Moths and Meadows Final Report
Patrick Waters, Sean McGregor, William Berk

Introduction

More than 580 species of Lepidoptera (Moths and Butterflies) can be found at the HJ Andrews Long Term Ecological Research (LTER) site. The population consists of several generalist herbivores and many species with highly specialized niches. The low difficulty of light-trap sampling provides an ideal method for tracking moth population diversity as an indicator of overall ecosystem health and change-type. Here we present the preliminary findings from three novel approaches to moth population analysis. William Berk performed meadow boundary sampling in order to describe the dispersal capabilities of moths. Patrick Waters built a metapopulation model to look at inter-meadow population dynamics and designed a classification algorithm for moth population habitat parameters. Sean McGregor worked with historical sampling to dynamically create Google Earth overlays and performed machine-learning analysis on the data.

Methods

Meadow Boundary Analysis - Selection of meadows to study was aided through the use of shapefiles on ArcGIS. One of the shapefiles depicted the meadows of HJA, as well as their size. Any meadow with an area larger than the average of all the HJA meadows was designated as "large." Meadows with areas smaller than average were considered "small." Another shapefile was created to illustrate the distances between meadows using lines. If a meadow was connected to other meadows with lines longer than average, it was considered "isolated." Meadows with shorter-than-average distance lines were considered "non-isolated."

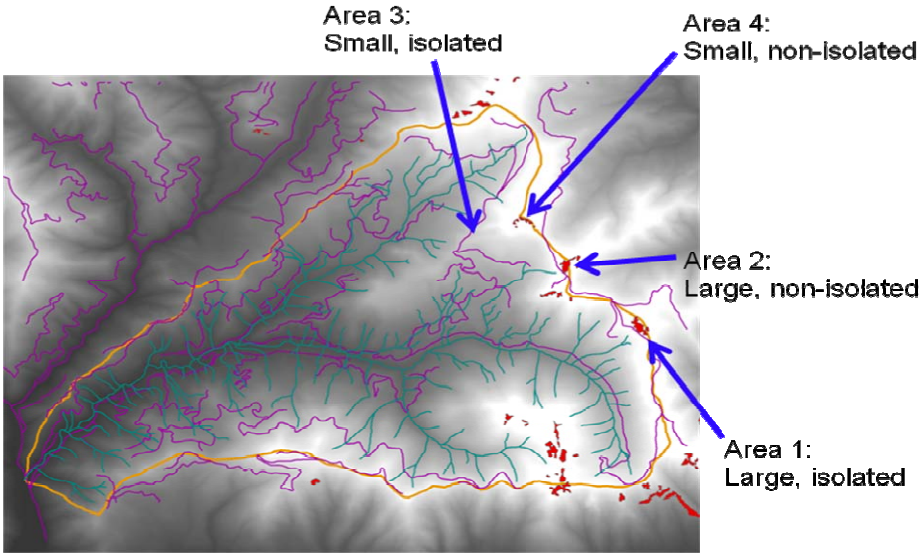

Fig. 1. The areas selected for study.

The intention was to study one area every night for four consecutive nights. Unfortunately, this was made impossible, due to periodic shortages and failures of equipment. As a result, Areas 1, 2, 3 and 4 were studied on July 31st, August 6th, August 7th and August 12th, respectively.

These meadows were studied through the use of three moth traps. One trap was set in the center of the meadow, one was set roughly 100 km into the meadow's neighboring forest and one was set on the border between the meadow and the forest. These traps attracted moths and other night fliers through the use of a UV lamp, powered by a wheelchair battery. Upon contact with the lamp, moths would fall into a bucket containing a pesticide strip. For each area studied, the traps would be set shortly before nightfall and would be collected early after daybreak the next morning. After each collection, the traps would be emptied into individual containers and the batteries would be recharged.

Moths collected in the traps were sorted, one trap at a time. The intention was to sort moths by species, but we did not have the time, resources or expertise to identify each moth with its proper scientific label. To that end, moths with identical body size, wing size and wing patterns were considered of the same species and given its own label of one or two letters. A logbook was kept to record how many moths of each species found in every individual trap. Also recorded were the traps' locations and when the samples were collected. In addition to moths, nine notable species of insect were found in the traps. Because these species were not moths, they were labeled as "Other," with individual numbers between one and nine.

