High-Dimensional Analysis of Butterfly Richness using unsupervised learning techniques

Natchira Chongsawad, Thomas Barrick, and Niharika Vijay Patil

Butterfly species richness serves as an important indicator of environmental change but modelling it remains a challenge due to the high dimensionality of influencing factors. This study applies unsupervised learning techniques, including Principal Component Analysis (PCA) and autoencoders, to reduce dimensionality and identify key variables affecting species distribution. A dataset comprising 13 features across 45 diverse locations was analysed, revealing significant correlations between climate variables, land use, and species richness. Clustering results from PCA and autoencoders showed distinct ecological patterns, with latitude, climate classification, and level of deforestation emerging as primary drivers.

Introduction.– Studying the richness of butterfly species is important because butterflies are highly sensitive to changes in their environment, and their migration patterns can indicate how climate change affects biodiversity. By understanding how different species are distributed, we can gain insights into broader ecological impacts [1]. However, modeling species richness is challenging due to the complex interactions among various factors like temperature, humidity, urbanization, and deforestation.

To tackle this complexity, unsupervised learning techniques such as principal component analysis (PCA) and autoencoders can be employed to reduce the dimensionality of the data, making it easier to spot patterns. This research aims to identify the key features that influence butterfly species richness and to examine how these species are distributed across different locations.

Dataset.–
Data Sources: The dataset used in this study consists of 13 features across 45 diverse locations, including large countries (Australia), cities (Hong Kong) and topological features (Mount Kinabalu), ensuring varied conditions affecting butterfly species.

The original dataset only had few features, so we increased the dimensionality of the feature space to capture a broader range of variables. Adding features is a balancing act as you must consider the potential of adding data that is relevant to butterfly richness with the risk of adding noise to the dataset. To ensure relevance, we selected features based on prior studies, referencing [2–12]. These variables include the average Annual Relative Humidity [13–19], Plant species [20–24], Köppen-Geiger Climate Classification [13, 25, 26], Number of species, Area, Latitude, Island, Metres from Sea Level, Percentage of landmass used for agriculture, Amount of forest lost in 2023, Percentage of urbanisation, and Population density per square kilometres [27–31]. The data comes from various sources that can be found in references [32–34].

Exploratory Data Analysis: The richness of butterfly species is influenced by various environmental factors. Generally, most regions exhibit low diversity, although there are some extreme cases. Species richness tends to decrease with increasing latitude, reaching its highest point near the equator. Climate zones significantly impact species counts, with some zones displaying more variation than others.

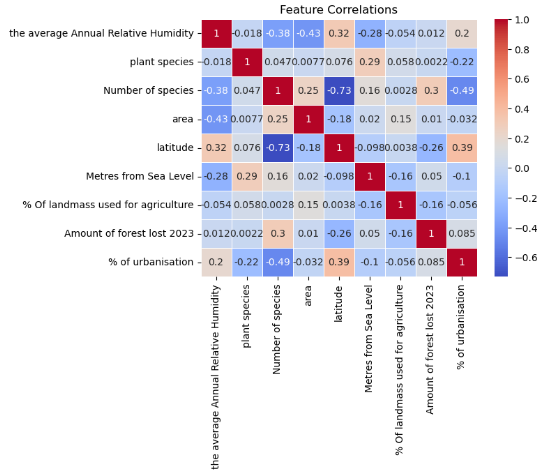

FIG. 1. Heatmap showing the relationships between various features and the number of species.

In Figure 1, a strong positive correlation between deforestation and the species richness was observed, which is contradicts what we expected based on previous studies. This is likely influenced by a confounding factor: areas with larger forest coverage tend to have greater biodiversity and are also more susceptible to deforestation driven by land use for timber. This incorrectly associates high deforestation with the richness of butterfly species, as both are influenced by the availability of forests.

Urbanization negatively impacts species richness, indicating that increased urban development results in a decline in butterfly species due to habitat loss and fragmentation. Humidity exerts a moderate negative influence, whereas forest loss appears to have a slight positive effect. The impact of agricultural land use is minimal, and larger areas generally support a greater number of species.

Methods.–

Data Preparation: Ensuring data quality is essential to obtain accurate results in machine learning. Considerable effort went into data imputation, normalization, and feature encoding [35]. We opted for mean imputation due to its efficiency, especially since most features are expressed as percentages, which minimizes the impact of outliers. Normalization plays a key role in methods like PCA and autoencoders, as it stops features with larger values from dominating others. We applied Z-score normalization to keep the scaling consistent, resulting in better performance and faster convergence.

