Distribution of Butterfly Species Associated with Environmental Factors in Sri Lanka

P.M.S.S Kumari[1],* and P.Wijekoon[1]

[1]*Department of Statistics and Computer Science, Faculty of Science, University of Peradeniya, Sri Lanka.*

ABSTRACT

The species diversity monitoring of butterflies in Sri Lanka is considered in this study under certain environmental factors. Species richness, and Shannon and Simpson's diversity indices were calculated to understand the variation of the distributions of butterfly species. Maximum and minimum diversity and richness were observed from Rathnapura and Puththalama districts in Sri Lanka, respectively. Based on the Diamond's assembly rules and Probabilistic models, it was noted that most of the butterflies were randomly distributed, and there was little predictable co-occurrence between species pairs. To study the distributional patterns of butterfly species with environmental factors, five different types of regression models were fitted by considering the occurrences of each species. The results clearly indicated that the distribution of butterfly species varies from species to species according to the different environmental factors. Further, the occurrence of most of the butterfly species depends on temperature and total rain fall. Prediction of species occurrences with respect to the environmental factors can be done by using the best fitted model of each species. The methodology and results of the study can be adapted to monitor the biodiversity of a certain area.

Keywords: Species occurrence, Butterfly distribution, Species diversity, Co-occurrence analysis, Environmental factors.

1. INTRODUCTION

The environmental factors play a vital role in the distribution of living organisms. The researchers have categorized the environmental factors into two main groups as abiotic and biotic. Biotic factors are the living parts of an environment, such as plants, animals and micro-organisms. All of the non-living parts in an ecosystem are considered as Abiotic. For example, water, light, radiation, temperature, humidity, atmosphere, and soil can be included as abiotic factors. Further, abiotic factors can be divided into two groups as climate conditions and topographical conditions that control the biodiversity, which is considered as variability among living organisms from all sources. Terrestrial, marine, and other aquatic ecosystems and the ecological complexes (this includes diversity within species, between species, and of ecosystems) were included to biodiversity. A common measure of biodiversity, called species richness, is the count of species in an area. There are an estimated 10 million species on the earth, which are considered as living parts of the ecosystem. Certain environmental factors contribute to increase or decrease this vast number of species. The species diversity monitoring of invertebrates is an efficient way to identify the biodiversity of a certain area. Among invertebrates, butterflies response rapidly and sensitively to climatic and habitat changes. Therefore, butterflies are increasingly recognized as an environmental indicator of changes in biodiversity (Maes and Dyck 2001[1]; Roy et al. 2001[2]).

The study sites of this research are located on the island of Sri Lanka, one of the most biologically diverse countries in Asia. Sri Lanka is listed as one of the biodiversity hotspots among the 25 hotspots of global importance (Myer et al..2000[3]; Brookes et al.,2002[4]). The total land area of Sri Lanka 64,740 km² and is the 25th largest island of the world ("Joshua Calder's World Island Info - Largest Islands of the World". Worldislandinfo.com.[5]). The central

* E-mail address: shamalisujeewa07@gmail.com

part of the southern half of the island is mountainous with heights more than 2,500 m. There are 25 administrative districts organized into 9 provinces. Due to the position of Sri Lanka, within the tropics between 5^0 -10^0 North latitude and between 79^0 to 82^0 East longitude, it endows the year-round warm weather and it is moderated by ocean winds and considerable moisture. The average low temperature ranges from 16 °C in Nuwara Eliya in the Central Highlands to 32 °C in Trincomalee on the northeast coast. The average yearly temperature falls between the ranges from 28 to 30 °C. The monsoon winds of the Indian Ocean and Bay of Bengal are caused for rainfall pattern in Sri Lanka. The mean annual rainfall varies from under 900mm in the driest parts (southeastern and northwestern) to over 5000mm in the wettest parts (western slopes of the central highlands), (Source: Department of Meteorology, Sri Lanka[6]). The island is traditionally divided into three climatic zones as dry, intermediate and wet zone, based on the seasonal rainfall. The wet zone receives high mean annual rainfall of over 2,500 mm, from the south-west monsoons (from April to June) and wet zone does not have any pronounced dry periods. Dry zone is composite from most of the east, southeast, and northern parts of the country, which receives between 1200 and 1900 mm of rain annually. Much of the rain falls in these areas are during the period from October to January, and the rest of the year there is a very little precipitation. The Intermediate zone of Sri Lanka is the area sandwiched between the Wet and Dry zones receiving a mean annual rainfall of 1750 to 2500 mm. This covers an area of about 1.2 million hectares of the country.

The varied climate conditions and topographical variations in Sri Lanka have contributed to creating rich species diversity per unit land area, and it has the highest species density for flowering plants, amphibians, reptiles and mammals in Asia having 4000 flowering plants, 107 freshwater fishes, 59 amphibians, 174 reptiles, 435 birds, 140 species of mammals and several thousand invertebrates.

In the butterfly conservation action plan enacted in 2014 in Sri Lanka (http://mmde.gov.lk/web/images/pdf/butterfly%20conservation%20action%20plan-%202014.pdf), 245 different butterfly species are identified. They belong to six families, Papilionidae – 15 species, Pieridae – 28 species, Lycaenidae – 84 species, Riodinidae – 1 species, Nymphalidae – 68 species, and Hesperiidae – 49 species, and this includes 20 species that are endemic to the island. Among the total butterfly species in Sri Lanka, 76 are nationally threatened (IUCN Sri Lanka, 2000). The major threats to butterflies in Sri Lanka include the destruction and degradation of habitats, air pollution, over-use of pesticides, over-exploitation for commercial trade and natural factors. The Butterfly Expert Group (established under the Ministry of Environment and Renewable Energy) has been selected provincial butterflies based on endemism, readily seen and being charismatic (Figure A1 in Appendix A). Most of the butterfly species in Sri Lanka are distributed island-wide, but differ in their relative abundance related to climatic zones. Although their populations vary according to the season, the distribution of population is somewhat stable throughout the year in Wet zone. Further, it was noted that butterflies usually migrate from Dry zone towards the Intermediate and Wet zones.

According to Samarasinghe et.al (1996)[7] and Gunathilake (2005)[8], butterfly distribution depends on the rainfall, temperature and vegetation environment factors in Sri Lanka. E.M.C.P. Edirisinghe ("Analysis of distribution of butterfly species in Sri Lanka", M.Sc. Project report, Post graduate institute of Science, University of Peradeniya, 2009, Unpublished results) has used the data collected in the National Conservation Review (NCR) conducted in 2000, and identified the effect of various environmental factors for distribution of butterfly species. This data set contains 204 plots in forests in Sri Lanka having 64 different butterfly species. In this study, climatic zones (dry, intermediate and wet), temperature and total rainfall were used as environmental factors, and multivariate techniques and logistic regression methods have been applied to identify natural grouping within species. Further, it was identified that the distribution of butterfly species in Sri Lanka is not homogeneous, and it depends on environmental factors (total rain fall, temperature and climatic zones). It was also noted that the species richness is changed according to the environmental factors.

In addition to the above environment factors, the butterfly species distribution may also depend on the wind speed and topographic conditions (area and elevation), and further, there may be a co-existence between the species pairs. Therefore, the main objectives of this study are to investigate the distribution patterns of butterfly species, examine the presence/absence of butterfly species based on environmental factors (temperature, rain fall, climatic zone, wind speed, land area and elevation), and to study the competition among butterfly species pairs when sharing the same area in Sri Lanka.

2. MATERIALS AND METHODS

2.1 Description of Data

2.1.1 <u>Butterfly species presence absence data</u>

Butterfly species presence absence data, collected from the National Conservation Review (NCR) in year 2000 using gradient-directed transect sampling within natural forests were used for this study. A total of 281 forests in Sri Lanka were considered in NCR to collect data except Northern Province. In 204 plots, it was noted 64 different butterfly species, and the presence of each butterfly species was taken within each plot. In the data cleaning process, 18 plots were eliminated based on the missing information, and 13 plots were eliminated since the border of district lies through the forest for which some forests belong to two or three districts. Then, 173 plots were selected for the analysis, and they were classified according to the districts where forests are located. After cleaning the data, it was noted that species presence/absence data of plots contains in 15 districts and seven administrative provinces in Sri Lanka (North-Central/ Uva/ Western/ Southern/ Central/ Sabaragamuwa and North-Western). Presence/absence data of each species in each district were used for this analysis.