Modeling rare moth species by a metapopulation method-To study the dynamics of a population of rare moths, we assume a set of rules that govern how the moth population will change in time. Each moth lives at one of several locations (we call the set of moths at a particular location a metapopulation). At each location, there is an abundance of moths of other species so that our rare species makes up an insignificant fraction of the total population. These moths all share the same food sources, predators, etc. The total population is in equilibrium. Time is divided into discrete generations. In each generation, each moth at location i starts as a caterpillar and has the following "lifecycle:"

1. To become a moth and successfully reproduce it must win a Bernoulli trial at probability p_i.
2. If it wins, it chooses a site to reproduce at. The chance of site j being chosen is T_{ij} (this mandates that $\Sigma_j T_{ij}=1$).
3. Each victorious moth begets r caterpillars at its chosen site.

Because the total population is in equilibrium, on average each moth is provides one replacement for its self in the next generation. Hence if P_i is the number of rare moths at the i^{th} location, and $P_i^{'}$ is the number of rare moths there on the next turn, then we may assume

$$E(P_i^{'}) = P_i ,$$

where $E(\mathrm{X})$ is the expectation value of the random variable X. Because the number of rare moths does not significantly contribute to the total number of moths, the probability p_i does not depend on the number of rare moths present. Thus each moth's Bernoulli trial is independent, and p_i depends only on i. So the number of caterpillars a given caterpillar a begets at the j^{th} site on the next turn is a random variable X_{aj} that takes the value r with probability $p_i T_{ij}$ and 0 otherwise. Then the number of caterpillars on the next turn at the j^{th} site is $\Sigma_a \mathrm{X}_{aj}$. The sum of Bernoulli trials is a binomial random variable, so

$$P_j' = r \cdot \Sigma_i Bin(P_i,\ p_i T_{ij}).$$

The expectation value of this is $r\Sigma_i P_i p_i T_{ij}$, so we know that $P_j = r\Sigma_i P_i p_i T_{ij}$. Thus for each j,

$$r\Sigma_i\, p_i T_{ij} = 1.$$

Null Hypothesis Anti-confidence method for analyzing moth habitat preference data-Suppose moth samples are taken from many sites with varying habitat characteristics such as altitude, or the presence of some other species. Suppose that from this data we wish to infer what species of moths prefer what types of habitat. We may partition the set of all possible habitats into subsets (habitat types), and count up which habitat type averages the most counts of our moth species (species *Z*) per sample. By partitioning the set of habitats, we mean that the habitat types are mutually exclusive. Suppose we have a set *S* of habitat types and we take S_i samples from each i^{th} habitat type. We find a total of *M* moths, and we find our species to be most prevalent in habitat type *x* at *A* counts per sample. If we wish to conclude that species *Z* prefers habitat type *x*, then we must ask: "How confident are we in this result?"

At this point we have two reasonable hypotheses: either species *Z* prefers habitat type *x*, or the null hypothesis—species *Z* doesn't prefer a particular habitat type but we found it prevalent in habitat type *x* by chance. We can calculate the probability that if that the null hypothesis were true, then at least one habitat type would average *A* moths per sample. If this probability is small, then we have high confidence in our result.

To find this probability we will first solve an easier counting problem: in how many ways can *M* moths be distributed to *N* samples such that no sample has more than *L* moths? We can solve this by finding a polynomial whose m^{th} coefficient is the number of ways *m* moths be distributed to *N* samples such that no sample has more than *L* moths. Consider the m^{th} coefficient of the polynomial given by the following product:

$$f(x) = m!\left(\frac{x^0}{0!}+\frac{x^1}{1!}+\frac{x^2}{2!}+\ldots+\frac{x^L}{L!}\right)\cdot\left(\frac{x^0}{0!}+\frac{x^1}{1!}+\frac{x^2}{2!}+\ldots+\frac{x^L}{L!}\right)\cdot \underset{\cdots}{N\ times} \cdot\left(\frac{x^0}{0!}+\frac{x^1}{1!}+\frac{x^2}{2!}+\ldots+\frac{x^L}{L!}\right)$$

To expand this product, we add up each way to multiply together one term from each sum. If we identify choosing the i^{th} term from the j^{th} sum with distributing *i* moths to the j^{th} sample, then we have a one to one correspondence between summands in the m^{th} coefficient of the expanded product and ways to choose numbers of moths to put in each sample such that no sample receives more than *L* moths. Since there are $m!/(n_1!n_2!...n_N!)$ ways to put $n_1,\ldots,n_N$ moths in boxes $n_1,\ldots,n_N$, it follows that the m^{th} coefficient of $f(x)$ is the number of ways to distribute *M* moths into *N* samples such that no sample has more than *L* moths. By the power rule for derivatives, we see that

$$g(M) = \frac{d^M}{dx^M}\prod_{i=1}^{N}\left(\frac{x^0}{0!}+\frac{x^1}{1!}+\frac{x^2}{2!}+\ldots+\frac{x^L}{L!}\right)\Bigg|_{x=0} = \frac{d^M}{dx^M}\left(\prod_{i=1}^{N}\sum_{j=0}^{L}\frac{x^j}{j!}\right)\Bigg|_{x=0}$$

also counts the number of ways to distribute *M* moths into *N* samples such that no sample has more than *L* moths.

