-Submitted-

`WhyWhere2.0`: An R package for modeling species distributions on big environmental data

David R.B. Stockwell*

May 30, 2015

Abstract

Previous studies have indicated that multi-interval discretization (segmentation) of continuous-valued attributes for classification learning might provide a robust machine learning approach to modelling species distributions. Here we apply a segmentation model to the *Bradypus variegatus* – the brown-throated three-toed sloth – using the species occurrence and climatic data sets provided in the niche modelling R package `dismo` and a set of 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 (AUC) on a k-fold validation of predictions of the segmented model and a third order generalized linear model (GLM). This paper also presents further advances in the `WhyWhere` algorithm available as an R package from the development site at http://github.com/davids99us/whywhere.

1 Introduction

Ecological niche models such as BIOCLIM [10] and generalised linear models (GLMs) [1] are based on the proposition that the response of the species to its environment is continuous and unimodal. Symmetry in the response has been associated with equilibrium of the species with its environment. However non-equilibrium due to ecosystem or climate disturbance and/or species mobility means a simple and symmetrical niche shape is most likely the exception and not the rule [9].

In contrast to climatic-based models, applied conservation studies tend to focus on habitat features, or habitat suitability indices (HSIs), based on the structure of the environment such as availability of nesting sites, physical barriers and resource sources. HSIs are expressive of finer scale forest and aquatic environments where climatic variables do not down-scale. HSIs are proximal causes of occurrence of species than more distal climate envelopes. Habitat

---

*All correspondence to be addressed to the author at david.r.stockwell@cqu.edu.au, Adjunct Researcher at Central Queensland University.

variables are often categorical factors and there are very many possible habitat factors that could determine each species. Integration of continuous and categorical models into a single model can be a challenge.

On the basis of these and other studies we hypothesized in [18] that segmented models could be used to fit arbitrary responses, would be at least as accurate as continuous models, and be of greater accuracy when mining large environmental data sets that contain mixed habitat and climatic variables. The second hypothesis was confirmed using the segmented modelling method called `WhyWhere` [18, 17] (WW1) and both are again addressed in this new version (WW2). Moreover, we present theoretical and practical advances of the algorithm enabling the goal of incorporating species distribution modelling into an artificial intelligence environment [2].

2 Methods

A package for species modelling in R called `dismo` includes a number of popular methods including MaxEnt, Bioclim, Domain, GLM, GAM, and RandomForest. This package implements elements of the SDM work flow originally developed in the GARP machine learning system [14], such as pseudo-absences, whereby a random sample of the environment is in lieu of true absence values. The `dismo` package contains bioclimatic variables from the WorldClim database [8] and terrestrial biome data on terrestrial ecoregions [11].

Other R packages that have made `WhyWhere` more accessible are the `raster` package for handling gridded spatial data [19] and a new package called `data.table` for fast aggregation of large data sets [3].

Any species modelling method must address three main stages, and problems arising at any of these stages can lead to poor results: (1) getting the environmental data into a uniform form for analysis (2) determining the best type of model to use to represent the response of the species (3) the interpretation of the results.

2.0.1 Environmental variables

As geographic variables come in varying extents, projections and resolutions, they must generally be unified by coregistrtion before modelling. But this entails additional processing and inefficiencies. Smaller resolution sets must be enlarged redundantly and information lost when contracted. Memory needs burgeon when coregistering many variables to the finest scale. An ideal method would utilize each data set at its native projection and resolution.

After coregistration of geographic layers, most satistical models input a 'wide' file format, i.e. with variables in columns and locations in rows. This adds an intermediate processing step where all environmental data must be then be held in memory, which can hit memory limitations at the prediction stage when the models are applied across the geographic space.

