-Manuscript Draft-

whywhere: An R package for ecological niche modelling on environmental big data

David R.B. Stockwell[*]

April 23, 2015

Abstract

This package is an R implemention of the WhyWhere data mining algorithm [6]. Developed for biodiversity modelling with big data issues in mind, the approach is simple and rigorous, efficient to compute, and provides accuracy equal to the best alternative approaches. It represents the way forward in utilize large spatial environmental data sets in a robust predictive and explanatory capacity.

1 Introduction

Niche modelling consists of developing models of environmental relationships from sparce point occurrence data such as sightings, museum records or other opportunisitc sources, and projecting these models into the 2D surface to produce a continuous map of the suitability of environments for the species. Such maps can then be used for a range of conservation purposes.

The geographic data sets used in niche modelling are diverse: characterised by a large number of themes, local or global coverag, and coarse to fine scales of resolution. As the size and availability of data sets expand there is a need to reexamine traditional approaches to data analysis and redesign the algorithms where necessary. A number of issues arise in developing a more streamlined, automated data modelling process:

1. Memory limitations: Most models are developed on an intermediate data stage in a 'wide' file format, i.e. with variables in columns and locations in rows. This is an addition step. All environmental data must be then be held in memory. Here we consider an algorithm needs only one data set at a time and so also supports model development on streaming spatial datasets.

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

2. Combining Variables: Higher dimensional models must first be coregistered in both projection and resolution before modelling. This adds additional processing. Either smaller data sets are enlarged which increases memory usage with duplicate cells, or larger data deleted which loses data. An ideal method would utilise the data at its native projection and resolution.

3. Mixed Types: Environmental data are either continuous (eg. temperature, rainfall) or categorical (eg. biome or soil type). Ideally a model would integrate both types, but very few do.

4. Non-linear responses: The response of species to environment is generally 'humped' around an ideal the environment may even be multi-modal. Thus models must be at least quadratic or have multiple forms, which often needs to be specified prior to analysis. An ideal method would be robust to non-linear response type.

5. Ecological interpretation: How does the structure of a multi-dimensional model embody the ecology of the species? Many model applications are based in other domains (eg. generalised linear model) without clear ecological interpretation.

The guide is arranged as follows. Section 2 describes background and related work on the criteria outlined above. Section 3 describes the algorithm and section 4 reports the results on a well known dataset on two environmental data sets: one a small set of 9 prepared spatial layers, and another a large set of 1000 global data-sets.

2 Related Work

A work flow for species modeling using spatial data and presence only records was developed in the GARP system [4]. Elements of this work flow such as random sampling of background points are now widely used. A newly developed package for species modelling (eg. MaxEnt, Bioclim, Domain, GLM, GAM, and RandomForest) has been collated in R `dismo`. Other R packages have made this modelling more accessible, particularly the `raster` package for handling gridded spatial data [7]. The R implementation of WhyWhere utilises a new package called `data.table` for fast aggregation of large data [1].

2.1 Arbitrary Distributions

In ecological niche theory a uni-modal or humped distribution of temperature or rainfall is the simplest viable distribution. However, these distributions are often skewed and can also be multimodal. In addition, many environmental variables are categorical variables such as soil and vegetation type. One approach to robust modelling is to provide a range of potenital response types [**?**].

The approach in WhyWhere is segmentation – to approximate non-linear responses with a discrete range. Both continuous and categorical variables can be handled in the same framework. The original WhyWhere [6] used a color quantization algorithm similar to the GIF format. The median cut algorithm attempts to assign equal numbers of points to a limited number of categories category [2] in the red, green, blue 3D space while retaining good appearance. Quantisation in WhyWhere supported a 3 variable conjunctive model. This is not a limitatio of the current version.

2.2 Evaluating Models

There are many ways to evaluate a model. The categorical nature of the predicted variable (ie. presences and absences) determined from models outputs in the range of zero to one, (eg. probablistic prediction of a rule, logistic model or descision tree) against a categorical value (presence or absence). The reciever operating curve (ROC) is widely used to compare such models, while the area under the curve of the receiver operating statistic (AUC) provides a single number estimate of model quality.

