Segmented or generalised regression models in predicting and explaining species distributions

David R.B. Stockwell[1]

[1]Centre for Intelligent and Networked Systems, Central Queensland University, Australia.

Email: d.stockwell@cqu.edu.au

Keywords: machine learning, big data, species distributions, fuzzy sets

Abstract. Previous studies have suggested that machine learning approaches utilizing large datasets of environmental variables might be an efficacious approach to species distribution modelling. In this study we demonstrate the application of a new version of `WhyWhere` against a generalized linear model on *Bradypus variegatus* (the brown-throated three-toed sloth) on two data sets: a 9 variable WorldClim climatic data sets and a 940 global data sets of mixed type on the Global Ecosystems Database. The primary measure of performance was the area under the curve of the receiver operating characteristic. We demonstrate that `WhyWhere` has superior accuracy at predicting Bradypus on both data sets and identifies a number of accurate, but completely different variables with the potential to provide unique insights into the ecology of the species not provided by the climatic variables. The implications of the two approaches for SDM are discussed. `WhyWhere` is available as an R package from the development site at http://github.com/davids99us/whywhere.

1 Introduction

Novel ecological niche models of all kinds have proven efficacy for modelling species distributions when adequate species occurrence data, environmental correlates and modelling methods are used [1]. Methods where the form the structure of the model resembles the unimodal response of the species to the environment such as BIOCLIM [2] and generalized linear models (GLMs) [3] are typically combined with climatic variables to represent the constraints of the climatic range of the species. In this case the use of bioclimatic variables such as the WorldClim database [4] is associated with high predictive accuracies.

At finer scales where there is little climatic variation, such as the study of rare and endemics, the species may be responding to habitat features with discrete structure such as nesting sites, physical barriers and resources. Thus models need to be constructed from a wide range of variables: soils, vegetation type and vertical development, and ecoregions. For example, *Bradypus variegatus* (the brown-throated three-toed sloth) is a largely arboreal mammal of the Amazon and Central America regions. While it crawls along the forest floor poorly, it does swim well [5]. While climate variables may provide correlations of its entire range, its proximal response may be linked to the type of forest vegetation, weather conditions associated with that forest such as cloudiness, or seasonal inundation known a flooded forest.

Thus it can be seen that many variables may have predictive value. Little is known about a great many species, raising concern that suboptimal selection of model structure and environmental variables produces inferior model results. Some points may also be errors or zoo locations which is typical of opportunistic data sets. A modelling system should be robust to errors, as the primary purpose is to provide systems that do not require advanced expertise. These concerns have driven the agenda of general purpose modelling systems which is a long term goal of using an artificial intelligence environment [6]. Moreover, the potential environmental data sets are in a variety of resolutions and projections, which can be very time consuming and error prone to rectify. It would be convenient for systems to handle this.

We hypothesized an approach in [7, 8] that a prototype system called WhyWhere (WW1) would address the above performance goals. WW1 used segmented models to generalization over the response of the species for both continuous and categorical models. The species modelling program MaxEnt addresses the issue of multiple response types by providing a number of potential response types: e.g. linear, quadratic categorical, logarithmic [9]. By comparison, the segmented model in `WhyWhere` provides a single unified approach by cutting up the range of the variable into discrete categorical factors. WW1 [7] used a median cut algorithm which assigns equal numbers of points to each color [10]. This has been shown to retain good visual appearance, but also has the statistical justification of minimizing the variance across the range, by minimizing the variance in each category.

Here we present the advances in algorithm structure, accuracy and explanatory capacity of a new implementation of the WhyWhere algorithm into new R package (WW2) and advances in the algorithmic flow. Models of *Bradypus variegatus* (the brown-throated three-toed sloth) in Amazonia use selected environmental variables in

the WorldClim climatic data sets provided in the R package `dismo` and 940 global data sets of mixed type on the Global Ecosystems Database. We show an improvement in accuracy from data-mining large numbers of environmental data sets compared with simple sets of climatic variables. It is demonstrated that the method would be robust to any non-linear continuous (e.g. temperature, rainfall) or categorical (e.g. biome or soil type).

2 Methods

2.1 Data

The `dismo` R package includes locations for *Bradypus variegatus* and a selection of WorldClim environmental variables in its distribution package. The Bradypus species data consist of 116 longitude and latitude records of occurrences in covering Central and South America. Table 1 lists the 9 bioclimatic variables from the WorldClim and one variable from the ecoregions databases [4, 11]. These are available at a common resolution of 0.5 degrees and projection of WGA84. The algorithm crops the environmental data to one degrees of the maximum extent the recorded locations.

