Using Feature Selection to Identify Features that are Most Important to Species Richness in Butterflies

Erin Aho, Albert Nyarko-Agyei, and Soham Talukdar

Butterflies are an important taxa, both ecologically and culturally, but populations have been declining globally despite conservation efforts. To improve understanding of where to focus conservation efforts we aim to identify the features that are the most informative of an area's species richness. 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. 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 for small sample sizes. 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: temperature seasonality, precipitation in the driest month, precipitation in the warmest quarter, max temperature in the warmest month, and the mean diurnal range . 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 among the most familiar group of all insects, approximately 17,200 species worldwide[1]. They have been prominently featured in cultural imagery going back at least two millennia.[2], and historically they have played an important part in developing our understanding of biodiversity[3].

Butterflies also have a vital role in their ecosystems, where they act as a pollinator, helping plants to reproduce and increasing floral genetic variation[4].

Their cultural and ecological importance makes it especially worrying that, despite conservation efforts, butterfly populations have been declining rapidly alongside most insect groups worldwide [5]. While it is known that this unprecedented rate of biodiversity loss is tied with human activities, such as deforestation and climate change[6–8], there has been limited success in predicting where best to focus future conservation efforts. 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.– 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 combines many relevant ecological features such as environmental stability, ecosystem productivity, and biological factors[9].

One feature which is sure to impact the number of species in a given region is the region's area. Regarding how the 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)[10]. 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, which could support a greater diversity of species. In contrast, the demographic explanation incorporates 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[11]. Investigating these hypotheses is beyond the scope of this paper, and while these hypotheses may disagree on the exact mechanism that drive the SAR, they agree $log(species\ count)$ correlates with $log(area)$, which is a key result we will make use of in our methodology.

Many factors which influence species viability can vary considerably even within small distances[12]. It is well known that total available energy (most commonly solar energy flux) is an important factor for the overall biodiversity in a region[13, 14]. 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[15]. It has also been shown that Butterfly communities follow Rapoport's Rule, and show a strong elevational gradient in species richness[16, 17].

Several studies show that agricultural land use can affect both habitat suitability and availability of species who would live in grasslands[18]. 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[18, 19]. Modern ecological research commonly investigates anthropogenic causes of biodiversity loss like urbanisation, deforestation, and global environmental change, with the aim of better understanding how human activity is affecting our planet so that we might mitigate our harmful impacts[6, 20–22]. 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, informed by our discussion above, and aim to find the most informative features for Butterflies specifically.

Method.–We were provided with a set of 45 regions, and the number of species found in each. We generated a dataset by taking the mean across our regions for each feature, using open-source GIS tools.

Based on the work in Species-Area relationships by Connor et al[10], we created our target variable from the given data by taking $log(species\,count)/log(area)$, which gives us an estimator for species richness that is not dependent on the area of a region. 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[23]. Our elevation, net solar flux and land coverage data are from the NASA Near Earth Observatory. We created our feature variables using open source GIS tools[24] to calculate zonal statistics, or a zonal histogram in the case of land coverage data over open source vectors for each region[25–28], 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 [29]. The full list of features used, and the sources for the rasters, is given in Appendix A.

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 the 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[30] found that Joint Mutual Information (JMI)[31] 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, X_k is a feature to select, Y is our target variable, I denotes the information shared, and β is a hyperparameter, which must be set experimentally. Using $\beta = 0$ would be equivalent to treating each feature as fully independent, and a larger value places focuses more on reducing inter-feature dependencies in the selected data.

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[30]. 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[32] and D. Homola [33] championed a kNN-based approach to estimating shared information, advancing work done by Kraskov et al in 2004[34]. This method avoids needing to bin continuous data by using k-nearest neighbour distances to estimate entropy instead, which we made use of in our implementation. For further details on how our implementation works, the code is available on Github [35].

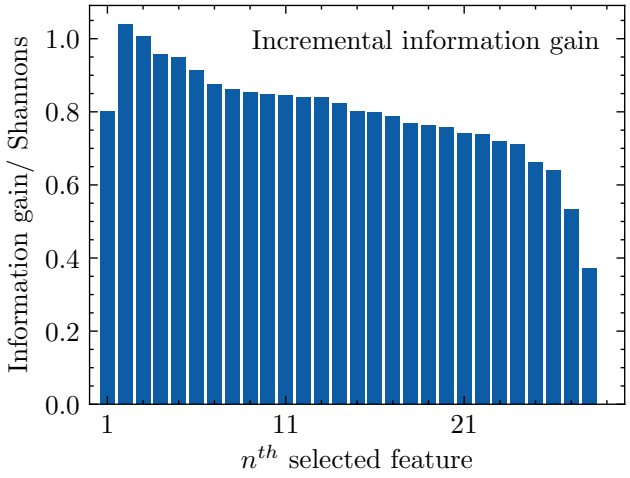
Results.–

Feature	Information Gain / shannons
Temp. seasonality	0.801
Precip. of driest month	1.04
Precip. of warmest quarter	1.01
Max temp. of warmest month	0.957
Diurnal temp. range	0.949
Precip. of wettest quarter	0.913
Annual precip.	0.876
Latitude	0.861
Annual mean temp.	0.854
Annual insolation	0.848

TABLE I: A table showing the first 10 sequentially selected features and their respective information gains

The result of the JMI feature selection algorithm highlighted a set of features that have an optimal information gain with the target variable of this paper, $log(species\ count)/log(area)$. The mean temperature seasonality 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. The full table of all sequentially selected features and their information gains are given in Appendix B.

