bioRxiv preprint doi: https://doi.org/10.1101/027839; this version posted September 30, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Pre-print available from bioRxiv doi: http://dx.doi.org/10.1101/027839

AUTOMATED DISCOVERY OF RELATIONSHIPS, MODELS AND PRINCIPLES IN ECOLOGY

Pedro Cardoso[1,2,*], Paulo A.V. Borges[2], José Carlos Carvalho[2,3], François Rigal[2], Rosalina Gabriel[2], José Cascalho[4,5], Luís Correia[5]

[1]*Finnish Museum of Natural History, University of Helsinki, P.O.Box 17 (Pohjoinen Rautatiekatu 13), 00014 Helsinki, Finland.*

[2]*CE3C – Centre for Ecology, Evolution and Environmental Changes / Azorean Biodiversity Group and Universidade dos Açores - Departamento de Ciências Agrárias, Rua Capitão João d'Ávila, 9700-042 Angra do Heroísmo, Açores, Portugal.*

[3]*Department of Biology, CBMA – Molecular and Environmental Centre, University of Minho, Braga, Portugal.*

[4]*NIDes - Núcleo de Investigação e Desenvolvimento em e-Saúde, Universidade dos Açores, Portugal*

[5]*BioISI – Biosystems and Integrative Sciences Institute, Faculdade de Ciências da Universidade de Lisboa, Campo Grande, 1749-016 Lisboa, Portugal.*

**Corresponding author. E-mail: pedro.cardoso@helsinki.fi, Tel: (+358) 294128854.*

ABSTRACT

Ecological systems are the quintessential complex systems, involving numerous high-order interactions and non-linear relationships. The most commonly used statistical modelling techniques can hardly reflect the complexity of ecological patterns and processes. Finding hidden relationships in complex data is now possible through the use of massive computational power, particularly by means of Artificial Intelligence (AI) methods, such as evolutionary computation.
Here we use symbolic regression (SR), which searches for both the formal structure of equations and the fitting parameters simultaneously, hence providing the required flexibility to characterize complex ecological systems. First, we demonstrate how SR can deal with complex datasets for: 1) modelling species richness; and 2) modelling species spatial distributions. Second, we illustrate how SR can be used to find general models in ecology, by using it to: 3) develop new models for the interspecific abundance-occupancy relationship; 4) develop species richness estimators; and 5) develop the species-area relationship and the general dynamic model of oceanic island biogeography. All the examples suggest that evolving free-form equations purely from data, often without prior human inference or hypotheses, may represent a very powerful tool for ecologists and biogeographers to become aware of hidden relationships and suggest general theoretical principles.

Keywords: abundance-occupancy relationship, artificial intelligence, evolutionary computation, genetic programming, species richness estimation, species-area relationship, species distribution modelling, symbolic regression.

bioRxiv preprint doi: https://doi.org/10.1101/027839; this version posted September 30, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Pre-print available from bioRxiv doi: http://dx.doi.org/10.1101/027839

INTRODUCTION

Ecology as a complexity science

Complexity is a term often used to characterize systems with numerous components interacting in ways such that their collective behaviour is difficult to predict, but where emergent properties give rise to, more or less simple but seldom linear, patterns (Box 1; Holland 1995; Mitchell 2011). Complexity science is therefore an effort to understand non-linear systems with multiple connected components and how "the whole is more than the sum of the parts" (Holland 1998). Biological systems probably are among the most complex (Sole & Goodwin 2000), and among them, ecological systems are the quintessential complex systems (Anand et al. 2010). These are composed of individuals, populations from different species, interacting and exchanging energy in multiple ways, furthermore relating with the physical environment at different spatial and temporal scales in non-linear relationships. As a consequence, ecology is dominated by idiosyncratic results, with most ecological processes being contingent on the spatial and temporal scales in which they operate, which makes it difficult to identify recurrent patterns, knowing also that pattern does not necessarily identify process (Lawton 1996; Dodds et al. 2009; Passy 2012). The most commonly used exploratory (e.g. PCA, NMDS) and statistical modelling techniques (e.g. linear and non-linear regression) can hardly reflect the complexity of ecological patterns and processes, often failing to find meaningful relationships in data. For ecological data, we require more flexible and robust analytical methods, which can eventually lead to the discovery of general principles and models.

Box 1 - Glossary of terms.

Artificial intelligence (AI) - A scientific field concerned with the automation of activities we associate with human thinking (Russell & Norvig 2010).

Big data - Very large amount of structured or unstructured data, hard to model with general statistical techniques but with the potential to be mined for information.

Complex system (CS) - A system in which a large network of components organize, without any central controller and simple although non-linear rules of operation, into a complex collective behaviour that creates patterns, uses information, and, in some cases, evolves and learns (Mitchell 2011).

General model - An equation that is found to be useful for multiple datasets, often but not necessarily, derived from a general principle. In most cases the formal structure of equations is kept fixed, while some parameters must be fitted for each individual dataset.

General principle - Refers to concepts or phenomenological descriptions of processes and interactions (Evans et al. 2013). May not have direct translation to any general model, but be a purely conceptual abstraction.

Genetic programming (GP) - A biologically-inspired method for getting computers to automatically create a computer program to solve a given problem (Koza 1992). It is a type of evolutionary algorithm, where each solution to be tested (individual in a population of possible solutions) is a computer program.

Pareto front - A curve connecting a set of best solutions in a multi-objective optimization problem. If several conflicting objectives are sought (e.g. minimize both error and complexity of formulas), the Pareto front allows visualizing the set of best solutions.

Symbolic regression (SR) - A function discovery approach for modelling of multivariate data. It is a special case of genetic programming, one where possible solutions are equations instead of computer programs.

bioRxiv preprint doi: https://doi.org/10.1101/027839; this version posted September 30, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Pre-print available from bioRxiv doi: http://dx.doi.org/10.1101/027839

General principles and models in ecology

The ultimate aim of any ecological principle is to provide a robust model for exploring, describing and predicting ecological patterns and processes regardless of taxon identity and geographic region (Lawton 1996; Dodds 2009). Finding a recurrently high goodness-of-fit for a model to an ecological pattern for most taxa and ecosystems is usually the most compelling evidence of a mechanistic process controlling that pattern. When general principles are translated into robust models, general statistical methods are mostly abandoned in favor of these. Some good examples of such models are the population growth model, the Lotka-Volterra model, the lognormal model of species abundance distributions (SADs), the interspecific abundance-occupancy relationship (IAOR), a variety of species richness estimators, the species-area relationship (SAR) and the general dynamic model of oceanic island biogeography (GDM) (Box 2). In all cases, general principles gave origin to general, widely applicable, equations mostly found by intellectual *tour de force*. Yet, they surely are only the tip of the iceberg, usually incorporating few of the variables increasingly available to ecologists and that could potentially explain such patterns.

Box 2 - Examples of general principles in ecology and of some of the respective statistical models.

General principle	Branch of ecology	General model	References
Population growth would be exponential if not limited by food availability, predation or other external factors. If this limitation is driven by resource availability (e.g. food), growth is capped by carrying capacity.	**Population dynamics**	**Logistic population growth** $\frac{\partial N}{\partial t} = \frac{rN(K-N)}{K}$ N = population size t = time r = rate of maximum population growth K = carrying capacity	Verhulst 1845
The fluctuations in predator abundances are strongly dependent on prey abundances and vice-versa, with a delay in time reflecting the delayed response of each species to the other species' abundance.	**Predator-prey interaction**	**Lotka-Volterra models** $\frac{\partial x}{\partial t} = Ax - Bxy$ $\frac{\partial y}{\partial t} = -Cy + Dxy$ x = abundance of prey y = abundance of predator A = growth rate of prey population B = death rate of prey due to interaction C = death rate of predator population D = growth rate of predator due to interaction	Lotka 1910, 1925; Volterra 1926
Natural communities are typically characterized by few common and many rare species.	**Species abundance distribution (SAD)**	**Probability function for log-normal distribution** $P(x) = \frac{1}{\sigma\sqrt{2\pi}x} e^{-(lnx-\mu)^2/2\sigma^2}$ x = log-normally	Preston 1948

bioRxiv preprint doi: https://doi.org/10.1101/027839; this version posted September 30, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Pre-print available from bioRxiv doi: http://dx.doi.org/10.1101/027839

		distributed abundance of species μ = average abundance σ = standard deviation	
Locally abundant species tend to be widespread while locally rare species tend to be narrowly distributed. That is, for a specific species assemblage, there is a positive interspecific abundance-occupancy relationship (Brown 1984).	**Interspecific Abundance-Occupancy Relationship (IAOR)**	**Linearization model** $logit(p) = a + b\log(\mu)$ **The exponential model** $p = 1 - e^{\alpha\mu^{\beta}}$ **The negative binomial distribution model** $p = 1 - (1 + \frac{\mu}{k})^{-k}$ p = occupancy μ = average abundance across sites k, α, β, a and b are constants.	Nachman 1981; He & Gaston 2000
Sampling complete communities usually is not instant. The few abundant species are sampled first and rare species slowly accumulate with time spent or samples added. This leads to asymptotic sampling accumulation curves and decreasing proportions of species being represented by few individuals (Colwell & Coddington 1994).	**Sampling theory**	**Clench function** $S_{obs} = \frac{aQ}{1 + bQ}$ $S^* = a/b$ **Negative exponential** $S = a(1 - e^{-bQ})$ $S^* = a$ **Rational function** $S = \frac{a + bx}{1 + cx}$ $S^* = a/b$ **Chao estimators** $S^* = S_{obs} + \frac{S_1^2}{2S_2}$ S_{obs} = observed richness S* = estimated richness S_1 = singletons or uniques S_2 = doubletons or duplicates Q = number of samples a, b = fitting parameters	Clench 1979 Chao 1984, 1987 Miller & Wiegert 1989 Ratkowski 1990 Soberón & Llorente 1993
All else being equal, larger areas support more species, as they usually have higher carrying capacity, habitat diversity, environmental heterogeneity and geographical barriers.	**Species-area relationship (SAR)**	**Power model** $S = c + A^z$ **Exponential model** $S = c + zlogA$ **Linear model** $S = c + zA$	Arrhenius 1920, 1921; Gleason 1922

bioRxiv preprint doi: https://doi.org/10.1101/027839; this version posted September 30, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Pre-print available from bioRxiv doi: http://dx.doi.org/10.1101/027839