	File	Variable	Source	Resolution..deg.	Type
1	bio1.grd	Annual Mean Temperature	BIOCLIM	0.50	real
2	bio12.grd	Annual Precipitation	BIOCLIM	0.50	real
3	bio16.grd	Precipitation of Wettest Quarter	BIOCLIM	0.50	real
4	bio17.grd	Precipitation of Driest Quarter	BIOCLIM	0.50	real
5	bio5.grd	Max Temperature of Warmest Month	BIOCLIM	0.50	real
6	bio6.grd	Min Temperature of Coldest Month	BIOCLIM	0.50	real
7	bio7.grd	Temperature Annual Range (BIO5-BIO6)	BIOCLIM	0.50	real
8	bio8.grd	Mean Temperature of Wettest Quarter	BIOCLIM	0.50	real
9	biome.grd	Terrestrial Ecoregions of the World	WWF	0.50	category

Table 1: The variables in the extttdismo package from WorldClim with resolution and type.

	File	Variable	Source	Res.deg.	Type
1	fnocwat.txt	Navy Terrain Data–Percent Water Cover	GED	0.17	real
2	lccld07.txt	Leemans and Cramer July Cloudiness (% Sunshine)	GED	0.50	real
3	i00an1.1.pgm	World Ocean Atlas 2001 - silicate at depth 0 metre	GED	1.00	real
4	etopo.pgm	Elevation from the National Geographic Data Center	GED	0.03	metres

Table 2: Selection of the 940 variables from the Global Ecosystems Database used in this study with resolution and type.

The second environmental dataset was the Global Ecosystems Database (GED) and consists of 940 global data sets of mixed type on the Global Ecosystems Database (selected on Table 2). These are all of global extent with a range of resolutions from 1 deg to 0.033... degrees and in the same global geographic projection. They were scaled to fit 0-255 values for the WW1 project in order to be more compact. This is no longer necessary and future studies will use the native datasets.

2.2 Algorithms

The flow of analysis for GLM and WW2 is shown on Table 3. The stages include data inputs, fitting, and evaluation. All occurrences are given as $< x, y >$ pairs of longitude and latitude. The data outputs of both methods include a probability estimate at each environmental data point. Other outputs include the respective models, other possible models, both singly and in combination, and maps of the probability of species in the region of interest, and so on.

The GLM requires a flat file in so-called 'wide' format. The wide format is where the variables are in columns and the data points are in rows. These data include the presence or absence and a listing of all environmental value(s):

$$< x, y, pa, v_1, ..., v_n > \tag{1}$$

GLM	WW2
1. Input long and lat occurrences	1. Input long and lat occurrences
2. Presample background points	<optional>
3. Input all env. vars	2. Start loop through list of env. vars.
5. Create flat file	2.1 Input env. variable
4. Fit additive model	2.2 Fit segmented model
5. Output model	2.3 <optional> Evaluate conjunction with best model so far
6. Evaluate model	3. Output ordered list of models evaluated

Table 3: The flow of analysis for GLMs and WW2 shows the stages of data inputs, fitting, and evaluation.

where x is longitude, y is latitude, pa is dependent variable the presence of absence of the species, and $v_1, ...v_n$ are the independent variables which are the values at each of the locations $< x, y >$. In the case of presence-only data, the environmental values are generally drawn from a sample B of the possible environmental data sets G, generally referred to as backgrounding. This was developed as a means to provide pseudo-absence data when none exists – such as in the case of museum data – and so allow the use of these techniques.

The default inputs to WW2 are the longitude and latitude of known locations and a set of environmental data files G_i that may or may not be coregistered. They must be able to be read into the `raster` package. The model in WW2 is based on the comparison of the frequency of environmental values in the environment G and the frequency of environmental values in the given locations S. The model is in effect a lookup table on segments in the environmental variable. As WW2 makes use of the frequency of the environmental values G and does not need a 'wide' file, even in multidimensional analysis, backgrounding is optional. The WW2 algorithm can predict on both presence-only and presence-absence data. Parameters to WW2 include **multi** identify the number of conjunctions of variables to be searched.

The data outputs of both methods is a probability estimate at each environmental data point. Other outputs include the respective models, other possible models, both singly and in combination, and maps of the probability of species in the region of interest, and so on.

The main loop of the algorithm develops a segmentation of the variables and a lookup table based on that segmentation. The output is a table listing each model tested ranked in decreasing order by the Chi-squared value. A standard Chi-squared test compares the counts of values of pixels in each segment of the overall environment(G), against the count of values in those pixels where the species occurs (S). A significant difference in the distribution is indicative of a strongly biased sample indicative of the response of the species to its environment.

In the current implementation of a multi-dimensional model, only the best variable is combined with each new variable using the fuzzy minimum of the predicted probabilities at each point. Alternative approaches to searching the space of conjunctions may be implemented in future. It is possible to monitor the progress with the plot option in a streaming work flow.