The approach in WW1 was to transform geographic variables into a compact image format and then to combine at most three variables into the image in the red, green and blue channels. While fast image processing packages could then be used for segmentation, the range of the variables were scaled between 0-255, and also did not obviate the need for coregistration of geographic data. One advance in WW2 is the manner of building multidimensional models. The key insight is that the evaluation of a combination of two or more variables can be performed on the predictions of the models, instead of being performed in the high-dimensional model model space; a higher dimensional model is not required.

For example, the response on a single variable is called a membership function (for reasons explained later). Prediction assigns a membership value to each data point in the training set, producing a vector of values in the range [0,1] which can be evaluated (using the ROC or AUC). The membership vectors for two single variables can be combined using a fuzzy AND operator to produce a new membership vector. This membership vector can then be evaluated (using the ROC or AUC), thus evaluating the performance of the conjunction without calculating the two-dimensional membership function. The combination of membership functions follows the AND, OR Zadeh operators [20]:

$$AND : x \wedge y = min(f(x), f(y)) \tag{1}$$

$$OR : x \vee y = max(f(x), f(y)) \tag{2}$$

This eliminates the need to develop and express a higher dimensional model, enabling a data mining approach where we can explore large databases for a parsimonious model of the species in analysis. Because of this we here focus on single variable models.

2.0.2 Choice of model structure

Ecological theory maintains that the response of species to the environment is generally 'humped' around an ideal, or restricted to a range of a variable (e.g. the range of temperature tolerances). Such a model must be at least quadratic to represent a unimodal response, and order three to incorporate skewness. An ideal method would be robust to any non-linear response type. Many environmental variables are categorical particularly soil and vegetation type. Ideally a modeling system integrates both types: continuous (e.g. temperature, rainfall) or categorical (e.g. biome or soil type) but very few do.

MaxEnt solves this problem by providing a range of potential response types [12]. The approach in `WhyWhere` is cutting up the range of the variable into discrete categorical factors. WW1 [18] used a color quantization algorithm in the GIF image format. WW2 segments a single continuous or categorical variables into open-closed intervals. For example, the output of the R cut function on the numbers 1..4 into 2 levels as open-closed intervals:

```
> cut(1:4,2)
```

```
[1] (0.997,2.5] (0.997,2.5] (2.5,4]     (2.5,4]
Levels: (0.997,2.5] (2.5,4]
```

The unique 4 categorical values could also be represented in the same open-closed syntax.

```
> cut(1:4,4)
```

```
[1] (0.997,1.75] (1.75,2.5]   (2.5,3.25]   (3.25,4]
Levels: (0.997,1.75] (1.75,2.5] (2.5,3.25] (3.25,4]
```

The cuts do not need to be uniformly spaced. The default method of determining the cut locations in WW1 followed the median cut algorithm which assigns equal numbers of points to a limited number of categories [7] in the red, green, blue 3D space. This has been shown to retain good visual appearance, but also has a statistical justification of minimizing the variance across the range, by minimizing the variance in each category. There are other methods of multi-interval discretization of continuous-valued attributes for classification learning [4] implemented in the `discretization` R package.

The calculation of the intensity of species' response is straightforward on a segmented variable. The locations where the species occurs can be thought of as a sampling of the environmental space with a count of values in the each environmental category (S), and expressed as a density by normalizing the sum over the values in the categories to one. The prior density of values in each environmental category is G. The change in the density from G to S indicates strength of the the response of the species to its environment in that category – the membership function M. A membership function is a fuzzy truth value as a function $f : \mathbb{R} \to [0, 1]$ from a numeric domain to the real unit interval. The membership function we used in WW1 and in WW2 for each category i in f is:

$$M_i = S_i/(S_i + G_i) \tag{3}$$

What we would like is the conditional probability of species being present given the environmental $P(S|G)$ for each category of G. Frequentest calculations give us $P(G|S)$. While $P(S|G)$ could be obtained using Bayes Theorem, it requires an estimate of probability of the occurrence of a species $P(S)$, which in opportunistic data is not well defined, and dependent on season, search effort and other uncontrolled variables. The best we can do is an approximate equivalence that holds under certain conditions (principally independence of variables) and has been shown to be sufficient and useful in modelling such relationships [15, 16]. Where β is a normalization factor:

$$P(S|E) = \beta P(E|S) \tag{4}$$

One may ask why not calculate S/G and not $S/(S + G)$? Because we are developing a model on a background set, developed from a random sample of points in G, this can result in points in S_i that are not in G_i. That is, the species occurs in environments that are not represented in the background set.

