Project Title: **“Species Distribution Modeling”**

Subject: **Data Mining and Data Warehousing**

Group Members: **Khwaja Zain – ur – Rehman (EP 20101020)**

Ali Ilyas (EP 20101007)

Hamza Hafeez (EP 20101016)

Muhammad Areeb (EP 20101024)

Degree Name: **MCS (Masters in Computer Science)**

Year: **2021**

Department Name: UBIT (Umaer Basha Institute of Technology)

University Name: **UOK (University of Karachi)**

Table of Contents

“Species Distribution Modeling”

1. Abstract:-

The vacuity of detailed environmental data, together with affordable and important computers, has fueled a rapid-fire increase in prophetic modeling of species environmental conditions and geographic distributions. For some species, detailed presence/ absence circumstance data are available, allowing the use of a variety of standard statistical ways. Still, absence data aren't available for utmost species. In this paper, we introduce the use of the maximum entropy system (Maxent) for modeling species geographic distributions with presence-only data. Maxent is a general- purpose machine literacy system with a simple and precise fine expression, and it has a number of aspects that make it well- suited for species distribution modeling. In order to probe the efficacity of the system, then we perform a international-scale case study using two Neotropical mammals a tableland species of idleness, Bradypus variegatus, and a small montane murid rodent, Microryzomys minutus. We compared Maxent prognostications with those of a generally used presence-only modeling system, the Inheritable Algorithm for Rule- Set Vaticination (GARP). We made prognostications on 10 arbitrary subsets of the circumstance records for both species, and also used the remaining points for testing. Both algorithms handed reasonable estimates of the species’ range, far superior to the shadowed figure maps available in field attendants. All models were significantly better than arbitrary in both binomial tests of elision and receiver operating characteristic (ROC) analyses. The area under the ROC wind (AUC) was nearly always advanced for Maxent, indicating better demarcation of suitable versus infelicitous areas for the species. The Maxent modeling approach can be used in its present form for numerous operations with presence-only datasets, and graces farther exploration and development.

2. Introduction:-

Modeling species' geographic distributions is an important problem in conservation biology. In this example we model the geographic distribution of two south american mammals given past observations and 14 environmental variables. Since we have only positive examples (there are no unsuccessful observations), we cast this problem as a density estimation problem and use the *OneClassSVM* as our modeling tool. The dataset is provided by Phillips et. al. (2005). *Species Distribution Modeling.* New York: Phillips. If available, the example uses *basemap* to plot the coast lines and national boundaries of South America.

The two species are:

- "*Bradypus variegatus*" , the Brown-throated Sloth.
- "*Microryzomys minutus*" , also known as the Forest Small Rice Rat, a rodent that lives in Peru, Colombia, Ecuador, Peru, and Venezuela.

1. Bradypus variegatus:-
    - Justification:-

      Bradypus variegatus is listed as Least Concern in view of its wide distribution including a large part of the Amazon forest, presumed large population, its occurrence in a number of protected areas, and because it is unlikely to be declining fast enough to qualify for listing in a threatened category.

    - Geographic Range Information:-

      Bradypus variegatus ranges from Honduras in the north, through southern Central America. In South America, it ranges from Colombia into western and southern Venezuela, and south into Ecuador, eastern Peru and Bolivia, into Brazil and northern Argentina (where it's now considered to be annihilated). Its distribution overlaps withB. torquatus in the central part of the Atlantic timber (Hirsch and Chiarello 2012). In Brazil, the species presently occurs in forested areas of the Amazon, Atlantic timber, and conceivably in the contact zones between these biomes and Cerrado. There are literal records ofB. variegatus in the Caatinga biome (Moraes-

      Barros unpublished data 2010). There are no verified records forB. variegatus in the Pantanal biome of Brazil, but the species might do in the contact zones between this biome and the Amazon timber to the north. Fresh field studies are necessary in order to duly define the current species distribution in the Cerrado, Caatinga and Pantanal.

- Threats Information:-

  It appears that there are no major threats to *B. variegatus* at the global level. Nevertheless, some subpopulations, especially in Colombia and the Atlantic Forest in Brazil, are declining due to deforestation leading to severe habitat degradation and fragmentation. The lowest levels of genetic diversity of the species were observed in the Atlantic Forest; they were similar to the levels observed in the Critically Endangered Bradypus pygmaeus (Silva 2013). Furthermore, they are hunted by local indigenous communities. Wild-caught individuals, especially offspring, are sold as pets to tourists in Colombia (Moreno and Plese 2006). This illegal trade is increasing and represents a cause of concern due to its impact on the wild populations. Mortality on roads also occurs.

- Use and Trade Information:-

  In Brazil, especially in the northeastern region and in the Amazon, and in Colombia the common sloth is hunted and sold in public markets as food, medicine, and as a pet species. In several touristic sites, *B. variegatus* is used by locals to entertain visitors.

2. Microryzomys minutes:-
   - Justification:-

     This species is listed as Least Concern in view of its wide distribution, presumed large population, and because it is unlikely to be declining at nearly the rate required to qualify for listing in a threatened category.

   - Geographic Range Information:-

     This species occurs from north Venezuela, through Colombia, Ecuador and Peru, to west central Bolivia (Musser and Carletob 2005). It has an altitudinal range of 1,500 to 4,000 m (Soriano et al. 1999, Tirira, in prep.).

   - Population Information:-

     This mouse is common throughout the range.

   - Habitat and Ecology Information:-

     This species inhabits in lower montane, subalpine forest (Musser and Carleton 2005) and paramo (B. Rivas pers. comm.). This species is terrestrial and arboreal, it is found in high mountain habitats, frequently near rocks, especially in cloud forest. Presumably it feeds on seeds and vegetation (Lord 1999).

- Threats Information:-

  Major threats are deforestation in some parts of its range (R. Anderson pers. comm.).

- Niche-based models from presence-only data:-

  We're interested in contriving a model of a species’environmental conditions from a set of circumstance points, together with a set of environmental variables that describe some of the factors that likely influence the felicity of the terrain for the species (Brown and Lomolino, 1998; Root, 1988). Each circumstance position is simply a latitude – longitude brace denoting a point where the species has been observed; similar georeferenced circumstance records frequently decide from samples in natural history galleries and herbaria (Ponder etal., 2001; Stockwell and Peterson, 2002a). The environmental variables in Civilians format all pertain to the same geographic area, the study area, which has been partitioned into a grid of pixels. The task of a modeling system is to prognosticate environmental felicity for the species as a function of the given environmental variables. A niche- grounded model represents an approximation of a species’ecological niche in the examined environmental confines. A species’ abecedarian niche consists of the set of all conditions that allow for its long-term survival, whereas its realized niche is that subset of the abecedarian niche that it actually occupies (Hutchinson, 1957). The species’ realized niche may be lower than its abecedarian niche, due to mortal influence, biotic relations (e.g.,inter-specific competition, predation), or geographic walls that have hindered disbandment and colonization; similar factors may help the species from inhabiting (or indeed encountering) conditions encompassing its full ecological eventuality (Pulliam, 2000; Anderson and Mart ´ ınez-Meyer, 2004). We assume then that circumstance points are drawn from source niche, rather than sink niche, which may contain a given species without having the conditions necessary to maintain the population without immigration; this supposition is less realistic with largely vagile taxa (Pulliam, 2000). By description, also, environmental conditions at the circumstance points constitute samples from the realized niche. A nichebased model therefore represents an approximation of the species’ realized niche, in the study area and environmental confines beingconsidered.However, we can not hope for any modeling algorithm to characterize the species’ full abecedarian niche the necessary information is simply not present in the circumstance points, If the realized niche and abecedarian niche don't completely coincide. This problem is likely aggravated when circumstance records are drawn from too small a geographic area. In a larger study region, still, spatial variation exists in community composition (and, hence, in the performing biotic relations) as well as in the environmental conditions available to the species. Thus, given sufficient slice trouble, modeling in a study region with a larger geographic extent is likely to increase the bit of the abecedarian niche represented by the sample of circumstance points (Peterson and Holt, 2003), and is preferable.