2.1.2 <u>Environmental Data</u>

Climatic data were obtained from meteoblue meteorological service created at the University of Basel, Switzerland, in cooperation with the U.S. National Oceanic and Atmospheric Administration and the National Centers for Environmental Prediction (https://www.meteoblue.com/)[9]. The meteoblue climate diagrams are based on 30 years (Since 1985 – 2015) of hourly weather model simulations. The simulated weather data have a spatial resolution of approximately 30 km. Average value of these data was considered as the usual whether condition in each of 15 districts. Topographical data (Elevation and area of districts) was obtained from ‘DistancesFrom’ web site, and the data was collected from satellite maps (http://www.distancesfrom.com)[10]. Altogether, six environmental variables (temperature, precipitation, wind speed, climatic zone, elevation and area of district) were considered in this analysis.

2.2 Statistical Techniques

<u>2.2.1 Identifying the patterns of Butterfly species distribution</u>

Species richness, and Shannon and Simpson’s diversity indices were calculated to study the distribution patterns of species in each district of Sri Lanka. To measure the species richness D, the Menhinick's index: $D = \frac{s}{\sqrt{N}}$ was used, where s equals the number of different species represented in the sample, and N equals the total number of individual species in the sample.

Shannon index (*H*) and Simpson’s index (*D*) are defined as $H = \sum (p_i)|ln\, p_i|$ and $D = \sum \frac{n_i(n_i - 1)}{N(N-1)}$ respectively, where, p_i is the proportion of the number of individuals in the population for species “*i*”, n_i is the number of individuals in species *i* and *N* is the total number of individuals in the community. Note that *D* is a measure of dominance, as *D* increases, diversity (in the sense of evenness) decreases.

Bray-Curtis dissimilarity matrix was used to identify the similarity and dissimilarity of occurrence of butterfly species within each district. To eliminate the “zero-truncated problem” from the species data, “Beals Smoothing” transformation was used and to provide some standard level for community decomposition data, "Hellinger Transformation" was applied. Transformed data was used for cluster analysis. Ward's clustering method was applied to combine the districts into groups based on the similarities of the community composition of butterfly species. Furthermore, correspondence analysis was used to ordinate species whose presence or absence is recorded at multiple districts.

2.2.2 Finding the Structure of natural butterfly communities

To find the coexistence, community structure and assembly, and the maintenance of biodiversity, the co-occurrence analysis was used. At fundamental level, two species are positively, negatively or randomly associated with one another. In this case, the data were analyzed by using assembly rule model and probabilistic model. Assembly rule model is applied to simulate data and probabilistic model is applied to the observed presence absence data matrix.

Assembly rule model is based on *C* Score (Co-occurrence indices), and it measures the degree to which species co-occur in the data matrix. The *C* score for species *i* and *j* is calculated for each pair of species and define as follows;

$$C_{ij} = (R_i - S)(R_j - S) \tag{1}$$

where R_i and R_j are the matrix row totals for species *i* and *j*, and *S* is the number of sites in which both species occur. The *C* score is the averaged of C_{ij} over all possible pairs of species in the data matrix.

Monte Carlo "null model" simulation is used to generate 1000 random data matrices similar to the observed dataset, and these random data matrices were created by using "sim9" algorithm (Gotelli et.al (2002)[11]). Each random data matrix has the same number of sites per species and the same number of species per site as in the real data matrix. The co-occurrence index was calculated for each of these random data matrices, and then the random data matrix which has an approximately similar index with compared to the observed data matrix was selected.

To identify whether there is an association between species pairs using the selected random data matrix, the following two tail test was used.

H_0: There is no association between species pairs

Vs.

H_1: There is an association between species pairs

In probabilistic model, data randomization is not required (Veech 2013)[12]. It uses combinatorics. The original combinatorics approach of Veech (2013) can be represented by the probability mass function of the hypergeometric distribution defined below:
The probability that the two species co-occur at exactly j number of sites is given by,

$$P_j = \frac{\binom{N_1}{j} \times \binom{N - N_1}{N_2 - j}}{\binom{N}{N_2}} \tag{2}$$

where for $j = 1$ to N_i sites (or samples),

$N_1 =$ number of sites where species 1 occurs

$N_2 =$ number of sites where species 2 occurs and

$N =$ total number of sites that were surveyed (where both species could occur)

This analysis is distribution-free, and the results can be interpreted and reported as *p*-values, without reference to a statistic.

Finally, association rule mining technique of apriori algorithm was applied to identify the most frequently occurred butterfly species sets in Sri Lanka. R software package, 'arules' was used for association rule mining.

2.2.3 Relationships among environmental factors and prevalence of butterfly species

First, the non-parametric approach of Classification and Regression Tree (CART) was used for each and every species as an alternative approach to nonlinear regression. The CART model is a binary tree, and CART is further pruned by reducing the errors. Then, the accuracy of the Pruned CART is given by the following equation:

$$Accuracy = \frac{\sum (Actually\_presence = \Pr edictive\_presence)}{Compaired\_number\_of\_observations} \quad (3)$$

Further, five different types of regression models (Binary Logistic, Bayesian Logistic, Ridge, Lasso and Polynomial) were fitted to study the distributional patterns of butterfly species based on environmental variables as predictor variables, and species presence/absence data as a binary (dependent) variable. Pairwise correlation coefficients were used to determine the relationship among environmental factors, and Variance inflation factors (VIF) were used to identify the multicolinearity among the predictor variables. If there is multicollinearity among the predictor variables, remedial measures have to be used to remove the multicollinearity before fitting the models. Before fitting the models, environmental variables were standardized to overcome the different scaling problem in variables measured at different scales.

The best Binary, Bayesian and Polynomial models were fitted by applying backward elimination method and Akaike Information Criterion (AIC). To validate the model assumptions, four diagnostic plots (Residual vs fitted plot, Normal Q-Q plot, Scale-location plot and Residual vs leverage plot) were used. Further, the best Ridge and Lasso models were identified using ten-fold-cross-validation method. Then, all five models were compared by using Receiver Operating Characteristic (ROC) Curves, and the best fitted model that describes the probability of occurrence of each species was selected.

3. RESULTS AND DISCUSSION

As indicated in Section 2, the data set contains the presence/absence data of 64 butterfly species for 15 districts, and six environmental variables, i.e. temperature (C^0), precipitation (mm), wind speed (kmh^{-1}), climatic zone, elevation (m) and area (km^2), related to each district.

3.1 Distributional patterns of butterfly species

Table 1 presents the species richness, Shannon and Simpson's diversity indices for a given district. According to the results, the maximum and the minimum number of butterfly species were observed in Rathnapura and Puththalama Districts, respectively. This finding is also tally with the Shannon and Simpson's diversity indices.

Table 1: Species richness and diversity indices of each district

District	Species Richness	Shannon Index	Simpson's Index
Puththalama	**7**	**1.945910**	**0.8571429**
Badulla	11	2.397895	0.9090909
Kurunegala	13	2.564949	0.9230769
Nuwara-Eliya	13	2.564949	0.9230769
Galle	14	2.639057	0.9285714
Kegalle	16	2.772589	0.9375000
Hambanthota	17	2.833213	0.9411765
Kandy	17	2.833213	0.9411765
Polonnaruwa	20	2.995732	0.9500000
Kaluthara	21	3.044522	0.9523810
Mathara	22	3.091042	0.9545455
Mathale	27	3.295837	0.9629630
Monaragala	30	3.401197	0.9666667
Anuradapura	32	3.465736	0.9687500
Rathnapuraya	**38**	**3.637586**	**0.9736842**

As described in section 2, Bray-Curtis dissimilarity matrix was used to identify the similarity and dissimilarity of occurrence of butterfly species within each district, and “Beals Smoothing” and “Hellinger” transformations were applied to transform species presence/absence data. Transformed data were used for Ward’s clustering method to identify different groups of districts. Figure. 1 shows the dendrogram for Species composition in each district based on Ward’s method. According to this figure, administrative districts were grouped into four different clusters.