Use of $S/(S+G)$ avoids a divide by zero error. The membership function is a proportional relationship which is sufficient to compare the efficiency of single variables.

There are many ways to evaluate skill of a model. The approach to evaluating the strength of the response in WW1 was significance with the Chi-squared test or a K-S test, however we find it more convenient to use the area under the curve of the receiver operating characteristic, or AUC. The receiver operating curve (ROC) is widely used to compare classification models, while the AUC provides a robust measure of skill. It is the probability that a model correctly classifies a random draw of a positive and negative example.

2.0.3 Ecological interpretation

How does the structure of a multi-dimensional model embody the ecology of the species? Many methods used in species modelling are based in other domains (e.g. linear regression) without clear ecological interpretation. For example, when inconsistent units such temperature and rainfall are combined in a generalized linear model, how are they to be interpreted?

Ecological principles such as competitive interactions [13] 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. Another law of ecology is Gauss's law of competitive exclusion. This is a proposition that two species competing for the same resource cannot coexist at constant population, due to effect of slight advantages magnified over generations. This is a fuzzy disjunction – a Zadeh OR. Thus fuzzy AND and OR can represent elements of established ecological theory. The approach of modelling with a logical expression of a small number of variables has utility in interpreting as ecological theory.

2.1 WhyWhere Algorithm

The first step is to 'presample' which when given a set of presence data locations, and a geographic file that serves as a mask, produces a combined list of the presence data and a randomly sampled list of sites of both presence and absence. The mask file defines the geographic extent for sampling the background data. If absences are available then they need not be generated.

The inputs to WW2 are: a `data.table` from presample with the longitude and latitude of known locations, and a list of environmental data files that may be read into the `raster` package. This file is then input to the main routine with a list of the geographic files. Parameters include **multi** for searching conjunctions of variables and **segment** to select the form of segmentation.

The algorithm proceeds by looping through the environmental variables and creating and evaluating a membership function on each one. A table of the variables so-far is retained. In the current implementation of a multi-dimensional model, only the best variable is combined with each new variable using the fuzzy minimum and the AUC recalculated. Alternative approaches to searching the space of conjunctions may be implemented in future. It is possible to monitor the progress with the plot option. This plots out the best model so far and prints out a list of the best models considered in a streaming work flow.

Listing 1: Listing of the main algorithm

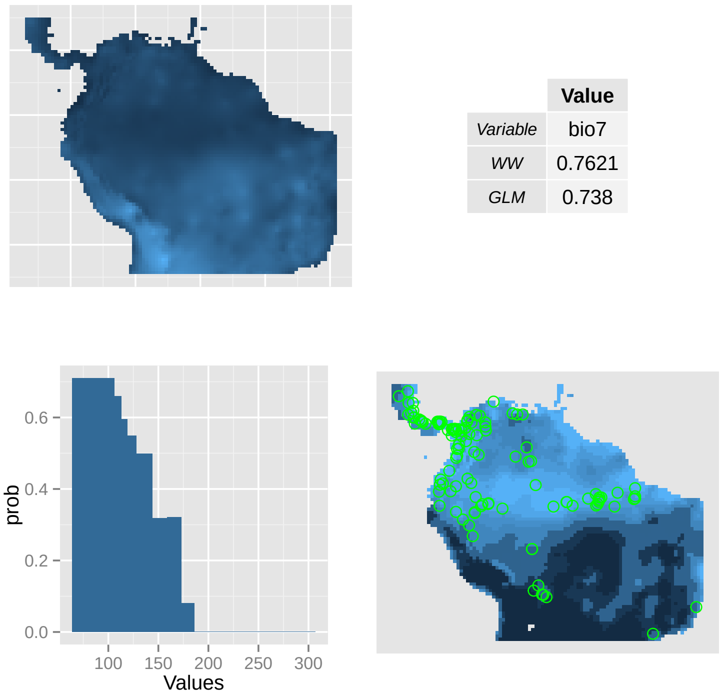
```
input locations
input a mask file
prepare background points and combine with presence points
for all environmental files do
 develop membership function for file
 insert AUC and variable name into ordered result
 test conjunct of this variable with best so far
 if result is better then insert into ordered result
output table of results
```