- Existing approaches for presence-only modeling:-

Many methods have been used for presence-only modeling of species distributions, and we only attempt here to give a broad overview of existing methods. Some methods use only presences to derive a model. BIOCLIM (Busby, 1986; Nix, 1986) predicts suitable conditions in a “bioclimatic envelope”, consisting of a rectilinear region in environmental space representing the range (or some percentage thereof) of observed presence values in each environmental dimension. Similarly, DOMAIN (Carpenter et al., 1993) uses a similarity metric, where a predicted suitability index is given by computing the minimum distance in environmental space to any presence record.

As a first step in the evaluation of Maxent, we chose to compare it with GARP, as the latter has recently seen extensive use in presence-only studies (Anderson, 2003; Joseph and Stockwell, 2002; Peterson and Kluza, 2003; Peterson and Robins, 2003; Peterson and Shaw, 2003 and references therein). While further studies are needed comparing Maxent with other widely used methods that have been applied to presence-only datasets, such studies are beyond the scope of this paper.

- Maxent:-

Maxent is a general- purpose system for making prognostications or consequences from deficient information. Its origins lie in statistical mechanics (Jaynes, 1957), and it remains an active area of exploration with an Annual Conference, Maximum Entropy and Bayesian Styles, that explores operations in different areas similar as astronomy, portfolio optimization, image reconstruction, statistical drugs and signal processing. We introduce it then as a general approach for presenceonly modeling of species distributions, suitable for all being operations involving presence-only datasets. The idea of Maxent is to estimate a target probability distribution by chancing the probability distribution of maximum entropy (i.e., that's utmost spread out, or closest to livery), subject to a set of constraints that represent our deficient information about the target distribution. The information available about the target distribution frequently presents itself as a set of real- valued variables, called “ features”, and the constraints are that the anticipated value of each point should match its empirical normal (average value for a set of sample points taken from the target distribution). When Maxent is applied to presence-only species distribution modeling, the pixels of the study area make up the space on which the Maxent probability distribution is defined, pixels with given species circumstance records constitute the sample points, and the features are climatic variables, elevation, soil order, foliage type or other environmental variables, and functions thereof.

3. Methods:-

- Maxent details:-

When approximating an unknown probability distribution, the question arises, what is the best approximation? E.T. Jaynes gave a general answer to this question: the best approach is to ensure that the approximation satisfies any constraints on the unknown distribution that we are aware of, and that subject to those constraints, the distribution should have maximum entropy (Jaynes, 1957). This is known as the maximum-entropy principle. For our purposes, the unknown probability distribution, which we denote π, is over a finite set X, (which we will later interpret as the set of pixels in the study area). We refer to the individual elements of X as points. The distribution π assigns a non-negative probability $\pi(x)$ to each point x, and these probabilities sum to 1. Our approximation of π is also a probability distribution, and we denote it $\pi\hat{}$. The entropy of $\pi\hat{}$ is defined as

$H(\pi\hat{}) = - x \in X \ \pi\hat{}(x) \ \ln \pi\hat{}(x)$

- A machine learning perspective:-

The maximum entropy principle has seen recent interest in the machine learning community, with a major donation being the development of effective algorithms for chancing the Maxent distribution (see Berger etal., 1996 for an accessible preface and Ratnaparkhi, 1998 for a variety of operations and a favorable comparison with decision trees). The approach consists of standardizing the constraints on the unknown probability distribution π in the following way. We assume that we've a set of known realvalued functions f1,., fn on X, known as " features" (which for our operation will be environmental variables or functions thereof). We assume further that the information we know about π is characterized by the prospects (pars) of the features under π. Then, each point fj assigns a real value fj (x) to each point x in X. The anticipation of the point fj under π is defined as x ∈ X π (x) fj (x) and denoted by π (fj). In general, for any probability distribution p and function f, we use the memorandum p (f) to denote the anticipation of f under p.

- A Maxent implementation for modeling species distributions:-

In order to make the Maxent method available for modeling species geographic distributions, we implemented an efficient algorithm together with a choice of feature types that are well suited to the task. Our implementation uses a sequential-update algorithm (Dud´ık et al., 2004) that iteratively picks a weight λj and adjusts it so as to minimize the resulting regularized log loss. The algorithm is deterministic, and is guaranteed to converge to the Maxent probability distribution. The algorithm stops when a user-specified number of iterations has been performed, or when the change in log loss in an iteration falls below a user-specified value (convergence), whichever happens first. As described in Section 2.1, Maxent assigns a nonnegative probability to each pixel in the study area. Because these probabilities must sum to 1, each probability is typically extremely small. Although these "raw" probabilities are an optional output, by default our software presents the probability distribution in another form that is easier to use and interpret, namely a "cumulative" representation. The value assigned to a pixel is the sum of the probabilities of that pixel and all other pixels with equal or lower probability, multiplied by 100 to give a percentage. The cumulative representation can be interpreted as follows: if we resample pixels according to the modeled Maxent probability distribution, then t% of the resampled pixels will have cumulative value of t or less. Thus, if the Maxent distribution $\hat{\pi}$ is a close approximation of the probability distribution π that represents reality, the binary model obtained by setting a threshold of t will have approximately t% omission of test localities and minimum predicted area among all such models (cf. the “minimal predicted area” evaluation measure of Engler et al. (2004)). This provides a theoretical foundation that aids in the selection of a threshold when a binary prediction is required.

- GARP:-

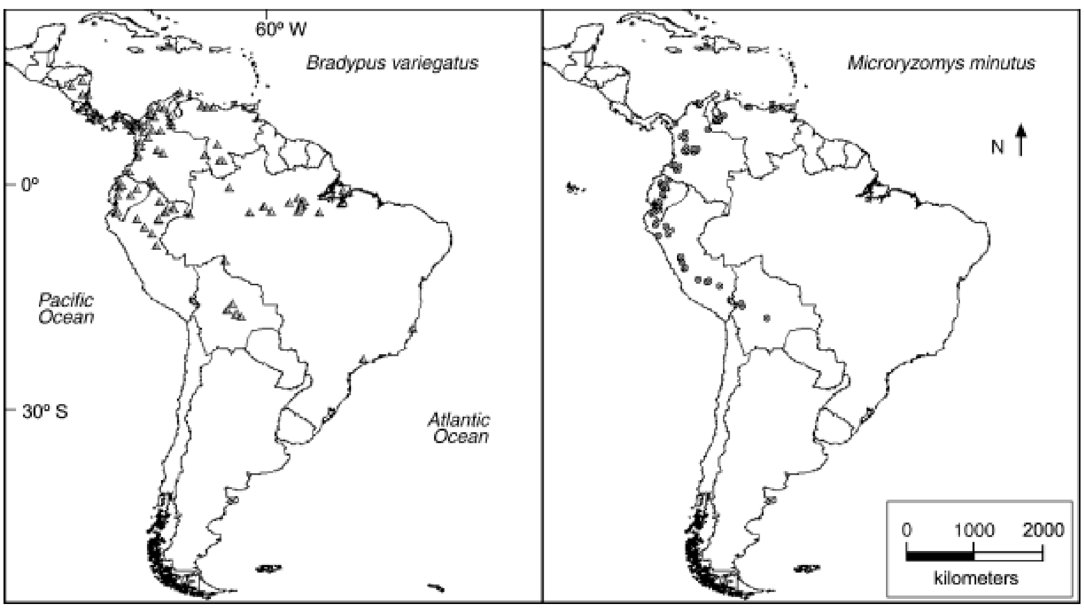
In its simplest form, GARP seeks a collection of rules that together produce a double vaticination. Positive rules prognosticate suitable conditions for pixels satisfying some set of environmental conditions; also, negative rules prognosticate infelicitous conditions. Rules are favored in the algorithm according to their significance (compared with arbitrary vaticination) grounded on a sample of 1250 presence pixels and 1250 background pixels, tried with relief. Some pixels may admit no vaticination, if no rule in the rule- set applies to them, and some may bear resolution of clashing prognostications. A inheritable algorithm is used to search heuristically for a good rule- set (Stockwell and Noble, 1992). There's considerable arbitrary variability in GARP prognostications, so we enforced the best-subset model selection procedure as follows, analogous to Peterson and Shaw (2003) and following the general recommendations of Anderson etal. (2003). First, we generated 100 double models, with pixels that didn't entered a vaticination interpreted as prognosticated absence, using GARP interpretation1.1.3 with dereliction values for its parameters (0.01 confluence limit, 1000 maximum duplications, and allowing the use of infinitesimal, range, negated range and logit rules). We also excluded all models with further than 5 natural elision (of training points). Still, they also constituted the stylish subset (this happed 4 out of 44 times, yielding stylish subsets with 5, If at most 10 models remained. In all other cases, we determined the median value of the prognosticated area of the remaining models, and named the 10 models whose awaited area was closest to the standard. Eventually, we combined the best-subset models to make a compound GARP vaticination, in which the value of a pixel was equal to the number of stylish-subset models in which the pixel was prognosticated present (0 – 10).