2.3 Response function

In a second or third order GLM the response of species is a 'humped' function on the range of the environmental variable. This function represents the falling frequency of the species around an ideal habitat, or restriction to a range of a variable (e.g. the range of temperature tolerances). While often second order or quadratic it must be at least order three to incorporate skewness. The range of the GLM is the odds of occurrence of the species at each environmental value where the species occurs.

In WW2 strength of a species' response is given by the frequency of environmental value in each segment of the environmental value where the segment is determined by a quantile. WW2 segments a single continuous or categorical variables into equal quantile open-closed intervals. The number of segments $2n$ is determined using the Freedman-Diaconis rule [12] for optimal binning of histograms. WW1 [7] used a median cut algorithm which assigns equal numbers of points to each color [10]. This has been shown to retain good visual appearance, but also has the statistical justification of minimizing the variance across the range, by minimizing the variance in each category.

$$G_i = (v_1, v_2]_1, ..., (v_{n-1}, v_n]_j, ..., (v_{2n-1}, v_2 n]_n \tag{2}$$

For a given environmental variable G_j and a segmentation, the tabulation of the counts in each of the environmental values allows calculation of the frequency of values in each segment j where the species occurs and over the prior frequency of environmental values. These frequencies are s_j and g_j respectively from which we calculate the odds ration or OR.

$$OR_j = s_j(1 - g_j)/g_j(1 - s_j) \tag{3}$$

The odds ratio is the increase in frequency of a species occurring in an environment over a specific range S, relative to the naturally occurring frequency of that environmental range in the region G. Given the odds ratio, the response is given as an expected probability $P : \mathbb{R} \to [0, 1]$ where

$$P(OR_j)_j = OR_j/(1 + OR_j) \tag{4}$$

While we would like to estimate the probability of species being present using Bayes' Theorem, it is dependent on season, search effort and other uncontrolled variables. The proxy above is sufficient and useful in most applications [13, 14].

2.4 Multivariate mode

Both algorithms estimate a probability value to each data point in the training set, producing a vector of values in the range [0,1] labelled with the *pa* presence absence variable in two values $0, 1$. While the GLM produces this from an additive model of the odds for each variable, WW2 produces a probability vectors only on single variables at a time. The two single variables are combined using a fuzzy AND Zadeh operator [15] to produce a new membership vector evaluated again using the ROC or AUC.

$$AND : x \wedge y = min(f(x), f(y)) \tag{5}$$

This eliminates the need to develop and express a higher dimensional model, even though the data is high dimensional, facilitating a data mining approach to very large databases.

The models for both GLM and 2 can be developed on a test training set and tested on a test set using any protocol. Here it is evaluated (using the ROC or AUC). The AUC gives the probability that a model correctly classifies a random draw of a positive and negative example and so is a type of accuracy. In order to avoid over fitting, a K-fold validation is used proceeds by sequentially holding back one fifth of the data each time for evaluation, and developing the model using the remaining four-fifths.

Analysis was on a Toshiba laptop and were well within the capacity of the machine. Both were implemented in the R language and all of the code for `WhyWhere` and for this study is available in R from the development site at http://github.com/davids99us/whywhere.

3 Results

3.1 Prediction on WorldClim Dataset

Predictions by WW2 of the brown-throated three-toed sloth (*Bradypus variegatus*) were made using 9 variable the WorldClim data set. The variable *bio*7 was selected which is a combination of *bio*7 of *bio*7 = the temperature annual range ($bio5 - bio6$) where *bio*5 = maximum temperature of the warmest month and *bio*6 = minimum temperature of the coldest month. Fig 3 shows the *bio*7 and *biome* variables, the model, and the the predicted distribution over the South American continent with occurrence points. The model lookup table for *bio*7 is shown in Table4. This contains ten segments with a width varying from 9 in the segment (108,117] and 115 in the segment (210,max.value]. The table also lists the odds ratio, which varies from 0 where environmental ranges are unoccupied by the species (N=0), to 3.94 in the range from (63,108] of the environmental variable. The probability that a species would be present in that environmental range was 0 to 0.8. The projection of these probabilities in the geographic space is shown in Figure 1D.

The area under the receiver operating curve (AUC) of the variables are listed on Table 5. The rank order of WW2 and GLM is similar which is expected as both are fitting similar types of unimodal response curves. The best WW2 model has an accuracy of 0.78 which exceeds the GLM accuracy of 0.75. The accuracy of WW2 on the other variables equals or exceeds the AUC accuracy of the GLM model. This is indicates that the WW2 model achieves a more exact fit to the actual response surface.

The mean accuracy of GLM and WW2 for modeling on the WorldClim data set using five-fold sampling model was 0.74 and 0.78 with a standard error of the mean around 0.01 (Table 6). There was no significant drop in accuracy between the training set and the test set indicating that overfitting is not occurring in these models.

