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions*, *17*(1), 43-57.

ESRI 2015. ArcGIS Desktop: Release 10.3.1. Redlands, CA: Environmental Systems Research Institute.

Gamfeldt, Lars, Helmut Hillebrand, and Per R. Jonsson. "Multiple functions increase the importance of biodiversity for overall ecosystem functioning." Ecology 89.5 (2008): 1223-1231.

"Gap Analysis Process." USGS. US Department of the Interior, 30 July 2015. Web. http://gapanalysis.usgs.gov/gap-analysis/process/

Gotelli, Nicholas J., and Robert K. Colwell. "Estimating species richness." Biological diversity: frontiers in measurement and assessment 12 (2011): 39-54.

Great Smoky Mountains National Park Biodiversity Database. 2015.

Hannah, L., et al. "Conservation of biodiversity in a changing climate." Conservation Biology 16.1 (2002): 264-268.

"Hemiptera - Bugs, Aphids, Cicadas." Insects and Their Allies. CSIRO, n.d. Web. http://www.ento.csiro.au/education/insects/hemiptera.html

Hooper, David U., et al. "Effects of biodiversity on ecosystem functioning: a consensus of current knowledge." Ecological monographs 75.1 (2005): 3-35.

Janzen, D. 2002. Biodiversity is us. ATBI Quarterly 3(3):3.

Jennings, M. D. (2000). Gap analysis: concepts, methods, and recent results*. Landscape ecology, 15(1), 5-20.

Kramer-Schadt, S., Niedballa, J., Pilgrim, J. D., Schröder, B., Lindenborn, J., Reinfelder, V., ... & Wilting, A. (2013). The importance of correcting for sampling bias in MaxEnt species distribution models. Diversity and Distributions, 19(11), 1366-1379.

Lees, Alexander C., and Stuart L. Pimm. "Species, extinct before we know them?." Current Biology 25.5 (2015): R177-R180.

Llorente, Jorge. "The use of species accumulation functions for the prediction of species richness." Conservation biology 7.3 (1993): 480-488.

Meyer, John. "Classification & Distribution." General Entomology. NC State University, 8 Apr. 2009. Web. https://www.cals.ncsu.edu/course/ent425/library/compendium/diptera.html

National Research Council (US) Committee on Noneconomic and Economic Value of Biodiversity. Perspectives on Biodiversity: Valuing Its Role in an Everchanging World. Washington (DC): National Academies Press (US); 1999. 2, What is Biodiversity?

Natural History Collections: Crustacea. The University of Edinburgh, Jan. 2007. Web. http://www.nhc.ed.ac.uk/index.php?page=24.25.312.330.363

Nichols, B. J., & Langdon, K. R. (2007). The Smokies all taxa biodiversity inventory: history and progress. Southeastern Naturalist, 6(sp2), 27-34.

Palmer, M.W. 1991. Estimating species richness: The second-order jackknife reconsidered. Ecology 72, 1512-1513.

Papeş, M., and Philippe Gaubert. "Modelling ecological niches from low numbers of occurrences: assessment of the conservation status of poorly known viverrids (Mammalia, Carnivora) across two continents." Diversity and distributions 13.6 (2007): 890-902.

Parker, C., & Bernard, E. (2006). The science approach to the Smokies ATBI. In The George Wright Forum (Vol. 23, No. 3, pp. 26-36).

Phillips, S. J., & Dudík, M. (2008). Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography*, *31*(2), 161-175.

Phillips, S. J., Dudík, M., Elith, J., Graham, C. H., Lehmann, A., Leathwick, J., & Ferrier, S. (2009). Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological Applications*, *19*(1), 181-197.

Phillips, S., Dudik, M. & Schapire, R., 2010, "Maxent Software, ver. 3.3.3e"

Proosdij, André SJ, et al. "Minimum required number of specimen records to develop accurate species distribution models." Ecography (2015).

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rodrigues, A. S., Akcakaya, H. R., Andelman, S. J., Bakarr, M. I., Boitani, L., Brooks, T. M., ... & Hoffmann, M. (2004). Global gap analysis: priority regions for expanding the global protected-area network. BioScience, 54(12), 1092-1100.