- Data sources:-

The brown-throated three-toed sloth Bradypus variegatus (Xenarthra: Bradypodidae) is a large arboreal mammal (3–6 kg) that is widely distributed in the Neotropics from Honduras to northern Argentina. It is found primarily in lowland areas but also ranges up to middle elevations. It has been documented in regions of deciduous forest, evergreen rainforest and montane forest, but is absent from xeric areas and non-forested regions (Anderson and Handley, 2001). Three other species are known in the genus.B. pygmaeusis endemic to Isla Escudo on the Caribbean coast of Panama, and two species have geographic distributions restricted to South America: B. tridactylus in the Guianan region and B. torquatus in the Atlantic forests of Brazil.

- Environmental variables:-

We examine the species' potential distributions in the Neotropics from southeastern Mexico to Argentina.



(23.55◦ N – 56.05◦ S, 94.8◦ W – 34.2◦ W), including the Caribbean from Cuba southward. The environmental variables fall into three categories: climate, elevation and potential vegetation. All variables are recorded at a pixel size of 0.05◦ by 0.05◦, yielding a 1212 × 1592 grid, with 648,658 pixels containing data for all variables. The climatic variables derive from data provided by the Intergovernmental Panel on Climate Change (IPCC; New et al., 1999). The original variables have a resolution of 0.5◦ by 0.5◦, and were produced using thin-plate spline interpolation based on readings taken at weather stations around the world from 1961 to 1990. They describe mean monthly values of various variables, which we processed to convert to ascii raster grid format, as required by GARP and Maxent. From these monthly data, we also created annual variables by averaging or taking the minimum or maximum as appropriate.

- Model building:-

We made 10 random partitions rather than a single one in order to assess the average behavior of the algorithms, and to allow for statistical testing of observed differences in performance (via Wilcoxon signed-rank tests). In addition, the algorithms were also run on the full set of occurrence localities, taking advantage of all available data to provide best estimates of the species' potential distributions for visual interpretation. The algorithms (Maxent and GARP) were run with two suites of environmental variables: first with only climatic and elevational data, and then with those variables plus potential vegetation. The reasons for treating potential vegetation separately are three-fold: (1) climatic and elevational data are readily available for the whole world (whereas potential vegetation is not), and we wished to determine whether good models can be created using uniformly available data. (2) The potential vegetation coverage is rather subjective, whereas the others are objectively produced from measured data. (3) Potential vegetation is the only categorical variable, and the potential existed for the algorithms to respond differently to categorical versus continuous data.

- Model evaluation:-

The first step in assessing the models produced by the two algorithms was to corroborate that both performed significantly better than arbitrary. For this purpose, we first used a (threshold-dependent) binomial test grounded on elision and prognosticated area. Still, it doesn't allow for comparisons between algorithms, as the significance of the test is largely dependent on prognosticated area. Indeed, comparison of the algorithms is made delicate by the fact that neither gives a double vaticination. Hence, we also used two relative statistical tests that employ veritably different means to overcome this complication. First, we employed a new thresholddependent system of model evaluation, which we name the " evened prognosticated area" test, whose purpose is to answer the following question at the generally used thresholds representing the axes of the GARP vaticination, how does Maxent compare? Second, we used (threshold-independent) receiver operating characteristic (ROC) analysis, which characterizes the performance of a model at all possible thresholds by a single number, the area under the wind (AUC), which may be also compared between algorithms.

4. Results:-

- Quantitative results:-

Both algorithms consistently produced predictions that were better than random. Using the simple threshold rule (Section 2.6.1), the binomial omission test was highly significant ($p < 0.001$, one-tailed) for both algorithms on all data partitions for each species (see Table 1 for details on runs with the climatic and elevational variables; results on the variable suite including potential vegetation were similar). For Maxent, the thresholds determined by the simple threshold rule ranged from 0.022 to 2.564 for B. variegatus and 0.543 to 3.822 for M. minutus. For GARP, the thresholds ranged from 1 to 7 for B. variegatus and 2 to 10 for M. minutus. In addition to statistical significance, omission rates were consistently low or zero, never exceeding 17% (Table 1). The results of the equalized predicted area test differed between the species (Tables 2 and 3). For B. variegatus, the omission rates of the two algorithms were lower for Maxent in 16 cases, equal in 15 cases, and lower for GARP in 9 cases. However, two-tailed Wilcoxon signed-rank tests did not reveal a significant difference in median omission rates for either threshold or either variable suite ($p = 0.178$ and 0.314 for thresholds of 1 and 10, respectively, with climatic and elevational variables; $p = 0.371$ and 0.155 for thresholds of 1 and 10, respectively, with addition of the potential vegetation variable). Maxent almost always had equal or lower omission than GARP for M. minutus (19 out of 20 models). The difference in median omission rates was significant at both thresholds on runs with climatic and elevational variables ($p = 0.036$ and $p = 0.014$ for thresholds of 1 and 10, respectively; two-tailed Wilcoxon signed-rank test). When the potential vegetation variable was added, the difference in median omission rates was highly significant for a threshold of 10, but not for a threshold of 1 ($p = 0.009$ and 0.345, respectively), largely because Maxent had greater omission than before on data partition 2, discussed below (Section 4.3).

Results of the threshold-dependent binomial tests of omission

Data partition	Maxent		GARP	
	Area	Omission rate	Area	Omission rate
Bradypus variegatus-1	0.51	0.03	0.41	0.11
B. variegatus-2	0.66	0	0.56	0.06
B. variegatus-3	0.80	0	0.61	0.03
B. variegatus-4	0.42	0.17	0.51	0
B. variegatus-5	0.75	0.03	0.57	0.06
B. variegatus-6	0.62	0	0.54	0
B. variegatus-7	0.59	0	0.53	0
B. variegatus-8	0.59	0.06	0.62	0
B. variegatus-9	0.69	0	0.66	0
B. variegatus-10	0.62	0.06	0.44	0.06
Average	0.626	0.034	0.545	0.031
Microryzomys minutus-1	0.03	0.11	0.06	0.15
M. minutus-2	0.04	0.11	0.06	0.15
M. minutus-3	0.03	0.11	0.07	0.15
M. minutus-4	0.04	0.04	0.08	0.04
M. minutus-5	0.03	0.04	0.06	0.15
M. minutus-6	0.04	0.15	0.06	0.11
M. minutus-7	0.05	0	0.09	0.07
M. minutus-8	0.04	0.04	0.10	0
M. minutus-9	0.03	0.07	0.10	0
M. minutus-10	0.03	0.11	0.08	0.07
Average	0.035	0.078	0.075	0.089

Area (proportion of the study area predicted) and extrinsic omission rate (proportion of the test localities falling outside the prediction) are given for each of 10 random data partitions for Maxent and GARP. For both *B. variegatus* and *M. minutus*, the binomial test was highly significant for all partitions ($p < 0.001$, one-tailed). Models were derived using the climatic and elevational variables for each random partition of occurrence records, and area and omission rates were calculated using simple threshold rules based on the training localities (see Section 2). The results for models made with the addition of the potential vegetation variable were similar but are not shown here (see Section 3). The omission rates should not be compared between algorithms, as they are strongly affected by differences in predicted area. The simple threshold rule used here for Maxent is not recommended for general use in practice; in this case, it gives too high a threshold for Maxent on *B. variegatus*-4, causing a high omission rate, and too low a threshold on *B. variegatus*-3, resulting in too much predicted area.

usually increased for both Maxent and GARP, with results significant or nearly so for both (p = 0.051 and 0.033, respectively, although performance was poorer for Maxent on data partition 2; see Section 4.3). While the differences in AUC values are very small, the changes may still be meaningful biologically. For example, the largest visual effect of adding potential vegetation for Maxent was to (correctly) exclude some non-forested areas from the prediction for B. variegatus (Section 3.2.2). However, because of the small geographic extent of those areas, the effect on AUC values was small.