		S = species richness A = area c = intercept z = rate of richness increase with area	
Oceanic islands, being of volcanic origin, have a life-cycle of birth (emergence), growth (continued volcanism), decline (erosion), and death (submergence). This ontogeny is reflected in each island's carrying capacity and diversity being hump-shaped when plotted against time (i.e. island maximum geological age).	**Island biogeography**	**General Dynamic Model of Oceanic Island Biogeography** $S = c + zlogA + xT + yT^2$ $logS = c + zlogA + xT + yT^2$ $logS = c + zlogA + xlogT + ylogT^2$ T = maximum age of island	Whittaker et al. 2008 Fattorini 2009 Steinbauer et al. 2013

Computing power applied to complex ecological systems

The automation of techniques for collecting and storing ecological and related data, with increasing spatial and temporal resolutions, has become one of the central themes in ecology and bioinformatics. Yet, automated and flexible ways to synthesise such complex and big data were mostly lacking until recently. Finding hidden relations within such data is now possible through the use of massive computational power. New computer-intensive methods have been developed or are now available or possible (e.g. Reshef et al. 2011) including in particular the broad field of Artificial Intelligence (AI) which has produced a variety of approaches. AI includes a series of evolution-inspired techniques, brought together in the sub-field of evolutionary computation, of which the most studied and well-known probably are genetic algorithms (GA; Holland 1975). Genetic programming (GP), namely in the form of symbolic regression (SR; Koza 1992), is a particular derivation of GAs that searches the space of mathematical equations without any constraints on their form, hence providing the required flexibility to represent complex systems as presented by many ecological systems. Contrary to traditional statistical techniques, symbolic regression searches for both the formal structure of equations and the fitting parameters simultaneously (Schmidt & Lipson 2009). Finding the structure of equations is especially useful to discover general models, providing general insights into the processes and eventually leading to the discovery of new and as yet undiscovered principles. Fitting the parameters provides insight into the specific data, and allow specific predictions.

So far, symbolic regression has seldom been used in ecology. Yet, successful examples include modelling of land-use change (Manson 2005; Manson & Evans 2007), effects of climate change on populations (Tung et al. 2009; Larsen et al. 2014), community distribution (Larsen et al. 2012, Yao et al. 2014), pollution effects on micro-organismal blooms (Muttil & Lee 2005; Jagupilla et al. 2015), deriving vegetation indices (Almeida et al. 2015) and using parasites as biological tags (Barret et al. 2005). SR has also been found to be very useful in many other fields, with results competitive with those produced by humans (see Koza 2010; Graham et al. 2013 for examples).

In this work we explain, test and demonstrate the usefulness of SR in uncovering hidden relationships within typical ecological datasets. First, we demonstrate how SR can deal with complex datasets, namely for: 1) modelling species richness; and 2) modelling species spatial distributions. Second, we illustrate how SR can be used to find general models in ecology, by using it to: 3) develop new models for the interspecific abundance-occupancy relationship (IAOR); 4) develop species richness

bioRxiv preprint doi: https://doi.org/10.1101/027839; this version posted September 30, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Pre-print available from bioRxiv doi: http://dx.doi.org/10.1101/027839

estimators; and 5) develop the species-area relationship (SAR) and the general dynamic model of oceanic island biogeography (GDM). We necessarily had to limit our analyses to a narrow number of examples, but countless other options could be used for demonstration purposes.

We compare the performance of SR to equivalent linear and non-linear model analyses using data from two systems, Macaronesia (mainly the Azores archipelago) and mainland Portugal, and two contrasting taxa, arthropods (mainly spiders) and bryophytes. These systems and taxa were chosen due to our familiarity with them and ability to immediately evaluate solutions derived from the SR analyses, a fundamental step with this approach (see discussion about the need for human inference). We should finally note that our objective is not to rigorously test every case in depth or advocate for the newly found models, as each of the case-studies certainly deserves a separate work using multiple datasets and covering several nuances of the different methods. Our objective is to compare the most commonly used, tested and validated tools for each case with SR, showing the latter's value and advantage over other tools in multiple situations when the objective is to unveil hidden relationships, models and principles in ecology.

METHODS

Symbolic regression

Symbolic regression (SR) may be used to model a response variable, usually numeric (e.g. richness or abundance of species) but possibly categorical (e.g. presence/absence), in function of several numerical, ordinal and/or categorical explanatory variables. SR works as a computational parallel to the evolution of species (Fig. 1), although a rather simplified and often liberal form of evolution (Correia 2010). A population of initial equations is generated randomly by combining different building blocks, such as the variables of interest (independent explanatory variables), algebraic operators (+, –, ÷, ×), analytic function types (exponential, log, power, etc.), constants and other ways to combine the data (e.g. Boolean or decision operators). Being random, these initial equations almost invariably fail, but some are slightly better than others. All are then combined through crossover ("sexual reproduction"), giving rise to new, on average, improved equations, with characteristics from both parents. The evolution towards our goal is guaranteed by equations with better fitness (e.g. higher r^2) having a higher probability of recombining and being parents of the next generation of equations. To avoid new equations being bounded by initially selected building blocks or quickly losing variability along the evolutionary process, a mutation step (acting on any building block) with a given, usually low, probability is added to the process after crossover. After multiple generations, an acceptable level of accuracy by some of the equations is often attained and the researcher stops the process.

bioRxiv preprint doi: https://doi.org/10.1101/027839; this version posted September 30, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Pre-print available from bioRxiv doi: http://dx.doi.org/10.1101/027839