Sala, Osvaldo E., et al. "Global biodiversity scenarios for the year 2100." science 287.5459 (2000): 1770-1774.

Scheffers, Brett R., et al. "What we know and don't know about Earth's missing biodiversity." Trends in ecology & evolution 27.9 (2012): 501-510.

Sharkey, M. J. (2001). The all taxa biological inventory of the Great Smoky Mountains National Park. *Florida Entomologist*, 556-564.

Scott, J. Michael, et al. "Gap analysis: a geographic approach to protection of biological diversity." Wildlife monographs (1993): 3-41.

Syfert, M. M., Smith, M. J., & Coomes, D. A. (2013). The effects of sampling bias and model complexity on the predictive performance of MaxEnt species distribution models. PloS one, 8(2), e55158.

The World Conservation Union. 2014. IUCN Red List of Threatened Species 2014.3. Summary Statistics for Globally Threatened Species. Table 1: Numbers of threatened species by major groups of organisms (1996–2014).

Ugland, Karl I., John S. Gray, and Kari E. Ellingsen. "The species–accumulation curve and estimation of species richness." Journal of Animal Ecology 72.5 (2003): 888-897.

United States. National Park Service (2016). "Natural Features & Ecosystems." National Parks Service. U.S. Department of the Interior, n.d. Web.

Waggoner, Ben. "Acari." Introduction to the Acari. UC Berkeley, 1997. Web. http://www.ucmp.berkeley.edu/arthropoda/arachnida/acari.html

Waggoner, Ben, and Brian Speer. "Hymenoptera." Introduction to the Hymenoptera. UC Berkeley, 1997. Web. http://www.ucmp.berkeley.edu/arthropoda/uniramia/hymenoptera.html

White, et al. "The science plan for the all taxa biodiversity inventory in Great Smoky Mountains National Park, North Carolina and Tennessee." Report to board of directors of Discover Life In America 15 (2000).

Wilson, E. O. (1999). The diversity of life. WW Norton & Company

Wisz, Mary Suzanne, et al. "Effects of sample size on the performance of species distribution models." Diversity and Distributions 14.5 (2008): 763-773.

VIII. Appendix:

a. R Code for Species Accumulation Curves:

```
##read data
mdat<-read.csv("Crustacean.csv",header=T,stringsAsFactors=F)

########################################
sites_u <-sort(unique(mdat$SiteCode))

specimenNum<-matrix(0,length(sites_u),2)
specimenNum[,1]<-sites_u
for(i in 1:length(sites_u)){
  specimenNum[i,2]<-length(which(mdat$SiteCode==sites_u[i]))
}

specimenNum_n<-as.numeric(specimenNum[,2])
specimenNum.cdf<-vector(length=length(specimenNum[,2]))
specimenNum.cdf[1]<-specimenNum[1,2]
for(i in 2:length(specimenNum[,2])){
  specimenNum.cdf[i]<-sum(specimenNum_n[1:i])
}

##find new species
foundSpecies<-unique(mdat$Species[which(mdat$SiteCode==sites_u[1])])
foundSpeciesN<-vector(length=length(sites_u))
foundSpeciesN[1]<-length(foundSpecies)

for(i in 2:length(sites_u)){
  tempSpecies<-unique(mdat$Species[which(mdat$SiteCode==sites_u[i])])
  newSpecies<-which(!tempSpecies%in%foundSpecies)
```

```
  foundSpeciesN[i]<-length(newSpecies)
  foundSpecies<-c(foundSpecies,tempSpecies[newSpecies])
}

foundSpeciesN.cdf<-vector(length=length(foundSpeciesN))
foundSpeciesN.cdf[1]<-foundSpeciesN[1]
for(i in 2:length(foundSpeciesN)){
  foundSpeciesN.cdf[i]<-sum(foundSpeciesN[1:i])
}
SpeciesIncidence <- write.csv(foundSpeciesN.cdf, "CrustaceanspeciesAccumulation.csv")
```