FIG. 1: A plot showing incremental information gain for each feature added to the selected feature set



Discussion.– As we can see from Fig 1, adding the second feature (precipitation in the driest month) added significantly more information than the first variable did alone. This shows that seasonality and the precipitation in the driest month together hold extra information about our target variable than either did separately. Precipitation is also known to be one of the most important environmental factors for butterflies[36], so seeing seasonal precipitation features be among the most informative in our dataset make sense.

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, but it was selected second last. This emphasises the difference between information gain and correlation, as the most informative feature may not be correlated well if the relationship is non-linear.

Among the features we looked at which were intended to investigate anthropogenic effects (PM 2.5 concentration & land usage percentages), PM 2.5 was the most informative, and was selected 17th. While it is interesting to note that our anthropogenic variables were not found to be particularly informative, this may be an artefact of our methodology. Due to the majority of our regions being large in area (country sized), and the nature of local human effects typically being short ranged (habitat destruction, air pollution, etc) [37], we would expect climatological variables to be more important at the wide scale of our regions. Further research is needed at with smaller scaled regions, and a larger number of observations, to better identify how butterfly species richness interacts with these features.

However we should also note that wide scale climate change brought about by human actions has already been shown to affect precipitation and temperature patterns[38, 39], and that butterflies in particular are sensitive to changes in temperature extremes that relate to seasonality[40]. Given that our most informative features on butterfly species richness are all rapidly changing year on year as a result of anthropogenic climate change[41], it is clear that butterfly species richness will be sensitive to future changes in these features.

A further limitation of our method is that we could not account for human biases in data collection. If more time is spent researching butterflies in a given area, then it is likely that more species will be found. While butterflies are one of the most studied taxa, there are predicted to still be many undiscovered species, particularly in remote areas. In turn, the robustness of species counting methods, such as counting along a transect, are also questioned by Kevin Gross et al[42]. While it is hard to account for these factors, the high uncertainty in the true number of species in an area may limit the accuracy of our results.

One must also consider the flaws in the JMI algorithm. Critics of the JMI algorithm raise that it is a greedy algorithm, which selects a locally optimal feature at each stage. This will often result in groups of features which have the highest mutual information with the target variable, but as noted by Gao et al, this method does not guarantee a globally optimal solution[43]. A further study could implement JMIM, an adaptation of the JMI algorithm proposed by Bennasar et al[44], which tackles this issue 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 last selected feature, and having an information gain of just 0.373 shannons, compared to 0.533 shannons for the penultimate feature. This follows the findings of Kalmar and Currie in Avian species[45]. 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. Future research should aim to verify our findings are replicable at different scales, and with more observations. We hope that our findings might help inform conservation efforts by combining them with climatic predictions to identify areas that are most at risk of biodiversity loss in butterflies.

[1] O. Shields, JOURNAL OF THE LEPIDOPTERISTS' SOCIETY **43**, 6 (1989).

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

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

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

[5] 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.

[6] P. J. White and J. T. Kerr, Global Ecology and Biogeography **16**, 290 (2007).

[7] 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).

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

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

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

[11] "The imprint of the geographical, evolutionary and ecological context on species–area relationships - Drakare - 2006 - Ecology Letters - Wiley Online Library," .

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

[13] 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).

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

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

[16] E. Fleishman, G. Austin, and A. Weiss, Ecology **79**, 2482 (1998).

[17] Sanchez-Roduigez and A. Baz, Journal of the Lepidopterist' Society, **49**, 192 (1995).

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

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

[20] J. Hill, C. Thomas, R. Fox, M. Telfer, S. Willis, J. Asher, and B. Huntley, Proceedings. Biological sciences / The Royal Society **269**, 2163 (2002).

[21] M. L. Munguira, in *Ecology and Conservation of Butterflies*, edited by A. S. Pullin (Springer Netherlands, Dordrecht, 1995) pp. 277–289.

[22] F. Sánchez-Bayo and K. A. G. Wyckhuys, Biological Conservation **232**, 8 (2019).

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

[24] QGIS Development Team, *QGIS Geographic Information System*, QGIS Association (2022).

[25] "Natural Earth Downloads - Free vector and raster map data at 1:10m, 1:50m, and 1:110m scales," .

[26] "GADM," .

[27] "Protected Planet | Kinabalu Park," .

[28] "Protected Planet | Región de Calakmul," .

