Identifying an optimal feature set to analyse species richness in Butterflies

Erin Aho, Albert Nyarko-Agyei, and Soham Talukdar

Butterflies are important taxa, both ecologically and culturally, but have been declining in population despite conservation efforts. To improve understanding of where to focus conservation efforts we aim to identify which features are the most informative in terms of species richness in an area. The species count of butterflies in 45 broad regions was analysed. Based on existing research, we decided on a set of 28 features, both bioclimatic and anthropogenic, which are known to affect species richness. We generated a dataset by taking the mean across our regions for each feature, using open-source GIS tools. After researching current feature selection methods, we decided to use Joint Mutual Information (JMI) as our selection criterion, due to the high amount of mutual information in many of our features, and the strong performance JMI has been shown to have form small data samples. After performing feature selection on our dataset using JMI, we were able to identify that the most informative set of 5 features is made up of: annual precipitation, isothermality, elevation, mean diurnal temperature range, precipitation of the warmest quarter, and the percent of urbanization. We discuss possible limitations of our approach, and areas further research is needed in order to better understand butterfly species richness.

Introduction.– Butterflies (Lepidoptera: Rhopalocera) are perhaps the most familiar group of all insects, with estimates of between 15,000 and 25,000 species worldwide. They have been prominently featured in cultural imagery going back at least two millennia.[1], and historically they have played an important part in developing our understanding of biodiversity[2].

Butterflies also have a vital role in their ecosystems, where they act as a pollinator, helping plants to reproduce and increasing floral genetic variation[3].

Their cultural and ecological importance makes it especially worrying that despite conservation efforts, butterfly populations have been declining rapidly alongside most insect groups worldwide [4]. While it is known that this unprecedented rate of biodiversity loss is tied with human activities, like deforestation and climate change, there has been limited success in predicting where best to focus future conservation efforts [5–7]. By finding which factors provide the most information about butterfly species richness, we aim to reduce this knowledge gap and help direct conservation efforts to where they are most needed.

Existing understanding.–

Regarding how number of species in a region varies with the area of the region, it has been observed that log(species count) correlates with log(area)[8]. Connor & McCoy (1979) discussed two main hypotheses explaining species-area relationships (SAR):(i) the habitat diversity hypothesis; and (ii) the demographic process hypothesis. The first hypothesis states that larger areas are more likely to contain more habitat types; this increase in habitat diversity causes the observed correlation. In contrast, the demographic explanation is process-based, incorporating the dynamic processes of dispersal, colonization, speciation and extinction at multiple spatial scales. Larger areas have higher probabilities of colonization and speciation and lower probabilities of extinction, fostering higher diversity[9].

The viability of a given species depends on a large number of factors, and can vary considerably on small scales, even across communities of the same species[10]. It is well known that total available energy (most commonly solar energy flux) is an important factor for the overall biodiversity in a region[11, 12]. For butterflies in particular, changes in volume of rainfall in local micro-climates has been shown to strongly affect the population of several species of butterflies[13]. It has also been shown that Butterfly communities follow Rapoport's Rule, and show a strong elevational gradient in species richness[14, 15].

Several studies show that agricultural land use can affect both habitat suitability and availability of species who would live in grasslands[16]. Urbanization is one of the most important causes of natural habitat loss and fragmentation, and is linked to decreased plant species diversity, reduced water quality, and increased air and soil pollution, which in turn can cause species diversity decline[16, 17].

Species richness (i.e., the number of species occurring in a given area) is commonly used as a health descriptor for a community, as it is combines many relevant ecological features such as environmental stability, ecosystem productivity, and biological factors[18]. Modern ecological research commonly investigates anthropogenic causes of biodiversity loss like urbanisation, deforestation, and global environmental change, in the aim of understanding how human activity is affecting our planet so that we might mitigate our harmful impacts. Against this background, understanding the relationship between species richness and environmental factors is particularly important.

Exactly which features are most important to species richness is not well understood, and differs for each taxa. We aim to help shrink this gap in information for butterflies by finding which features are the most important to track so that we can better understand how the impacts of anthropogenic habitat disruption will affect butterflies moving forward. We have thus created a broad set of features based on the factors which are important to biodiversity, and aim to find the most informative features for Butterflies specifically.

Method.– In this study, we aim to identify the most informative features in predicting species richness in Butterflies. We were provided with a set of 45 regions, and the number of species found in each.

Based on the work in Species-Area relationships by Connor et al[8], we created our target variable from the given data by taking $log(speciescount)/log(area)$, which we will refer to as log(species richness). From the findings Based on the studies discussed, we identified the importance of accounting for climatic variation, elevation, available energy, and human influences. To account for climatic variability, we used the unitary 19 bioclimatic indices which account for the main climate based physiological constraints for biodiversity[19]. Our elevation, net solar flux and land coverage data are from the NASA Near Earth Observatory[][20]. We created our feature variables using open source GIS tools to calculate zonal statistics, or a zonal histogram in the case of land coverage data, and rasters for each of the features we wanted to investigate. The output from the GIS feature extraction, as well as any preprocessing done before inputting the features into our model, is available on Github [24]

As we are aiming to create a set of features that are easily interpreted and applied to new data, we decided to use feature selection over principal component analysis (PCA). Within feature selection, we made use of a filter method instead of a wrapper method to limit risk of over-fitting, which could be quite prominent due to the large number of features in total (28) when compared to the total number of regions we had observations for (45).