Results of the equalized predicted area tests of omission for *B. variegatus* and *M. minutus* produced with Maxent and GARP using the climatic and elevational variables

Data partition	GARP threshold = 1			GARP threshold = 10		
	Area	Maxent omission	GARP omission	Area	Maxent omission	GARP omission
B. variegatus-1	0.59	0.03	0.03	0.27	0.17	0.17
B. variegatus-2	0.56	0.03	0.06	0.34	0.11	0.17
B. variegatus-3	0.61	0.06	0.03	0.33	0.09	0.17
B. variegatus-4	0.63	0.14	0	0.40	0.17	0.06
B. variegatus-5	0.67	0.03	0	0.36	0.11	0.26
B. variegatus-6	0.69	0	0	0.29	0.14	0.11
B. variegatus-7	0.74	0	0	0.31	0.03	0.14
B. variegatus-8	0.69	0	0	0.33	0.17	0.11
B. variegatus-9	0.72	0	0	0.36	0.06	0.11
B. variegatus-10	0.61	0.06	0.03	0.34	0.14	0.17
Average	0.652	0.034	0.014	0.333	0.120	0.149
M. minutus-1	0.12	0	0.07	0.06	0.04	0.15
M. minutus-2	0.10	0	0.07	0.06	0.04	0.15
M. minutus-3	0.16	0	0.04	0.07	0.07	0.15
M. minutus-4	0.17	0	0.04	0.08	0.04	0.04
M. minutus-5	0.12	0	0.07	0.06	0	0.15
M. minutus-6	0.12	0	0.04	0.06	0.07	0.11
M. minutus-7	0.16	0	0	0.09	0	0.07
M. minutus-8	0.17	0	0	0.09	0	0
M. minutus-9	0.17	0	0	0.09	0	0.04
M. minutus-10	0.18	0	0	0.08	0	0.07
Average	0.146	0	0.033	0.073	0.026	0.093

Area (proportion of the study area predicted by GARP with the indicated threshold) and extrinsic omission rate (proportion of test localities falling outside the prediction) for each algorithm are given for each random partition of occurrence records under two threshold scenarios. Thresholds were set for the extremes of the GARP predictions: any GARP model predicting presence (GARP threshold = 1) and all 10 GARP models predicting presence (GARP threshold = 10). To allow for direct comparison of omission rates between the algorithms, thresholds were then set for each Maxent model to yield a binary prediction with the same area as the corresponding GARP prediction.

The ROC curves for the two algorithms showed distinct patterns, evident in the curves for the first random data partition for each species, for models made using climatic and elevational variables (Fig. 3). In the case of M. minutus, the performance of Maxent was better across the entire spectrum: for any given omission rate, Maxent achieved that rate with a lower false positive rate (1–specificity, which is almost identical to proportional predicted area, see Section 2). The results with B. variegatus were more complex. There is a point where the ROC curves for the two algorithms intersect, corresponding to a sensitivity of 0.83 (omission rate of 0.17) and a false positive rate of 0.27. At that point, therefore, the performance of the two algorithms was the same. A small component of the higher AUC for Maxent was due to the lower omission rate it achieved to the right of that point. However, most of Maxent's higher AUC occurred to the left of that point, where many test localities fell in small areas very strongly predicted by Maxent. In contrast, GARP did not differentiate environmental quality to the left of that point, as all pixels there were predicted by all 10 of the bestsubset models. Results for other data partitions were roughly similar (not shown).

- Visual interpretation:-

The affair format differs dramatically between Maxent and GARP, so watch must be taken when making comparisons between them. Maxent produces a nonstop vaticination with values ranging from 0 to 100, whereas GARP, as used then, yields a separate compound vaticination with integer values from 0 to 10. Visual examination of the Maxent prognostications for both species indicated that a low threshold was applicable, and in general terms, pixels with prognosticated values of at least 1 may be interpreted as a reasonable approximation of the species' implicit distribution. This is in concordance with the thresholds attained in Section3.1.1, and the theoretical anticipation that the elision rate for a thresholded accretive vaticination will be roughly equal to the threshold value (see Section2.2). For GARP, visual examination suggested a advanced threshold in the range 5 – 10 was applicable for approaching the species' implicit distribution. In the ensuing sections, we interpret the models under this frame.

- Models derived from climatic and elevational variables:-

When using the full set of occurrence localities for each species, the two algorithms produced broadly similar predictions for the potential geographic distribution of B. variegatus (Fig. 4). For this species, both algorithms indicated suitable conditions throughout most of lowland Central America, wet lowland areas of northwestern South America, most of the Amazon.

Results of the equalized predicted area tests of omission for *B. variegatus* and *M. minutus* produced with Maxent and GARP using the climatic, elevational and potential vegetation variables

Data partition	GARP threshold = 1			GARP threshold = 10		
	Area	Maxent omission	GARP omission	Area	Maxent omission	GARP omission
B. variegatus-1	0.57	0.03	0.03	0.28	0.20	0.23
B. variegatus-2	0.58	0	0.06	0.29	0.11	0.29
B. variegatus-3	0.67	0	0.03	0.33	0.14	0.11
B. variegatus-4	0.67	0	0	0.42	0.06	0.11
B. variegatus-5	0.67	0.03	0.03	0.36	0.14	0.17
B. variegatus-6	0.71	0	0	0.28	0.17	0.17
B. variegatus-7	0.74	0	0	0.33	0.06	0.20
B. variegatus-8	0.67	0	0	0.34	0.20	0.17
B. variegatus-9	0.78	0	0	0.39	0.03	0.06
B. variegatus-10	0.67	0	0	0.36	0.14	0.17
Average	0.672	0.006	0.014	0.337	0.126	0.169
M. minutus-1	0.12	0	0.04	0.06	0.04	0.15
M. minutus-2	0.11	0.11	0.04	0.06	0.15	0.19
M. minutus-3	0.13	0	0.04	0.07	0.04	0.15
M. minutus-4	0.15	0	0.04	0.08	0.04	0.04
M. minutus-5	0.12	0	0.07	0.06	0	0.15
M. minutus-6	0.14	0	0	0.05	0.04	0.11
M. minutus-7	0.16	0	0.04	0.08	0	0.07
M. minutus-8	0.16	0	0	0.08	0	0.04
M. minutus-9	0.16	0	0	0.08	0	0.07
M. minutus-10	0.17	0	0	0.07	0	0.04
Average	0.142	0.011	0.026	0.070	0.030	0.100

Area (proportion of the study area predicted by GARP with the indicated threshold) and extrinsic omission rate (proportion of test localities falling outside the prediction) for each algorithm are given for each random partition of occurrence records under two threshold scenarios. Thresholds were set for the extremes of the GARP predictions: any GARP model predicting presence (GARP threshold = 1) and all 10 GARP models predicting presence (GARP threshold = 10). To allow for direct comparison of omission rates between the algorithms, thresholds were then set for each Maxent model to yield a binary prediction with the same area as the corresponding GARP prediction.

basin, large areas of Atlantic forests in southeastern Brazil, and most large Caribbean islands in the study area. The species was generally predicted absent from high montane areas, temperate areas in southern South America, and non-forested areas of central Brazil (e.g., cerrado). The algorithms differed in their predictions for non-forested savannas in northern South America. The composite GARP model indicated the species' potential presence there, but Maxent excluded some non-forested savannas in Venezuela (llanos) and the Guianas.