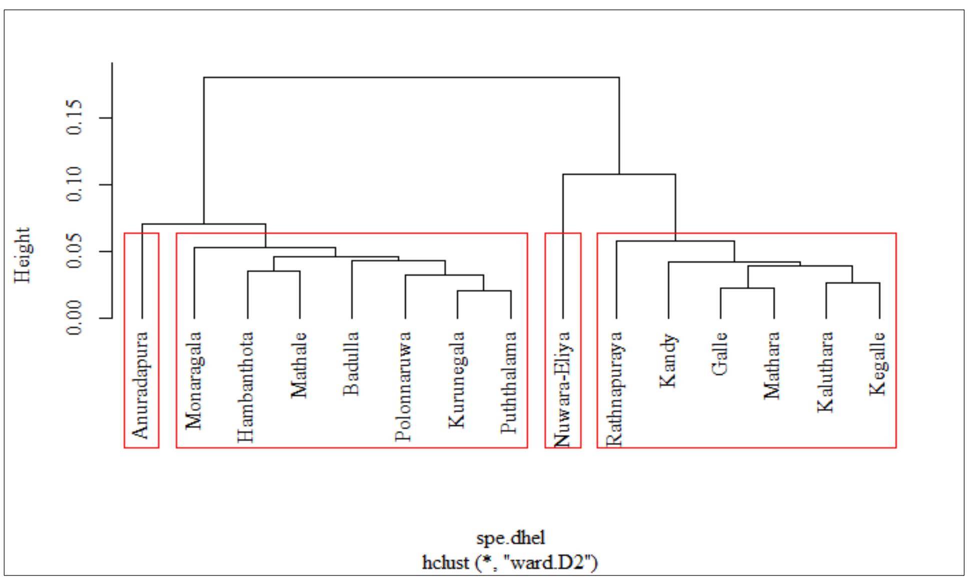

Figure 1: Cluster dendrogram for species composition in each district based on ward’s method

Further, combining both environmental data and species presence/absence data (After applying "Beals Smoothing" and "Hellinger" transformations) were used to identify the similarity and dissimilarity of occurrence of butterfly species within each district, Figure 2 shows the dendrogram for combined data in each district based on ward's method.

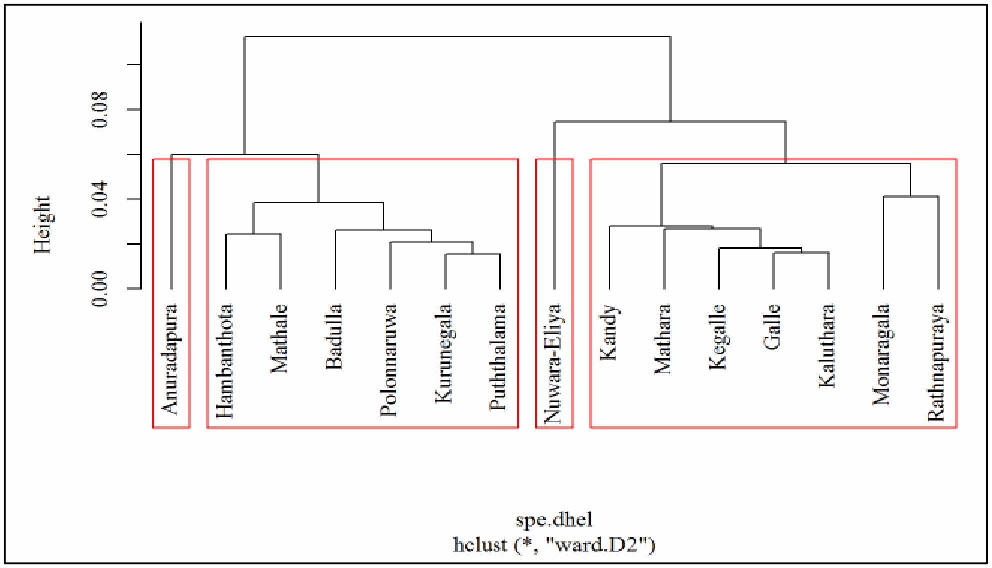

Figure 2: Cluster dendrogram for combined data in each district based on ward's method

When comparing Figure. 1 and Figure. 2 of each cluster dendograms, it is clear that the same grouping is present in four clusters even after adding environmental data for species composition data. This indicates that the districts which have approximately similar weather conditions are clustered together, and it similarly affects to the species presence/ absence data.
To understand the above clustering results further, the correspondence analysis was applied for both transformed species presence/absence data and combined data. Ordinate plots were drawn to identify the different groups of districts.Figure 3 and Figure 4 show ordinate plots based on transformed species presence/absence data, and combined data having both environmental data and species presence/absence data in each district, respectively. DCA1 and DCA2 represent the first two Detrended Correspondence Analysis axes, respectively. Ordinate plots confirm the results obtained by the dendrograms, and it further indicates that the presence of butterfly species behaves according to the weather conditions.

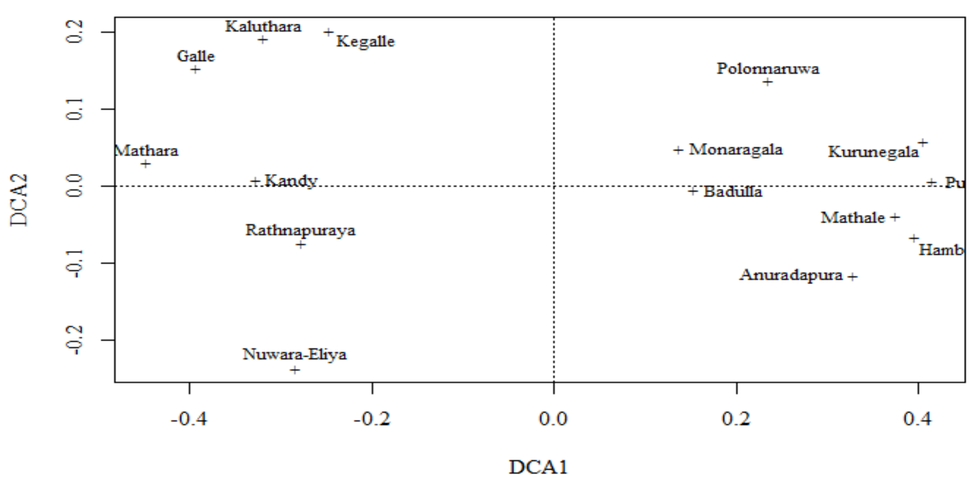

Figure 3: Ordinate plot based on transformed species presence/absence data

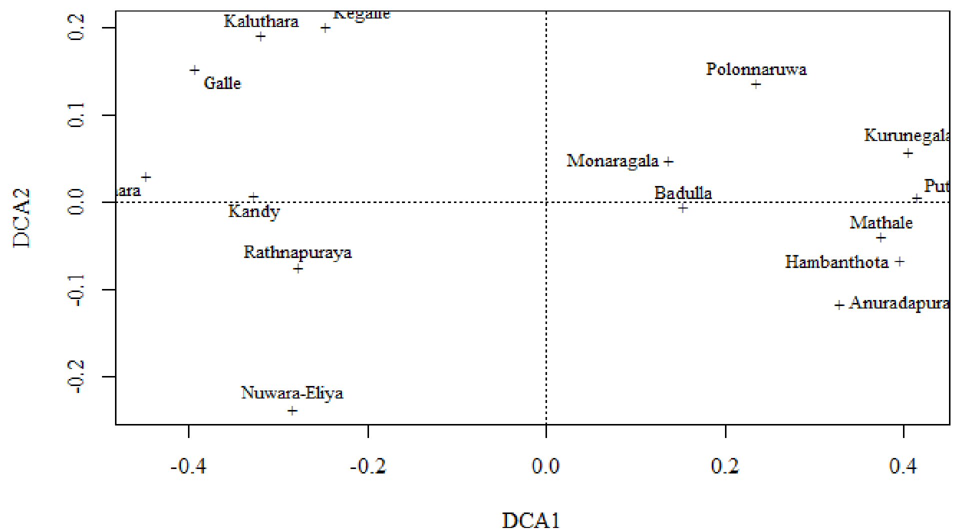

Figure 4: Ordinate plot of combining both environmental data and transformed species presence/absence data

3.2 Structure of natural butterfly communities

The two methods, assembly rule model and probabilistic model, described in section 2.2.2 were used to understand whether there exists any co-occurrence between butterfly species.

