1 Title: Reconsidering environmental diversity (ED) as a biodiversity surrogacy strategy.

2

3 Authors: Ian Engelbrecht¹, Mark Robertson², Marnus Stoltz³, Johan W. Joubert⁴.

4

5 Affiliations: 1. Department of Zoology and Entomology, University of Pretoria, Pretoria, South Africa.

6 ianicus.za@gmail.com. 2. Centre for Invasion Biology, Department of Zoology and Entomology,

7 University of Pretoria, Pretoria, South Africa. 3. Department of Mathematics and Applied Mathematics,

8 University of Pretoria, Pretoria, South Africa. 4. Department of Industrial and Systems Engineering,

9 University of Pretoria, Pretoria, South Africa.

Abstract:

Environmental Diversity (ED) has been proposed as a potential biodiversity surrogate for use in spatial biodiversity planning. The concept is based on the premise that differences in species composition between areas are correlated with differences in environmental conditions. If this is true then sites selected to represent the full range of environmental conditions in an area should provide good coverage of species diversity. This is potentially significant for conservation planning because environmental data are more easily available while species distribution data are often incomplete, but empirical tests of ED have been inconclusive. We explored the value of ED as a biodiversity surrogate using species distribution data from South Africa, focusing on how implementation affects its performance. We used empirical distribution datasets for birds, butterflies, frogs, and reptiles as well as artificial species distributions and modeled distributions for butterflies. Our results show that the form of the underlying optimization model and the number of environmental variables have a significant effect on the performance of ED while the choice of heuristic procedures for selecting sites has no significant effect. We conclude that ED can perform well as a biodiversity surrogacy strategy if it is implemented correctly.

Keywords: species assemblage, conservation planning, simulated annealing, p -median, maximum diversity, Generalized Dissimilarity Model.

29 Introduction:

30
31 Environmental Diversity, or ED, has been proposed as a strategy for prioritising areas for biodiversity
32 conservation or inclusion in protected area networks (Faith and Walker 1996). The underlying concept is
33 that if a set of areas is selected to provide good coverage of the environmental conditions in the region
34 of interest it will also provide good coverage of the biodiversity of the region. This is an appealing
35 prospect because environmental data are easier to obtain than biodiversity data and because the
36 relationship between biodiversity and environmental conditions is intuitive for most ecologists.
37 Comprehensive environmental datasets derived from climate monitoring stations or remote sensing are
38 readily and cheaply available for most parts of the world (Beier et al. 2015b), whereas comprehensive
39 species distribution datasets are often lacking, particularly in the biodiverse tropics (Pimm 2000) and for
40 species-rich taxa like arthropods (e.g. Oliver and Beattie 1996, New 1998, Ponder and Lunney 1999,
41 Ward and Lariviere 2004). Due to these shortcomings in available biodiversity data, spatial conservation
42 planning is traditionally based on taxa for which relatively good distribution data are available, typically
43 vascular plants and vertebrate groups (e.g. Lombard et al. 1999, Fairbanks et al. 2001, Cowling et al.
44 2003, Reyers 2004). In a spatial biodiversity planning framework these taxa act as surrogates for overall
45 biodiversity (Margules and Pressey 2000, Reyers et al. 2002, Rodrigues and Brooks 2007). However,
46 there is often limited congruence in the spatial patterns of biodiversity between different taxa which
47 limits the efficacy of taxon based surrogates (e.g. van Jaarsveld et al. 1998). While conservation
48 prioritization methods that use these surrogates are well developed and find wide application (Margules
49 and Pressey 2000, Sarkar et al. 2006), effective surrogacy strategies based on environmental data would
50 be valuable for conservation planning projects (Faith and Walker 1996, Beier et al. 2015a).

51 Faith and Walker (1996) described the Environmental Diversity (ED) strategy as a two-stage
52 process for selecting or prioritising sites for conservation using environmental data as surrogate for

biodiversity (Table 1). The first stage is to create a distance matrix representing environmental dissimilarities between all pairs of sites in the region of interest and the second is to use an optimization procedure to select a set of sites that is representative of the total range of environmental conditions. They proposed the p -median optimization model as the most appropriate description of this problem in mathematical terms. The p -median procedure originates from operations research and is typically used for identifying sites for facilities or services such as warehouses, schools or hospitals, to service a group of users in an urban area (Hakimi 1965, Reese 2006). The locations of the facilities are termed supply points and the user locations are the demand points. Sites for facilities are selected from the set of potential supply points so that the travel distance of any one user to their nearest facility is minimized. In the simplest form of ED, distances in the ordination space replace the geographic distances typically used in p -median problems (see Reese (2006) or Mladenović et al. (2007) for a formal definition of the p -median problem) with the result that each selected site is 'representative' of the range of environmental conditions of covered by the demand points allocated to it. Distances can also be modified to include weightings for species richness or to represent community turnover more directly where biotic data are available (see below).

Faith and Walker (1996) noted that the location of sites selected by p -median procedures would be affected by the underlying distribution of sites relative to one another in environmental space with more sites selected in regions with greater densities of demand points (Fig. 1). In order to maximize coverage of environmental conditions for a given number of sites they proposed a modification to the p -median procedure that uses artificial demand points spread uniformly across environmental space to avoid this problem. They termed this 'continuous' ED, and used the term 'discrete' ED to refer to the classical case where sites simultaneously represent demand points and potential supply points in the p -median procedure.

Despite its conceptual appeal, evaluations of ED using empirical species distribution datasets have been inconclusive about its effectiveness for representing biodiversity (Ferrier and Watson 1997, Araújo et al. 2001, 2004, Hortal et al. 2009). ED appears to select sites that represent species equally or only marginally better than random selections, which has led to recommendations that it should not be considered as a substitute for other area prioritization methods until the factors governing its performance are better understood. ED is based on the assumptions that species show symmetric, unimodal responses to environmental gradients represented in ordination space (Faith and Walker 1996, Faith 2003, Faith et al. 2004), and that species distributions are randomly located in that space (Faith and Walker 1996). Araújo et al. (2003, 2004) argued that the poor performance of ED could be attributed to departures of real species distributions from these assumptions. Species can show a range of response types across environmental gradients (Oksanen and Minchin 2002) and species richness may vary systematically across gradients (Hawkins et al. 2003, Rodríguez et al. 2005, Field et al. 2009). Araújo et al. (2001) found that sites selected using actual species distribution data and the principle of complementarity were clustered in environmental space rather than spanning it, which they interpreted to mean that certain habitats harbour disproportionate levels of diversity and that factors other than environmental conditions influence species turnover. For example, frog and reptile diversity was found to be concentrated in the warmer parts of their study region and much of this diversity was missed by ED selections that sampled the total range of environmental conditions of Europe in a uniform fashion. They also showed that sites selected using ED tended to exclude marginal species (species at their range edges in the study region), alien species, and species with limited dispersal capacity. These species are unlikely to occupy all environmentally suitable areas within a region and they proposed that the performance of ED would also be related to the degree to which species distributions are at equilibrium with environmental conditions.