b. Python Script for Maxent Species Distribution Modeling Script:

```
#
#-------------------------------------------------------------
# Author:     Micah Jasny (micah.jasny@duke.edu)
# Created:    12/10/2015
# ArcGIS Version:   10.3.1
# Python Version:   2.7
# Maxent Version:   3.3.3k
#-------------------------------------------------------------
#

import arcpy, sys, os, csv, glob
from arcpy.sa import *
arcpy.env.overwriteOutput = 1

# get the location to the maxent jar file
maxent = arcpy.GetParameterAsText(0)

# Get the csv file to model
csvFile = arcpy.GetParameterAsText(1)

# Identifying location of environmental predictor variables
climatedataFolder = arcpy.GetParameterAsText(2)

# Setting output folder location
outputFolder = arcpy.GetParameterAsText(3)

# Setting model type to logistic
optOutputFormat = "logistic"

# setting output type to ascii
optOutputFileType = "asc"

# get Maxent options from user
optResponseCurves = "true"
optPredictionPictures = "true"
```

```
optJackknife = "true"
optSkipifExists = "false"
supressWarnings = "false"

# Naming output file containing species
speciesInput = os.path.basename(csvFile)
speciesInput = speciesInput[0:-4]

# create a location within output folder for outputs (makes it easier to clear the output folder for the user)
newOutputFolder = outputFolder + "\\" + speciesInput
if os.path.isdir(newOutputFolder):
   arcpy.AddMessage(" ")
else:
   os.mkdir(newOutputFolder)

# create the maxent command
myCommand = "java -mx2048m -jar \"" + maxent + "\" -e \"" + climatedataFolder + "\""
myCommand += " -t disturbance -t soiltype -t understoryveg -t vegtype"
myCommand += " -s \"" + csvFile + "\" -o \"" + newOutputFolder + "\""
myCommand += " outputformat=" + optOutputFormat.lower() + " outputfiletype=" +
optOutputFileType.lower()


# add options
if optResponseCurves == 'true':
   myCommand += " -P"
if optPredictionPictures == 'true':
   myCommand += " pictures=true"
else:
   myCommand += " pictures=false"
if optJackknife  == 'true':
   myCommand += " -J"
if optSkipifExists  == 'true':
   myCommand += " -S"
if supressWarnings  == 'true':
   myCommand += " warnings=false"
else:
   myCommand += " warnings=true"

# finish the command
myCommand += " -a"

# add a message
arcpy.AddMessage("Starting Maxent")
arcpy.AddMessage(myCommand)

# execute the command
```

```
result = os.system(myCommand)

if (result == 0):
   arcpy.AddMessage(" ")
   arcpy.AddMessage("Finished")
else:
   arcpy.AddMessage(" ")
   arcpy.AddError("Error Running Maxent")

#change directory to output location
os.chdir(os.path.abspath(newOutputFolder))
#Create dictionary linking species name with balance training omission logistic threshold

WrkDir = os.path.abspath(newOutputFolder)
#locate all Maxent ascii files
for file in glob.glob("*.asc"):
   #convert ascii layers to raster
   input = WrkDir + "\\""" + file
   file = file[:-4]
   OutName = file[0:9]
   rootfolder = os.getcwd()
   output = WrkDir + "\\""" + OutName + "cont"
   arcpy.ASCIIToRaster_conversion(input, output, "FLOAT")
Rastlist = []
for SDM in glob.glob("*cont"):
   Rastlist.append(SDM)
#combine thresholded raster into a community map
CellStats = arcpy.sa.CellStatistics(Rastlist, "MEAN")
CommRastName = WrkDir + "\\""" + "CommunityRast"
CellStats.save(CommRastName)
#define spatial reference of community raster to be the same as the data and environmental predictor
variables
sr = arcpy.SpatialReference("NAD 1927 UTM Zone 17N")
arcpy.DefineProjection_management(CommRastName, sr)
```

c. R code used to compare species found wihin and beyond median trail and road distances

```
dat<- read.csv("TrailDistComparison.csv", header=TRUE)
head(dat)
Near <- unique(dat$CrustNear)
Far <- unique(dat$CrustFar)
length(Near)
length(Far)
NearSpecies <- Near[which(!Near%in%Far)]
length(NearSpecies)
FarSpecies <- Far[which(!Far%in%Near)]
length(FarSpecies)
```