We also evaluated some alternative means of segmenting the response function: an even distribution of cuts over the range of the variable, distribution by quantile frequency, and an entropy optimizing method [16]. The quantile

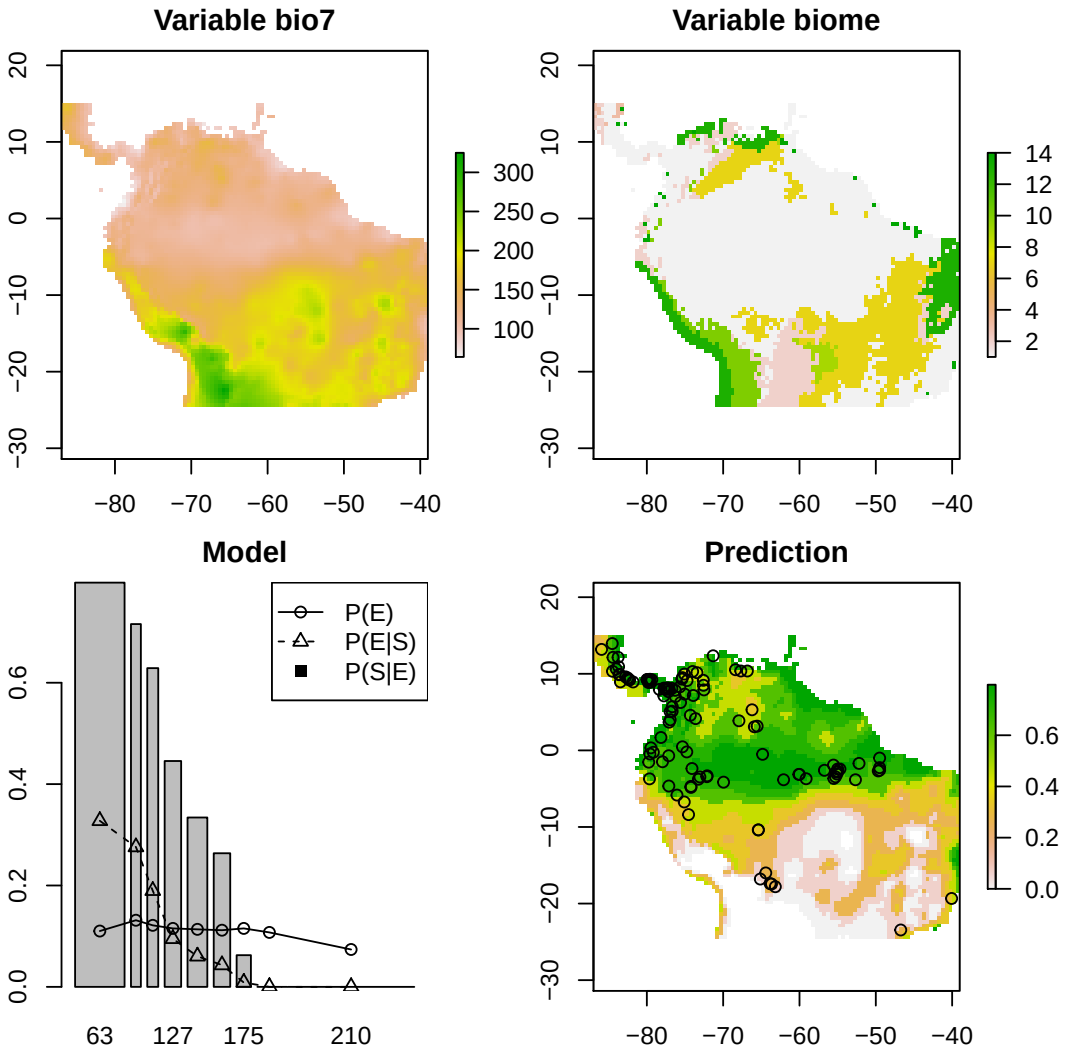

Figure 1: Output of WW2 on the Bradypus data set. (A - upper left) the best variable, (B - upper right) the biome variable, (C - lower left) the response function, and (D - lower right) the predicted probability distribution with the occurrence points as circles.