Faith (2003, 2011) and Faith et al. (2004) argued that the apparent poor performance of ED may have been the result of flawed implementation of the underlying optimization procedure for selecting sites, or of inadequacies in the methods used to quantify environmental dissimilarities between sites. They advocated for combining available species and environmental data to maximize the performance of the ED, suggesting Generalized Dissimilarity Modeling (GDM, Ferrier et al. 2007, see Faith and Ferrier 2002) as one way for doing so. They also indicated that environmental dissimilarity matrices could be weighted to cater for variable species richness to improve the performance of the ED procedure. Beier and Albuquerque (2015) have subsequently shown that ED provides significant improvements over random site selections, using data from the same regions and for the same taxa as the study by Araújo et al. (2001) as well as others, when the continuous form of ED is properly implemented.

Arponen et al. (2008) developed a strategy called Maximization of Complementary Richness (MCR) which makes use of predicted turnover from GDM models as well as predicted species richness to select sites in a way that mimics the principle of complementarity by using these ‘emergent’ measures of biodiversity (Ferrier 2002) and found that this approach was significantly better than ED in a simulation study. However, their implementation of ED was based on the same optimization procedure as previous empirical studies and may have suffered similar shortcomings, and may not provide an accurate representation of the capability of the relative abilities of these two methods as a result. Also, Faith (2016) argued that MCR is largely conceptually similar to ED, and that differences in the model degrade its performance rather than improve it.

ED may be a valuable surrogacy strategy for conservation planning but implementation issues need to be separated from criticism of the underlying assumptions, and a suite of empirical studies is required to understand the merits and shortcomings of the methodology. The aim of the study presented here was to assess the performance of ED using several new empirical datasets from South Africa and to investigate how various aspects of implementation might affect its performance. South

Africa has a rich history of biodiversity conservation (Balmford 2003) and regional scale species surveys (Robertson et al. 2010). Using species distribution datasets from this region, as well as artificial and modeled species distribution data, our aims were to assess: 1) how the choice of the optimization model affects the performance of ED, 2) whether the form of heuristic used to select sites is an important consideration in implementing ED, 3) the effect of weighting distances using species richness, 4) whether including additional environmental variables affects the performance of ED, and 5) whether using GDM models to combine species and environmental data to produce distance measures improves ED selections.

Methods:

Empirical data:

We used species distribution datasets for birds, reptiles, amphibians, and butterflies in South Africa for this assessment (Harrison et al. 1997, 2008, Minter et al. 2004, Mecenero et al. 2013, Bates et al. 2014). These datasets have been developed from extensive, formalized atlas projects coordinated largely by the Animal Demography Unit of the University of Cape Town and the South African National Biodiversity Institute, and are the most comprehensive faunal species distribution datasets available for the country. The datasets include sighting records, museum specimen records, and photographic records submitted by the general public to an online Virtual Museum (<http://vmus.adu.org.za>). The data were extracted on the 26th February 2013 and include 6 428 082 records for 791 bird species, 355 062 records for 657 butterfly species, 42 238 records for 122 frog species, and 131 325 records for 468 reptile species.

The assessment was conducted at the quarter degree square (QDS) scale to provide a common baseline for comparison across different datasets. QDSs are a standardized grid of areas approximately

20 km x 20 km in size and commonly used in large scale biogeographic analyses in South Africa. Only QDSs that fell largely within the boundary of South Africa were included in the analysis, resulting in 1816 QDSs in total. The data for each taxon were used to create a community matrix representing presence and absence of each species within each QDS.

Atlas data are generally the result of non-systematic survey effort and exhibit varying degrees of geographic sampling bias and completeness (Robertson et al. 2010, Yang et al. 2013). To assess the performance of ED without these confounding factors we simulated species geographic ranges in the same region and at the same scale as the atlas data. Artificial species ranges were based on an environmental envelope concept. Between one and four environmental variables were selected randomly from a climatic dataset (see Distance Matrices below) and used as the basis for a species envelope. Suitability thresholds were specified for each variable by selecting a median value from within the range of the variable and a proportion of between 0.1 and 1.0 of the total range of the variable around the median. QDS's that fell within the thresholds for the selected variables were then identified as suitable for the species. We repeated this procedure to create 500 artificial species ranges. Only distributions that covered 0.1% (i.e. 181 QDSs) to 90% of all QDSs were included in the dataset. While this process is entirely random our results showed patterns of species richness that conform to real species richness patterns for many taxa in South Africa (i.e. higher richness in the moister, more topographically diverse eastern and southern parts of the country), and the resulting range-size frequency distribution was strongly right-skewed. This conforms with general range-size frequency distributions for real taxa (Gaston 2003) and neutral community assembly processes (Hubbell 2001). This approach also avoids circularity in the initial ED assessment, as different combinations of variables are used for each species and these are not the same as the variables used to create environmental distances for ED selections.

In addition to this artificial dataset we used modeled species distributions for butterflies based on the butterfly atlas data used (Mecenero et al. 2015). The modeling procedure used environmental correlates for species occurrence to predict total distributions while simultaneously correcting for sampling biases in the atlas data. This procedure produced probabilities of occurrence in each QDS for each species, which we transformed to presence and absence by finding the probability threshold that contained 90% of observed occurrence records. While the artificial distribution dataset was based on randomly generated distributions, the modeled butterfly distribution dataset would presumably include the effects of additional factors such as biogeographic and evolutionary history, ecological effects, and disequilibrium with environmental conditions, at least to some degree.