- Addition of potential vegetation variable:-

  The two algorithms responded else to the addition of the implicit foliage variable (Fig. 5). The Maxent vaticination with implicit foliage forB. variegatus was generally analogous to the original one, but now indicated infelicitous conditions for the species in the champaigns of Colombia and Venezuela and in othernon-forested areas in Bolivia and Brazil. On the negative, the compound GARP vaticination with implicit foliage included was veritably analogous to the original vaticination, still indicating suitable environmental conditions for the species innon-forested areas of Colombia, Venezuela, Guyana, Brazil, Paraguay and Bolivia.

Results of threshold-independent receiver operating characteristic (ROC) analyses for *B. variegatus* and *M. minutus* produced with Maxent and GARP using the climatic and elevational variables (left) and climatic, elevational and potential vegetation variables (right)

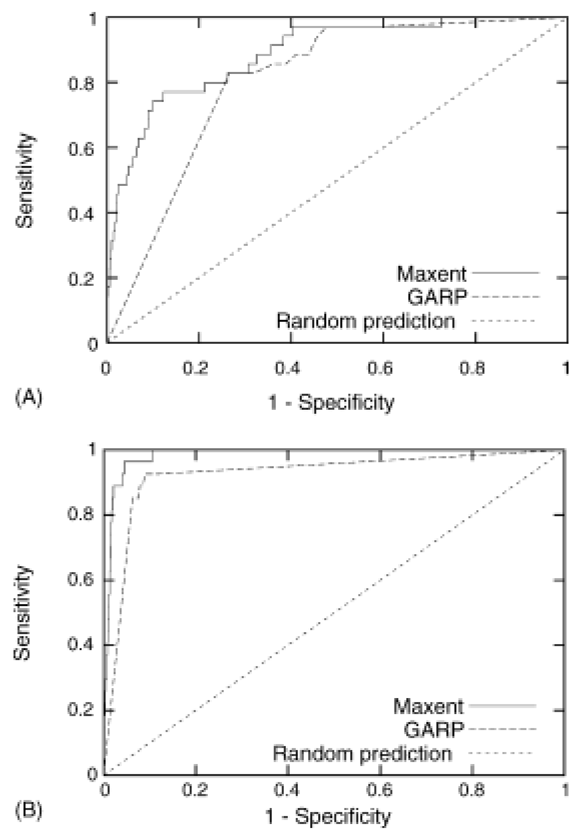
Data partition	Without potential vegetation			With potential vegetation		
	Maxent AUC	GARP AUC	*p*	Maxent AUC	GARP AUC	*p*
B. variegatus-1	0.889	0.807	<0.01	0.879	0.793	<0.01
B. variegatus-2	0.892	0.765	<0.01	0.899	0.769	<0.01
B. variegatus-3	0.872	0.779	0.01	0.887	0.790	<0.01
B. variegatus-4	0.819	0.789	0.51	0.858	0.757	<0.01
B. variegatus-5	0.868	0.740	<0.01	0.885	0.753	<0.01
B. variegatus-6	0.881	0.818	<0.01	0.868	0.812	0.03
B. variegatus-7	0.902	0.812	<0.01	0.919	0.784	<0.01
B. variegatus-8	0.839	0.807	0.34	0.829	0.786	0.13
B. variegatus-9	0.903	0.794	<0.01	0.897	0.784	<0.01
B. variegatus-10	0.866	0.779	0.01	0.879	0.769	<0.01
Average	0.873	0.789		0.880	0.780	
M. minutus-1	0.985	0.926	0.01	0.986	0.946	0.02
M. minutus-2	0.987	0.931	0.02	0.932	0.943	0.75
M. minutus-3	0.985	0.938	<0.01	0.987	0.939	<0.01
M. minutus-4	0.983	0.938	<0.01	0.984	0.941	<0.01
M. minutus-5	0.988	0.926	0.02	0.990	0.926	0.01
M. minutus-6	0.983	0.947	0.05	0.986	0.966	<0.01
M. minutus-7	0.989	0.950	<0.01	0.988	0.936	<0.01
M. minutus-8	0.988	0.954	<0.01	0.989	0.956	<0.01
M. minutus-9	0.989	0.952	<0.01	0.990	0.955	<0.01
M. minutus-10	0.985	0.955	<0.01	0.987	0.961	<0.01
Average	0.986	0.942		0.982	0.947	

For each random partition of occurrence records, the area under the ROC curve (AUC) is given for Maxent and GARP, as well as the probability of the observed difference in the AUC values between the two algorithms (under a null hypothesis that the true AUCs are equal). All AUC values for both algorithms were significantly better than a random prediction ($p < 0.0001$; individual p values not shown). AUC values are given to three decimal places to reveal small changes under addition of the potential vegetation coverage.

5. Discussion and conclusions:-

- Statistical tests:-

Both algorithms constantly performed significantly better than arbitrary, and Maxent constantly achieved better results than GARP. Thresholddependent binomial tests (Table 1) showed low elision of test points and significant prognostications for both algorithms across the board. The evened prognosticated area test generally indicated better performance for Maxent onM. minutus, but the test didn't descry a significant difference between the two algorithms forB. variegatus (Tables 2 and 3). Threshold-independent ROC analysis also showed significantly better-thanrandom performance for both algorithms. The area under the ROC wind (AUC) was significantly advanced for Maxent on nearly all data partitions for both species (Table 4). Use of the categorical implicit foliage variable (in addition to the nonstop climatic and elevational variables) generally bettered performance for both algorithms onM. minutus and for Maxent onB. variegatus, but the changes had limited statistical significance, probably due to the small quantum of data.



- Biological interpretations:-

Both algorithms produced reasonable prognostications of the implicit distribution for B. variegatus. The areas prognosticated by 5 – 10 GARP models were analogous geographically to those areas prognosticated with a value of at least 1 (out of 100) for Maxent. Although important exploration addressing the issue of operationally determining an optimal threshold remains for both algorithms, these thresholds produce good charts of the species' implicit distributions (areas of suitable environmental conditions). In particular, the models perform far superior to the shadowed figure maps available in standard field attendants, (e.g., Eisenberg and Redford, 1999; Emmons, 1997), and in digital compendiums of species ranges designed for use in conservation biology and macroecological studies (Patterson etal., 2003). Utmost strikingly, the models rightly indicate an extensive region of infelicitous environmental conditions forB. variegatus in thenon-forested cerrado of Brazil, whereas the shadowed figure charts indicate nonstop distribution for the species from Amazonian timbers to littoral Atlantic timbers. Although GARP has the capacity to consider categorical variables, the addition of the implicit foliage variable didn't amend the scarcities seen in the original compound GARP vaticination forB. variegatus. In discrepancy, Maxent successfully integrated this fresh information. This is most apparent in close-up images inFig.4.2, where GARP (inaptly) prognosticated suitable conditions for the species in thenon-forested champaigns of Colombia and Venezuela.