3.2.1 <u>Assembly Rule model using Simulation method</u>

The following results were obtained by using assembly rule model for species presence/absence data and testing the respective hypotheses as stated in Section 2.2.2. Figure 5 illustrates the simulated (left panel, blue) and the observed (right panel, red) presence/absence data matrix of butterfly species, and these figures are graphical representations of randomness of species presence/absence. Here, data are portrayed as a grid with colored cells (species presences) and empty cells (species absences). These two matrices have approximately equal distributions, and the plots indicate that the most of the species pairs are randomly distributed.

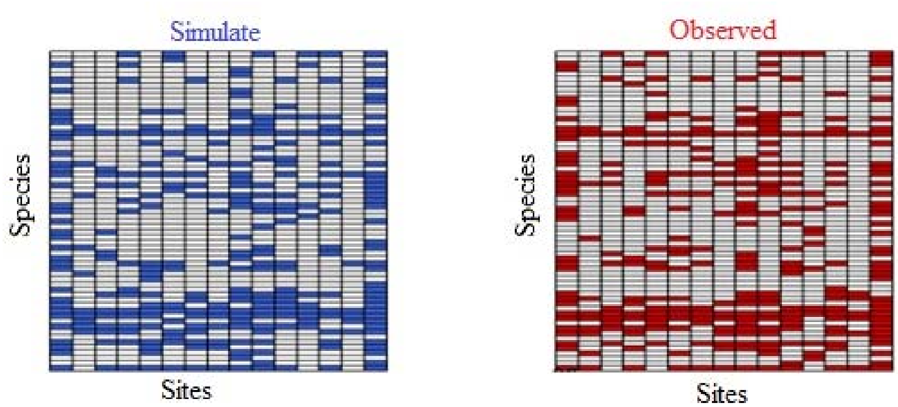

Figure 5: Selected Simulated Matrix and Original Data Matrix

Table 2 shows the inferential results related for checking the co-occurrence between butterfly species. According to Table 2, the observed C_score index of 3.8907 and the mean simulation index of 3.9068 are approximately similar, and that indicates the observed distribution and the simulated random distribution are the same. Also, the standardized effect size of -0.4739 indicates the standardized difference between original data matrix and the simulated data matrix. The null hypothesis, i.e. there is no association between species pairs is not rejected at 5% significance level since both lower-tail (*P*=0.324) and upper-tail (*P*=0.681) p-values are greater than 0.05. Further, the observed index falls within 95% confidence interval, which indicates that there is enough evidence to say that the species pairs are randomly distributed at 5% significance level.

Table 2: Summary statistics of Assembly rule model

95% CI (1-tail)		95% CI (2-tail)		Lower-tail P-value	Upper-tail P-value
Lower	**Upper**	**Lower**	**Upper**		
3.8546	3.9683	3.8410	3.9798	0.324	0.681

Observed Index	Mean of Simulated Index	Variance of Simulated Index	Standardized Effect Size (SES)
3.8907	3.9068	0.001146	-0.47391

According to the above results, the butterfly species are mostly randomly associated and there isn’t such a large competition to their co-existence. However, to understand these co-occurrence patterns further, the probabilistic model was applied.

3.2.2 <u>Results based on probabilistic model</u>

Figure 6 was drawn based on the results of the probabilistic model, and it produces a visualization of all of the pairwise combinations of species and their co-occurrence signs (positive or negative). The plot trims out any species that do not have any significant negative or positive associations and orders the remaining species starting from those with the most negative interactions to those with the most positive interactions.

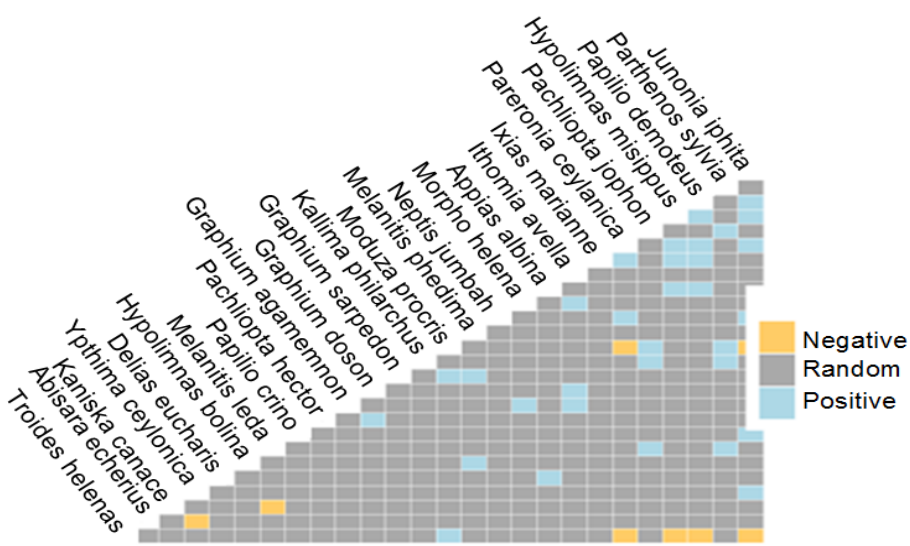

Figure 6: Graphical Representation of Species Co-occurrence Matrix

According to the results of this method, 1255 species pairs were eliminated out of 2016 species pairs since a threshold value was set in the algorithm of probabilistic model (refer R package "co-occur"). Any species pairs that are expected to share at least 1 site will be filtered in this elimination process, and finally 761 species pairs were in the data set to apply the co-occurrence classification.

Table 3 presents the classification results of the probabilistic model and it shows that among 761 species pairs only 43 is unclassifiable, and most of the classifiable species pairs have 'truly random' associations, since the random component of the model is 678. Percentage of non-random species pairs is 5.3%. Also, the significant non-random associations were mostly positive (32 positive compared to 8 negative).

Table 3: Summary statistics of probabilistic model

Species	Sites	Positive	Negative	Random	Unclassifiable	Non-random (%)
64.0	15.0	32.0	8.0	678.0	43.0	5.3

Table 5 contains a list of 40 significantly co-occurred species pairs based on the above results. Table 4 gives the descriptions of variables used in Table 5.

For a given two species in a dataset, the probl ≤ 0.05 (or probg ≥ 0.05) suggests that the corresponding two species are negatively (positively) associated. Therefore eight species combinations which are in bold in Table 5 are negatively associated. This indicates that when the probability of occurrence of one species is high the other species is low. Also remaining 32 species combinations in Table 5 are positively associated, which implies that the probability of occurrence of both species vary in the same direction. According to the results based on the probabilistic model, it is clear that most of the butterfly species combinations in the selected data set show a random co-occurrence, and there is no such large competition for co-existence among butterfly species.

Table 4: Definitions of column names of table 5

Field name	Field definition
obsco	Observed number of sites having both species
probco	Probability that both species occur at a site
expco	Expected number of sites having both species
probl	Probability that the two species would co-occur at a frequency less than the observed number of co-occurrence sites if the two species were distributed randomly (independently) of one another
probg	Probability of co-occurrence at a frequency greater than the observed frequency
sp1	If species names were specified in the community data matrix this field will contain the supplied name of species 1 in the pairwise comparison
sp2	The supplied name of species 2 in the pairwise comparison

Table 5: Significantly co-occurred species combinations

obs co	probco	expc o	probl	probg	sp1	sp2
3	**0.320**	**4.8**	**0.04396**	**1.00000**	***Abisara echerius***	***Pareronia ceylanica***
4	0.124	1.9	1.00000	0.02564	*Appias albina*	*Delias eucharis*
4	0.107	1.6	1.00000	0.01099	*Appias albina*	*misippus*
4	0.107	1.6	1.00000	0.01099	*Appias albina*	*Papilio demoteus*