Assessing implementation of ED:

We assessed several approaches for selecting sites using environmental dissimilarities to compare how implementation details affected their performance in recovering species diversity. We drew a distinction between the class of optimization model used to represent the goal of ED and the underlying heuristic computational procedure used to solve those models. We implemented both the ‘discrete’ and ‘continuous’ forms the p -median optimization procedure as well as a modified form of the maximum diversity problem (Marti et al. 2013). The maximum diversity approach is described by a model where the sum of distances between them is maximized (the formal mathematical notation is provided in Marti et al. 2013). The optimization model does not include demand points as is the case with p -median models. In initial investigations we found that the standard form of maximum diversity created clusters of sites that were widely dispersed from one another in ordination space. We thus developed an alternative model that produces a more even spread of sites through ordination space by using the inverse of the square of distances between sites. This model was tested on several geographic distance

matrices as well as environmental matrices (see Fig. 1) and found to produce an outcome very similar to the 'continuous' p -median model. Our modified maximum diversity model is:

$$\text{minimise} \quad \sum_{i=1}^{n-1} \sum_{j=i+1}^n \frac{1}{w_i w_j d_{ij}^2}$$

subject to:

$$\sum_i^n x_i = m,$$

$$x_i \in \{0,1\}$$

where n is the total number of sites available for inclusion in the set of selected sites, i and j are the indices of individual sites, w_i and w_j are optional weighting factors for sites i and j respectively, d is the dissimilarity measure between pairs of sites, x is an allocation variable indicating whether a site has been selected or not, and m is the total number of sites that must be selected.

We used climate data from the Climond global dataset (Kriticos et al. 2012) to produce environmental distance matrices for the ED procedures. This dataset consists of 35 climatic variables covering temperature, rainfall, solar radiation, and seasonality conditions, at a 10' resolution. We selected five variables to match those used by Araújo et al. (2001) as closely as possible. They were: average temperature in the warmest and coldest quarters, moisture index in the driest and wettest quarters, and annual precipitation. We also included elevation, which was obtained from GTOPO30 digital elevation model with 1 km resolution (Gesch et al. 1999). These six variables were resampled to match the extent, placement and resolution the QDSs in South Africa using ArcGIS v10 (ESRI 2013). We used principal components analysis to reduce collinearity amongst the variables and created an ordination space from the first two components, which were identified as significant using the broken stick method (Jackson 1993), and which accounted for 73% of the variation of all six variables included.

To create a distance matrix for the ‘discrete’ p -median procedure we calculated inter-site Euclidean distances using the site scores for each QDS along the first two principal components. For the ‘continuous’ form of ED we created an artificial grid of demand points by dividing the total range of site scores for each principal component into nine equal intervals and placing demand points at the intersections of the end points of those intervals. Thus the artificial demand point grid contained 100 equally spaced points. Euclidian distances between these demand points and the QDSs were calculated to produce the distance matrix. In each matrix potential supply points were represented by rows, and demand points by columns. Our method for creating artificial demand points for ‘continuous’ ED differed slightly to that described in the literature in that we did not use a convex hull to limit the grid to the portion of environmental space occupied by the QDSs (see Faith et al. 2004: Fig. 1). However, our results suggested that our approach was sufficient for producing site selections that spanned the environmental space adequately. For the modified maximum diversity procedure we used the same distance matrix as for the ‘discrete’ p -median procedure.

Finding solutions to optimization problems requires the use of heuristic computation procedures for practical reasons (Table 1). Heuristics can be broadly divided into iterative and simultaneous procedures. Iterative procedures add or remove sites from a solution (called greedy or stingy procedures, respectively) until a set number of sites is obtained. At each step, the site that produces the best improvement in the optimization model is added or removed. Simultaneous procedures start with a set number of sites, often chosen at random, and exchange selected and unselected sites to find improved solutions. Each exchange is influenced by the locations of all of the other sites in the set and therefore the most profitable swap can be found for each exchange and the set is improved as a whole. Incremental procedures can produce solutions that are as good as solutions from simultaneous procedures but because each addition or removal occurs without knowledge of later changes they can also produce very poor final solutions (Underhill 1994). We used one greedy site

selection heuristic and two simultaneous heuristics (Table 2) to assess whether the form of the heuristic had any significant effect on the performance of ED.

We used the Species Accumulation Index as a framework for assessing the performance of ED for recovering species diversity (SAI, Ferrier 2002, Rodrigues and Brooks 2007, Table 1). We ran site selections for 100 samples sizes (i.e. the number of sites to be selected) between 2 and 600 sites (representing up to approximately 30% of the total area) and fitted Weibull functions (Flather 1996) constrained to pass through the origin. Optimal site selections were obtained using a greedy algorithm that adds sites sequentially based on the maximum number of new species that will be represented with each site added. It is important to represent variability in the performance of site selection procedure to get an indication of the likelihood that it will perform well in practical applications where only a single selection is performed. The most appropriate measure of variation for the SAI remains to be identified (Rodrigues and Brooks 2007) so we used the median as the measure of central tendency with the interquartile range and the 5th and 95th percentiles to indicate variation in the performance of ED. The results are presented graphically using boxplots where values above the median represent upper values for site selections using ED compared to lower values for random site selections, and vice versa for values below the median. Curves were fitted to each percentile using nonlinear quantile regression with the R package quantreg (Koenker 2015) and areas under the curves were found with integration.

Improvements to ED:

The first stage of our analysis was intended to assess how the implementation details of ED affected its performance. Using the lessons learned there we then investigated how the ED strategy might be optimized for application in conservation planning by including additional environmental variables or using GDM models to produce distance matrices that better represent community turnover. We created

distance matrices using principal components analysis with all of the environmental variables in the Climond dataset, as well as altitude, latitude and longitude. Broken stick analysis showed four significant components accounting for 92% of the variation in original variables. We used these four variables to calculate Euclidean distances between sites for the ED procedure.