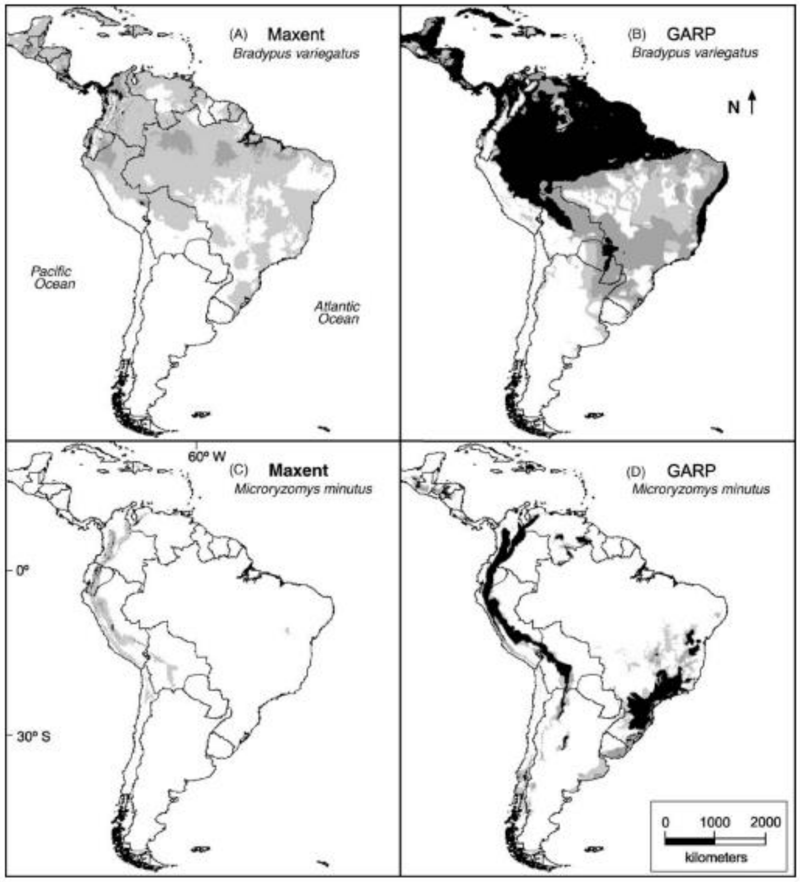


Fig. 4. Predicted potential geographic distributions for *B. variegatus* (top) and *M. minutus* (bottom) made using all occurrence records and the climatic and elevational variables. Results are given for Maxent (left) and GARP (right). Four colors are used to indicate the strength of the prediction for each individual map pixel. Maxent produces a continuous prediction with values ranging from 0 to 100; the values are depicted here using white = [0,1); pale grey = [1,34); dark grey = [34,66); black = [66,100]. The best-subsets selection procedure employed here for GARP yields a discrete prediction with integer values from 0 to 10, depicted here using white = 0; pale grey = 1–4; dark grey = 5–9; black = 10. The strength of the predictions thus cannot be compared directly. All areas with a Maxent prediction of 1 or greater likely represent suitable environmental conditions for the species; in contrast, areas with a GARP prediction of 5–10 probably indicate suitable conditions (see Section 3.2).

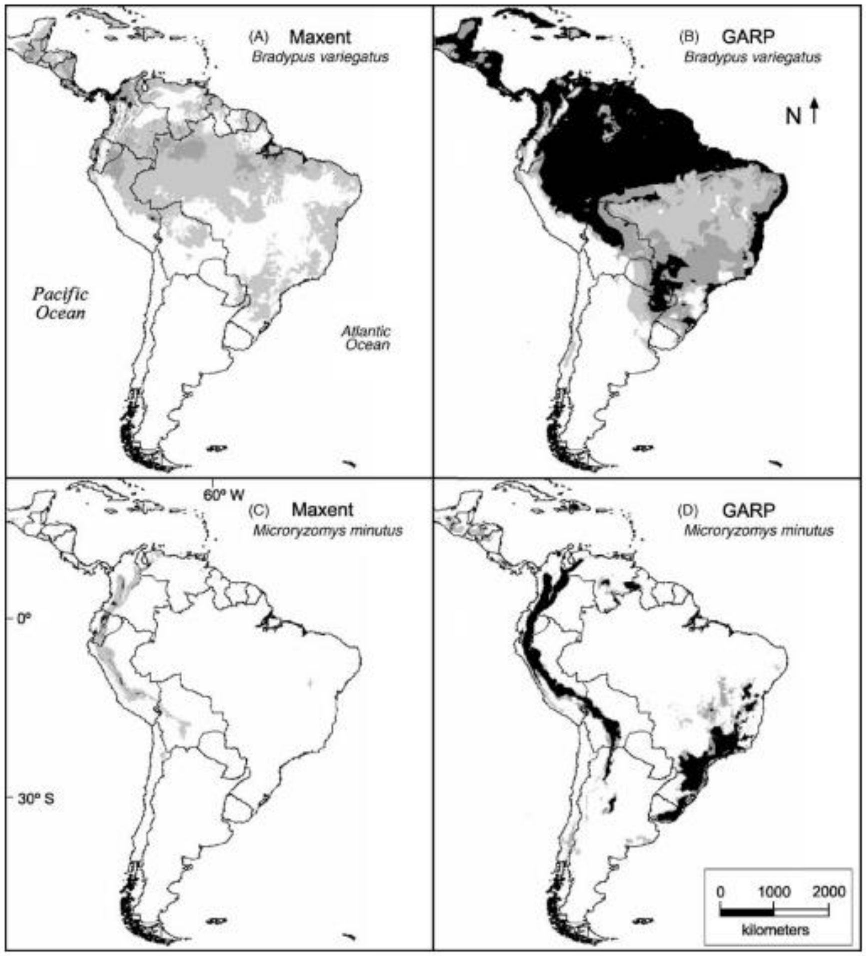


Fig. 5. Predicted potential geographic distributions for *B. variegatus* (top) and *M. minutus* (bottom) made using all occurrence records and climatic, elevational and potential vegetation variables. Results are given for Maxent (left) and GARP (right). Four colors are used to indicate the strength of the prediction for each map pixel. Maxent produces a continuous prediction with values ranging from 0 to 100; the values are depicted here using white = [0,1); pale grey = [1,34); dark grey = [34,66); black = [66,100]. The best-subsets selection procedure employed here for GARP yields a discrete composite prediction with integer values from 0 to 10, depicted here using white = 0; pale grey = 1–4; dark grey = 5–9; black = 10. The strength of the predictions thus cannot be compared directly. All areas with a Maxent prediction of 1 or greater likely represent suitable environmental conditions for the species; in contrast, areas with a GARP prediction of 5–10 probably indicate suitable conditions (see Section 3.2).