3	0.080	1.2	1.00000	0.04396	*Pachliopta jophon*	*Kallima philarchus*
4	0.133	2.0	0.99800	0.04695	*Pachliopta jophon*	*Moduza procris*
6	0.240	3.6	1.00000	0.01678	*Pachliopta jophon*	*Melanitis leda*
6	0.160	2.4	1.00000	0.00020	*Pachliopta jophon*	*Parthenos sylvia*
8	0.391	5.9	1.00000	0.02564	*Coryllus avellana*	*Graphium agamemnon*
8	0.356	5.3	1.00000	0.00699	*Coryllus avellana*	*Graphium doson*
8	0.391	5.9	1.00000	0.02564	*Coryllus avellana*	*Morpho helena*
10	0.538	8.1	0.99927	0.03297	*Graphium agamemnon*	*Morpho helena*
10	0.533	8.0	1.00000	0.02198	*Graphium doson*	*Pachliopta hector*
5	0.200	3.0	1.00000	0.04196	*Graphium sarpedon*	*Moduza procris*
7	0.280	4.2	1.00000	0.00559	*Graphium sarpedon*	*Melanitis phedima*
6	0.218	3.3	0.99984	0.00886	*Hypolimnas bolina*	*Melanitis phedima*
3	0.080	1.2	1.00000	0.04396	*Hypolimnas misippus*	*Ixias marianne*
6	0.160	2.4	1.00000	0.00020	*Hypolimnas misippus*	*Papilio demoteus*
1	**0.213**	**3.2**	**0.03497**	**0.99860**	***Hypolimnas misippus***	***Troides helenas***
6	0.240	3.6	1.00000	0.01678	*Hypolimnas misippus*	*Junonia iphita*
5	0.160	2.4	0.99980	0.01099	*Hypolimnas misippus*	*Pareronia ceylanica*
3	0.080	1.2	1.00000	0.04396	*Kallima philarchus*	*Parthenos sylvia*
5	0.178	2.7	1.00000	0.01865	*Moduza procris*	*Troides helenas*
4	0.133	2.0	0.99800	0.04695	*Moduza procris*	*Parthenos sylvia*
1	**0.200**	**3.0**	**0.04695**	**0.99800**	*Moduza procris*	***Junonia iphita***
0	**0.133**	**2.0**	**0.04196**	**1.00000**	*Moduza procris*	***Pareronia ceylanica***
3	0.080	1.2	1.00000	0.04396	*Ixias marianne*	*Papilio demoteus*
3	0.080	1.2	1.00000	0.04396	*Ixias marianne*	*Pareronia ceylanica*
6	0.240	3.6	1.00000	0.01678	*Melanitis leda*	*Parthenos sylvia*
1	**0.200**	**3.0**	**0.04695**	**0.99800**	*Melanitis leda*	***Kaniska canace***
5	0.200	3.0	1.00000	0.04196	*Neptis jumbah*	*Junonia iphita*
4	0.133	2.0	0.99800	0.04695	*Neptis jumbah*	*Pareronia ceylanica*
8	0.400	6.0	0.99800	0.04695	*Papilio crino*	*Junonia iphita*
1	**0.213**	**3.2**	**0.03497**	**0.99860**	***Papilio demoteus***	***Troides helenas***
6	0.240	3.6	1.00000	0.01678	*Papilio demoteus*	*Junonia iphita*
5	0.160	2.4	0.99980	0.01099	*Papilio demoteus*	*Pareronia ceylanica*
2	**0.320**	**4.8**	**0.00559**	**1.00000**	***Troides helenas***	***Junonia iphita***
1	**0.213**	**3.2**	**0.03497**	**0.99860**	***Troides helenas***	***Pareronia ceylanica***
6	0.240	3.6	1.00000	0.01678	*Junonia iphita*	*Pareronia ceylanica*
9	0.480	7.2	1.00000	0.04396	*Junonia iphita*	*Pareronia ceylanica*

3.2.3 Association rule mining technique results

Association rule mining technique was used with two parameters of minimum support count (=8) and minimum confidence (= 90%) to discover the frequently occurring species set. The minimum support count indicates that out of all 15 districts, any butterfly species occur in 8 districts or more were considered as frequently occurring species. According to Table 6, eight butterfly species (*Euploea core* -SP17, *Ithomia avella* -SP23, *Graphium agamemnon* -SP25, *Graphium doson* -SP27, *Papilio polymnestor* -SP53, *Papilio polytes* -SP54, *Pachliopta hector* -SP56, *Morpho helena* -SP59) were identified as the frequently occurring butterfly species in each district, and there is a strong association among these eight species.

Table 6: Summary of strong association rules

Occurred Species Set	Dependent Species	support	confidence
{SP17,SP23,SP25,SP27,SP54,SP56,SP59}	=> {SP53}	0.5333333	1.0000000
{SP23,SP25,SP27,SP53,SP54,SP56,SP59}	=> {SP17}	0.5333333	1.0000000
{SP17,SP23,SP25,SP27,SP53,SP56,SP59}	=> {SP54}	0.5333333	1.0000000
{SP17,SP23,SP25,SP27,SP53,SP54,SP59}	=> {SP56}	0.5333333	1.0000000
{SP17,SP23,SP27,SP53,SP54,SP56,SP59}	=> {SP25}	0.5333333	1.0000000
{SP17,SP23,SP25,SP27,SP53,SP54,SP56}	=> {SP59}	0.5333333	1.0000000
{SP17,SP23,SP25,SP53,SP54,SP56,SP59}	=> {SP27}	0.5333333	1.0000000

3.3 Environmental factors that affect for prevalence of butterfly species

Before fitting the models, pruned CART was generated for every species as the Non-parametric method to find the environmental factors that affect for prevalence of butterfly species.

Figure 7 illustrates the Pruned CART Tree for the species *Hypolimnas bolina*, and the first value which is inside the shapes indicate the presence (1) or absence (0) of that relevant species and the second value represent the percentage of presence or absence of *Hypolimnas bolina* species. Here species presence/absence was considered as the dependent variable and environmental variables were considered as the independent variables. According to the pruned CART tree, three variables (zone, elevation and total rain fall (TRF)) are identified as the best predictive variables for *Hypolimnas bolina* species, and it has a 20% chance of not living in the intermediate zone. Also, when elevation is less than 12m from the sea level, *Hypolimnas bolina* species has a 20% chance of not living in other zones (wet and dry zones). If elevation is greater than 12m and total rain fall (TRF) is less than 534mm, then there is a 7% chance of not living in wet and dry zones. If TRF is greater than 534mm, then there is a 53% of chance of living of *Hypolimnas bolina* species in wet and dry zones. After getting those pruned values, accuracy of this CART was checked by using actual presence data of *Hypolimnas bolina* species in butterfly conservation action plan 2014 (APPENDIX B, Table B1 and B2). Accuracy of this pruned CART was calculated by using equation 3 stated in section 2, which is 0.556. This value indicates that the prediction accuracy of this pruned CART is only 55.6%. Therefore it is important to fit logistic regression models to each species to get more accurate results.

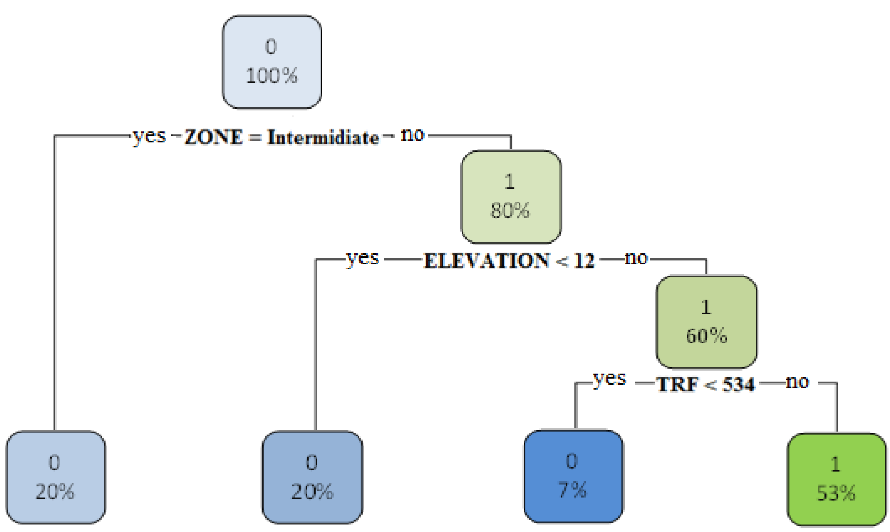

Figure 7: Pruned CART Tree for *Hypolimnas bolina* Species

Before fitting Binary and Bayesian logistic regression, Ridge and Lasso regression models, and 2nd order polynomial model, it is necessary to understand the association between environmental variables. Pearson correlation coefficients and VIF values were used to identify pairwise correlations and multicollinearity, respectively.