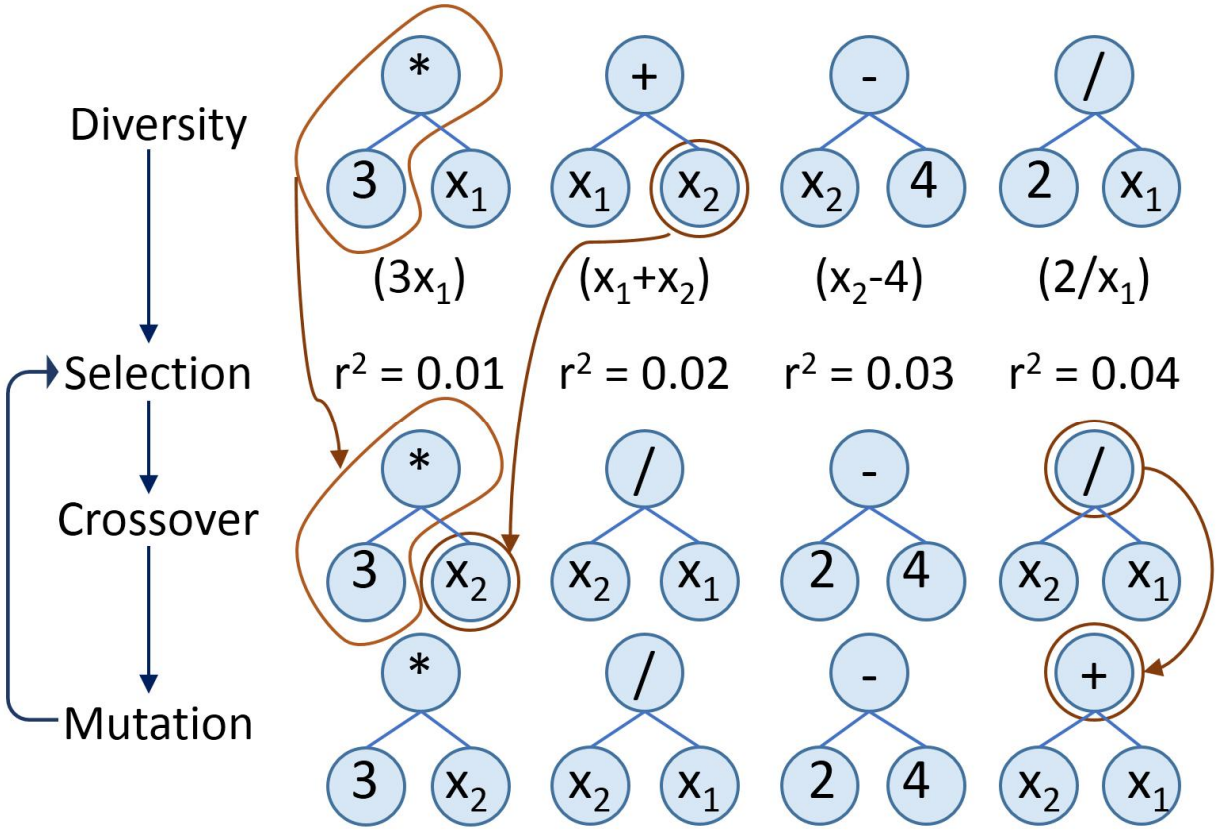

Fig. 1 – Schematic representation of the symbolic regression workflow. The basic representation is a parse-tree where building blocks such as variables (in this case: x_1, x_2), parameters (integers or real numbers) and operators (e.g. +, –, ×, ÷) are connected forming functions (in parenthesis under the first line of trees). Initial equations are generated by randomly linking different building blocks. Equations are combined through crossover, giving rise to new equations with characteristics from both parents (arrows linking the first and second rows of trees). Equations with better fitness (e.g. r^2) have higher probabilities of recombining. To avoid loss of variability, a mutation step is added after crossover (arrows linking the second and third rows of trees). After multiple generations, evolution stops and a set of free-form equations best reflecting the input data is found.

For this work we used the software Eureqa (Schmidt & Lipson 2014). For each run, the software outputs a list of equations along an error/complexity Pareto front, with the most accurate equation for each level of complexity being shown (Fig. 2). The Pareto front often presents an "elbow", where near-minimum error meets near-minimum complexity. The equation in this inflection is closer to the origin of both axes and is a good starting point for further investigation – if both axes are in comparable qualitative scales. Often, however, this inflection point is not obvious. In such cases, using indices that positively weight accuracy and negatively weight complexity, such as Akaike's Information Criterion and its modification for finite sample sizes (AIC and AICc respectively, Akaike 1974) may be warranted (Burnham & Anderson 2002; Johnson & Omland 2004).

bioRxiv preprint doi: https://doi.org/10.1101/027839; this version posted September 30, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Pre-print available from bioRxiv doi: http://dx.doi.org/10.1101/027839

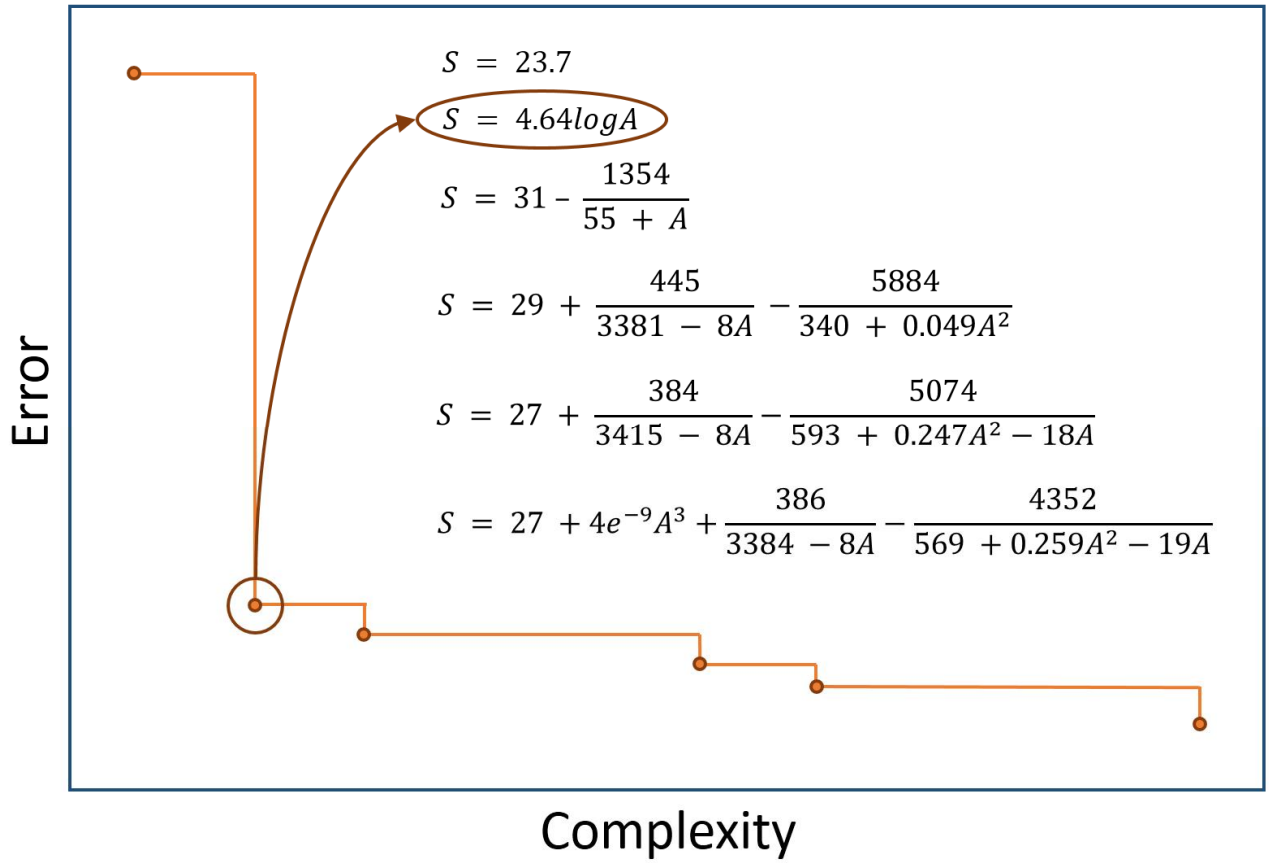

Fig 2. – Example of a Pareto front depicting error vs. complexity in the case of a symbolic regression search of the best species–area relationship for native spiders in the Azores (Portugal). The second formula is clearly the most promising, with both high accuracy (low error) and low complexity.

Case-studies

Modelling species richness

Modelling and mapping the species richness (or other diversity descriptors) of high diversity taxa at regional to large scales is often impossible without some kind of extrapolation from sampled to non-sampled sites. This is usually done correlating environmental or other variables to richness in known sites and estimating the expected richness in the entire region of interest. Here, we used an endemic arthropod dataset collected in Terceira Island, Azores. Fifty-two sites were sampled using pitfall traps for epigean arthropods (more details in Cardoso et al. 2009), 13 in each of four land-use types: natural forest, exotic forest, semi-natural pasture and intensively managed pasture. This dataset was randomly divided into training and test data, each including 26 sites (50%). Using the training dataset, we tried to explain and predict species richness per site using as independent variables elevation, slope, annual average temperature, annual precipitation and an index of disturbance (Cardoso et al. 2013).

As the response variable was count data, the most common way to approach this question is through Generalized Linear Models (GLM) with a Poisson error structure with log link. We used the package MuMIn (Barton 2015) and the R environment (R Development Core Team 2015) for multi-model inference based on AICc values, using all variables plus all possible interactions. For the SR search we used only algebraic and analytic operators (+, –, ÷, ×, log, power), in this and all examples below, so that outputs could be most easily interpreted. The r^2 goodness of fit was used as the fitness measure for evolving equations. As there was no clearly best formula, AICc was used to choose a single equation along the Pareto front (Appendix S1). Both r^2 and AICc were used to compare GLM with SR on the test dataset. Here and in subsequent analyses, all models with a ΔAICc value < 2 (the

bioRxiv preprint doi: https://doi.org/10.1101/027839; this version posted September 30, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Pre-print available from bioRxiv doi: http://dx.doi.org/10.1101/027839

difference between each model's AICc and the lowest AICc) were considered as receiving equal statistical support.

Modelling species distributions

Because most sites remain non-sampled for most species, species distribution modelling (SDM) is widely used to fill gaps in our knowledge on individual species distributions. One of the general statistical methods used for SDM is logistic regression, i.e., GLM with a with binomial error structure and logit link. This method, although easy to apply, usually does not perform well compared with other methods (e.g. Elith et al. 2006). Among the multiple alternatives, the principle of maximum entropy (Maxent; Phillips et al. 2006) has been found to be particularly robust (Elith et al. 2006) and the existence of a user friendly software package (https://www.cs.princeton.edu/~schapire/maxent/) has contributed to its widespread use during the latter decade.