We show the results for the brown-throated three-toed sloth (*Bradypus variegatus*) that is documented and modeled in `dismo`. The environmental data consist of 9 environmental files from the WorldClim data set covering the South American continent. Figure 1 shows (A) the best variable, (B) the AUC of the best model, (C) the membership function, and (D) the predicted distribution with the presence points. The highest rated variable in the Bradypus data set is $bio7$ = the temperature annual range ($bio5 - bio6$) where $bio5$ = maximum temperature of the warmest month and $bio6$ = minimum temperature of the coldest month. The result is very similar to the results from GARP and MaxEnt in [12]. Table 2 lists the results for all variables.

Table 1: The AUC of environmental variables on the Bradypus data: WW is WhyWhere and GLM is a third order generalised linear model.

	name	WW	GLM
1	bio7	0.76	0.74
2	bio12	0.73	0.70
3	bio6	0.72	0.67
4	bio17	0.70	0.67
5	bio16	0.68	0.67
6	biome	0.67	0.65
7	bio5	0.63	0.61
8	bio1	0.63	0.59
9	bio8	0.59	0.57
10	limit	0.00	0.00

The membership function shown graphically in Figure 1.C is represented internally as a lookup table, shown in Table2. The prior distribution ($G1$) is the blue line, and the distribution of presences ($G2$) is the red line. The membership function is shown by the grey bars ($G1/(G1 + G2)$). Note the almost uniform distribution of background classes in the variable width quantile cuts.



	Value
Variable	bio7
WW	0.7621
GLM	0.738

Figure 1: Output of WW2 on the Bradypus data set. (A) the best variable, (B) the AUC of the best model, (C) the membership function, and (D) the predicted distribution with the presence points in green.

Table 2: The lookup table for the membership function on the Bradypus data. Each row is a segment of the range of the variable (factors). The density of points in each segment in the background (g1) and the points where species occur (g2) is used to calculate the odds ration, and then membership in each segment. The width is the size of the segment interval

	factors	levels	g1	g2	odds	prob	width
1	(64,106]	1	0.13	0.31	2.44	0.71	42.00
2	(106,113]	2	0.11	0.22	1.94	0.66	7.00
3	(113,119]	3	0.11	0.16	1.47	0.59	6.00
4	(119,128]	4	0.10	0.12	1.21	0.55	9.00
5	(128,144]	5	0.10	0.10	1.00	0.50	16.00
6	(144,159]	6	0.09	0.04	0.47	0.32	15.00
7	(159,173]	7	0.09	0.04	0.47	0.32	14.00
8	(173,186]	8	0.10	0.01	0.09	0.08	13.00
9	(186,200]	9	0.08	0.00	0.00	0.00	14.00
10	(200,307]	10	0.09	0.00	0.00	0.00	107.00

3 Experiments

3.1 K-fold Validation

K-fold validation provides a robust evaluation of the accuracy of a method on independent data. The data was prepared by generating a data set using presample, and then labeling each row from one to 5. K-fold validation proceeds by sequentially holding back one fifth of the data each time for evaluation, and developing the model using the remaining four-fifths. Table 3 shows the results of the five-fold validation.