Choosing a selection criterion is non-trivial, and decades of trying theoretical and heuristic approaches has produced a large number of options. Brown et al[21] found that Joint Mutual Information (JMI)[22] had the best overall performance in their framework across a number of small sample size datasets. Given our small dataset, we decided to use JMI as our selection criterion.

JMI is an extension of the Mutual Information Feature Selection (MIFS) criterion, proposed by Battiti in 1994:

$$J_{mifs}(X_k) = I(X_k;Y) - \beta \sum_{X_j \in S} I(X_k;X_j)$$

where S is the set of currently selected features, I denotes the information shared, and β is a hyperparameter, which must be set experimentally. Using $\beta = 0$ would be equivalent to $J_{mim}(X_k)$, treating each feature as independent, and a larger value places more emphasis on reducing inter-feature dependencies.

Yang and Moody (1999) and Meyer et al. (2008) proposed JMI in an effort to reduce the need to train hyperparameters, while still minimising redundancy in a selected feature set[21]. The JMI score for a feature X_k given an already selected set of features S is:

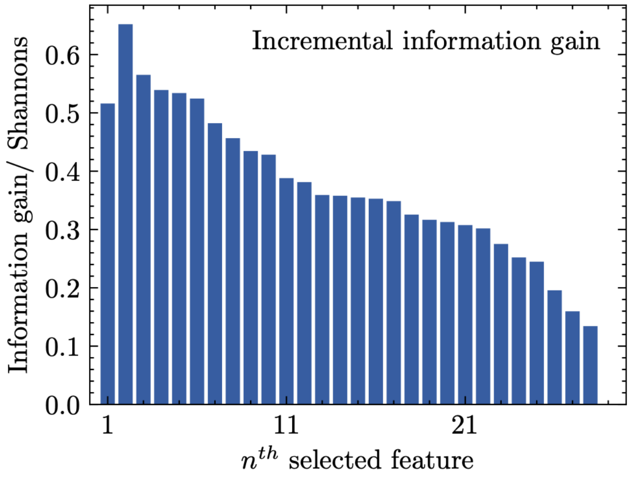
$$J_{jmi}(X_k) = \sum_{X_j \in S} I(X_k X_j;Y)$$

One major difficulty in implementing a information based feature selection method is that entropy (and therefore information) for continuous variables is not as simple to calculate as it is for discrete variables. B. Ross[23] and D. Homola [?] championed a kNN approach to estimating shared information, which we made use of in our implementation. For further details on how our implementation works, the code is available on Github [?].

Results.–

Top 6 Features	Information Gain
Mean annual precipitation	0.516
Mean isothermality	0.652
Elevation	0.539
Mean Diurnal Range	0.507
Precipitation of warmest quarter	0.499
Mean min temp of coldest month	0.457

The result of the JMI feature selection algorithm highlighted a set of features that have optimal information gain with the target variable of this paper, Log(Species Density). The mean annual precipitation was the first feature selected by the algorithm. This means that when we took each feature and the target variable, this feature had the highest amount of information gain.



Discussion.– The second added feature was isothermality, and worth noting is that adding this second feature has a greater information gain than the first feature alone. This shows that isothermailty and precipitation together hold extra information about our target variable than either did separately. Isothermality is also known to be one of the most important environmental factors for butterflies [25]. After adding the second feature, successive additions result in less information gain. This is indicative of increasing amounts of redundancy in our dataset. It should be noted that each feature is still increasing the total amount of information within our selected set, but that some of the information is already accounted for in other variables already selected. (E.g mean temperature in the summer has a large amount of redundancy with annual max temperature)

It is interesting to consider the correlation of the variables and the species density along with their order of selection. From the original list of features, the mean minimum temperature of the coldest month had the highest correlation with the species density at 0.572 but it was only the sixth selected feature. The next highest correlated feature was mean annual precipitation and this was the first feature to be selected by the JMI algorithm. This emphasises the difference between information gain and correlation, as the most informative feature may not be correlated well if the relationship is nonlinear.

A limitation of our method is that we could not account for human biases. For a given area, if more time is spent researching butterflies in that area, then it is likely that more species will be found. While butterflies are one of the most studied taxa, there is predicted to still be many undiscovered species, particularly in remote areas. At the very least, the robustness of species counting methods, such as counts along a transect, are question by Kevin Gross et al [26]. It is hard to account for these factors, and the high uncertainty in the true number of species in an area is may limit the accuracy of our results.