To develop GDMs for each taxon we first selected a subset of relatively well sampled sites for each empirical dataset. Our procedure for identifying these sites was based on the assumption that differences in species richness between environmentally similar sites (sites that are close together in ordination space) are a reflection of sampling bias rather than actual differences in species richness (Robertson and Barker 2006). Therefore sites with the highest species richness relative to their neighbours in ordination space should be the best sampled and most representative of community composition. We used a moving window to identify well sampled sites in ordination space. The moving window had dimensions of 0.1 times the length of the principal component axes and moved over the ordination space in step sizes of 0.5 times the window size. At each step the site with the highest richness within the window was identified. All the sites that were identified by the moving window were then plotted along the first two ordination axes using a grey scale to indicate their relative species richness values and a final subset was chosen by visual inspection. For artificial species and modeled butterflies we used 200 randomly selected sites as there is effectively no sampling bias in these datasets. GDM's were produced in R using the `gdm.fit` function in the `GDM_R Distribution_Pack_V1.1` available at <https://sites.google.com/site/gdmsoftware/>, using the default input parameters. All environmental variables in the Climond dataset, and geographic distance between QDSs, were included as predictors. The GDM procedure identifies a subset of significant environmental variables, and produces plots of predicted community turnover rates as a function of each of those variables. We inspected these plots to identify variables that had relatively low effects on turnover and removed them from the models by running the modeling procedure again with the remaining variables. In all cases this

improved the fit of the models. The GDMs were then used to predict community dissimilarity for all pairs of QDSs for use in the ED procedure.

Finally we assessed whether weighting predicted community dissimilarity from the GDM models with predicted species richness had any effect on performance. We used the same set of well sampled sites identified for each taxon and modeled richness as a function of environmental conditions using generalized linear models with a Poisson error distribution and log link function. Richness was predicted for all QDSs using the resulting models and distances between sites were weighted multiplicatively using the results.

We applied the modified maximum diversity procedure with a greedy site selection heuristic to each of the distance matrices using the same numbers of sites as we did for the earlier analysis. We used the SAI to evaluate the results but only present median values as relative performance of the different methods was of primary interest in this case, and because the degree of variation in each case was similar to the previous analysis.

Results:

Our analysis of the importance of the form of the optimization problem describing ED and the heuristic used to select sites showed that in all cases ED outperformed random site selections with median SAI values ranging from 0.1 to 0.6 (Fig. 2). In all cases except modeled distributions for butterflies the ‘continuous’ p -median and modified maximum diversity clearly outperformed ‘discrete’ p -median selections. ‘Continuous’ p -median and modified maximum diversity produced similar results in all cases. There appeared to be no significant effect of the heuristic applied with greedy and interchange procedures producing similar results for each class of optimization problem. The largest SAI values (> 0.5) were obtained for the artificial species distribution. The lowest values were for reptiles with values

of 0.10 and 0.13 using the discrete p -median and 0.26 – 0.31 for the ‘continuous’ p -median or modified maximum diversity.

The boxplots presented in Figure 2 give an indication of the likelihood of a single ED selection recovering fewer species than a single random selection for a particular sample size. In several cases the whiskers for the SAI extend below zero, indicating that the worst ED selections recover fewer species than the best random site selections. Where lower quartiles extend below zero (corresponding with median SAI values closer to zero) a larger proportion of ED selections perform poorly compared to random site selections. For all taxa except modeled butterflies, the interquartile range of SAI for the ‘continuous’ p -median and modified maximum diversity were above zero indicating a relatively low probability of random selections outperforming ED selections.

Table 3 shows the relative performance of ED selections when more environmental variables are considered and when environmental and species data are combined using GDMs. In all cases using predicted dissimilarities from GDM models or distances derived from the larger set of variables produced large increases in SAI values compared with selections based on six environmental variables. Median SAI values ranged from 0.44 for reptiles to 0.81 for frogs. SAI values for selections using predicted distances from GDMs were lower than selections using all raw environmental variables in all cases except reptiles, where they were comparable. The results for selections based on distances predicted from GDMs together with predicted richness were inconsistent. SAI values were much lower for artificial species and modeled butterflies when richness was combined with predicted distances, but the effect was minor in other taxa. The values for the deviance explained by the GDMs ranged between 69% and 93% and transformed intercept values were between 0.01 and 0.24. These values suggest good models of compositional dissimilarity for the input data. There was no significant correlation between deviance explained for the GDMs and the SAI values for ED selections based on those models (Spearman rank correlation coefficient: 0.14, $P = 0.80$).

Discussion:

Our goal with this study was to evaluate how ED performs as a surrogate for biodiversity in systematic conservation planning, and specifically how technical aspects of its implementation may influence its performance. Our results show that the most important consideration for implementation of ED is the choice of underlying optimization problem. Our results show the ‘continuous’ form of ED outperforms the ‘discrete’ form for recovering species diversity, supporting the arguments put forward by Faith (2003), Faith et al. (2004) and Faith (2011) that poor performance of ED in previous studies may be attributable to incorrect choice of the underlying model . Site selections must be independent of varying densities of sites in environmental space, which is not the case with ‘discrete’ ED. Using ‘continuous’ ED minimizes redundancy in the set of environmental conditions represented in selected sites (Fig. 2). Since the objective of ED is to maximize coverage of environmental conditions, and not servicing demand from a set of users as in the sense of the classical p -median location problems, demand must be formulated accordingly to produce uniformly distributed sets of sites using p -median selection procedures. Our new implementation of ED using a modified form of maximum diversity arguably represents a better choice of underlying optimization model than p -median location allocation. Artificial demand points are not required as in ‘continuous’ ED, so an ordination space does not need to be created in cases where environmental distances are calculated directly. Selected sites interact with each other in the selection procedure independently of the relative locations of unselected sites and so selections are not influenced by varying densities of sites in environmental space. Given that this approach produces very similar results to ‘continuous’ ED we propose that it should be the method of choice for ED applications.

Our comparison of the various heuristic procedures assessed showed that the form of the heuristic applied does not significantly affect performance. Importantly, simultaneous heuristic procedures are not essential. This means that the choice of heuristic can focus on practical considerations such as ease of implementation, computation speed, and specific requirements of the problem at hand. For example, the greedy site selection procedure can be used in situations where the order in which sites will be added to a protected area network is an important consideration (Margules and Pressey 2000). In conservation planning projects additional protected areas are likely to be added to a protected area network incrementally and the number of protected areas that will ultimately be created will always be uncertain. Real world conservation planning is more likely to proceed in a 'greedy' fashion than a theoretically more optimal, 'simultaneous' fashion (Pressey et al. 1996), so greedy procedures, and optimization models that perform well, may present some benefits.