Table 3: Results of a five-fold validation of Bradypus model. WW is WW2 and GLM is generalised linear regression. the suffix tr is the AUC on the training set and te is the AUC on the test set.

	name	WWtr	GLMtr	WWte	GLMte
1	bio7	0.77	0.73	0.77	0.78
2	bio7	0.77	0.74	0.74	0.73
3	bio7	0.76	0.73	0.74	0.78
4	bio7	0.76	0.73	0.78	0.77
5	bio7	0.77	0.74	0.74	0.75
6	mean	0.77	0.73	0.75	0.76
7	sd	0.00	0.00	0.02	0.02

The accuracies were similar on test and training sets and between the WW2 and the GLM models, and the best variable *bio*7 was chosen consistently. The performance of WW2 not dissimilar to the GLM.

3.2 Multi-dimensional option

Table 4 lists the results for multidimensional models developed by combining the prediction of two or more variables with a fuzzy AND operator and then evaluating the resulting AUC.

Table 4: The AUC of environmental variables on multi-dimensional prediction of the Bradypus data: WW is WhyWhere and GLM is a third order generalised linear model.

	name	WW	GLM
1	bio7.biome	0.85	0.90
2	bio7	0.78	0.75
3	bio12	0.73	0.70
4	bio17	0.73	0.68
5	bio6	0.72	0.67
6	bio12.bio6	0.72	0.71
7	bio12.bio17	0.72	0.68
8	bio16	0.68	0.66
9	bio12.bio16	0.68	0.66
10	biome	0.67	0.65
11	bio5	0.64	0.61
12	bio1	0.62	0.59
13	bio8	0.59	0.57
14	bio12.bio5	0.52	0.72
15	bio7.bio8	0.49	0.71
16	limit	0.00	0.00

The best result is a combination of *bio*7 and *biome* variables (the AUC of WW=0.7969 and GLM=0.906). The variable *biome* is a categorical variable expressing ecosystem type, and so more like a habitat variable than a numeric climatic variable.

3.3 Alternate segmentation

We also evaluated some alternative methods of segmenting the response function, shown on Table 5: an even distribution of cuts over the range of the variable, distribution by quantile frequency, and an entropy optimizing method. The even method segments the variable evenly over the range. The quantile method segments the variable so that as far a possible each segment contains an equal number of data.

The number of segments is determined using the Freedman-Diaconis rule [5] for optimal binning of histograms. The entropy method uses the R package `discretization` and the routine cutPoints() that perform cuts for the Minimum Description Length Principle. This analysis is not a comprehensive evaluation of discretization method, but serves to validate the performance of the quantile approach in comparison of some readily available alternatives.

Table 5: Results of a five-fold validation of these data with other segmentation approaches: even, quantile and entropy.

	name	WWtr	GLMtr	WWte	GLMte	segment
1	mean	0.77	0.73	0.75	0.76	even
2	mean	0.77	0.73	0.75	0.76	quantile
3	mean	0.75	0.72	0.71	0.72	entropy

The segmentation methods that performs best on the test set are the even and quantile methods, and their performance is similar on the training set. The entropy method performed less well. This provides support for the quantile approach that was used in WW1, although further testing of segmentation approaches may lead to improvements.

3.4 Prediction on World Data

To demonstrate the system on a larger data set we use the Global Ecosystems Database, a set of 940 global data sets of environmental variables, previously prepared and used in the WW1. The multi-agency distribution of the original CD includes many groups of variables listed in [17] including satellite greening, monthly temperature and rainfall and many others in a range of different resolutions in raster and vector formats.