One must also consider the flaws in the JMI algorithm. In particular, it is a greedy algorithm which selects the locally optimal feature at each stage, which results in should result in groups of features that have the highest mutual information with the target variable, but this approach ”lacks theoretical guarantees” according to Gao et al [27]. A further study could implement JMIM, which tackles this issue of local maxima by conducting more searches on combinations of features, and compare the findings to our results.

An interestingly unimportant feature is that of the Island boolean tag, which did not show significant information gains, being the 19th of 28 features selected, and only added 0.316 Shannons to the dataset when selected. This follows the findings of Kalmar and Currie in Avian species [28]. Kalmar and Currie found that most variation in species density is not context-specific, and varied similarly on continents and islands according to the same bioclimatic indices. Our findings seem to agree, and find that whether an area is an island or not does not add significant information to the dataset, and so is not particularly relevant to the species richness in butterflies.

Conclusion.–

Overall, we have been successful in reducing the original 28 features to a small subset of informative features for butterfly species richness. These findings can be used to help identify areas to focus conservation efforts on. However, they highlight that the list of features that affect butterfly richness is complex and that many connections exist between these variables.

[1] M. Dicke, American Entomologist **46**, 228 (2000).

[2] P. J. DeVries, in *Encyclopedia of Biodiversity (Second Edition)*, edited by S. A. Levin (Academic Press, Waltham, 2001) second edition ed., pp. 650–661.

[3] M. Ghazanfar, M. F. Malik, M. Hussain, R. Iqbal, and M. Younas, Journal of Entomology and Zoology Studies , 5 (2016).

[4] D. L. Wagner, E. M. Grames, M. L. Forister, M. R. Berenbaum, and D. Stopak, Proceedings of the National Academy of Sciences **118** (2021), 10.1073/PNAS.2023989118.

[5] P. J. White and J. T. Kerr, Global Ecology and Biogeography **16**, 290 (2007), _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1466-8238.2007.00298.x.

[6] G. A. Montgomery, R. R. Dunn, R. Fox, E. Jongejans, S. R. Leather, M. E. Saunders, C. R. Shortall, M. W. Tingley, and D. L. Wagner, Biological Conservation **241**, 108327 (2020).

[7] P. J. Clark, J. M. Reed, and F. S. Chew, Urban Ecosystems **10**, 321 (2007).

[8] E. Connor and E. McCoy, Encyclopedia of Biodiversity **5**, 397 (2001).

[9] S. Drakare, J. J. Lennon, and H. Hillebrand, Ecology letters **9**, 215 (2006).

[10] R. L. Brown, L. A. Jacobs, and R. K. Peet, en*eLS*, (2007), 10.1002/9780470015902.a0020488.

[11] B. A. Hawkins, R. Field, H. V. Cornell, D. J. Currie, J.-F. Guégan, D. M. Kaufman, J. T. Kerr, G. G. Mittelbach, T. Oberdorff, E. M. O'Brien, E. E. Porter, and J. R. G. Turner, Ecology **84**, 3105 (2003), _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1890/03-8006.

[12] D. J. Currie, The American Naturalist **137**, 27 (1991), publisher: The University of Chicago Press.

[13] N. F. Haneda and P. B. Panggabean, IOP Conference Series: Earth and Environmental Science **394**, 012041 (2019).

[14] E. Fleishman, G. T. Austin, and A. D. Weiss, Ecology **79**, 2482 (1998), https://doi.org/10.1890/0012-9658(1998)079[2482:AETORS]2.0.CO;2.

[15] SANCHEZ-RODRIGUEZ and A. Baz, Journal of the Lepidopterist' Society, **49**, 192 (1995).

[16] E. Öckinger and H. G. Smith, Oecologia **149**, 526 (2006).

[17] O. Tzortzakaki, V. Kati, M. Panitsa, E. Tzanatos, and S. Giokas, Landscape and Urban Planning **183**, 79 (2019).

[18] Y. S. Park, R. Céréghino, A. Compin, and S. Lek, Ecological Modelling **160**, 265 (2003).

[19] R. Hijmans, S. Cameron, J. Parra, P. Jones, and A. Jarvis, International Journal of Climatology **25**, 1965 (2005).

[20] For this paper,.

[21] G. Brown, A. Pocock, M.-J. Zhao, and M. Luján, Journal of Machine Learning Research **13**, 27 (2012).

[22] H. H. Yang and J. Moody, , 7 (2012).

[23] B. C. Ross, PLoS ONE **9** (2014), 10.1371/JOURNAL.PONE.0087357.

[24] Https://github.com/E-Aho/Butterfly-Modeling.

[25] N. Rueda-M, F. C. Salgado-Roa, C. H. Gantiva-Q, C. Pardo-Díaz, and C. Salazar, Frontiers in Ecology and Evolution **9** (2021).

[26] K. Gross, E. J. Kalendra, B. R. Hudgens, and N. M. Haddad, Population Ecology **49**, 191 (2007).

[27] S. Gao, G. V. Steeg, and A. Galstyan, arXiv:1606.02827 [cs, stat] (2016), arXiv: 1606.02827.

[28] A. Kalmar and D. J. Currie, Ecology **88**, 1309 (2007).