	ID	value	N	count	cuts	width	odds	prob
1	1.00	3270.00	38	527.00	63.00	45.00	3.94	0.80
2	2.00	1008.00	32	629.00	108.00	9.00	2.52	0.72
3	3.00	1215.00	22	581.00	117.00	10.00	1.69	0.63
4	4.00	2010.00	11	552.00	127.00	15.00	0.80	0.45
5	5.00	2709.00	7	543.00	142.00	18.00	0.50	0.33
6	6.00	2505.00	5	535.00	160.00	15.00	0.36	0.26
7	7.00	2353.00	1	552.00	175.00	13.00	0.07	0.06
8	8.00	4367.00	0	514.00	188.00	22.00	0.00	0.00
9	9.00	22524.00	0	353.00	210.00	115.00	0.00	0.00

Table 4: The lookup table for bio7 on Bradypus data. Each row is a segment of the range of the variable (factors) containing information on the counts in that segment, from which are calculated the odds and the probability.

	files	AUC.WW2	AUC.GLM
1	bio7.grd	0.78	0.75
2	bio12.grd	0.75	0.71
3	bio17.grd	0.71	0.68
4	bio6.grd	0.70	0.67
5	bio16.grd	0.69	0.66
6	biome.grd	0.67	0.66
7	bio5.grd	0.62	0.61
8	bio8.grd	0.60	0.57
9	bio1.grd	0.59	0.58

Table 5: The AUC of the variables on the WorldClim dataset from GLM and 2 respectively.

and even segmentation were marginally superior to the entropy method, although further work would be needed on optimal segmentation methods which is beyond the scope of this paper.

3.2 Prediction on GED Dataset

The top 10 variables out of 940 variables in GED are listed in Table 7. The top variables were *lccld*08 Leemans and Cramer August Cloudiness (percentage Sunshine), *lccprc*09 Leemans and Cramer September Precipitation (mm/month) and *lccld*08 Leemans and Cramer July Cloudiness (percentage Sunshine). The lesser variable are a mixed bag of soil soil particle size classifications at various horizons ($wrcla-$), a soil classification ($whsoil$), and $fnocwat$: Navy Terrain Data - Percent Water Cover.

Note that the accuracy of WW2 on the best variable is 0.84 and greater than GLM at 0.7 in this case. Unlike the WorldClim variables the rank accuracy of WW2 and GLM differs. In case of $fnocwat$ the accuracy of GLM is 0.83 which exceeds the accuracy of WW2 of 0.73. Note that the accuracy of WW2 of 0.84 greatly exceeds the accuracy of be best model on the WorldClim dataset of 0.74.

Two of these alternative variables *lccld*08 and $fnocwat$. are pictured on Figure 2. Note that the predicted regions are quite different. The areas of highest predicted probability for *lccld*08 are in Central America, while the $fnocwat$ variable identifies areas on the river system of special significance. These predictions are different again from the distribution of the best WorldClim variable *bio*7 that highlighted Amazonia region.

The receiver operating characteristic for WW2 and GLM models on a selection of variables are shown on Figure 3. The curves serve to show the benefit of one variable over another based on the area under the curve.

4 Discussion

Niche model based analysis, such as the application of a higher order generalized linear model on a small set of climate variables, is the standard approach to predicting the distributions of species from occurrence records. The results of this trial showed that the single variable segmented model in WW2 on the 9 variable climate data in the `dismo` distribution has superior accuracy to a third order GLM by 0.77 to 0.74 with a standard error of 0.01. The results for accuracy on the 940 variable Global Ecosystems Database show greater accuracy of 0.84 can be achieved for WW2 and 0.83 for GLM on non-climatic variables: *lccld*08 (cloudiness) and $fnocwat$ (percent water cover)