Table 6: The top ten variables in a single variable WhyWhere model of Bradypus using the 940 variarable World dataset. AUC is the WW AUC and BAUC is the AUC from GLM.

	name	AUC	BAUC
1	fnocwat	0.85	0.80
2	srzsoil	0.85	0.68
3	wrzsoil	0.85	0.58
4	wrroot	0.84	0.71
5	lcprc08	0.84	0.74
6	lwerr05	0.84	0.76
7	wrcla01	0.84	0.57
8	wrcla03	0.84	0.54
9	SALINITY_ANN_AVG	0.84	0.68
10	wrsil02	0.83	0.65

Table 6 lists the top 10 variables identified by the algorithm in predicting Bradypus on the World dataset. Out of 940 variables, the top variables were $fnocwat$: Navy Terrain Data - Percent Water Cover. The next three variables are soil classifications – $srzsoil$: Staub and Rosenzweig Zobler Soil Units, and Webb et al Soil Particle Size Properties Zobler Soil Types. The fifth variable is climatic: Leemans and Cramer August Precipitation (mm/month) which corresponds to the dry season in Amazonia. Note that the accuracy of the GLM is less than the WW2 in this case.

It might be inferred from this limited study that habitat features have greater predictive power than climatic variables over the region of distribution for this species. By way of interpretation, *Bradypus variegatus* does leave the trees in search of food and while it crawls along the forest floor poorly, it does swim well [6]. Its distribution may be closely linked to the flooded forest (defined as a seasonal inundation of the forest floor) facilitating access to other trees for food. The identification of soil classification may be indicative of the soils supporting a the flooded forest ecosystem.

Habitat variables are proximal causes of species which necessarily produce higher accuracy than the more distal climatic variables. It is well known that habitat features are crucial in identifying suitable areas of land for the conservation of threatened species, and due to the proximal relationship should be a more important determinant of species decline than distal factors such as climate change. This is not a definitive examination of determinants of Bradypus, but serves to illustrate the potential expositions available from this approach.

4 Discussion

This study evaluated the accuracy of a segmented model and algorithm in an updated version of `WhyWhere` against a generalized linear model, and also modelling species response to climate variables and accuracy on a large data set containing mixtures of continuous and categorical environmental data.

The accuracies were similar on the Bradypus data set and `WhyWhere` was superior on the large World dataset, attributed to selection of the best WW2 variables and handling of categorical as well as continuous variables. The benefit of WW2 are more accurate species prediction, and potential insight into proximal cause of the species occurrences. The results verify the findings of the previous version of the WhyWhere, showing progress that could be made in modelling species response to the environment by using segmented models of few variables mined from large databases of environmental variables.

The recoding of `WhyWhere` into R has greatly improved the programs' utility. Refinements to the algorithm reduce the steps in the species modelling workflow and support more efficient higher dimensional models using novel fuzzy set operators.

It is interesting to note that that the two variables identified in the multivariable mode are habitat variables (*biome*) and a climate variable (*bio*7). We hypothesize that climate and habitat factors are independent causal factors that together determine species distribution, and that the multi-dimensional WW2 can correctly identify such independent determinants of species response, yielding a parsimonious explanation of the species' response to its environment.

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.

References

[1] 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.

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

[3] M Dowle, T Short, S Lianoglou, and A Srinivasan. data.table: Extension of data.frame. 10 2014.

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

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

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

[7] 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.

[8] 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.

[9] Leo Joseph and David Stockwell. Climatic modeling of the distribution of some pyrrhura parakeets of northwestern south america with notes on their systematics and special reference to pyrrhura caeruleiceps todd, 1947. *Ornitologıa Neotropical*, 13:1–8, 2002.

[10] 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.

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

[13] 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.

[14] D Stockwell and D Peters. The GARP modelling system: problems and solutions to automated spatial prediction. *International Journal Of Geographical Information Science*, 13(2):143–158, 1999.

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

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

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

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

[19] Robert J. Hijmans & Jacob van Etten. raster: Geographic analysis and modeling with raster data, 2012. R package version 2.0-12.

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