We modelled the potential distribution of two endemic Azorean species in Terceira Island: the rare forest click-beetle *Alestrus dolosus* (Coleoptera, Elateridae) and the abundant but mostly forest restricted spider *Canariphantes acoreensis* (Araneae, Linyphiidae). The software Maxent was set to default settings. Given the intrinsic differences between methods, we had to use different background datasets. Maxent used the environmental maps of the islands with a resolution of 100 m, from where it extracted pseudo-absences. We then converted the probabilistic potential distribution maps to presence/absence using the maximum value of training sensitivity plus specificity as the threshold (as recommended by Liu et al. 2005). Logistic regression and SR used only presence/absence data from the 52 sampled sites. We used the package MuMIn (Barton 2015) and the R environment (R Development Core Team 2015) for multi-model inference of logistic regression based on AICc values. In the SR run, to reach a binary classification, a step function was included, so that the algorithm was looking for equations where positive and negative values were converted to presence and absence, respectively. Absolute error, reflecting the number of incorrect classifications, was used as the fitness measure. As inflection points of the Pareto fronts were clear, the best SR formula for each species was chosen based on them (Appendix S1). In all cases only the training data (26 sites) were used for running the models. Logistic regression, Maxent and SR were compared in their performance for predicting presence and absence of species on the 26 test sites using sensitivity, specificity and the True Skill Statistic - TSS (Alouche et al. 2006).

Developing new models for the interspecific abundance-occupancy relationship (IAOR)

There is a general positive relationship between distribution (or occupancy) and mean local abundance of species (Brown 1984; Gaston et al. 1997). This positive relationship is due to the fact that as species increase locally in abundance, they tend to occupy more sites, generating a positive interspecific abundance-occupancy relationship (IAOR) that can be explained by several ecological processes (see Gaston et al. 1997 for a review). Different models were proposed to describe the IAOR (Box 2). The ones found to be more general are the linearization, exponential and negative binomial distribution models (Box 2).

Our objective was to rediscover these relationships or find new models that outperformed them. The training dataset was from the same project as above, including 101 native and introduced arthropod species sampled in 52 sites. Abundances of each species per site were recorded. As an independent test dataset we used bryophytes sampled in Terceira Island, Azores, including 92 species sampled in 19 sites (see Gabriel & Bates, 2005). The dependent variable was either the proportion of sites occupied by each species or its logit transformation. The independent variable was either the average abundance of each species over all sites or the log of this value. The r^2 value was used as the fitness measure. The best equations found by SR were chosen based on the inspection of the Pareto front (Appendix S1), looking also for interpretability of the different models. The best models were then compared with the existing models using AICc, both with training and testing data, using the R package BAT (Cardoso et al. 2015).

bioRxiv preprint doi: https://doi.org/10.1101/027839; this version posted September 30, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Pre-print available from bioRxiv doi: http://dx.doi.org/10.1101/027839

Developing species richness estimators

Several asymptotic functions have been used to estimate species richness (Soberón & Llorente 1993). Among the proposed equations, three are often used (Box 2): the Clench function (Clench 1979), the negative exponential function and the rational function (Ratkowski 1990). These models are usually the best performing asymptotic functions for estimating unseen diversity, although often far from performing as well as non-parametric estimators (Cardoso et al. 2008a, b).

Our objective was to rediscover or eventually find asymptotic models that would outperform them. Two independent datasets were used for training and testing. Both were data resulting from exhaustive sampling for spiders in 1ha plots, performed by 8 collectors during 320 hours of sampling in a single hectare using five different methods. The training dataset was from a mixed forest in Gerês (northern Portugal) and the test dataset was from a *Quercus* forest in Arrábida (southern Portugal) (see Cardoso et al. 2008a, b for details).

Randomized accumulation curves for both sites were produced using the R package BAT (Cardoso et al. 2015: the package also includes both datasets). The true diversity of each site was calculated as the average between different non-parametric estimators (Chao 1 and 2, Jackknife 1 and 2). Because the sampled diversity in the training dataset reached a very high completeness but we wanted to simulate typically very incomplete sampling, datasets with 10, 20, 40, 80 and 160 randomly chosen samples were extracted and used, in addition to the complete 320 samples dataset, as independent runs in SR. Squared error was used as the fitness measure. Additionally, we imposed a strong penalty to non-asymptotic functions, although these were still used in the search process to optimize it (simply rejecting such functions would result in a much slower search). The true diversity value was not used for the training in any way, the SR process was blind to it and only looking for the best fitting asymptotic functions to the different accumulation curves. All asymptotic functions were compared in their accuracy when fitted to the test dataset. The weighted and non-weighted scaled mean squared errors implemented in BAT (Cardoso et al. 2015) were used as accuracy measures.

Developing the species-area relationship (SAR) and the general dynamic model of oceanic island biogeography (GDM)

One of the most studied examples of SARs is their application to island biogeography (ISAR; e.g. Darlington 1957; MacArthur & Wilson 1967). The shape of ISARs has been modelled by many functions, but three of the simplest seem to be preferred in most cases (Triantis et al. 2012). The most commonly used was the first to be proposed, the power model (Arrhenius 1920, 1921). Almost as longstanding and ubiquitous is the exponential model (Gleason 1922), and in many cases the simplest linear model may also be verified (Box 2).

The general dynamic model of oceanic island biogeography was proposed to account for diversity patterns within and across oceanic archipelagos as a function of age and area of the islands (Whittaker et al. 2008). Several different equations have been found to describe the GDM model, extending the different SAR models with the addition of a polynomial term using island age and its square (TT^2), depicting the island's ontogeny. The first to be proposed was an extension of the exponential model (Box 2; Whittaker et al. 2008), the power model extensions following shortly after (Fattorini 2009; Steinbauer et al. 2013).

Our objective was to test if we could re-discover and eventually refine existing models for the ISAR and GDM from data alone. We used the Azores and Canary Islands spiders (Appendix S2, Table S1; Cardoso et al. 2010) as training data in Eureqa. To independently test the generality of models arising from spider data, we used bryophyte data from the same archipelagos (Appendix S2, Table S1; Aranda et al. 2014). The area and maximum time since emergence of each island were used as explanatory variables and the native species richness per island as the response variables. The r^2 value was used as the fitness measure. It should be noted that due to the young age of its islands, the Azores have been found to not provide good fit to the GDM, contrary to the Canary Islands (Whittaker et al.

bioRxiv preprint doi: https://doi.org/10.1101/027839; this version posted September 30, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Pre-print available from bioRxiv doi: http://dx.doi.org/10.1101/027839

2008; Cardoso et al. 2010). The best SAR and GDM equations found by SR were chosen based on the inspection of the Pareto front (very clear in all cases, Appendix S1), but looking also for interpretability of the models. These were then compared with the existing models using AICc, both with training and testing data, using the R package BAT (Cardoso et al. 2015).

RESULTS

Modelling species richness

For Terceira Island arthropod richness, the model selected by GLM was:

$$S = e^{5.381 + 0.003432H - 0.001904P - 0.05257D}$$

($r^2 = 0.744$, AICc = 30.793), where H = altitude, P = precipitation and D = disturbance. Yet, the GLM model seems to be overfitting to the training data, as the results with the test data were considerably worse ($r^2 = 0.146$, AICc = 63.672). The SR results performed worse than GLM with the training data, with the formula chosen according to AICc being:

$$S = 0.673 + (8.696 - 0.002P)^{0.006H - 2.461}$$

($r^2 = 0.641$, AICc = 43.050). However, the SR equation performs considerably better than GLM with the test data ($r^2 = 0.289$, AICc = 62.354), revealing a higher generality of this formula. Despite the completely different structure, both GLM and SR formulas indicate similar importance of variables, with a positive effect of altitude and negative effect of precipitation on species richness. However, the disturbance index features only in the GLM solution.

Modelling species distributions

The potential distribution models are relatively similar for *Canariphantes* but show marked differences for *Alestrus* (Fig. 3). Symbolic regression outperforms both other models for *A. dolosus* and is as good as Maxent for *C. acoreensis*, with both outperforming LR (Table 1). The SR models are not only the best, presenting maximum values for TSS, but are also the easiest to interpret. *Alestrus* is predicted to have adequate environmental conditions in all areas above 614m elevation, being restricted to pristine native forest. *Canariphantes* can potentially be present in all areas with disturbance values below 41.3, occurring not only in native forest but also in adjacent semi-natural grassland and humid exotic forest. The LR and Maxent models used a large number of explanatory variables for *Alestrus*, yet performed worse on the test data than did SR. Expert opinion of PAVB, based on decades of fieldwork by our team, supports the TSS rankings.

bioRxiv preprint doi: https://doi.org/10.1101/027839; this version posted September 30, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Pre-print available from bioRxiv doi: http://dx.doi.org/10.1101/027839