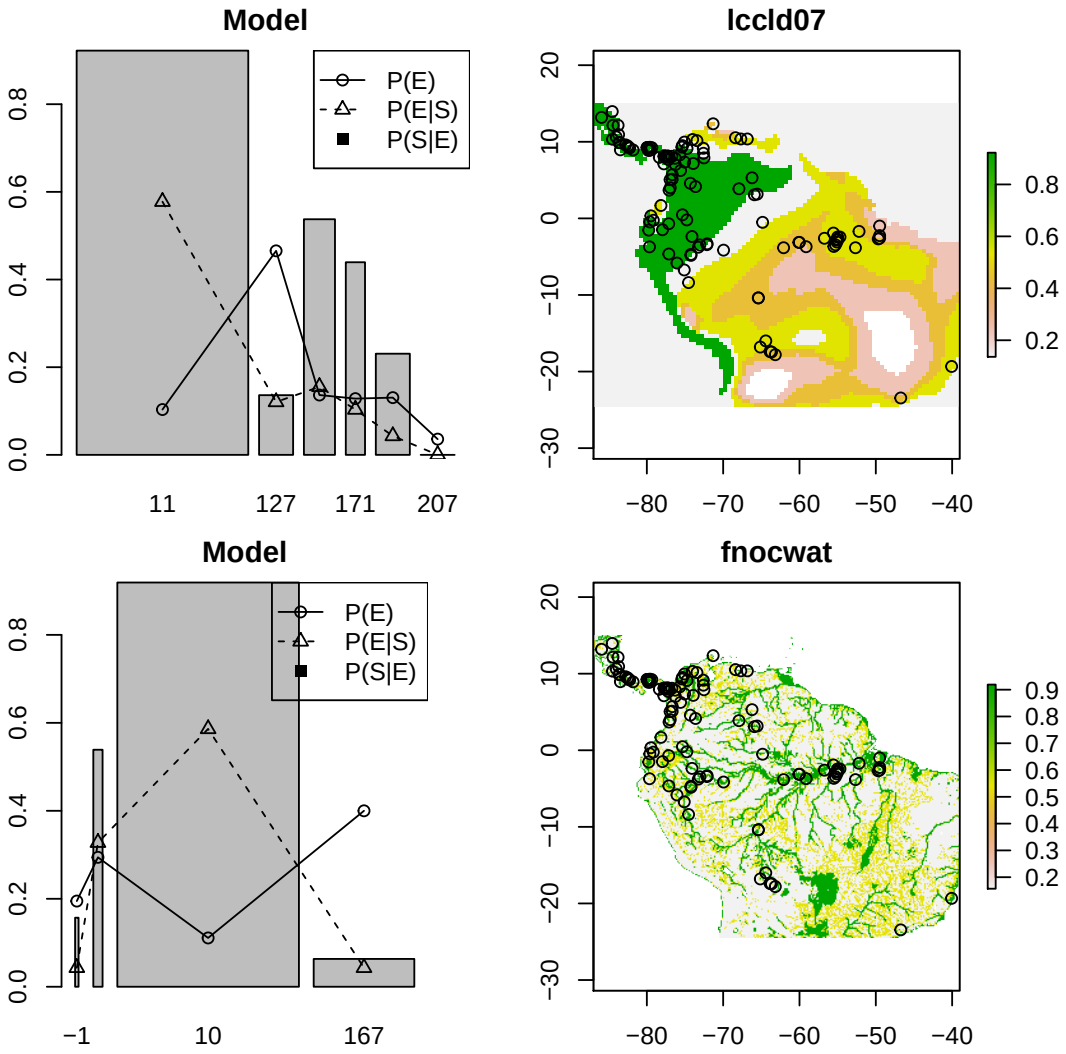

Figure 2: Models and predictions of two alternative variables from the GED dataset lccld08 and fnocwat.