- Spatial context of errors:-

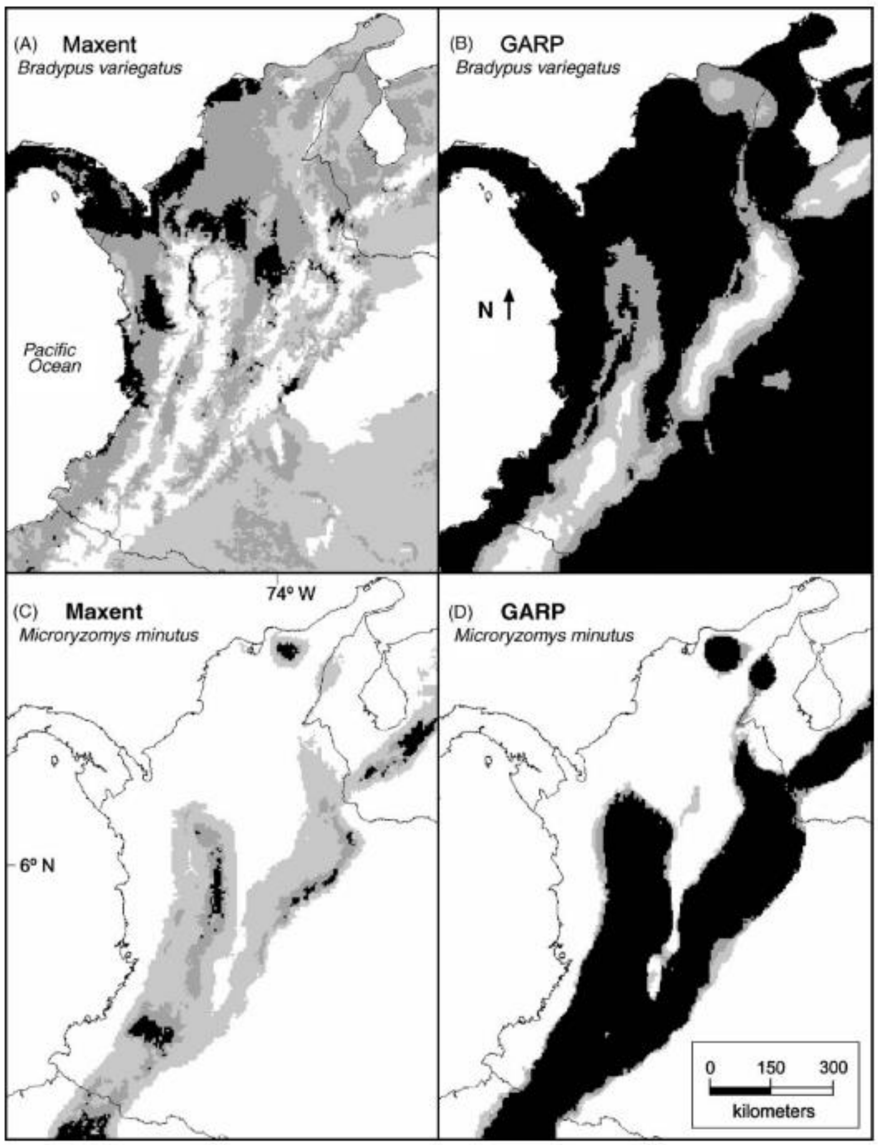
The performance of Maxent on M. minutus when the potential vegetation variable was used warrants some discussion. The AUC for the second random data partition was notably lower than for the other partitions, and for the model run on the same partition without potential vegetation. Most of the occurrence localities for the species are contained in the “tropical and subtropical moist broadleaf forest” and “tropical and subtropical dry broadleaf forest” classes of potential vegetation. However, two of them fall within the “montane grasslands” class (the species indeed can inhabit this vegetation type in mosaic habitats along the ecotone with forested regions below; Carleton and Musser, 1989). For data partition 2, both of those latter two localities fell in the test dataset (i.e., not the training set). Accordingly, Maxent’s prediction strongly avoided the “montane grasslands” class. The pixels corresponding to those two test localities thus had very low predicted value, bringing down the AUC for that partition. This is an artifact of under-regularization. More regularization for categorical features would allow some prediction in classes with no presence records, especially if the total number of presence records is small (Haffner et al., in preparation, and implemented in later versions of Maxent). The behavior of Maxent is in fact reasonable in this case, as the training data do not cover the range of vegetation classes that the species can inhabit. Furthermore, it is better than the statistics would suggest, as the occurrence localities falling in montane grasslands both lie on the border with pixels of one of the other two classes inhabited by the species, and are therefore close to highly predicted areas. Their omission should thus be penalized less than other test localities(Fielding and Bell, 1997). Indeed, smoothing the prediction by twice applying a simple 3 × 3 smoothing convolution with the following weights as a low-pass filter(Jensen, 1996)

$$\begin{matrix} \mathbf{0.05} & \mathbf{0.05} & \mathbf{0.05} \\ \mathbf{0.05} & \mathbf{0.6} & \mathbf{0.05} \\ \mathbf{0.05} & \mathbf{0.05} & \mathbf{0.05} \end{matrix}$$

increases the AUC to 0.98 for that partition, which is in line with those of the other random partitions, and causes very little visible change to the prediction. Such post-processing may be of general utility when spatial error is known to exist in the data, for example due to errors in site localities or boundaries of polygons representing categorical variables.

- Advantages of Maxent:-

Maxent exhibits a number of essential advantages (see Section 1). In addition, visual examination of the models indicates two farther possible advantages. In these analyses, areas prognosticated by 5 – 10 of the bestsubset GARP models generally showed a reasonable vaticination of the species'geographic ranges (see over). Utmost of those areas were prognosticated by all 10 models. In discrepancy, the Maxent vaticination is nonstop, and within those areas suitable for each species, it further distinguishes between those with a hardly (but sufficiently) strong vaticination versus those with decreasingly stronger prognostications. This represents an important advantage for Maxent, and explains part of its constantly advanced AUC values. The AUC for GARP could potentially be bettered by trying to increase the resolution at the left end of the ROC wind, videlicet by creating further original double GARP models (say 1000) and choosing a larger stylish subset (say 100). We tried this for both species using all circumstance points and all variables, and plant that the prognostications were nearly unchanged (in comparison to a stylish subset of 10 out of 100 models). We also note that indeed with 100 total models, GARP was formerly testing the limits of the computers we used (recycling all 22 datasets produced nearly 20 GB of affair, compared with 285 MB for Maxent). Piecemeal from affair size, the computational conditions of the two algorithms were analogous in this study; GARP equaled1.95 h to produce a single vaticination (best-subset compound deduced from 100 models), compared with2.27 h for Maxent, both on an 850 MHz Pentium 3 processor. Latterly performances of Maxent available on the website use a briskly algorithm (Haffner and Phillips, in medication); Version1.8.1 takes a aggregate of 70 min to reuse all 22 datasets on the below- mentioned computer, or 20 min on a newer3.2 GHz Intel Xeon computer.



- Future work:-

Much work can be done to refine the use of Maxent for modeling species geographic distributions. Research should determine the number of occurrence localities needed to make an adequate prediction, and to determine how much regularization is needed to avoid overfitting when the number of occurrence localities is small; preliminary results regarding these issues are presented by Dud´ık et al. (2004) and Phillips et al. (2004). Regarding the quality of the inputs to Maxent, the effect of non-uniform sampling of species localities should be also investigated, building on Zadrozny (2004), with an eye to estimating and limiting the impact of sampling bias (Reddy and Davalos, 2003) ´ . For illustration, selection of background points taking into account which spots have been tried (rather than simply at arbitrary) can meliorate the goods of slice bias in some cases (Zaniewski etal., 2002). As described in Section4.3, smoothing a vaticination may be a useful general system of reducing the negative impact of spatial crimes in points and environmental variables. Also, before modeling the species' conditions, smoothing could also be applied to any variables that are suspected of having spatial crimes, but it's far from a complete approach to error operation. Another possibility, which may ameliorate performance indeed in the absence of crimes in the input data, would be to use bilinear (rather than nearest-neighbor) interpolation to gain values for the environmental variables at the training points. Therefore, training points near the boundary between two pixels would admit a combination of the values of the two pixels; for categorical variables, training points veritably near to the boundary between two classes would have partial class in both classes. Alternately, rather than using a double point to represent class in each class, a nonstop point representing distance from the class could be used.