One hot encoding was used for any categorical data. This ensures that data is usable while maintaining the relationship between the variables as non-ordinal.

Principal Component Analysis (PCA): PCA is a technique designed to reduce the dimensionality of a dataset by identifying principal components that explain the greatest variance within the data. These components are linear combinations of the original variables and are ranked in descending order of their explained variance. In other words, PCA minimizes the dimensionality of the dataset while preserving the most critical information, enhancing the interpretability of the data and facilitating more efficient analysis [36].

Autoencoders: Autoencoders are a type of neural network that compress data into a lower-dimensional representation in the latent space, which is located in the hidden layers of the network [37]. The encoder takes the input and compresses it into this latent space, while the decoder works to reconstruct the original data from it. The model aims to minimize reconstruction loss, ensuring that the latent space retains the most important information. Using activation functions like sigmoid allows the model to capture non-linear relationships present in the dataset.

K-Means Clustering: K-means is an unsupervised learning algorithm that clusters data points based on their similarities, helping to uncover hidden patterns for further analysis. In this study, we apply K-means to the feature space created by PCA and autoencoders to study its relationship with butterfly species richness. The process begins with the random placement of k centroids in the feature space, where k is set to 3 to categorize the data into high, medium, and low richness groups. Data points are then assigned to the nearest centroid, and the centroids are adjusted iteratively until they reach stability [38].

Results.–

Dimensionality Reduction Results: In Figure 2, the first Principal Component (PC) accounts for 33.74% of the variance, while the second PC explains 19.08%, together making up 52.82% of the total variance. By utilizing the first 8 Principal Components, we can capture nearly 100% of the variance, indicating that these 8 features—Climate Classification, Latitude, Area, Amount of forest lost, Humidity, Percent of land mass used for agriculture, Meters from sea level, and Plant species—adequately represent the data.

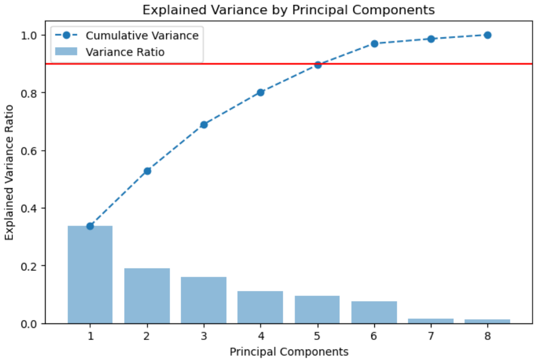

FIG. 2. The bar chart represents the variance ratio of each principal component, while the dashed line shows the cumulative variance. The red line indicates the 90 percent threshold.

Introducing the urbanization feature, even though it correlates with species richness, leads to a decrease in the explained variance of the first principal components. This implies that urbanization might add some noise. Nevertheless, in the autoencoder model we will see that, urbanization proves to be quite significant, highlighting its importance to the data structure that PCA does not fully capture.

Clustering Insights:

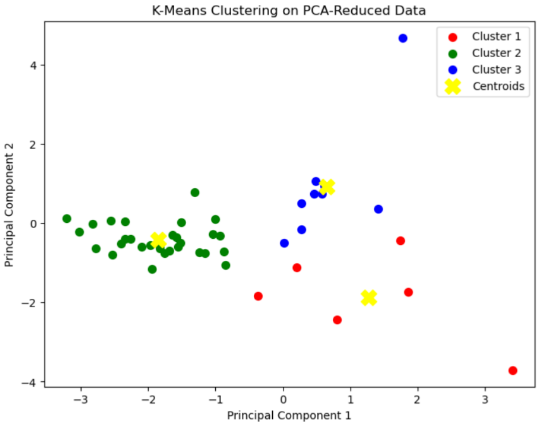

FIG. 3. The scatter plot shows K-Means clustering on PCA-reduced data, with points colored by clusters (red, green, blue) and yellow crosses marking centroids.