According to the Pearson correlation coefficients wind speed and precipitation are strongly negative correlated (r = -0.81) and elevation and average temperature are also strongly negative correlated (r = -0.88). Precipitation and average temperature are fairly negative correlated (r = -0.67), and elevation and precipitation are fairly positive correlated (r = 0.69). VIF values of variables of climatic zone, average temperature, precipitation, wind speed, elevation and area of districts are 1.75, 7.44, 4.99, 3.86, 7.46 and 1.57 respectively. Since all VIF values are less than 10, there is no multicollinearity among these variables.

Five type of models, Binary and Bayesian logistic regression, Ridge and Lasso regression models, a 2nd order polynomial model were fitted for each species. A less predictive ability was observed when fitting the Binary and Bayesian logistic regression models of some of the species. In Binary, Bayesian and polynomial logistic regression analysis, backward elimination method and AIC values were used to select the best model and four diagnostic plots (Residual vs fitted plot, Normal Q-Q plot, Scale-location plot and Residual vs leverage plot) of residuals were used to validate the model assumptions. Ridge and Lasso regression models were fitted to reduce the multicollinearity problem, if exists, between the variables, and 2nd order polynomial model was fitted to each species to catch the non-linear behavior of the models. For *Neptis jumbah*, it was noted that the polynomial regression model satisfied the model assumptions rather than the binary and Bayesian regression model. Finally, ROC values of all five models were obtained, and these values and ROC curves for *Neptis jumbah* are given in Table 7 and Figure 8, respectively.

Table 7 shows the five type of best models for *Neptis jumbah*. Here, Y represents the Species presence absence, and X_1, X_2, X_3, X_4, X_5, X_6 represent environmental variables Zones, Average Temperature, Total Rain Fall, Wind Speed, Elevation and Area, respectively. According to Table 7 and Figure 8, the highest ROC value for *Neptis jumbah* is for the 2nd order polynomial model. Therefore polynomial model is the best fitted model to predict the occurrence of *Neptis jumbah*. Similarly, the best fitted model of each butterfly species was identified. According to the results, the presence/absence of most of the butterflies can be modeled using Binary logistic model and Polynomial model. The best model for *Papilio crino*, *Delias eucharis*, *Ithomia avella*, *Hypolimnas bolina*, *Morpho helena* and *Neptis jumbah* butterfly species was only the polynomial model. Predicted probabilities were calculated from the best model of each species to determine the occurrence of each species in each district. The models with best predictive ability for all the species were included in APPENDIX B (Table B3).

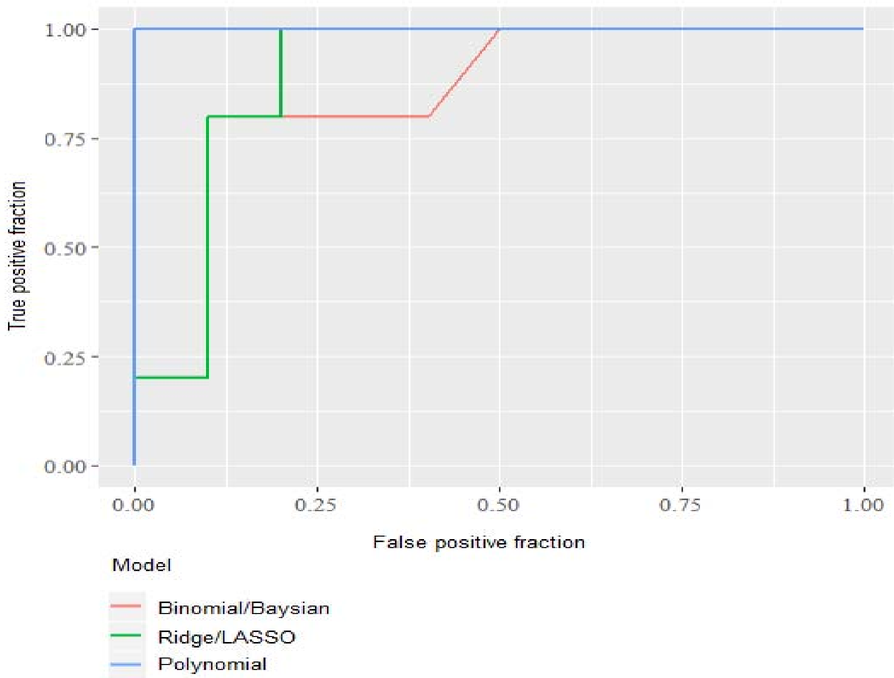

Figure 8: ROC curves of five models for *Neptis jumbah* species

Table 7: Five types of best fitted models for *Neptis jumbah* species

Regression Model	Best Model	ROC Value
Binary Logistic	Y= -2.245 -3.641X_3	0.85
Bayesian Logistic	Y= -1.3888 -2.1756 X_3	0.85
Ridge	Y= -0.75094064 + 0.17745520X_1 +0.09259067X_2 -0.17943172X_3 +0.16526079X_4 -0.04207682X_5 +0.20869408X_6	0.90
Lasso Logistic	Y= -0.89115527 +0.35767264X_1 -0.67311653X_3 +0.07192323X_4 +0.47576260X_6	0.90
Polynomial	Y= -21.34 +931.05X_2 -285.36X_2^2 +67.32X_4 +108.81X_4^2 +912.13X_5 -232.85X_5^2	**1.00**

3.4 Butterfly species analyzer

Based on the analysis, a web application called BUSA (Butterfly species analyzer) was created by using shiny package in R (Link: https://shamali.shinyapps.io/shiny-app/) which acts as a statistical software tool. It has a user friendly interface, and can perform the statistical analysis as a menu driven software package. Distributions of species, environmental factors that affect for prevalence of species in the ecosystem, and structure of natural butterfly communities with the competition among butterfly species can be mainly analyzed by using this application. Most of the statistical tools that we use to analyze the species data are included in this web application. Although this web tool mainly aims for analyzing occurrence of butterfly species, it can also be used for any other species occurrence data set in the same data format. In future work this will be improved as a tool for analyzing any other species occurrence data set in the same format.

4. CONCLUSION

According to the results, it was revealed that the distribution of butterfly species is not homogenous in different administrative districts of Sri Lanka. Four different groups of districts were identified having similar environment factors, which show similar butterfly species presence/absence. Distribution of butterfly species varies from species to species according to the different types of environmental factors. There were fewer species combinations which are non-randomly (negatively or positively) distributed, and most of the butterflies are randomly associated. Hence, there is no such a large competition to their co-existence or to share the same area. There was a strong association among eight butterfly species (*Euploea core*, *Ithomia avella*, *Graphium agamemnon*, *Graphium doson*, *Papilio polymnestor*, *Papilio polytes*, *Pachliopta hector*, *Morpho helena*) which are frequently occurred as a group. Presence of most of the butterfly species depend on average temperature and total rain fall. Further, it was noted that there is high butterfly species diversity in Rathnapura, Anuradapura and Monaragala districts. However, the occurrence of butterfly species in Puththalama, Badulla, Kurunegala and Nuwara-Eliya districts is less. This study further indicates that it is easy to launch projects to conserve butterflies in Sri Lanka by identifying the distributional pattern of butterfly species according to the environmental conditions.

ACKNOWLEDGEMENT

We would like to thank Prof. N. Gunathilake, Department of Botany, University of Peradeniya and National Conservation Review (NCR), Sri Lanka for providing Species presence/absence dataset.

COMPETING INTERESTS

Authors have declared that no competing interests exits. The authors report no conflicts of interest. The authors alone are responsible for the content and writing of this article.

AUTHORS' CONTRIBUTIONS

Author 1* designed the study, performed the statistical analysis, wrote the protocol, and wrote the first draft of the manuscript. Author 1 supervised the project and checked the manuscript. All authors read and approved the final manuscript."