Although mean SAI values were always significantly greater than zero, they were lowest for reptiles where distributions are known to be strongly affected by dispersal limitations (Williams et al. 2000, Crnobrnja-Isailovic 2007, Qian 2009, Qian and Ricklefs 2012). This might lend some support to the propositions by Araújo et al. (2001, 2003, 2004) and Hortal et al. (2009) that the performance of ED will be determined by the degree to which species distributions are influenced by biotic and biogeographic factors and consequently their degree of disequilibrium with environmental conditions. However, higher SAI values were obtained for butterflies and frogs where these factors also play a role, and lower values were obtained for birds where dispersal limitation is likely less than it would be for butterflies, and certainly so for frogs. Why would ED selections perform better for butterflies and frogs but not for reptiles or birds? We found no relationship between the deviance explained by the GDMs for each taxon, which might indicate the degree to which distributions are determined by environmental factors, or the range-size frequency distribution for each taxon, and performance of ED (Fig. 3). While this certainly presents an opportunity for further research it might suggest that the performance of ED will

depend on idiosyncrasies of the taxon under consideration, the dataset used for evaluation, and factors related to the specific region of interest.

Understanding the importance of the other factors we investigated on the performance of ED is more complex than the effect of the optimization problem and heuristic used to solve it (Table 3). We found the greatest increase in ED performance when we included the total set of 36 environmental variables with geographic coordinates to produce distances for ED selections. Whether this improvement was the result of increased variation represented by the distance matrix and thus a greater capacity for discrimination between sites in the selection process, or inclusion of the important variables influencing turnover patterns in biodiversity in the region, is difficult to determine. Beier and Albuquerque (2015) used different numbers of variables for the different regions their empirical datasets came from but didn't specifically examine the relationship between the number of variables and the performance of ED. They selected the individual variables that were most correlated with each significant principal component axis for their analysis. Our results suggest that using principal components derived from all available environmental data as well as geographic distance to produce distance matrices is also a viable approach.

The results we obtained from using GDMs to combine environmental and species data for ED selections were disappointing, and do not support the suggestions by Faith (2003), Araújo et al. (2003), and Faith et al. (2004) that this would be the ideal approach for using ED as a surrogacy strategy. However, these results must be interpreted with caution. Theoretically, ED selections based on distances that accurately represent compositional turnover between sites should result in very high rates of species recovery. The original intent of ED was to use environmental turnover between sites as a surrogate for species compositional turnover. GDMs refine this by providing predictions of compositional dissimilarity based on environmental conditions and other variables such as geographic distance in a rigorous statistical framework, and so should produce distances that better reflect turnover

than distances derived from environmental data directly. The problem is that GDM predicts dissimilarity only, whereas maximising species diversity relies on the concept of complementarity of species composition between sites (Arponen et al. 2008). Complementarity is a function of both turnover and species richness. Arponen et al. (2008) showed that including species richness was essential to maximize species diversity from site selections when using distances derived from GDM. However, when we combined species richness with our GDM results performance decreased in almost all instances. This might be because site selections were biased towards poorly sampled regions in the empirical datasets thus giving a poor indication of real performance of this approach, but the same results were obtained for the artificial species dataset where this problem does not exist. This suggests a fundamental underlying difference between using raw environmental dissimilarities as a surrogate for turnover in biodiversity, and using measures that represent compositional turnover between assemblages directly. Our approach appears to work well with environmental distances, while alternative models may be more appropriate when using metrics of community dissimilarity and species richness.

Our analysis was performed at a relatively coarse spatial scale to match the scale of the available distribution datasets. We did not specifically assess the impact of scale on the performance of ED. Also, conservation planning projects are likely to take place at finer spatial scales than the QDS scale. The impact of scale on the performance of ED needs to be addressed more specifically if ED is to be used in practical applications. We expect that species accumulation curves will be flatter at finer spatial resolutions as a consequence of the species-area relationship, and that the abundance frequency distribution of the focus taxon might influence the shape of the curve too. The absolute performance of ED in these situations will depend on their influence on random site selections, and also on the appropriateness of the chosen environmental variables at the scale of interest (Leibold et al. 2004, Williams et al. 2012).

Concluding remarks:

Conservation planning is a complex process that requires much more than simply identifying sets of sites with complimentary biodiversity. Other considerations include the spatial configuration of those sites, probabilities of species occurrence or persistence at selected sites, inclusion of ecological processes, the distribution and nature of threats to biodiversity, available resources for conservation, and others (e.g. Rouget et al. 2003, 2006, Wilson et al. 2005, 2009, Naidoo et al. 2006). Representing biodiversity remains central to any systematic conservation planning initiative though, and the results of our assessment suggest that the ED strategy has the potential to play a valuable role in this process. The best SAI values we obtained are comparable with SAI values for taxon based surrogates and are higher than values for environmental surrogates reported in a review of surrogate effectiveness by Rodrigues and Brooks (2007), making ED an important competitor for other surrogacy based approaches. Our results compliment the recent study by Beier and Albuquerque (2015) which also showed significant potential for ED as a surrogacy strategy. Further research is needed using a range of empirical datasets for different taxa, different geographic regions and different spatial scales to fully understand the value of ED as a surrogacy strategy for conservation planning. In addition, clarification of the conceptual terminology used in further research on ED is required (Table 1). It is important to draw a distinction between the optimisation model, the computational procedure used to find solutions for that model, and the nature of the distance matrix that is being used. There remains scope for improvement in all three of these components of the ED strategy.

Acknowledgements:

449 Les Underhill, Rene Navarro, and Michael Brooks provided access to the species distribution data
450 managed by the Animal Demography Unit of the University of Cape Town. A large number of Citizen
451 Scientist volunteers deserve recognition for their contribution to these datasets. Silvia Mecenero and
452 Colin Beale kindly provided access to their butterfly distribution models. Glen Manion provided
453 information on the Generalized Dissimilarity Modeling package used. Rafa Marti and Micael Gallego
454 provided a program for testing maximum diversity selections. Joaquín Hortal, Anni Arponen, Daniel
455 Faith, Alistair Mathie and three anonymous reviewers provided suggestions for improvement of the
456 manuscript. Mark Robertson is supported by the DST-NRF Centre for Invasion Biology. Ian Engelbrecht
457 was supported for a large proportion of this work by the Gauteng Department of Agriculture and Rural
458 Development. Roumen Anguelov provided assistance with developing heuristics and inputs on the
459 manuscript.