As illustrated in Figure 3, PCA clustering uncovers distinct ecological patterns across various climate zones. Cluster 1 encompasses temperate regions such as Finland and Germany, Cluster 2 includes tropical locations like Cuba and Sri Lanka, and Cluster 0 represents transitional areas like Spain and Australia. The principal components underscore the influence of latitude, climate classification, area, forest loss, and humidity on species richness in these regions.

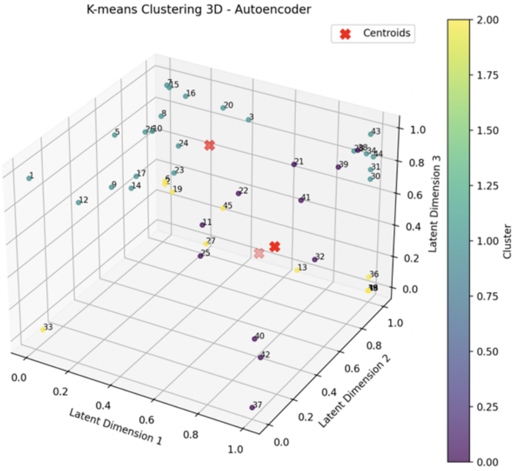

FIG. 4. The plot visualizes K-Means clustering applied to the latent space of an autoencoder, with points colored by cluster assignment. Red crosses indicate centroids, highlighting the cluster centers in the three-dimensional latent space.

By compressing the data into a latent space of three dimensions, the autoencoder captured intricate non-linear patterns that PCA could not. This model was able to achieve a loss of 0.70 after 4500 epochs.

Figure 4 highlights that the autoencoder clustering identified several distinct groups: Cluster 1 comprises countries with high species richness (over 1000 species), primarily found in tropical and subtropical areas like Tanzania and Cameroon. Cluster 2 features moderate diversity (ranging from 78 to 925 species) and includes more temperate countries such as Austria. Cluster 3 is made up of insular regions like Malta, Ireland, and the UK, which exhibit low species diversity (generally under 100 species) due to their limited land area.

Feature Importance: To understand the importance of features on butterfly species richness, we analyze the principal component loadings. PC1 is influenced by Köppen-Geiger Climate Classification, area, and latitude, reflecting the role of climate and geography. PC2 is driven by deforestation, humidity, and area, emphasizing the effects of forest loss and climate factors.

Feature importance in the autoencoder is measured through feature ablation, where changes in feature values are analyzed based on reconstruction loss. Results show that plant species, elevation, forest loss, and urbanization are key factors affecting the latent space. While urbanization correlates with species richness, its inclusion in PCA seems to lead to a decrease in explained variance and silhouette scores, suggesting it may introduce complexities or noise that obscure the principal components. In contrast, autoencoders effectively capture its significance. Additionally, islands were found to reduce biodiversity.

Conclusions.– This study delves into the complex task of modeling butterfly species richness, which serves as an important indicator of climate change. By employing PCA and autoencoders, we were able to pinpoint significant features and ecological patterns associated with global climate zones. Key factors such as latitude, Köppen-Geiger classification, area, and humidity emerged as vital, with autoencoder clustering indicating a higher richness in tropical regions and a decrease in diversity on islands. Nonetheless, there are limitations to consider. The data reflects only a single time point, which restricts our understanding of long-term trends. K-means clustering operates under the assumption of spherical clusters and is particularly sensitive to outliers; exploring alternative methods like DBSCAN or hierarchical clustering could reveal more nuanced patterns. Additionally, broadening the dataset beyond 44 countries would improve the generalizability across different climate zones.

[1] J. Sunde et al. Century-long butterfly range expansions in northern europe depend on climate, land use, and species traits. *Nature Communications*, 6(1), June 2023.

[2] R. L. H. Dennis C. Carbone R. J. Curtis, T. M. Brereton and N. J. B. Isaac. Butterfly abundance is determined by food availability and is mediated by species traits. *Journal of Applied Ecology*, 52(6):1676–1684, September 2015.

[3] F. Maihoff L. Bofinger J. M. Kerner, J. Krauss and A. Classen. Alpine butterflies want to fly high: Species and communities shift upwards faster than their host plants. *Ecology*, 104(1), November 2022.

[4] P. Gros W. Ulrich D. Rödder, T. Schmitt and J. C. Habel. Climate change drives mountain butterflies towards the summits. *Scientific Reports*, 11(1):14382, July 2021.