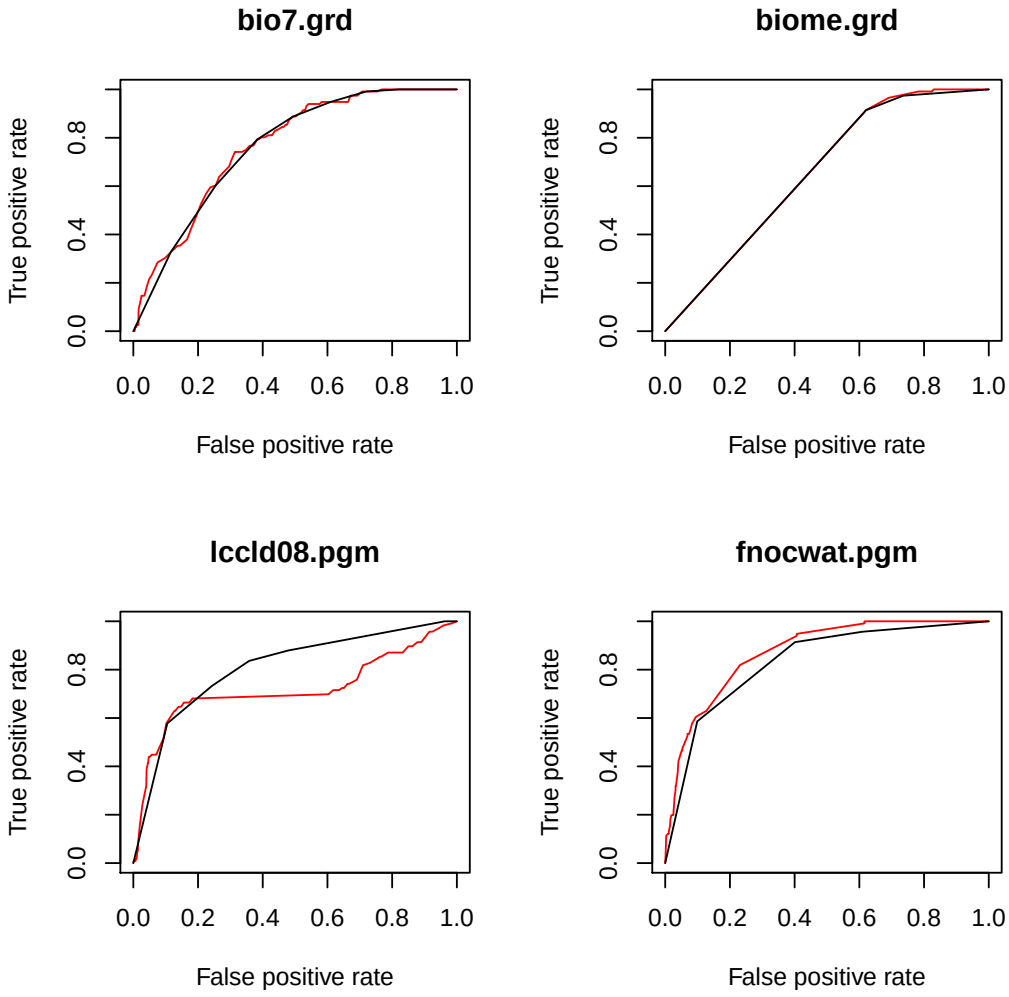

Figure 3: Receiver operating curves are shown for GLM (red) and WW2 (black) on selected variables.

	file	GLMtrain	GLMtest	WWtrain	WWtest
1	bio7	0.74	0.73	0.78	0.76
2	bio7	0.74	0.74	0.78	0.79
3	bio7	0.74	0.75	0.78	0.79
4	bio7	0.74	0.76	0.78	0.81
5	bio7	0.74	0.73	0.78	0.76
6	mean	0.74	0.74	0.78	0.78
7	s.e.	0.00	0.01	0.00	0.01

Table 6: The AUC of the 5-fold validation test of GLM and WW2 models on the WorldClim data.

	files	AUC.WW2	AUC.GLM
1	lccld07.pgm	0.85	0.69
2	lccld08.pgm	0.85	0.71
3	lcprc09.pgm	0.79	0.65
4	whsoil.pgm	0.77	0.76
5	wrcla03.pgm	0.74	0.69
6	wrcla04.pgm	0.72	0.70
7	wrcla06.pgm	0.72	0.72
8	fnocwat.pgm	0.72	0.82
9	wrcla05.pgm	0.72	0.71
10	malbsp.pgm	0.69	0.75

Table 7: The AUC of the variables on the GED dataset from GLM and WW2.

variables respectively. Thus the superior accuracy of WW2 on the small climate dataset is not necessarily due to a general superiority of model structure of WW2 over GLM; the large increase in accuracy can be attributed to the search over a larger domain of environmental correlates. These results confirm the results previously that the data mining approach pursued in WW2 leads to large increases in accuracy of species distribution modelling [7, 8].

For exploration into the determinants of species distributions, WW2 on WorldClim data selected similar variables to those reported for from GARP and MaxEnt in [9]. Application of WW2 the large GED identified a number of surprising non-climatic correlates as potential proximal determinants of the species distribution. The overall benefits of the WhyWhere approach will be driven primarily by the enhancements from this explanatory function.

The most accurate variable for WW2 on GED was *lccld*08 Leemans and Cramer August Cloudiness (percentage Sunshine) (WW2=0.84 and GLM=0.71). The most accurate variable for GLM on the GED data was the $fnocwat$ variable percentage water cover (GLM=0.84 and WW2=0.73). The contrasting outcome on *lccld*08 is probably due to the capacity of the segmented model to fit the unusual bimodal response curve (Figure 2 Model). This conclusion is supported by the the shapes of the receiver operating curves, with a match of the initial portions, and a dip around the segment labeled 127 (Figure 3). On examination of the ROC curve for $fnocwat$ the GLM exceeds WW2 over the range except for contact at the points of inflection, which suggests the that superiority of GLM may be attributed to the continuity of GLM model. The range of soil clay fraction variables $wrcla3-6$ selected is unexpected and further study would be needed to determine why these soil variables have so strong a relationship.

Austin and Meyers [3] maintained that ecological niche modelling should be performed on unimodal GLMs or beta functions in order to properly represent the ecological constraints of niche theory. However on the objective assessment of performance, a WW2 model of Bradypus on GED data has demonstrated a contradiction whereby the response to the most accurate variable *lccld*08 is two humped. The ecological relevance of this is not clear, although one may hypothesize the existence of two separate populations of sloth. Our study confirms the value of relaxation of the constraints of ecological theory will lead to improvements in prediction and explanation of species distribution as compared with application of conventional models to conventional climate variables.

The efficacy of WW2 in this study is combined with reduced preparation of data into flat files. This in turn allows handling geographic variables of varying extents, projections and resolutions, without the need for unification or coregistration before modelling. Apart from the time lost, memory needs burgeon when co-registering many variables to the finest scale, and information is lost when contracting fine data to coarse scales. The recoding of `WhyWhere` into R has greatly improved the programs' utility as has the `raster` package.