Now suppose instead that we have N habitat types, and each i^{th} habitat type is sampled n_i times. How many ways are there to distribute M moths to the samples so that no habitat type receives more than L moths? If we rename the samples in $g(M)$ habitat types, we get the number of ways to distribute M moths into N habitat types such that no habitat type has more than L moths. Since we are counting the number of ways to distribute the moths to samples, each time we distribute a moths to a habitat type, we must multiply the number of ways to do this by the number of samples taken from that habitat type. The function

$$h(M) = \frac{d^M}{dx^M}\left(\prod_{i=1}^{N} \sum_{j=0}^{L} \frac{(n_i x)^j}{j!} \right)\Bigg|_{x=0}$$

these ways of distributing moths.

We can now adjust the maximum number of moths allowed in each habitat type so that each habitat type averages at most A moths per sample. The function

$$F(M) = \frac{d^M}{dx^M}\left(\prod_{i=1}^{N} \sum_{j=0}^{\lfloor n_i A \rfloor} \frac{(n_i x)^j}{j!} \right)\Bigg|_{x=0}$$, where $\lfloor n_i A \rfloor$ is the greatest integer less than $n_i A$,

counts the number of ways to distribute M moths to samples, where each sample is in one of N habitat types, there are N_i samples in each i^{th} habitat type, and no habitat type may average more than A moths per sample.

We can now solve the problem that we originally posed. If our moths have no preference of habitat type, then each moth independently chooses a sample, each sample being equally likely. The number of samples is

$$\sum_{i=1}^{N} n_i ,$$

So the number of ways to distribute M moths to these samples is

$$\left(\sum_{i=1}^{N} n_i \right)^M .$$

Since distribution of moths to samples is equally likely, the probability that some habitat type averages more than A moths per sample is

$$P = 1 - \frac{\dfrac{d^M}{dx^M}\left(\prod_{i=1}^{N} \sum_{j=0}^{\lfloor n_i A \rfloor} \frac{(n_i x)^j}{j!} \right)\Bigg|_{x=0}}{\left(\sum_{i=1}^{N} n_i \right)^M} .$$

P is the chance that random data would furnish an average of at least A moths per sample to some habitat type by coincidence; the lower P is, the more confident we are in our hypothesis. We call P the null hypothesis anti-confidence of our data (NHA).

Data Visualization- Dr. Jeff Miller's HJ Andrews moth abundance data have more than 25,000 observations. The data have been maintained in an Excel spreadsheet, but most machine-learning applications have their own input formats and Excel is not

suitable for generating many formats. To facilitate the generation of machine-learning formats we imported the data into a MySQL database according to the schema in appendix A. Storing the dataset in a database allowed us to plug the moth data into the Django web framework. Web frameworks are ideally suited for generating Keyhole Markup Language (KML). KML is the Geospatial Consortium's XML standard for data presentation. Google Earth renders KML on a 3-dimensional dynamic map. Four KML maps are currently being generated from the database (see appendix A).

Machine Learning- We applied two machine-learning algorithms to the moth database. The apriori algorithm comes from the world of supermarket basket analysis. Each sample was treated as a "basket" where each species of moth present in the sample is in the basket. Apriori then looked at frequently co-occurring species to build rules of inference according to thresholds of support and confidence. Support is the percentage of baskets that have all the species in the inference rule and confidence is the percentage of baskets that have a consequent conditional on the antecedent.

The second machine-learning algorithm we applied was Latent Dirichlet Allocation (LDA). LDA mines "topics" consisting of "word" (moth species) generation probabilities. Each sample then has a probability of being generated under a topic. We used these probabilities to generate image overlays for Google Earth, interpolating the probabilities according to a Gaussian distribution.

Results & Discussion

Meadow Boundary Sampling-In the end, 360 different species were collected. Though further work is necessary to properly identify and group these samples, some trends are immediately apparent. For example, consider the following table:

	M1	B1	F1	M2	B2	F2	M3	B3	F3	M4	B4	F4
Other2	10	11	22	20	22	14	9	4	5	3	6	9

Fig. 2. The occurrence of species "Other 2" in all of the traps set out during the study. M1 stands for the trap in the meadow of Area 1. B1 stands for the trap at the border of Area 1, F1 stands for the trap at the forest of Area 1, M2 stands for the trap at Area 2, and so on.