[5] P. Gros T. Schmitt J. C. Habel, M. Teucher and W. Ulrich. Land use and climate change affects butterfly diversity across northern austria. *Landscape Ecology*, April 2021.

[6] A. Traveset Jesús Sánchez-Dávila and P. Colom. Effects of food availability on butterfly diversity and network specialization across altitudinal levels in a mediterranean landscape. *Biodiversity and Conservation*, 33(1):239–256, November 2023.

[7] C. Tassin de Montaigu and D. Goulson. Factors influencing butterfly and bumblebee richness and abundance in gardens. *Science of The Total Environment*, 908:167995, 2024.

[8] A. Hoskins. *Butterflies of the World.* New Holland Publishers, 2018.

[9] R. I. Vane-Wright and P. R. Ackery. *The Biology of Butterflies.* Academic Press, 1984.

[10] W. B. Watt C. L. Boggs and P. R. Ehrlich. *Butterflies: Ecology and Evolution Taking Flight.* University of Chicago Press, 2003.

[11] A. H. Hannan A. A. Alqarni and A. A. Owayss. Butterfly diversity in saudi arabia: distribution and phenology. *Saudi Journal of Biological Sciences*, 28(1):103–109, January 2021.

[12] S. Herrando C. Stefanescu and F. Páramo. Butterfly species richness in the north-west mediterranean basin: the role of natural and human-induced factors. *Journal of Biogeography*, 31(6):905–915, June 2004.

[13] WeatherOnline. Estonia climate, 2025. [Online].

[14] C. Tiwari H. Gupta and S. Diwakar. Butterfly diversity and effect of temperature and humidity gradients on butterfly assemblages in a sub-tropical urban landscape. *Tropical Ecology*, 60:150–158, April 2019.

[15] D. Gutiérrez and R. J. Wilson. Intra- and interspecific variation in the responses of insect phenology to climate. *Journal of Animal Ecology*, 90:248–259, January 2021.

[16] World Data. World data - average relative humidity, 2025. [Online].

[17] World Weather Online. Weather averages - paraná, tocantins, brazil, 2025. [Online].

[18] Statista. Annual average relative humidity measured, 2025. [Online].

[19] United Nations. Relative humidity data, 2025. [Online].

[20] Á. L. Viloria J. R. Ferrer-Paris, A. Sánchez-Mercado and J. Donaldson. Congruence and diversity of butterfly-host plant associations at higher taxonomic levels. *PLoS ONE*, 8(5), May 2013.

[21] University of South Florida. Florida plant atlas. [Online].

[22] Biodb. Plants per country - biodb, 2025. [Online].

[23] World Rainforests. World rainforests, 2025. [Online].

[24] Visit Calakmul. Nature in calakmul - visit calakmul, 2025. [Online].

[25] Wikipedia. Köppen climate classification. [Online].

[26] B. L. Finlayson M. C. Peel and T. A. McMahon. Updated world map of the köppen-geiger climate classification. *Hydrology and Earth System Sciences*, 11:1633–1644, 2007.

[27] J. Heliölä-J. Pöyry J. Mellado J. Ekroos V. Hyyryläinen I. Vähä-Piikkiö M. Kuussaari, M. Toivonen and J. Tiainen. Butterfly species' responses to urbanization: differing effects of human population density and built-up area. *Urban Ecosystems*, 24(4):515–527, 2020.

[28] Worldometers. World population, 2025. [Online].

[29] City Population. City population - statistics and maps, 2025. [Online].

[30] World Population Review. World population review - population of countries and cities, 2025. [Online].

[31] United Nations. World urbanization prospects, 2018. [Online].

[32] AtlasBig. Countries by average elevation, 1970. [Online].

[33] Trading Economics. Agricultural land (% of land area) by country, 2025. [Online].

[34] Vizzuality. Search — global forest watch, 2025. [Online].

[35] H. Bowne-Anderson. *The Unreasonable Importance of Data Preparation*. O'Reilly Media, March 2020.

[36] A. Maćkiewicz and W. Ratajczak. Principal components analysis (pca). *Computers & Geosciences*, 19(3):303–342, March 1993.

[37] Y. Wang, H. Yao, and S. Zhao. Auto-encoder based dimensionality reduction. *Neurocomputing*, 184:232–242, April 2016.

[38] X. Huang and W. Su. An improved k-means clustering algorithm. *Journal of Networks*, 9(1), January 2014.