The efficacy of this approach should be coupled with the capacity to perform a range of dynamic analysis such as range changes, rare species, and competitive interactions. More complex expressions of species distribution models in the R package `WhyWhere` are planned, as are improvements in computational efficiency and testing

over a greater range of higher resolution environmental data. Future work will characterize the performance of alternative multi-variate models, comparing the multiple variable model using the fuzzy set minimum function for combining variables, against of the additive polynomial model of the GLM. Ecological principles such as competitive interactions [17] tend to be in the form of logical expressions. The principle of Liebig's Law of the Minimum states that growth is controlled by the scarcest necessary resource. This is logically a fuzzy conjunction of limiting factors – a Zadeh AND. Very few modelling methods have made this explicit connection with a basis in ecological theory. Thus studies intent on discovering the determinants of species distribution will benefit most from the explicit logical structure of WhyWhere. `WhyWhere` is available as an R package from the development site at http://github.com/davids99us/whywhere.

References

[1] Jane Elith, Catherine H Graham, Robert P Anderson, Miroslav Dudik, Simon Ferrier, Antoine Guisan, Robert J Hijmans, Falk Huettmann, John R Leathwick, Anthony Lehmann, Jin Li, Lucia G Lohmann, Bette A Loiselle, Glenn Manion, Craig Moritz, Miguel Nakamura, Yoshinori Nakazawa, Jacob M Overton, Townsend A Peterson, Steven J Phillips, Karen Richardson, Ricardo Scachetti-Pereira, Robert E Schapire, Jorge Soberon, Stephen Williams, Mary S Wisz, and Niklaus E Zimmermann. Novel methods improve prediction of species distributions from occurrence data. *Ecography*, 29(2):129–151, April 2006.

[2] H Nix. A biogeographic analysis of Australian Elapid snakes. In R Longmore, editor, *Atlas of Australian Elapid Snakes*, volume 8, pages 4–15. Australian National University, 1986.

[3] M Austin and J Meyers. Current approaches to modelling the environmental niche of eucalypts: Implication for management of forest biodiversity. *Forest Ecology And Managment*, 85(1-3):95–106, 1996.

[4] Robert J Hijmans, Susan E Cameron, Juan L Parra, Peter G Jones, and Andy Jarvis. Very high resolution interpolated climate surfaces for global land areas. *International journal of climatology*, 25(15):1965–1978, 2005.

[5] Virginia Hayssen. Bradypus variegatus (pilosa: Bradypodidae). *Mammalian Species:*, 42(1):19–32, 2010.

[6] S M Davey and D R B Stockwell. Incorporating Wildlife Habitat Into An Ai Environment - Concepts, Theory, And Practicalities. *AI Applications*, 5(2), 1991.

[7] David R B Stockwell. Improving ecological niche models by data mining large environmental datasets for surrogate models. *Ecological Modelling*, 192:188–196, 2006.

[8] David R B Stockwell. *Ecological Niche Modeling: Ecoinformatics in Application to Biodiversity*, chapter 7. CRC Press, 2006.

[9] Steven J Phillips, Miroslav Dudík, and Robert E Schapire. Maxent software for species habitat modeling, 2007.

[10] Paul Heckbert. Color image quantization for frame buffer display. In *SIGGRAPH '82: Proceedings of the 9th annual conference on Computer graphics and interactive techniques*, pages 297–307, New York, NY, USA, 1982. ACM Press.

[11] David M Olson, Eric Dinerstein, Eric D Wikramanayake, Neil D Burgess, George VN Powell, Emma C Underwood, Jennifer A D'amico, Illanga Itoua, Holly E Strand, and John C Morrison. Terrestrial ecoregions of the world: A new map of life on earth a new global map of terrestrial ecoregions provides an innovative tool for conserving biodiversity. *BioScience*, 51(11):933–938, 2001.

[12] David Freedman and Persi Diaconis. On the histogram as a density estimator:l 2 theory. 57(4):453–476–, 1981.

[13] D R B Stockwell. Lbs - Bayesian Learning-System For Rapid Expert System-Development. *Expert Systems With Applications*, 6(2), 1993.

[14] D R B Stockwell. Generic predictive systems: An empirical evaluation using the Learning Base System (LBS). *Expert Systems With Applications*, 12(3), 1997.

[15] L.A. Zadeh. Fuzzy sets. *Information and Control*, 8(3):338 – 353, 1965.

[16] U. M. Fayyad and K. B. Irani. Multi-interval discretization of continuous-valued attributes for classification learning. *Artificial intelligence*, 13:1022âĂŞ1027, 1993.

[17] Víctor Sánchez-Cordero, David Stockwell, Sahotra Sarkar, Hawoei Liu, Christopher R Stephens, and Joaquín Giménez. Competitive interactions between felid species may limit the southern distribution of bobcats lynx rufus. *Ecography*, 31(6):757–764, 2008.