460

461 References:

462 Araújo, M. B., Densham, P., Humphries, C., 2003. Predicting species diversity with ED: the quest for
463 evidence. *Ecography* 26, 380–383.

464 Araújo, M. B., Densham, P. J., Williams, P. H., 2004. Representing species in reserves from patterns of
465 assemblage diversity. *J. Biogeog.* 31, 1037–1050.

466 Araújo, M. B., Humphries, C. J., Densham, P. J., Lampinen, R., Hagemeyer, W. J. M., Mitchell-Jones, A. J.,
467 Gasc, J. P., 2001. Would environmental diversity be a good surrogate for species diversity?
468 *Ecography* 24, 103–110.

469 Arponen, A., Moilanen, A., Ferrier, S., 2008. A successful community-level strategy for conservation
470 prioritization. *J. Appl. Ecol.* 45, 1436–1445.

471 Balmford, A., 2003. Conservation planning in the real world: South Africa shows the way. *Tr. Ecol. Evol.*
 472 18, 435–438.

473 Bates, M., Branch, W., Bauer, A., Burger, M. J., Alexander, G., De Villiers, M., 2014. Atlas and Red List of
 474 the Reptiles of South Africa, Lesotho and Swaziland. *Suricata* 1. South African National Biodiversity
 475 Institute, Pretoria.

476 Beier, P., Albuquerque, F., 2015. Environmental diversity as a surrogate for species representation. *Cons.*
 477 *Biol.* 29, 1401–1410.

478 Beier, P., Hunter, M. L., Anderson, M., guest editors, 2015a. Special section: Conserving Nature's Stage.
 479 *Cons. Biol.* 29, 613–617.

480 Beier, P., Sutcliffe, P., Hjort, J., Faith, D. P., Pressey, R. L., Albuquerque, F., 2015b. A review of selection-
 481 based tests of abiotic surrogates for species representation. *Cons. Biol.* 29, 668–679.

482 Cowling, R. M., Pressey, R. L., Rouget, M., Lombard, A. T., 2003. A conservation plan for a global
 483 biodiversity hotspot—the Cape Floristic Region, South Africa. *Biol. Cons.* 112, 191–216.

484 Crnobrnja-Isailovic, J., 2007. Cross-section of a refugium: genetic diversity of amphibian and reptile
 485 populations in the Balkans. In Ferrand, N. and Weiss, S., (Eds). *Phylogeography of Southern*
 486 *European Refugia*. Springer. pp 327–337.

487 Eddelbuettel, D., Francois, R., 2011. Rcpp: Seamless R and C++ Integration. *J.Stat. Software* 48, 1–14.

488 ESRI, 2013. ArcGIS Desktop: Release 10.2. Environmental Systems Research Institute, Redlands,
 489 California.

490 Fairbanks, D. H. , Reyers, B., van Jaarsveld, A. S., 2001. Species and environment representation:
 491 selecting reserves for the retention of avian diversity in KwaZulu-Natal, South Africa. *Biol. Cons.* 98,
 492 365–379.

493 Faith, D. P., 2016. Using phylogenetic dissimilarities among sites for biodiversity assessments and
 494 conservation. In Pellens, R. and Grandcolas, R. (Eds). *Biodiversity Conservation and Phylogenetic*
 495 *Systematics: preserving our evolutionary heritage in an extinction crisis*. Springer, Dordrecht.

496 Faith, D. P. 2003., Environmental diversity (ED) as surrogate information for species-level biodiversity.
 497 *Ecography* 26, 374–379.

498 Faith, D. P., 2011. Attempted tests of the surrogacy value of the ED environmental diversity measures
 499 highlight the need for corroboration assessment of surrogacy hypotheses. *Ecol. Ind.* 11, 745–748.

500 Faith, D. P., Ferrier, S., 2002. Linking beta diversity, environmental variation, and biodiversity
 501 assessment. Worked example available at [http://australianmuseum.net.au/document/linking-](http://australianmuseum.net.au/document/linking-beta-diversity-environmental-variation-and-biodiversity-assessment)
 502 [beta-diversity-environmental-variation-and-biodiversity-assessment](http://australianmuseum.net.au/document/linking-beta-diversity-environmental-variation-and-biodiversity-assessment).

503 Faith, D. P., Ferrier, S., Walker, P. A., 2004. The ED strategy: how species-level surrogates indicate
 504 general biodiversity patterns through an “environmental diversity” perspective. *J. Biogeog.* 31,
 505 1207–1217.

506 Faith, D. P., Walker, P. A., 1996. Environmental diversity: on the best-possible use of surrogate data for
 507 assessing the relative biodiversity of sets of areas. *Biodiv. Cons.* 5, 399–415.

508 Ferrier, S. 2002., Mapping spatial pattern in biodiversity for regional conservation planning: where to
 509 from here? *Syst. Biol.* 51, 331–363.

510 Ferrier, S., Drielsma, M., Manion, G., Watson, G., 2002. Extended statistical approaches to modelling
511 spatial pattern in biodiversity in northeast New South Wales. II. Community-level modelling. *Biodiv.*
512 *Cons.* 11, 2309–2338.

513 Ferrier, S., Manion, G., Elith, J., Richardson, K., 2007. Using generalized dissimilarity modelling to analyse
514 and predict patterns of beta diversity in regional biodiversity assessment. *Div. Distr.* 13, 252–264.

515 Ferrier, S., Watson, G., 1997. An evaluation of the effectiveness of environmental surrogates and
516 modelling techniques in predicting the distribution of biodiversity. Environment Australia.

517 Field, R., Hawkins, B. A., Cornell, H. V., Currie, D. J., Diniz-Filho, J. A. F., Guégan, J.F., Kaufman, D. M.,
518 Kerr, J. T., Mittelbach, G. G., Oberdorff, T., O'Brien, E. M., Turner, J. R. G., 2009. Spatial species-
519 richness gradients across scales: a meta-analysis. *J. Biogeog.* 36, 132–147.