The locations where the species occurs can be thought of as a sampling of the environmental space. For a given data set - presence and absence - there is a count of values in the each category ($G1$). There is a lesser count of values in the sample of presence points ($G2$). The change in the distribution of counts from $G1$ to $G2$ indicates strength of the the response of the species to its environment - called the membership function M. A membership function describes a fuzzy truth value as a function $f : \mathbb{R} \to [0, 1]$ from a variable V to the real unit interval $[0, 1]$.

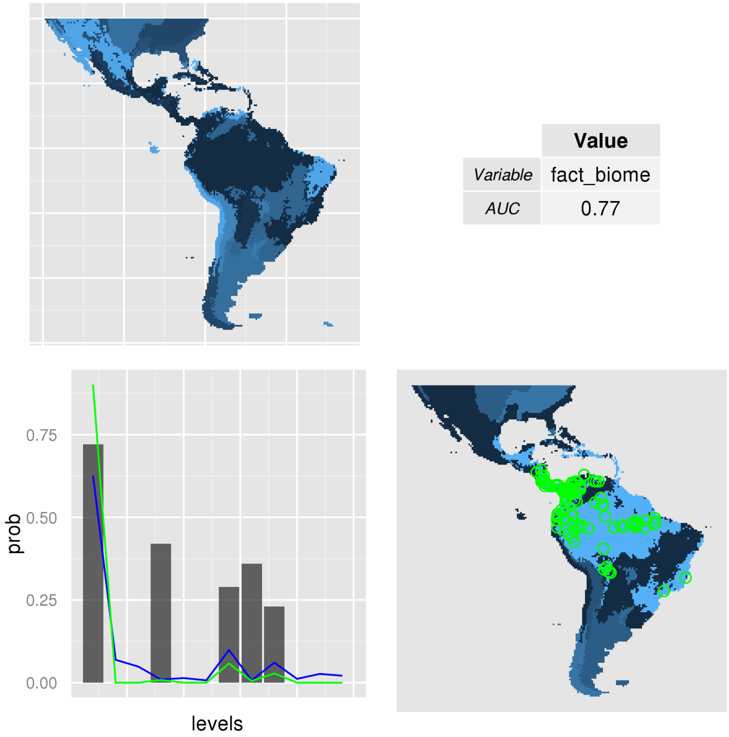
$$M = G1/G2 \tag{1}$$

Table 1 lists and example lookup table.

	factors	levels	g1	g2	prob
1	(-191,-9.83]	1	166	0	0.00
2	(-9.83,65]	2	169	0	0.00
3	(65,120]	3	141	112	0.79
4	(120,150]	4	164	28	0.17
5	(150,171]	5	174	67	0.39
6	(171,187]	6	183	88	0.48
7	(187,198]	7	151	61	0.40
8	(198,204]	8	198	157	0.79
9	(204,207]	9	192	143	0.74
10	(207,215]	10	109	85	0.78
11	(215,219]	11	165	105	0.64
12	(219,235]	12	187	154	0.82

Table 1: These are the results

```
> source("../R/ww2.R")
> files <- list.files(path="/home/davids99us/data/dismo",pattern='grd', full.names=TRUE )
> file <- paste(system.file(package="dismo"), '/ex/bradypus.csv',sep='')
> Pres <- fread(file,  header=T,sep=",")
> Pres$species=NULL
> result=ww(Pres,files,plot=FALSE,multi=TRUE,trim=FALSE)
> plot.ww(result)
```



	Value
Variable	fact_biome
AUC	0.77

Figure 1: WhyWhere predicted distribution of the Bradypus data set.

In any sense, M is similar to the conditional probability of species being present given the environmental value falls in a category. While many models attempt to represent the full probability using Bayes Theorem for example, the probability of the occurrence of a species $P(S)$ in opportunistic data is not well defined, and not generally of interest and dependent on season, search effort and other uncontrolled variables. The membership function is a proportional relationship which is sufficient for AUC to compare models.