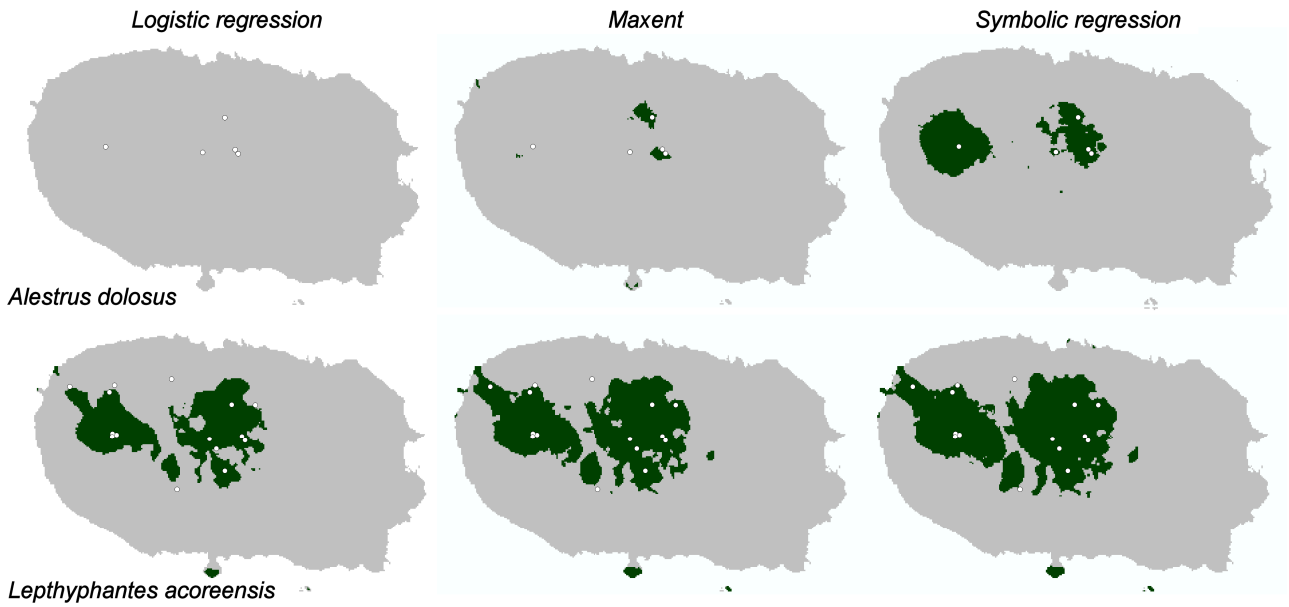

Fig. 3 – Observed locations (white dots) and predicted distribution (dark areas) of two endemic arthropod species in the island of Terceira (Azores, Portugal) using three different modelling methods.

Table 1 – Species distribution models for two endemic arthropod species on the island of Terceira (Azores, Portugal) and respective accuracy statistics on an independent test dataset (TSS = True Skill Statistic. *H* = altitude, *Sl* = slope, *T* = average annual temperature, *P* = annual precipitation and *D* = disturbance index. The step function in symbolic regression converts positive values inside parentheses to presence and negative values to absence. Best values in bold.

Model	Formula	Sensitivity	Specificity	TSS
Alestrus dolosus				
Logistic regression	$\frac{1}{1+e^{-(8469-0.432P-540.7T)}}$	0	**1**	0
Maxent	Uses all variables but *Sl*, main is *D* (contribution = 74.1%)	0.5	**1**	0.5
Symbolic regression	$step\ (H - 614)$	**1**	0.75	**0.75**
Canariphantes acoreensis				
Logistic regression	$\frac{1}{1+e^{-(3.617-0.103D)}}$	0.667	**0.7**	0.367
Maxent	Uses only *D* (contribution = 100%)	**0.833**	0.65	**0.483**
Symbolic regression	$step\ (41.3 - D)$	**0.833**	0.65	**0.483**

Developing new models for the interspecific abundance-occupancy relationship (IAOR)

For the Azorean arthropod training data, the best fitting model among the 3 selected from the literature (both highest r^2 and lowest AICc) for the IAOR was the exponential model (Table 2). The first SR run, using raw occupancy and abundance values, discovered a simpler, yet not as powerful model:

$$p = \frac{\mu}{a + \mu}$$

bioRxiv preprint doi: https://doi.org/10.1101/027839; this version posted September 30, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Pre-print available from bioRxiv doi: http://dx.doi.org/10.1101/027839

where *a* is a fitting parameter. This model is in fact a special case of the asymptotic Clench function used for estimating species richness (Box 2), where the asymptote, representing maximum occupancy, is 1. The SR model for IAOR represents a case in which a small initial increase of abundance causes a large increase in occupancy. The second SR run, using logit(p) and log(μ) rediscovered the linearization model (Appendix S1). When all four models were applied to the testing dataset the results were different, with the SR model being worst (Table 2). For bryophytes, the best model follows a negative binomial, suggesting that resource limitation for the mean local abundance works differently for bryophytes and arthropods.

Table 2 – Interspecific abundance-occupancy relationship (IAOR) models for Azorean taxa. p = proportion of sites occupied by species, μ = average abundance across all sites. Best values in bold.

Model	Formula	r^2	AICc
Arthropods (training)			
Linearization	$logit(p) = -1.129 + 0.721 * \log(\mu)$	0.875	-563.890
Exponential	$p = 1 - e^{-0.266\mu^{0.682}}$	**0.894**	**-580.438**
Negative Binomial	$p = 1 - (1 + \frac{\mu}{0.198})^{-0.198}$	0.809	-523.608
SR	$p = \frac{\mu}{3.629 + \mu}$	0.863	-557.291
Bryophytes (testing)			
Linearization	$logit(p) = -0.856 + 0.548 * \log(\mu)$	0.486	-310.971
Exponential	$p = 1 - e^{-0.374\mu^{0.372}}$	0.486	-310.958
Negative Binomial	$p = 1 - (1 + \frac{\mu}{0.231})^{-0.231}$	**0.489**	**-313.535**
SR	$p = \frac{\mu}{3.096 + \mu}$	0.314	-286.539

Developing species richness estimators

For spiders in Gerês, one asymptotic model was found by SR in all six training datasets (Appendix S1):

$$S = \frac{aQ}{b + Q}$$

where *a* and *b* are fitting parameters. This model is in fact the Clench model with a different formulation (Box 2), where the asymptote is *a*. A second, slightly more complex but better fitting, model was found for datasets with 40 or more samples:

$$S = \frac{c + aQ}{b + Q}$$

where *c* is a third fitting parameter. The asymptote is again given by the value of *a* (Fig. 4). This model is similar to the rational function (Box 2). It was found to outperform the Clench and negative exponential for both the training and testing datasets (Table 3).

bioRxiv preprint doi: https://doi.org/10.1101/027839; this version posted September 30, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Pre-print available from bioRxiv doi: http://dx.doi.org/10.1101/027839

Table 3 – Comparison of three asymptotic equations used to estimate spider species richness in two forest sites intensively sampled in Portugal (see Box 2 for formulas). Raw accuracy is the scaled mean squared error considering the entire observed accumulation curve (each formula was fitted to the curves using 4 to 320 samples) and weighted accuracy is this value weighted by the sampling effort at each point in the curve (where effort is the ratio between number of individuals and observed species richness). Note that lower values (in bold) are better as they reflect the deviation from a perfect estimator.

Model	Raw accuracy	Weighted accuracy
Gerês (training)		
Observed	0.113	0.037
Clench	0.055	0.018
Negative exponential	0.115	0.049
Rational function	**0.045**	**0.012**
Arrábida (testing)		
Observed	0.103	0.031
Clench	0.038	0.010
Negative exponential	0.092	0.037
Rational function	**0.032**	**0.008**

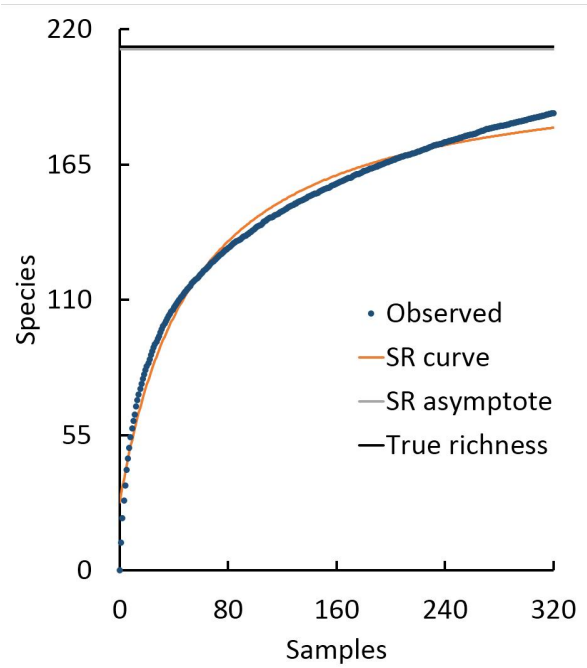

Fig. 4 – Accumulation curve for spider sampling in Gerês (Portugal) and the result of searching for the best fitting asymptotic formula using symbolic regression (SR).

Developing the species-area relationship (SAR) and the general dynamic model of oceanic island biogeography (GDM).

For the Azorean spiders, the best fitting previous model (both highest r^2 and lowest AICc) for the ISAR was the exponential model (Table 4). The SR run discovered roughly the same model, indicating, however, that the intercept (*c* term) was adding unnecessary complexity. A similar ranking of models was verified for bryophytes in the same region, revealing the robustness of the new model.

bioRxiv preprint doi: https://doi.org/10.1101/027839; this version posted September 30, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Pre-print available from bioRxiv doi: http://dx.doi.org/10.1101/027839

Table 4 - Species area relationship (SAR) models for Azorean taxa and General Dynamic Models (GDM) of oceanic island biogeography for Canarian taxa. S = native species richness, A = area of the island and T = maximum time of emergence. Best models are indicated in bold.