REFERENCES

[1]. Maes, D., & Van Dyck, H. (2001). Butterfly diversity loss in Flanders (north Belgium): Europe's worst case scenario?. *Biological conservation*, *99*(3), 263-276. (https://doi.org/10.1016/S0006-3207(00)00182-8)

[2]. Roy, D. B., Rothery, P., Moss, D., Pollard, E., & Thomas, J. A. (2001). Butterfly numbers and weather: predicting historical trends in abundance and the future effects of climate change. *Journal of Animal Ecology*, *70*(2), 201-217. (**https://doi.org/10.1111/j.1365-2656.2001.00480.x**)

[3]. Myers, N., Mittermeier, R. A., Mittermeier, C. G., Da Fonseca, G. A., & Kent, J. (2000). Biodiversity hotspots for conservation priorities. *Nature*, *403*(6772), 853.

[4]. Brooks, T. M., Mittermeier, R. A., Mittermeier, C. G., Da Fonseca, G. A., Rylands, A. B., Konstant, W. R., ... & Hilton-Taylor, C. (2002). Habitat loss and extinction in the hotspots of

biodiversity. *Conservation biology*, *16*(4), 909-923. (**https://doi.org/10.1046/j.1523-1739.2002.00530.x**)

[5]. Calder, J. (2009). *100 Largest Islands of the World*. Worldislandinfo.com. Retrieved from http://www.worldislandinfo.com/LARGESTV1.html.

[6]. *Department of meteorology, Sri Lanka*. Meteo.gov.lk. Retrieved from http://www.meteo.gov.lk/index.php?lang=en.

[7]. Samarasinghe, M. D. P., Paranagama, P., & Veediyabandara, S. (1996). Survey of the butterfly fauna of Udawalawa National Park. In *Proceedings of International Forestry and Environment Symposium*. (https://doi.org/10.31357/fesympo.v0i0.1223)

[8]. Gunatilleke, I. A. U. N., Gunatilleke, C. V. S., & Dilhan, M. A. A. B. (2005). Plant biogeography and conservation of the southwestern hill forests of Sri Lanka. *The Raffles Bulletin of Zoology*, *12*(1), 9-22.

[9]. *Weather Ann Arbor- meteoblue website,* Switzerland. Accessed 20 Feb 2017.
Available: https://www.meteoblue.com/

[10]. Distancesfrom.com. *Distance Calculator* | *Distance From* | *Find distance between cities*. Accessed 22 Feb 2017.
Available: http://www.distancesfrom.com.

[11]. Gotelli, N. J., & McCabe, D. J. (2002). Species co-occurrence: a meta-analysis of JM Diamond's assembly rules model. *Ecology*, *83*(8), 2091-2096. (**https://doi.org/10.1890/0012-9658(2002)083[2091:SCOAMA]2.0.CO;2**)

[12]. Veech, J. A. (2013). A probabilistic model for analysing species co-occurrence. *Global Ecology and Biogeography*, *22*(2), 252-260. (**https://doi.org/10.1111/j.1466-8238.2012.00789.x**)

APPENDIX A

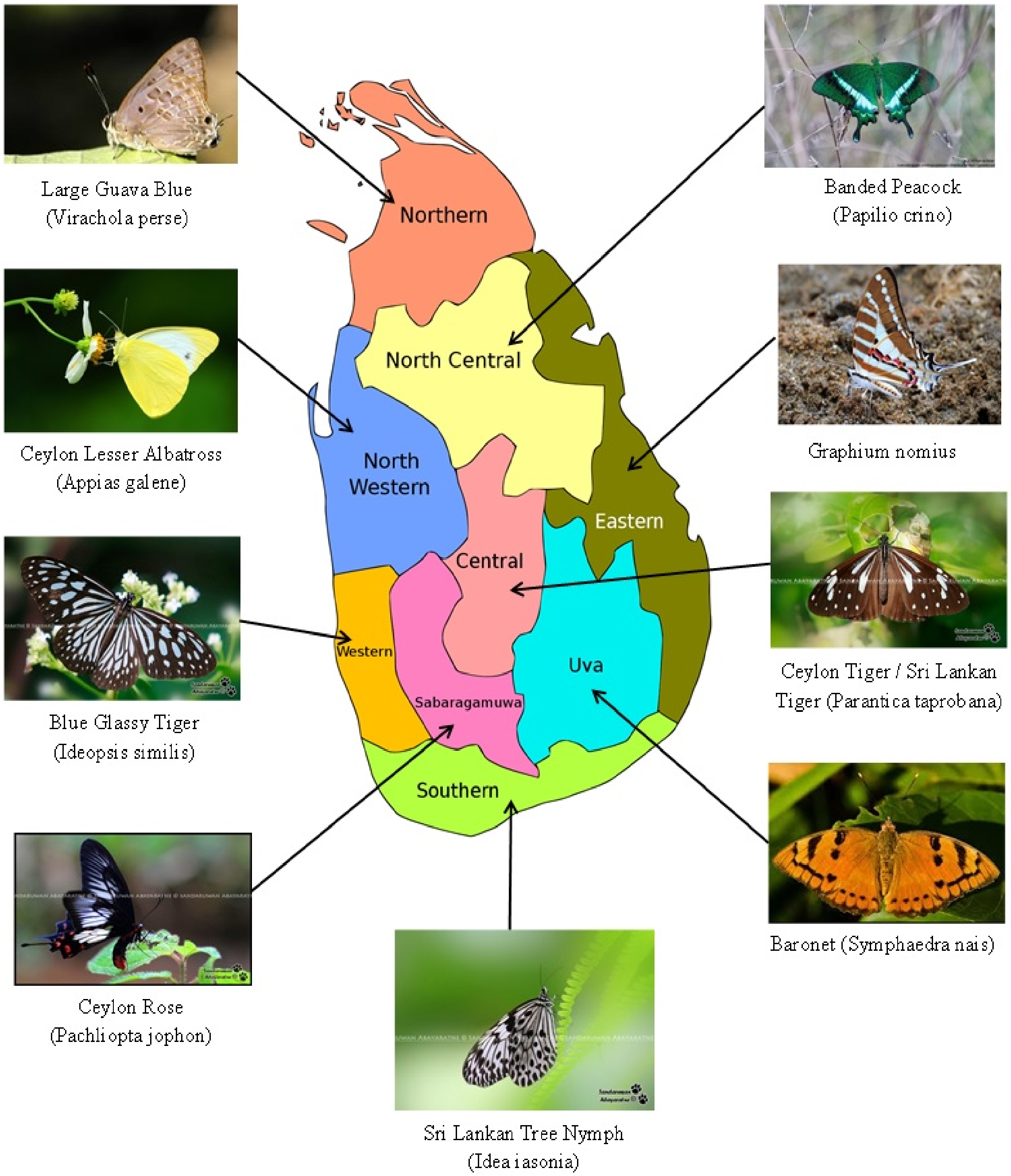

Figure A1: Provincial Butterflies of Sri Lanka

APPENDIX B

Table B1: Presence data of *Hypolimnas bolina* species in butterfly conservation action plan 2014

District	Zone	TRF	Elevation	ATM	WS	Area	*Hypolimnas bolina*
Anurahapura	Dry	1284.6	91.52	27.37	15.25	7179	1
Badulla	Intermediate	2062.82	661.49	23.47	7.08	2861	1
Galle	Wet	2427.58	8.31	27.37	8.42	1652	1
Hambanthota	Dry	1049.6	13.57	29.11	15.33	2609	1
Kurunegala	Intermediate	2197.18	123.05	24	12.83	4816	1
Nuwara-Eliya	Wet	1905.3	1893.45	16.52	7.75	1741	1
Polonnaruwa	Dry	1822.38	50.99	28.56	15.75	3293	1
Puththalama	Dry	1143.76	5.75	27.92	17.08	3072	1
Rathnapuraya	Wet	3749.2	42.07	27.72	7.92	3275	1

Table B2: Predicted presence/absence of *Hypolimnas bolina* species by using Pruned CART

District	Presence or Absence of ***Hypolimnas bolina*** Species	
	In Actual data set	**By using Pruned CART**
Anurahapura	1	1
Badulla	1	0
Galle	1	0
Hambanthota	1	1
Kurunegala	1	0
Nuwara-Eliya	1	1
Polonnaruwa	1	1
Puththalama	1	0
Rathnapuraya	1	1