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

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

[31] H. H. Yang and J. E. Moody, *NIPS*, (1999).

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

[33] D. Homola, "MIFS," (2016).

[34] A. Kraskov, H. Stoegbauer, and P. Grassberger, Physical Review E **69**, 066138 (2004), arXiv: cond-mat/0305641.

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

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

[37] C. He, Z. Liu, J. Tian, and Q. Ma, Global Change Biology **20**, 2886 (2014).

[38] K. E. Trenberth, Climate Research **47**, 123 (2011).

[39] S. Ashton, D. Gutiérrez, and R. J. Wilson, Ecological Entomology **34**, 437 (2009).

[40] S. S. Bauerfeind and K. Fischer, Population Ecology **56**, 239 (2014).

[41] S. R. Loarie, P. B. Duffy, H. Hamilton, G. P. Asner, C. B. Field, and D. D. Ackerly, Nature **462**, 1052 (2009), bandiera_abtest: a Cg_type: Nature Research Journals Number: 7276 Primary_atype: Research Publisher: Nature Publishing Group.

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

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

[44] M. Bennasar, Y. Hicks, and R. Setchi, Expert Systems with Applications **42** (2015), 10.1016/j.eswa.2015.07.007.

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

[46] S. E. Fick and R. J. Hijmans, International Journal of Climatology **37**, 4302 (2017).

[47] "Solar Insolation (1 month) | NASA," (2022), publisher: NASA Earth Observations (NEO).

[48] "Net Radiation (1 month) | NASA," (2022), publisher: NASA Earth Observations (NEO).

[49] "Land Cover Classification (1 year) | NASA," (2022), publisher: NASA Earth Observations (NEO).

[50] A. van Donkelaar, R. V. Martin, M. Brauer, N. C. Hsu, R. A. Kahn, R. C. Levy, A. Lyapustin, A. M. Sayer, and D. M. Winker, Environmental Science & Technology **50**, 3762 (2016).

[51] "Topography | NASA," (2022), publisher: NASA Earth Observations (NEO).

Appendix A: Features Used

The 28 features examined in our method were the following:

- Latitude
- Island (boolean referring to if the region is an island or not)
- Bioclimatic Variables (averaged across region)[46]:
    1. Annual Mean Temperature
    2. Mean Diurnal Range
    3. Isothermality (i.e relative scale of temperature change from day/night to temperature change across year)
    4. Temperature Seasonality
    5. Maximum temperature of warmest month
    6. Minimum temperature of coldest month
    7. Temperature annual range
    8. Mean temperature of wettest quarter
    9. Mean temperature of driest quarter
    10. Mean temperature of warmest quarter
    11. Mean temperature of coldest quarter
    12. Annual precipitation
    13. Precipitation of wettest month
    14. Precipitation of driest month
    15. Precipitation Seasonality
    16. Precipitation of wettest quarter
    17. Precipitation of driest quarter
    18. Precipitation of warmest quarter
    19. Precipitation of coldest quarter
- Total Annual Insolation, Dec 2020 - Nov 2021[47]
- Annual Net Flux, Dec 2020 - Nov 2021[48]
- Percent Urbanisation, defined as percent land coverage that is characterized as:
    - Urban and Built Up[49]
- Percent Developed, defined as percent land coverage that belongs to any of:
    - Cropland
    - Urban and Built up
    - Croplands and Vegatation[49]
- Percent Naturally Vegetated, defined as percent land coverage that belongs to any of:
    - Evergreen Needleleaf Forest
    - Evergreen Broadleaf Forest
    - Deciduous Needleleaf Forest
    - Deciduous Broadleaf Forest
    - Mixed Forest
    - Closed Shrublands

  - Open Shrublands
  - Woody Savannas
  - Savannas
  - Grasslands
  - Permanent Wetlands[49]
- Particulate Matter < 2.5 µm [50]
- Elevation[51]

Appendix B: Full Results Table

Sequential features selected	Information Gain / shannons
Mean temperature seasonality	0.801
Mean precipitation of driest month	1.04
Mean precipitation of warmest quarter	1.01
Mean max temperature of warmest month	0.957
Mean diurnal temperature range	0.949
Precipitation of wettest quarter	0.913
Annual precipitation	0.876
Latitude	0.861
Annual mean Temperature	0.854
Annual insolation	0.848
Isothermality	0.845
Temperature annual range	0.839
Mean temperature of wettest quarter	0.839
Precipitation of driest quarter	0.823
Mean temperature of driest quarter	0.802
Precipitation of wettest month	0.798
PM 2.5 Concentration	0.789
Precipitation seasonality	0.767
Net Flux	0.764
Percent Developed	0.758
Precipitation of coldest quarter	0.743
Mean temperature of coldest quarter	0.739
Mean temperature of warmest quarter	0.718
Percent Urbanized	0.711
Elevation	0.661
Percent Naturally Vegetated	0.639
Minimum temperature of coldest month	0.533
Island	0.373

TABLE II: A table showing the each sequentially selected features and their respective information gains