Model	Formula	r^2	AICc
SAR Azorean Spiders (training)			
Power	$S = 13.379 * A^{0.438}$	0.642	32.505
Exponential	$S = 0.549 + 4.538 \log A$	**0.780**	28.102
Linear	$S = 19.357 + 0.017A$	0.435	36.604
Exponential (SR)	$S = 4.641 \log A$	**0.780**	**23.319**
SAR Azorean Bryophytes (testing)			
Power	$S = 181.625 * A^{0.803}$	0.666	78.085
Exponential	$S = - 27.824 + 57.114 \log A$	**0.728**	76.208
Linear	$S = 196.215 + 0.259A$	0.617	79.295
Exponential (SR)	$S = 51.889 \log A$	0.722	**71.617**
GDM Canarian Spiders (training)			
Whittaker	$S = -185.589 + 41.732\log A + 17.776T - 1.022T^2$	0.873	110.350
Fattorini	$\log S = 2.585 + 0.281\log A + 0.157T - 0.009T^2$	0.941	105.025
Steinbauer	$\log S = 3.367 + 0.098\log A + 1.502\log T - 0.454\log T^2$	0.814	113.007
SR	$S = 42.283 + 0.051A + 17.379T - T^2$	**0.952**	**61.505**
GDM Canarian Bryophytes (testing)			
Whittaker	$S = -176.599 + 66.602\log A + 21.361T - 1.620T^2$	**0.773**	**125.214**
Fattorini	$\log S = 4.544 + 0.137\log A + 0.126T - 0.009T^2$	**0.803**	**124.217**
Steinbauer	$\log S = 5.136 + 0.017\log A + 1.063\log T - 0.382\log T^2$	0.612	128.963
SR	$S = 192.660 + 0.075A + 20.702T - 1.576T^2$	**0.785**	**124.841**

For the Canary Islands, the best model for spiders was a linear function of area:

$$S = 75 + 0.047A$$

(r^2 = 0.364, AICc = 65.631). Although it is easy to interpret, the explained variance is relatively low. The SR run reached a much higher explanatory power:

$$S = 112 + (-1.002A)$$

(r^2 = 0.806, AICc = 57.320). In this case though, the model is over-fitting to the few available data (7 data points), as this function fluctuates with small differences in area creating a biologically indefensible equation. The reason the ISAR is hard to model for the Canary Islands spiders is because we were missing the major component Time (Cardoso et al. 2010). This is depicted by the GDM, of which the best of the current equations was found to be the power model described by Fattorini (2009, Table 4). Nevertheless, using SR we were able to find an improved, yet undescribed, model (Table 4). This represents a general model expanding the linear SAR:

$$S = c + zA + xT - yT^2$$

The novelty in this model relies on the linear relation between richness and area, as all previous GDM model formulations represented log or power relations (Box 2). When tested with Canarian bryophytes, this new formulation is almost as good as the power model (Table 4).

bioRxiv preprint doi: https://doi.org/10.1101/027839; this version posted September 30, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Pre-print available from bioRxiv doi: http://dx.doi.org/10.1101/027839

DISCUSSION

Symbolic regression has several advantages over other, commonly used, statistical and machine learning techniques: (1) numerical, ordinal and categorical variables are easily combined; (2) redundant variables are usually eliminated in the search process and only the most important are retained if anti-bloat measures (intended to reduce the complexity of equations) are used; (3) the evolved equations are human-readable and interpretable; and (4) solutions are easily applied to new data. Using SR we were able to "distil" free-form equations and models that not only consistently outperform but are more intelligible than the ones resulting from rigid methods such as GLM or "black-boxes" such as Maxent. We were also able to re-discover and refine equations for estimating species richness based on sampling curves and the IAOR, ISAR and GDM from data alone.
All the examples presented in this work suggest that evolving free-form equations purely from data, often without prior human inference or hypotheses, may represent a yet unexplored but very powerful tool for ecologists and biogeographers, allowing the finding of hidden relationships in data and suggesting new ideas to formulate general theoretical principles. Each of the case studies here presented is now being further developed and thoroughly tested with extensive datasets covering different taxa, regions and spatial/temporal scales. This will be a crucial step before any conclusions can be reached. Yet, this approach seems extremely promising.

From particular relations to general principles

The usual way to find relationships, models and principles in science is through observation, construction and testing of hypotheses and eventually reaching a conclusion, either accepting or rejecting hypotheses. Other sciences such as physics rarely rely on general statistical inference methods such as linear regression for hypothesis testing. The complexity of ecology made such methods an imperative in most cases (despite the exceptions mentioned in the Introduction). The method now presented not only allows the discovery of relationships specific to particular datasets, but also the finding of general models, globally applicable to multiple systems of particular nature, as we tried to exemplify. As mentioned, SR is designed to optimize both the form of the equations and the fitting parameters simultaneously. The fitting parameters usually are specific to each dataset, but the form may give clues towards some general principle (e.g. all archipelagos will follow an ISAR even if each archipelago will have its own c and z values). Although this aspect has not been explored in this study, we suggest two ways of finding general principles.
First, as was hinted by our estimators' example, one may independently analyse multiple datasets from the same type of systems. From each dataset, one or multiple equations may arise. Many of these will be similar in form even if the fitting parameters are different. Terms repeated in several equations along the Pareto front or with different datasets tend to be meaningful (Schmidt & Lipson 2009). We may then try to fit the most promising forms to all datasets optimizing the fitting parameters to each dataset and look for which forms seem to have general value over all data.
Second, one may simultaneously analyse multiple datasets from the same type of systems but with a change to the general SR implementation. Instead of optimizing both form and fitting parameters, the algorithm may focus on finding the best form, with fitting parameters being optimized during the evaluation step of the evolution for each dataset independently. This parameter optimization could be done with standard methods such as quasi-newton or simplex (Wright & Nocedal 1999). To our knowledge, this approach has yet to be implemented, but it would allow finding general models and possibly principles, independently of the idiosyncrasies of each dataset.

bioRxiv preprint doi: https://doi.org/10.1101/027839; this version posted September 30, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Pre-print available from bioRxiv doi: http://dx.doi.org/10.1101/027839

The need for human inference

Our results show that an automated discovery system can identify meaningful relationships in ecological data. Yet, some equations might be very accurate but overfit the data. This was certainly the case of our Canary Island spider SAR model, which was found uninterpretable even if the correlation was very high, and possibly true of our IAOR Clench-like model, although further testing is needed. As with any relationship finding, either automated or human, correlation does not imply causation and spurious relationships are not only possible but probable given complex enough data. Other equations might be oversimplifications of reality. The quest for simple models may, however, prevent us from finding more complex yet more general models (Evans et al. 2013).

Although the method here presented is automated, it is part of a collaborative human–machine effort. The possibility of exploiting artificial intelligence working together with human expertise can be traced back to Engelbart (1962), where the term "augmented intelligence" was coined to designate such collaboration. It has been subsequently developed and extended to teamwork involving one or more artificial intelligence agents together with one or (many) more humans, in diverse domains such as robotic teams (Yanco et al. 2004) or collective intelligence for evolutionary multi-objective optimization (Cinalli et al. 2015).

In ecological problems, human knowledge may play a fundamental role: 1) in the beginning of the process, when we must select input variables, building blocks and SR parameters; and 2) in the interpretation and validation of equations. The choice of equations along a machine-generated Pareto front should also take advantage of human expert knowledge to identify the most interesting models to explain the data. In fact, and in contrast with inflexible methods, in big data problems there is no standard way to model data, making human expertise arguably more necessary. The researcher might then decide to disregard, accept or check equation validity using other methods.

A priori knowledge

To some extent, it is possible to select *a priori* the type of models the algorithm will search for by selecting the appropriate variables and building blocks. For example, if we know beforehand that area is the main driver of richness on islands but have no or few clues on how area and richness relate, we may start by running SR with area as the only explanatory variable, even if other variables are available.

Another way to take advantage of previous knowledge is to use as part of the initial population of equations some, possibly simpler, equations we know are related with the problem. For example, when searching for the GDM we could have given the algorithm multiple forms of the ISAR to seed the search process. This should be complemented with random equations to create the necessary variation for evolution. Seeding the search allows us not only to lessen the computation time but also to simplify the interpretation of resulting equations, as these might be related with the initial ones. Seeding might, however, cause too fast a convergence on not so novel equations, so it might be useful to run both approaches in parallel (seeded and non-seeded).