By equation (3), $\text{Accuracy of the Pruned CART} = \frac{5}{9} = 0.556$

657 **Table B3: Best fitted models of each butterfly species and Districts of presence**

Species Name	Family	Best Model(s) (ROC value = 1.00)	Districts of Present
Phocides polybius	Hesperiidae	Binary Logistic, Polynomial	Hambanthota
Tagiades japetus	Hesperiidae	Binary Logistic, Polynomial	Rathnapura
Graphium sarpedon	Papilionidae	Binary Logistic, Polynomial	Anuradapura, Kaluthara, Kandy, Kegalle, Mathale, Mathara, Monaragala, Nuwara-Eliya, Rathnapura
Zerynthia polyxena	Papilionidae	Binary Logistic, Polynomial	Anuradapura, Rathnapura
Papilio polytes	Papilionidae	Binary Logistic, Bayesian Logistic, Polynomial	All districts except Nuwara-Eliya
Graphium nomius	Papilionidae	Binary Logistic, Polynomial	Anuradapura, Mathale
Pachliopta jophon	Papilionidae	Binary Logistic, Polynomial	Galle,Kaluthara,Kegalle, Mathara, Monaragala, Rathnapura
Troides helenas	Papilionidae	Binary Logistic, Polynomial	Galle,Kaluthara, Kandy, Kegalle, Mathara, Nuwara-Eliya,Polonnaruwa, Rathapura
Pachliopta hector	Papilionidae	Binary Logistic, Polynomial	All districts except Kegalla, Nuwara-Eliya, Puththalama
Graphium doson	Papilionidae	Binary Logistic, Polynomial	All districts except Hambanthota,Kandy,Kegalle,Nuwara-Eliya,Puththalama
Papilio demoteus	Papilionidae	Binary Logistic, Polynomial	Anuradapura,Hambanthota,Kurunegala,Mathale,Monaragala,Rathnapura
Papilio crino	Papilionidae	Polynomial	All districts except Galle,Kegalle,Kurunegala,Mathara,Nuwara-Eliya
Papilio clytia	Papilionidae	Binary Logistic, Polynomial	Kauthara,Kandy,Mathale,Mathara
Graphium antiphates	Papilionidae	Binary Logistic, Polynomial	Anuradapura,Kaluthara
Graphium agamemnon	Papilionidae	Binary Logistic, Lasso, Polynomial	All districts except Badulla,Hambanthota,Nuwara-Eliya,Puththalama
Cepora nadina	Pieridae	Binary Logistic, Ridge, Polynomial	Monaragala,Polonnaruwa,Rathnapura
Ixias marianne	Pieridae	Binary Logistic, Polynomial	Anuradapura,Hambanthota,Monaragala
Appias lyncida	Pieridae	Binary Logistic, Bayesian Logistic, Polynomial	Anuradapura
Delias eucharis	Pieridae	Polynomial	Anuradapura, Badulla, Polonnaruwa,Rathnapura
Pareronia ceylanica	Pieridae	Binary Logistic, Lasso, Polynomial	Anuradapura,Hambanthota, Kurunegala,Mathale, Monaragala,Polonnaruwa
Eurema blanda	Pieridae	Binary Logistic, Polynomial	Anuradapura,Mathale
Appias albina	Pieridae	Binary Logistic, Polynomial	Anuradapura, Hambanthota, Mathale, Rathnapura
Arhopala amantes	Lycaenidae	Binary Logistic, Polynomial	Mathale

Cheritra freja	Lycaenidae	Binary Logistic, Polynomial	Monaragala
Lampides lacteata	Lycaenidae	Binary,Bayesian Logistic, Polynomial	Anuradapura
Abisara echerius	Riodinidae	Binary Logistic, Polynomial	Galla,Kaluthara, Mathale, Mathara, Nuwara-Eliya, Rathnapura
Polyura athamas	Nymphalidae	Binary Logistic, Polynomial	Mathara, Rathnapura
Ithomia avella	Nymphalidae	Polynomial	Anuradapura,Galle,Kurunegala,Mathale,Mathara,Monaragala,Rathnapura
Hypolimnas bolina	Nymphalidae	Polynomial	Anuradapura,Hambanthota,Kandy,Mathara,Monaragala,Nuwara-Eliya,Rathnapura
Kaniska canace	Nymphalidae	Binary,Bayesian Logistic, Polynomial	Badulla,Kandy,Mathale,Nuwara-Eliya,Rathnapura
Vanessa cardul	Nymphalidae	Binary,Bayesian Logistic, Polynomial	Nuwara-Eliya
Ypthima ceylonica	Nymphalidae	Binary,Bayesian Logistic, Polynomial	All districts except Galle,Kaluthara,Nuwara-Eliya
Lethe drypetis	Nymphalidae	Binary,Bayesian Logistic, Polynomial	Nuwara-Eliya
Lethe dynsate	Nymphalidae	Binary Logistic, Lasso, Polynomial	Nuwara-Eliya, Rathnapura
Vindula erota	Nymphalidae	Binary Logistic, Polynomial	Mathara, Rathnapura
Cupha erymanthis	Nymphalidae	Binary Logistic, Polynomial	Mathale, Monaragala
Morpho helena	Nymphalidae	Polynomial	All districts except Badulla,Hambanthota,Kandy,Puththalama
Pantoporia hordonia	Nymphalidae	Binary Logistic, Polynomial	Polonnaruwa
Junonia iphita	Nymphalidae	All five model	Anuradapura,Badulla,Hambanthota,Kurunegala,Mathale,Monaragala,Polonnaruwa, Puththalama,Rathnapura
Neptis jumbah	Nymphalidae	Polynomial	Anuradapura,Mathale,Monaragala,Polonnaruwa,Puththalama
Melanitis leda	Nymphalidae	Binary Logistic, Polynomial	Anuradapura,Galla,Hambanthota,Kaluthara,Kegalla,Mathara,Monaragala,Polonnaruwa,Rathnapura
Libythea lepita	Nymphalidae	Binary Logistic, Bayesian, Polynomial	Nuwara-Eliya
Mycalesis mineus	Nymphalidae	Binary Logistic, Polynomial	Rthnapura
Hypolimnas misippus	Nymphalidae	Binary Logistic, Polynomial	Anuradapura, Hambanthota, Kurunegala, Mathale, Monaragala, Rathnapura
Euthalia nais	Nymphalidae	Binary Logistic, Polynomial	Monaragala,Rathnapura
Cethosia nietneri	Nymphalidae	Binary Logistic, Polynomial	Kaluthara, Polonnaruwa
Rohana parisatis	Nymphalidae	Binary Logistic, Polynomial	Rathnapura
Euploea phaenareta	Nymphalidae	Binary Logistic, Polynomial	Hambanthota,Kaluthara,Kandy,Mathale, Rathnapura

Melanitis phedima	Nymphalidae	Binary Logistic, Polynomial	Anuradapura,Kaluthara,Kandy,Mathara,Monaragala,Nuwara-Eliya, Rathnapura
Kallima philarchus	Nymphalidae	Binary Logistic, Polynomial	Mathara, Monaragala, Rathnapura
Moduza procris	Nymphalidae	Binary Logistic	Kaluthara,Kandy,Kegalle,Mathara,Rathnapura
Lethe rohria	Nymphalidae	Binary Logistic, Polynomial	Badulla, Kandy
Charaxes solon	Nymphalidae	Binary Logistic, Bayesian, Polynomial	Anuradapura
Euploea sylvester	Nymphalidae	Binary Logistic, Polynomial	Kegalle
Parthenos sylvia	Nymphalidae	Binary Logistic, Polynomial	Galle,Kegalle,Mathara, Monaragala,Rathnapura
Cirrochroa thais	Nymphalidae	Binary Logistic, Polynomial	Anuradapura, Kaluthara, Kegalle,Mathale, Monaragala, Polonnaruwa, Rathnapura
Loxura atymmus	Lycaenidae	Binary Logistic, Ridge, Polynomial	Galle, Kandy, Monaragala, Rathnapura
Geitoneura klugil	Nymphalidae	Binary Logistic, Bayesian, Polynomial	Anuradapura

658