520 Flather, C. H., 1996. Fitting species-accumulation functions and assessing regional land use impacts on
521 avian diversity. *J. Biogeog.* 23, 155–168.

522 Gaston, K., 2003., *The Structure and Dynamics of Geographic Ranges*. Oxford University Press.

523 Gesch, D. B., Verdin, K. L., Greenlee, S. K., 1999. New land surface digital elevation model covers the
524 Earth. *Eos, Trans. Am. Geophys. Un.* 80, 69–70.

525 Hakimi, S. L. 1965. Optimum distribution of switching centers in a communication network and some
526 related graph theoretic problems. *Op. Res.* 13, 462–475.

527 Harrison, J., Allan, D., Underhill, L., Herremans, M., Tree, A., Parker, V., Brown, C., 1997. *The Atlas of*
528 *Southern African Birds*. BirdLife South Africa.

529 Harrison, J., Underhill, L., Barnard, P., 2008. The seminal legacy of the Southern African Bird Atlas
530 Project. *S. Afr. J. Sci.* 104, 82–84.

531 Hawkins, B. A., Field, R., Cornell, H. V., Currie, D. J., Guégan, J. F., Kaufman, D. M., Kerr, J. T., Mittelbach,
532 G. G., Oberdorff, T., O’Brien, E. M., Porter, E. E., Turner, J. R. G., 2003. Energy, water, and broad
533 scale geographic patterns of species richness. *Ecol.* 84, 3105–3117.

534 Hortal, J., Araújo, M. B., Lobo, J. M., 2009. Testing the effectiveness of discrete and continuous
535 environmental diversity as a surrogate for species diversity. *Ecol. Ind.* 9, 138–149.

536 Hromkovič, J., 2010. *Algorithmics for Hard Problems: Introduction to Combinatorial Optimization,*
537 *Randomization, Approximation, and Heuristics.* Springer.

538 Hubbell, S. 2001., *The Unified Neutral Theory of Biodiversity and Biogeography.* Princeton University
539 Press, Princeton, New Jersey.

540 Van Jaarsveld, A., Freitag, S., Chown, S., Muller, C., Koch, S., Hull, H., Bellamy, C., Kruger, M., Endrody-
541 Younga, S., Mansell, M., Scholtz, C., 1998. Biodiversity assessment and conservation strategies.
542 *Science* 279, 2106–2108.

543 Jackson, D. 1993., Stopping rules in principal components analysis: comparison of heuristic and
544 statistical methods. *Ecology* 74, 2204–2214.

545 Koenker, R., 2015. quantreg: Quantile Regression. R package version 5.11. Available at [http://cran.r-](http://cran.r-project.org/package=quantreg)
546 [project.org/package=quantreg](http://cran.r-project.org/package=quantreg).

547 Kriticos, D. J., Webber, B. L., Leriche, A., Ota, N., Macadam, I., Bathols, J., Scott, J. K., 2012. CliMond:
 548 global high-resolution historical and future scenario climate surfaces for bioclimatic modelling.
 549 Met. Ecol. Evol. 3, 53–64.

550 Lombard, A. T., Hilton-Taylor, C., Rebelo, A. G., Pressey, R. L., Cowling, R. M., 1999. Reserve selection in
 551 the Succulent Karoo, South Africa: coping with high compositional turnover. Pl. Ecol. 142, 35–55.

552 Margules, C. R., Pressey, R. L., 2000. Systematic conservation planning. Nature 405, 243–53.

553 Marti, R., Gallego, M., Duarte, A., Pardo, E., 2013. Heuristics and Metaheuristics for the Maximum
 554 Diversity Problem. J. Heur. 19, 591–615.

555 Mecenero, S., Altwegg, R., Colville, J., Beale, C., 2015. Roles of spatial scale and rarity on the relationship
 556 between butterfly species richness and human density in South Africa. PLoS One 10, e0124327.

557 Mecenero, S., Ball, J., Edge, D., Hamer, M., Henning, G., Kruger, M., Pringle, E., Terblanche, R., Williams,
 558 M., 2013. Conservation Assessment of Butterflies of South Africa, Lesotho and Swaziland: Red Rist
 559 and Atlas. Saftronics (Pty) Ltd., Johannesburg & Animal Demography Unit, Cape Town.

560 Minter, L., Burger, M., Harrison, J., Braak, H., Bishop, P., and Kloepfer, D., 2004. Atlas and Red Data Book
 561 of the Frogs of South Africa, Lesotho and Swaziland. Smithsonian Institution, Washington, D.C.

562 Mladenović, N., Brimberg, J., Hansen, P., Moreno-Pérez, J. A., 2007. The p -median problem: A survey of
 563 metaheuristic approaches. Eur. J. Op. Res. 179, 927–939.

564 Murray, A. T., Church, R. L., 1996. Applying simulated annealing to location-planning models. J. Heur. 2,
 565 31–53.

566 Naidoo, R., Balmford, A., Ferraro, P. J., Polasky, S., Ricketts, T. H., Rouget, M., 2006. Integrating
567 economic costs into conservation planning. *Tr. Ecol. Evol.* 21, 681–687.

568 New, T. 1998., *Invertebrate surveys for conservation*. Oxford University Press, Oxford, U.K.

569 Oksanen, J., Minchin, P. R., 2002. Continuum theory revisited: What shape are species responses along
570 ecological gradients? *Ecol. Mod.* 157, 119–129.

571 Oliver, I., Beattie, A., 1996. Designing a cost-effective invertebrate survey : a test of methods for rapid
572 assessment of biodiversity. *Ecol. Appl.* 6, 594–607.

573 Pearl, J., 1984. *Heuristics: Intelligent Search Strategies for Computer Problem Solving*. Addison-Wesley.

574 Pimm, S., 2000. Conservation connections. *Tr. Ecol. Evol.* 15, 262–263.

575 Ponder, W., Lunney, D., 1999. *The other 99%: the conservation and biodiversity of invertebrates*. Royal
576 Zoological Society of New South Wales.

577 Pressey, R. L., Possingham, H. P., Margules, C. R., 1996. Optimality in reserve selection algorithms: When
578 does it matter and how much? *Biol. Cons.* 76, 259–267.

579 Qian, H. 2009., Global comparisons of beta diversity among mammals, birds, reptiles, and amphibians
580 across spatial scales and taxonomic ranks. *J. Syst. Evol.* 47, 509–514.