Fine-tuning the process

The number of options in SR is immense. Population size is positively correlated with variability of models and how well the search space is explored, but might considerably slow the search. Mutation rates are also positively correlated with variability, but rates that are too high might prevent the algorithm converging on the best models. The fitness measure depends on the specific problem and the type of noise expected, with r^2 goodness of fit being best in many cases but, e.g., AUC being most suited for classification problems, or logarithmic error being best for cases with numerous outliers.

bioRxiv preprint doi: https://doi.org/10.1101/027839; this version posted September 30, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Pre-print available from bioRxiv doi: http://dx.doi.org/10.1101/027839

Data can also be weighted, for example according to the precision of each data point, with less precise measurements being down-weighted when fitness is calculated.
The number of generations to let the search run is entirely dependent on the problem complexity and time available. Often the algorithm reaches some equation that makes immediate sense to the researcher and the process can be immediately stopped for further analysis of results. Sometimes several equations seem to make sense but are not entirely convincing, in which case several indicators can be used as a stop rule, such as high values of stability and maturity of the evolution process (Schmidt & Lipson 2014).
The speed with which evolution occurs is extremely variable, depending on factors including the complexity of the relationships, having the appropriate variables and building blocks (which requires some *a priori* knowledge on the system) and the level of noise in the data. Fortunately, as mentioned before, the process is easily adaptable to parallel computing, as many candidate functions can be evaluated simultaneously, allowing the use of multiple cores and even computer clusters to speed the search of equations.
Finally, as the search is not deterministic, different equations may be found in different runs of the algorithm with exactly the same data and parameters. Repeating the same search multiple times is a good way to further explore the search space and test the consistency of results.

Caveats and alternatives

The SR approach is fully data-driven. This means it requires high-quality data if meaningful relationships are to be found. Also, it makes no a priori assumptions, so the final result might make no (obvious) sense, leading to spurious inferences, particularly if data are scarce or poor-quality, or if the right building blocks are not provided. Additionally, SR suffers from the same limitations of evolutionary algorithms in general. In many cases the algorithm may get stuck in local minima of the search space, requiring time (or even a restart with different parameters) to find the global minimum. This means there is no guarantee that the solutions provided are the best for each problem or even any way to know how good the solution is compared with the optimal.
Many data mining techniques are regarded, and rightly so, as “black boxes”. Neural networks are certainly such a case, as relationships between variables are mostly impractical to interpret in a direct way. This is also the case for maximum entropy models (Maxent) as implemented in species distribution modelling. SR is transparent in this regard, as variables are related through human-interpretable formulas. This is particularly important if the goal is to find equations with both predictive and explanatory power, building the bridge between finding the pattern and explaining the driving process, or if a general principle is to be suggested.

The automation of science?

Many applications exist for this approach, from systems biology to cosmology. In ecology (and ecological biogeography), probably the most complex of sciences, this and related techniques might be particularly relevant, as we tried to demonstrate. SR modelling can be a powerful addition to theoretical and experimental ecology, even if new conceptual hypotheses have to be created to accommodate the new equations. Such models could even be the only available means of investigating complex ecological systems when experiments are not feasible or datasets get too big/complex to model, using traditional statistical techniques.
This kind of techniques has led several authors to talk about the “automation of science” (King et al. 2009), where computers are able to advance hypotheses, test them and reach conclusions in largely unassisted processes. Yet, as mentioned above, human inference is still needed, if not more so, with these techniques. Only, as Graham et al. (2013) put it “We still make discoveries, but as the complexity of data increases, we need machine intelligence to help us guide towards an insight”. The

bioRxiv preprint doi: https://doi.org/10.1101/027839; this version posted September 30, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Pre-print available from bioRxiv doi: http://dx.doi.org/10.1101/027839

SR potential as an exploratory step, to be reasoned alongside and proven with other methods is also exciting. The resulting formulas will help researchers to focus on initially imperceptible but interesting relationships within datasets and help guide the process of hypothesis creation. Yet, we reiterate, the final word still depends entirely on human reasoning.

ACKNOWLEDGEMENTS

We thank Robert Whittaker, Stano Pekár and Otso Ovaskainen for comments on earlier versions of the manuscript. PAVB and FR were partly funded by the project FCT-PTDC/BIA-BIC/119255/2010 - "Biodiversity on oceanic islands: towards a unified theory". LC was partially funded by UID/MULTI/ 04046/2013, from FCT/MCTES/PIDDAC, Portugal.

REFERENCES

Akaike, H. (1974) A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19: 716-723.

Allouche, O., Tsoar, A. & Kadmon, R. (2006) Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology*, 43: 1223-1232.

Almeida, J., Santos, J.A., Miranda, W.O., Alberton, B., Morellato, L.P.C. & Torres, R.S. (2015) Deriving vegetation indices for phenology analysis using genetic programming. *Ecological Informatics*, 26: 61-69.

Anand, M., Gonzalez, A., Guichard, F., Kolasa, J. & Parrott, L. (2010) Ecological systems as complex systems: challenges for an emerging science. *Diversity*, 2: 395-410.

Aranda, S.C., Gabriel, R., Borges, P.A.V., Santos, A.M.C., Azevedo, E.B., Patiño, J., Hortal, J. & Lobo, J.M. (2014) Geographical, temporal and environmental determinants of bryophyte species richness in the Macaronesian islands. *PloS One*, 9: e101786.

Arrhenius, O. (1920) Distribution of the species over the area. *Meddelanden fran Vetenskapsakadmiens Nobelinstitut*, 4: 1-6.

Arrhenius, O. (1921) Species and area. *Journal of Ecology*, 9: 95-99.

Barrett, J., Kostadinova, A. & Raga, J.A. (2005) Mining parasite data using genetic programming. *Trends in Parasitology*, 21: 207-209.

Barton, K. (2015) MuMIn: Multi-Model Inference. R package version 1.13.4. http://cran.r-project.org/package=MuMIn.

Brown, J.H. (1984) On the relationship between abundance and distribution of species. *American Naturalist*, 124: 255-279.

Burnham, K.P. & Anderson, D.R. (2002) *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd edn. Springer, New York, NY.

Cardoso, P., Gaspar, C., Pereira, L.C., Silva, I., Henriques, S.S., Silva, R.R. & Sousa, P. (2008a) Assessing spider species richness and composition in Mediterranean cork oak forests. *Acta Oecologica*, 33: 114-127.

Cardoso, P., Scharff, N., Gaspar, C., Henriques, S.S., Carvalho, R., Castro, P.H., Schmidt, J.B., Silva, I., Szüts, T., Castro, A. & Crespo, L.C. (2008b) Rapid biodiversity assessment of spiders (Araneae) using semi-quantitative sampling: a case study in a Mediterranean forest. *Insect Conservation and Diversity*, 1: 71-84.

Cardoso, P., Lobo, J.M., Aranda, S.C., Dinis, F., Gaspar, C. & Borges, P.A.V. (2009) A spatial scale assessment of habitat effects on arthropod communities of an oceanic island. *Acta Oecologica*, 35: 590-597.

bioRxiv preprint doi: https://doi.org/10.1101/027839; this version posted September 30, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Pre-print available from bioRxiv doi: http://dx.doi.org/10.1101/027839

Cardoso, P., Arnedo, M.A., Triantis, K.A. & Borges, P.A.V. (2010) Drivers of diversity in Macaronesian spiders and the role of species extinctions. *Journal of Biogeography*, 37: 1034-1046.

Cardoso, P., Rigal, F., Fattorini, S., Terzopoulou, S. & Borges, P.A.V. (2013) Integrating landscape disturbance and indicator species in conservation studies. *PLoS One*, 8: e63294.

Cardoso, P., Rigal, F. & Carvalho, J.C. (2015) BAT - Biodiversity Assessment Tools, an R package for the measurement and estimation of alpha and beta taxon, phylogenetic and functional diversity. *Methods in Ecology and Evolution*, 6: 232-236.

Chao, A. (1984) Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics*, 11: 265-270.

Chao, A. (1987) Estimating the population size for capture-recapture data with unequal catchability. *Biometrics*, 43: 783-791.

Cinalli, D., Martí, L., Sanchez-Pi, N., & Garcia, A.C.B. (2015) Collective preferences in evolutionary multi-objective optimization: techniques and potential contributions of collective intelligence. In *Proceedings of the 30th Annual ACM Symposium on Applied Computing*, ACM, pp. 133-138.

Clench, H. (1979) How to make regional lists of butterflies: some thoughts. *Journal of the Lepidopterists' Society*, 33: 216-231.

Colwell, R.K. & Coddington, J.A. (1994) Estimating terrestrial biodiversity through extrapolation. *Philosophical Transactions of the Royal Society of London - Biological Sciences*, 345: 101-118.

Correia, L. (2010) Computational evolution: taking liberties. *Theory in Biosciences*, 129: 183-191.

Darlington, P.J. (1957) *Zoogeography: the geographical distribution of animals*. Wiley, New York, NY.

Dodds, W.K. (2009) *Laws, theories and patterns in ecology*. University of California Press, Berkeley, CA.

Elith, J., Graham, C.H., Anderson, R.P., Dudik, M., Ferrier, S., Guisan, A., Hijmans, R.J., Huettmann, F., Leathwick, J.R., Lehmann, A., Li, J., Lohmann, L.G., Loiselle, B.A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J.M.M., Peterson, A.T., Phillips, S.J., Richardson, K., Scachetti-Pereira, R., Schapire, R.E., Soberón, J., Williams, S., Wisz, M.S. & Zimmermann, N.E. (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, 29: 129-151.