Appendix

Modeling distribution of species 'bradypus variegatus' using Python:-

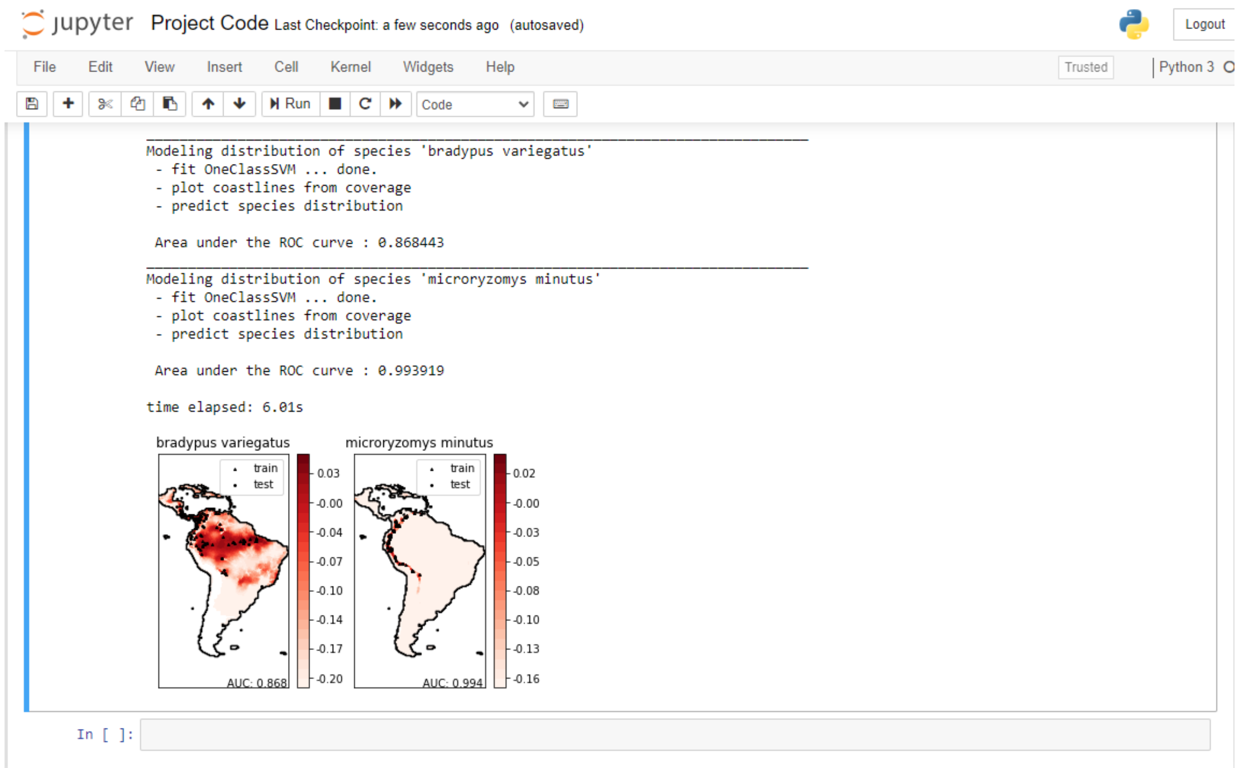
- Input:-

```
from time import time

import numpy as np
import matplotlib.pyplot as plt

from sklearn.utils import Bunch
from sklearn.datasets import fetch_species_distributions
from sklearn import svm, metrics

# if basemap is available, we'll use it.
# otherwise, we'll improvise later...
try:
    from mpl_toolkits.basemap import Basemap

    basemap = True
except ImportError:
    basemap = False


def construct_grids(batch):
    """Construct the map grid from the batch object

    Parameters
    ----------
    batch : Batch object
        The object returned by :func:`fetch_species_distributions`

    Returns
    -------
    (xgrid, ygrid) : 1-D arrays
        The grid corresponding to the values in batch.coverages
    """
    # x,y coordinates for corner cells
    xmin = batch.x_left_lower_corner + batch.grid_size
    xmax = xmin + (batch.Nx * batch.grid_size)
    ymin = batch.y_left_lower_corner + batch.grid_size
    ymax = ymin + (batch.Ny * batch.grid_size)

    # x coordinates of the grid cells
    xgrid = np.arange(xmin, xmax, batch.grid_size)
    # y coordinates of the grid cells
    ygrid = np.arange(ymin, ymax, batch.grid_size)

    return (xgrid, ygrid)


def create_species_bunch(species_name, train, test, coverages, xgrid, ygrid):
    """Create a bunch with information about a particular organism

    This will use the test/train record arrays to extract the
    data specific to the given species name.
    """
    bunch = Bunch(name=" ".join(species_name.split("_")[:2]))
    species_name = species_name.encode("ascii")
    points = dict(test=test, train=train)

    for label, pts in points.items():
        # choose points associated with the desired species
        pts = pts[pts["species"] == species_name]
        bunch["pts_%s" % label] = pts

        # determine coverage values for each of the training & testing points
        ix = np.searchsorted(xgrid, pts["dd long"])
        iy = np.searchsorted(ygrid, pts["dd lat"])
        bunch["cov_%s" % label] = coverages[:, -iy, ix].T

    return bunch
```

```python
def plot_species_distribution(
    species=("bradypus_variegatus_0", "microryzomys_minutus_0")
):
    """
    Plot the species distribution.
    """
    if len(species) > 2:
        print(
            "Note: when more than two species are provided,"
            " only the first two will be used"
        )

    t0 = time()

    # Load the compressed data
    data = fetch_species_distributions()

    # Set up the data grid
    xgrid, ygrid = construct_grids(data)

    # The grid in x,y coordinates
    X, Y = np.meshgrid(xgrid, ygrid[::-1])

    # create a bunch for each species
    BV_bunch = create_species_bunch(
        species[0], data.train, data.test, data.coverages, xgrid, ygrid
    )
    MM_bunch = create_species_bunch(
        species[1], data.train, data.test, data.coverages, xgrid, ygrid
    )

    # background points (grid coordinates) for evaluation
    np.random.seed(13)
    background_points = np.c_[
        np.random.randint(low=0, high=data.Ny, size=10000),
        np.random.randint(low=0, high=data.Nx, size=10000),
    ].T

    # We'll make use of the fact that coverages[6] has measurements at all
    # land points.  This will help us decide between land and water.
    land_reference = data.coverages[6]

    # Fit, predict, and plot for each species.
    for i, species in enumerate([BV_bunch, MM_bunch]):
        print("_" * 80)
        print("Modeling distribution of species '%s'" % species.name)

        # Standardize features
        mean = species.cov_train.mean(axis=0)
        std = species.cov_train.std(axis=0)
        train_cover_std = (species.cov_train - mean) / std

        # Fit OneClassSVM
        print(" - fit OneClassSVM ... ", end="")
        clf = svm.OneClassSVM(nu=0.1, kernel="rbf", gamma=0.5)
        clf.fit(train_cover_std)
        print("done.")

        # Plot map of South America
        plt.subplot(1, 2, i + 1)
        if basemap:
            print(" - plot coastlines using basemap")
            m = Basemap(
                projection="cyl",
                llcrnrlat=Y.min(),
                urcrnrlat=Y.max(),
                llcrnrlon=X.min(),
                urcrnrlon=X.max(),
                resolution="c",
            )
            m.drawcoastlines()
            m.drawcountries()
        else:
            print(" - plot coastlines from coverage")
            plt.contour(
                X, Y, land_reference, levels=[-9998], colors="k", linestyles="solid"
            )
            plt.xticks([])
```

```
            plt.yticks([])

        print(" - predict species distribution")

        # Predict species distribution using the training data
        Z = np.ones((data.Ny, data.Nx), dtype=np.float64)

        # We'll predict only for the land points.
        idx = np.where(land_reference > -9999)
        coverages_land = data.coverages[:, idx[0], idx[1]].T

        pred = clf.decision_function((coverages_land - mean) / std)
        Z *= pred.min()
        Z[idx[0], idx[1]] = pred

        levels = np.linspace(Z.min(), Z.max(), 25)
        Z[land_reference == -9999] = -9999

        # plot contours of the prediction
        plt.contourf(X, Y, Z, levels=levels, cmap=plt.cm.Reds)
        plt.colorbar(format="%.2f")

        # scatter training/testing points
        plt.scatter(
            species.pts_train["dd long"],
            species.pts_train["dd lat"],
            s=2 ** 2,
            c="black",
            marker="^",
            label="train",
        )
        plt.scatter(
            species.pts_test["dd long"],
            species.pts_test["dd lat"],
            s=2 ** 2,
            c="black",
            marker="x",
            label="test",
        )
        plt.legend()
        plt.title(species.name)
        plt.axis("equal")

        # Compute AUC with regards to background points
        pred_background = Z[background_points[0], background_points[1]]
        pred_test = clf.decision_function((species.cov_test - mean) / std)
        scores = np.r_[pred_test, pred_background]
        y = np.r_[np.ones(pred_test.shape), np.zeros(pred_background.shape)]
        fpr, tpr, thresholds = metrics.roc_curve(y, scores)
        roc_auc = metrics.auc(fpr, tpr)
        plt.text(-35, -70, "AUC: %.3f" % roc_auc, ha="right")
        print("\n Area under the ROC curve : %f" % roc_auc)

    print("\ntime elapsed: %.2fs" % (time() - t0))


plot_species_distribution()
plt.show()
```

- Output:-



- GitHub Link:-

https://github.com/ZainRehman-1/Species-Distribution-Modeling.git