2.3 Single Variables

The approach to evaluating the strength of the response in the original WhyWhere was significance with the Chi-squared test or a K-S test. However Chi2 doesn't work in evaluating the multivariate case, and there is another approach.

The output of the membership function for a single variable is vector of memberships for each location. We can combine the membership vectors for two variables using a fuzzy AND to produce a new membership vector. This can be evaluated with the AUC. The AND, OR operators on probability are Zadeh operators:

$$AND : x \wedge y = min(f(x), f(y))OR : x \wedge y = max(f(x), f(y)) \qquad (2)$$

Using this approach eliminates the need to develop and apply a high dimensional model to data.

2.4 Ecological Models

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 established ecological theory.

3 WhyWhere Algorithm

The inputs to `WhyWhere` are: a `data.table` with the longitude and latitude of known locations, and a list of environmental data files that may be read into the raster package. Parameters include **multi** for searching conjunctions of variables, **split** for split testing on train and test sets, **trim** to trim the spatial variables to the range of the location points.

Figure 1.A shows the highest rated variable in the Bradypus dataset, listed in panel Figure 1.B. The lookup table is illustrated as a bar graph in Figure 1.C with categorical ranges listed on the **x** axis. The prior distribution ($G1$) is the blue line, and the distribution of presences ($G2$) is the dashed line. The membership function is the grey bars ($G1/G2$). Note the almost uniform distribution of background classes due to the quantile cuts. The predicted distribution of Bradypus is on Figure 1.D.

The current algorithm implements a beam search in which a conjunction of each new variable is tested with the best variable so far. Alternaitve approaches to searching the space of conjunctions may be implemented in future.