Engelbart, D. (1962) Augmenting human intellect: a conceptual framework. *Summary Report AFOSR-3233*, Stanford Research Institute, Menlo Park, CA.

Evans, M.R., Grimm, V., Johst, K., Knuuttila, T., Langhe, R., Lessells, C.M., Merz, M., O'Malley, M.A., Orzack, S.H., Weisberg, M., Wilkinson, D.J., Wolkenhauer, O. & Benton, T.G. (2013) Do simple models lead to generality in ecology? *Trends in Ecology and Evolution*, 28: 578-583.

Fattorini, S. (2009) On the general dynamic model of oceanic island biogeography. *Journal of Biogeography*, 36: 1100-1110.

Gabriel, R. & Bates, J. W. (2005) Bryophyte community composition and habitat specificity in the natural forests of Terceira, Azores. *Plant Ecology*, 177: 125-144.

Gaston, K.J., Blackburn, T.M. & Lawton, J.H. (1997) Interspecific abundance-range size relationships: an appraisal of mechanisms. *Journal of Animal Ecology*, 66: 579-601.

Gleason, H.A. (1922) On the relation between species and area. *Ecology*, 3: 158-162.

Graham, M.J., Djorgovski, S.G., Mahabal, A.A., Donalek, C. & Drake, A.J. (2013) Machine-assisted discovery of relationships in astronomy. *Monthly Notices of the Royal Astronomical Society*, 431: 2371-2384.

He, F.L. & Gaston, K.J. (2000) Estimating species abundance from occurrence. *American Naturalist*, 156: 553-559.

Holland, J.H. (1975) *Adaptation in Natural and Artificial Systems*. University of Michigan Press, MI.

Holland, J. (1995) *Hidden Order: How Adaptation Builds Complexity*. Basic Books, NY.

Holland, J. (1998) *Emergence: From Chaos to Order*. Basic Books, NY.

bioRxiv preprint doi: https://doi.org/10.1101/027839; this version posted September 30, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Pre-print available from bioRxiv doi: http://dx.doi.org/10.1101/027839

Jagupilla, S.C.K., Vaccari, D.A., Miskewitz, R., Su, T.-L. & Hires, R.I. (2015) Symbolic regression of upstream, stormwater, and tributary *E. coli* concentrations using river flows. *Water Environment Research*, 87: 26-34.

Johnson, J.B. & Omland, K.S. (2004) Model selection in ecology and evolution. *Trends in Ecology and Evolution*, 19: 101-108.

King, R.D., Rowland, J., Oliver, S.G., Young, M., Aubrey, W., Byrne, E., Liakata, M., Markham, M., Pir, P., Soldatova1, L.N., Sparkes, A., Whelan, K.E. & Clare, A. (2009) The automation of science. *Science*, 324: 85-89.

Koza, J.R. (1992) *Genetic Programming: on the Programming of Computers by Means of Natural Selection*. MIT Press, Cambridge, MA.

Koza, J.R. (2010) Human-competitive results produced by genetic programming. *Genetic Programming and Evolvable Machines*, 11: 251-284.

Larsen, P.E., Field, D. & Gilbert, J.A. (2012) Predicting bacterial community assemblages using an artificial neural network approach. *Nature Methods*, 9: 621-625.

Larsen, P.E., Cseke, L.J., Miller, R.M. & Collart, F.R. (2014) Modeling forest ecosystem responses to elevated carbon dioxide and ozone using artificial neural networks. *Journal of Theoretical Biology*, 359: 61-71.

Lawton, J.H. (1996) Patterns in Ecology. *Oikos*, 75: 145-147.

Liu, C., Berry, P.M., Dawson, T.P. & Pearson, R.G. (2005) Selecting thresholds of occurrence in the prediction of species distributions. *Ecography*, 28: 385-393.

Lotka, A.J. (1910) Contribution to the theory of periodic reaction. *Journal of Physical Chemistry*, 14: 271-274.

Lotka, A.J. (1925) *Elements of Physical Biology*. Williams and Wilkins, Baltimore, MD.

MacArthur, R.H. & Wilson, E.O. (1967) *The Theory of Island Biogeography*. Princeton University Press, NJ.

Manson, S.M. (2005) Agent-based modelling and genetic programming for modelling land change in the Southern Yucatan Peninsular Region of Mexico. *Agriculture, Ecosystems and Environment*, 111: 47-62.

Manson, S.M. & Evans, T. (2007) Agent-based modeling of deforestation in southern Yucatan, Mexico, and reforestation in the Midwest United States. *Proceedings of the National Academy of Sciences*, 104: 20678-20683.

Miller, R.I. & Wiegert, R.G. (1989) Documenting completeness, species–area relations, and the species–abundance distribution of a regional flora. *Ecology*, 70: 16-22.

Mitchell, M. (2011) *Complexity: a Guided Tour*. Oxford University Press, NY.

Muttil, N. & Lee, J.H.W. (2005) Genetic programming for analysis and real-time prediction of coastal algal blooms. *Ecological Modelling*, 189: 363-376.

Nachman, G. (1981) A mathematical model of the functional relationship between density and spatial distribution of a population. *Journal of Animal Ecology*, 50: 453-460.

Passy, S.I. (2012) A hierarchical theory of macroecology. *Ecology Letters*, 15: 923-934.

Phillips, S.J., Anderson, R.P. & Schapire, R.E. (2006) Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, 190: 231-259.

Preston, F.W. (1948) The commonness, and rarity, of species. *Ecology*, 29: 254-283.

R Development Core Team (2015) R: A Language and Environment for Statistical Computing v3.1.3. R Foundation for Statistical Computing, Vienna. ISBN 3-900051-07-0. http://www.R-project.org.

Ratkowski, D.A. (1990) *Handbook of Nonlinear Regression Models*. Marcel Dekker, NY.

Reshef, D.N., Reshef, Y.A., Finucane, H.K., Grossman, S.R., McVean, G., Turnbaugh, P.J., Lander, E.S., Mitzenmacher, M. & Sabeti, P.C. (2011) Detecting novel associations in large data sets. *Science*, 334: 1518-1524.

bioRxiv preprint doi: https://doi.org/10.1101/027839; this version posted September 30, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Pre-print available from bioRxiv doi: http://dx.doi.org/10.1101/027839

Russel, S. & Norvig, P. (2010) *Artificial Intelligence: a Modern Approach*. Pearson Education Inc., NJ.

Schmidt, M. & Lipson, H. (2009) Distilling free-form natural laws from experimental data. *Science*, 324: 81-85.

Schmidt, M. & Lipson, H. (2014) Eureqa (Version 1.12.0 beta). Available from http://www.eureqa.com/

Soberón J. & Llorente, J. (1993) The use of species accumulation functions for the prediction of species richness. *Conservation Biology*, 7: 480-488.

Sole, R. & Goodwin, B. (2000) *Signs of Life: How Complexity Pervades Biology*. Basic Books, NY.

Steinbauer, M.J, Klara, D., Field, R., Reineking, B. & Beierkuhnlein, C. (2013) Re-evaluating the general dynamic theory of oceanic island biogeography. *Frontiers of Biogeography*, 5: 185-194.

Triantis, K.A., Guilhaumon, F. & Whittaker, R.J. (2012) The island species–area relationship: biology and statistics. *Journal of Biogeography*, 39: 215-231.

Tung, C.-P., Lee, T.-Y., Yang, Y.-C.E. & Chen, Y.-J. (2009) Application of genetic programming to project climate change impacts on the population of Formosan Landlocked Salmon. *Environmental Modelling & Software*, 24: 1062-1072.

Verhulst, P.-F. (1845) Recherches mathématiques sur la loi d'accroissement de la population. *Nouvelles Mémoires de l'Academie Royale des Sciences et Belles-Lettres de Bruxelles*, 18: 1-41.

Volterra, V. (1926) Variazioni e fluttuazioni del numero d'individui in specie animali conviventi. *Memoria della Regia Accademia Nazionale dei Lincei*, 2: 31-113.

Whittaker, R.J., Triantis, K.A. & Ladle, R.J. (2008) A general dynamic theory of oceanic island biogeography. *Journal of Biogeography*, 35: 977-994.

Wright, S.J. & Nocedal, J. (1999) *Numerical Optimization*, vol. 2. Springer, NY.

Yanco, H., Drury, J. & Scholtz, J. (2004) Beyond usability evaluation: analysis of human-robot interaction at a major robotics competition. *Human-Computer Interaction*, 19: 117-150.

Yao, M., Rui, J., Li, J., Dai, Y., Bai, Y., Hedĕnec, P., Wang, J., Zhang, S., Pei, K., Liu, C., Wang, Y., He, Z., Frouz, J. & Li, X. (2014) Rate-specific responses of prokaryotic diversity and structure to nitrogen deposition in the *Leymus chinensis* steppe. *Soil and Biochemistry*, 79: 81-90.