Though Other 2 is not a species of moth, it deserves special mention as the only species to appear in every single trap set out. In Areas 1 and 2, particularly, Other 2 is more prominent than most other moths in the area. In our opinion, this merits further study into the identification and attributes of Other 2, as well as its feeding habits and interactions with moths.

	M1	B1	F1	M2	B2	F2	M3	B3	F3	M4	B4	F4
AO	29	2		4		1						
AP	26	7		1							1	

Fig. 3. The occurrence of species "AO" and "AP" in all of the traps set out during the study. M1 stands for the trap in the meadow of Area 1. B1 stands for the trap at the border of Area 1, F1 stands for the trap at the forest of Area 1, M2 stands for the trap at Area 2, and so on.

The moth species AO and AP show an overwhelming preference for the large and isolated Meadow 1. AO has a decent presence in the large, non-isolated Meadow 2 as

well. It is possible that there is something else about Meadow 1 that attracts so many of these species. In any case, it is safe to assume that if Meadow 1 becomes completely forested, these species may face extinction in HJA.

	M1	B1	F1	M2	B2	F2	M3	B3	F3	M4	B4	F4
GY					1		25		1			

Fig. 4. The occurrence of species "GY" in all of the traps set out during the study. M1 stands for the trap in the meadow of Area 1. B1 stands for the trap at the border of Area 1, F1 stands for the trap at the forest of Area 1, M2 stands for the trap at Area 2, and so on.

As species AO and AP show preference to Meadow 1, GY appears extraordinarily dependent on Meadow 3. Again, it is quite possible that if Meadow 3 shrunk out of existence, species GY could cease to exist in this forest.

Moth Metapopulation-We were unable to use metapopulation models to make any interesting predictions. Because there are many parameters to infer and the data available did not include the population of moths at the sample location, we were unable to calibrate our model.

Data Visualization-The KML produced by the web framework is verbose by nature, but optimizations could be put in place to reduce the network load below the 7MB of two of the files. Also, placemark icons could be generated according to genera or another grouping factor.

Machine Learning-The apriori algorithm produced numerous association rules, but varying emergence times prevent the algorithm from producing valuable rules. Future work should look at ways of aggregating samples across years to compensate for temporal factors.

We ran LDA for 5, 10, 20, 30, 40, and 50 topics. The information for each run is available at atlasoflife.com. Expert etymologists have several things to look at when examining the output from this algorithm. The visual output for Google Earth allows for anyone to flip through each topic geographically. It is easy to determine the habitat type (meadow vs woods), the aspect, and elevation to look for commonalities among the higher probability samples. The file "model-final.twords" reports the top 20 species probabilities for each topic. Together these files will help an etymologyst determine whether LDA can create meaningful topics.

Conclusions

Our work produced many starting points. Additional boundary sampling may provide statistically significant evidence of species woodland dispersal capability. By linking the moth database to an environmental factor database, future researchers can apply the box algorithm to factors influencing the presence of moth species. It is unclear whether LDA yields useful information at this point, but with thorough analysis by Jeff Miller, the topics may provide a valuable classification tool for habitats.

Appendix A

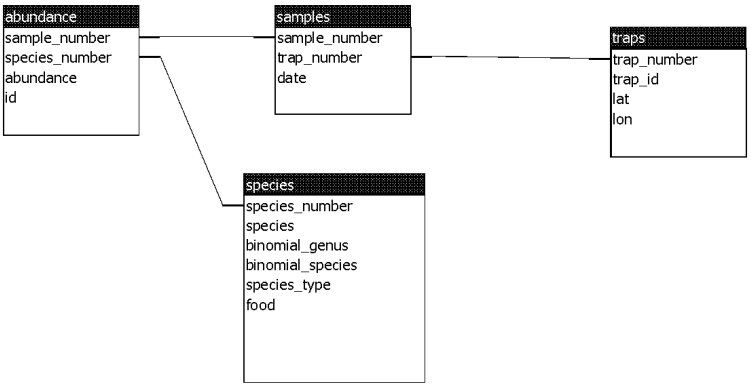

Table 1. kml Layers	
Species Name	http://atlasoflife.com/kml/topics/binomial_species
Genus	http://atlasoflife.com/kml/topics/binomial_genus
Trap Number	http://atlasoflife.com/kml/topics/trap_number
Sample Number	http://atlasoflife.com/kml/topics/sample_number