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
```

The mask file defines the geographic extent for the sampling the background data. If absences are available then they need not be generated. On looping through the environmental variables, a combination of the variables with the next best variables by applying the minimum of the item probability vectors and recalculating the AUC. It is possible to monitor the progress of `WhyWhere` with the plot option. This plots out the best model sofar and prints out a list of the best models considered. This protocol would also support a streaming work flow.

4 Experiments

4.1 Local Data

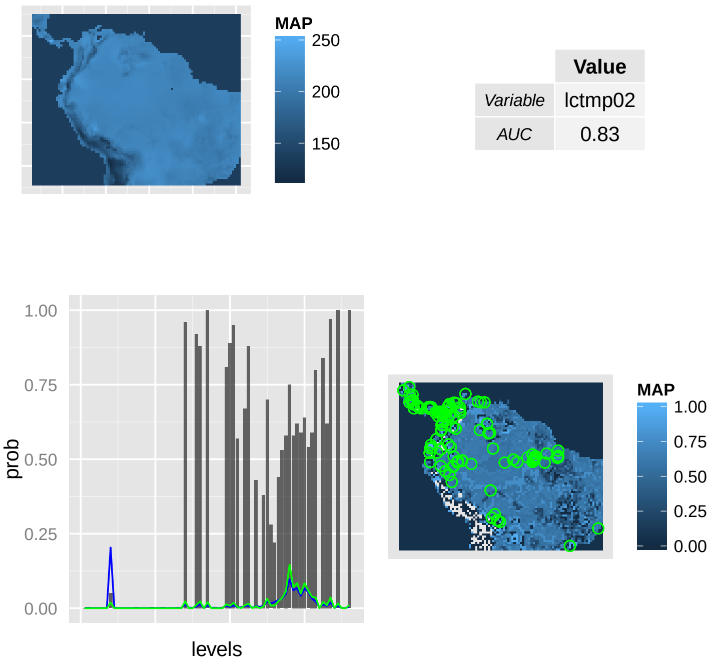
Data points for the feeding brown-throated three-toed sloth (**Bradypus variegatus**) are documented and applied to models in `dismo`. The environmental data consist of 9 environmental files covering the South American continent.

Figure 1 shows (A) the top variable, (B) the AUC of the top models, (C) the lookup table that is the basis of the model, and (D) the predicted distribution with the presence points plotted. In this distribution the result is very similar to the results from GARP and MaxEnt in paper [**?**]. Table 2 lists the results for all variables.

Table 3 shows the AUC of these data with other algorithms in [3]. The same protocol was used to ensure comparability. The models that perform best include geographic models but these are not in a split test protocol.

4.2 World Data

Figure 2 shows the Bradypus data applied to a dataset of 940 layers of world extent. These contain many groups of variables listed in [5] including satellite greening, monthly temperature and rainfall and many others. They are also of different resolution ranging from x to topo data sets with resolutions.



	Value
Variable	lctmp02
AUC	0.83

Figure 2: `WhyWhere` predicted distribution of using world data set.

	name	AUC	file
1	bio1.bio12.bio6	0.80	/home/davids99us/data/dismo/bio6.grd
2	bio1.bio12	0.78	/home/davids99us/data/dismo/bio12.grd
3	bio1	0.75	/home/davids99us/data/dismo/bio1.grd
4	fact_biome	0.75	/home/davids99us/data/dismo/fact_biome.grd
5	bio6	0.70	/home/davids99us/data/dismo/bio6.grd
6	bio8	0.70	/home/davids99us/data/dismo/bio8.grd
7	bio1.bio12.bio6.bio8	0.69	/home/davids99us/data/dismo/bio8.grd
8	bio7	0.64	/home/davids99us/data/dismo/bio7.grd
9	bio1.bio12.bio16	0.62	/home/davids99us/data/dismo/bio16.grd
10	bio1.bio12.bio17	0.62	/home/davids99us/data/dismo/bio17.grd
11	bio12	0.59	/home/davids99us/data/dismo/bio12.grd
12	bio5	0.56	/home/davids99us/data/dismo/bio5.grd
13	bio16	0.54	/home/davids99us/data/dismo/bio16.grd
14	bio1.bio12.bio6.fact_biome	0.53	/home/davids99us/data/dismo/fact_biome.grd
15	bio17	0.50	/home/davids99us/data/dismo/bio17.grd
16	bio1.bio12.bio6.bio7	0.50	/home/davids99us/data/dismo/bio7.grd
17	bio1.bio12.bio5	0.49	/home/davids99us/data/dismo/bio5.grd
18	limit	0.00	

Table 2: These are the results

Table 4 ists the top 5 results from the algorithm. We note the accuracy of these top variables is higher than shown in Table 2.

5 Conclusion and Further Work

The WhyWhere package provides a useful approach to exploring the interaction between spatial data and point locations. The changes to the original WhyWhere method implemented in this package can bring greater utility while maintaining the verifyable increases in accuracy from the big data approach. For example, by enabling processing on cloud based data sets WhyWhere will permit analysis of data sets that were not possible before. The new method reduces few steps in the whole work flow of data processing by simplifying development of more complex models. The R package `WhyWhere` is a useful packages to fit, plot and test empirical species as a conjunction of response functions. More complex logical expressions are planned, as are improvements in computational efficiency and access to cloud data.

References

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

	Method	AUC
1	Geographic	0.90
2	geoIWD	0.89
3	MaxEnt	0.87
4	Circles	0.84
5	Decision_Trees	0.83
6	WhyWhere	0.83
7	GLM	0.81
8	Mahalanobis	0.80
9	GARP	0.78
10	Convex_Hulls	0.74
11	Domain	0.73
12	Bioclim	0.66
13	SVM	0.62
14	voroniHull	0.50

Table 3: These are the comparative results

	name	AUC	file
1	lctmp02	0.83	/home/davids99us/data/Terrestrial/lctmp02.pgm
2	mev8710	0.83	/home/davids99us/data/Terrestrial/mev8710.pgm
3	lctmp01	0.82	/home/davids99us/data/Terrestrial/lctmp01.pgm
4	lctmp03	0.82	/home/davids99us/data/Terrestrial/lctmp03.pgm
5	mev8706	0.82	/home/davids99us/data/Terrestrial/mev8706.pgm

Table 4: These are the results

[2] 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.

[3] R.J. Hijmans, S. Phillips, J. Leathwick, and J. Elith. dismo: Species distribution modeling with r., 2011.

[4] 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.

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

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

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