581 Qian, H., Ricklefs, R. E., 2012. Disentangling the effects of geographic distance and environmental
582 dissimilarity on global patterns of species turnover. *Glob.Ecol.Biogeog.* 21, 341–351.

583 R Core Team, 2013. *R: A language and environment for statistical computing*. R Foundation for Statistical
584 Computing, Vienna, Austria.

585

586 Reese, J. 2006., Solution methods for the p -median problem: An annotated bibliography. Networks 48,
587 125–142.

588 Resende, M. G. C., Werneck, R. F., 2007. A fast swap-based local search procedure for location problems.
589 An. Op. Res. 150, 205–230.

590 Reyers, B. 2004., Incorporating anthropogenic threats into evaluations of regional biodiversity and
591 prioritisation of conservation areas in the Limpopo Province, South Africa. Biol. Cons. 118, 521–
592 531.

593 Reyers, B., Wessels, K., van Jaarsveld, A., 2002. An assessment of biodiversity surrogacy options in the
594 Limpopo Province of South Africa. Afr. Zool. 37, 185–195.

595 Robertson, M., Barker, N., 2006. A technique for evaluating species richness maps generated from
596 collections data. S. Afr. J. Sci. 102, 77–84.

597 Robertson, M. P., Cumming, G. S., Erasmus, B. F. N., 2010. Getting the most out of atlas data. Div. Distr.
598 16, 363–375.

599 Rodrigues, A. S. L., Brooks, T. M., 2007. Shortcuts for biodiversity conservation planning: the
600 effectiveness of surrogates. An. Rev. Ecol. Evol. Syst. 38, 713–737.

601 Rodríguez, M. Á., Belmontes, J. A., Hawkins, B. A., 2005. Energy, water and large-scale patterns of reptile
602 and amphibian species richness in Europe. Acta Oecologica 28, 65–70.

603 Rouget, M., Cowling, R. M., Lombard, A. T., Knight, A. T., Kerley, G. I. H., 2006. Designing large-scale
604 conservation corridors for pattern and process. *Cons. Biol.* 20, 549–561.

605 Rouget, M., Cowling, R. M., Pressey, R. L., Richardson, D. M., 2003. Identifying spatial components of
606 ecological and evolutionary processes for regional conservation planning in the Cape Floristic
607 Region, South Africa. *Div. Distr.* 9, 191–210.

608 Sarkar, S., Pressey, R. L., Faith, D. P., Margules, C. R., Fuller, T., Stoms, D. M., Moffett, A., Wilson, K. A.,
609 Williams, K. J., Williams, P. H., Andelman, S., 2006. Biodiversity conservation planning tools:
610 present status and challenges for the future. *An. Rev. Env. Resources* 31, 123–159.

611 Underhill, L. G. 1994., Optimal and suboptimal reserve selection algorithms. *Biol. Cons.* 70, 85–87.

612 Ward, D. F., Lariviere, M. C., 2004. Terrestrial invertebrate surveys and rapid biodiversity assessment in
613 New Zealand: Lessons from Australia. *N. Zeal. J. Ecol.* 28, 151–159.

614 Williams, P., Humphries, C., Araújo, M., Lampinen, R., Hagemerijer, W., Gasc, J., Mitchell-Jones, T., 2000.
615 Endemism and important areas for representing European biodiversity: a preliminary exploration
616 of atlas data for plants and terrestrial vertebrates. *Belg. J. Ent.* 2, 21–46.

617 Wilson, K. A., Carwardine, J., Possingham, H. P., 2009. Setting conservation priorities. *Ann. N. York Ac.*
618 *Sci.* 1162, 237–264.

619 Wilson, K., Pressey, R. L., Newton, A., Burgman, M., Possingham, H., Weston, C., 2005. Measuring and
620 incorporating vulnerability into conservation planning. *Env. Man.* 35, 527–543.

621 Yang, W., Ma, K., Kreft, H., 2013. Geographical sampling bias in a large distributional database and its
622 effects on species richness-environment models. *J. Biogeog.* 40, 1415–1426.

624
 625 Figure 1: Ordination space derived from Principal Components Analysis of six environmental variables
 626 showing the locations of all existing sites (grey) with examples of 50 sites selected using different
 627 methods. Black squares are the locations of selected sites. Ordination represents 83% of the variation in
 628 the six original environmental variables summarized by the first two component axes.

Figure 2: Performance of ED using different optimization models and heuristic computational procedures using the Species Accumulation Index (SAI, Table 1). Heuristics are presented in Table 2. Horizontal bars indicate median SAI values, boxes indicate the interquartile range, and whiskers indicate the 0.05 and 0.95 percentiles. Values above zero indicate higher levels of species recovery by ED than random site selection. Boxplots provide an indication of the range of potential performance of ED. Abbreviations for optimization models are follows: discr – ‘discrete’ p -median; cont – ‘continuous’ p -median; maxdisp – modified maximum diversity (dispersion).

Table 1: Important concepts regarding Environmental Diversity (ED) and its application as a biodiversity surrogacy strategy.

Concept	Description	References
Environmental Diversity (ED)	A strategy for selecting areas for conservation based on measures of environmental dissimilarity between sites, with an underlying optimization model and a computational procedure to find solutions to that model.	Faith and Walker (1996), elaborated in Faith et al. (2004)
p -median location allocation	A methodology from Operations Research used for determining the optimal locations of a set of facilities to supply a demand from a set of users. The model is specified by an objective function which minimizes the sum of travel distances between each user and their nearest facility in the set. The model includes constraints that ensure that each user is allocated to a facility, and that the solution contains exactly p facilities. Faith and Walker proposed p -median location allocation as the appropriate optimization model for Environmental Diversity.	Hakimi (1965), Reese (2006), Mladenović et al. (2007)
Maximum diversity (location allocation)	A methodology for locating facilities that are maximally spread out from one another across an area. The objective function maximizes the sum of distances between facilities and their neighbours. The same model can be applied in a range of other scenarios (Marti et al. 2013)	Marti et al. (2013), Engelbrecht and Robertson (in prep)

Discrete and continuous ED	<p>ED can take one of two forms based on the way in which demand points are defined for the p-median procedure. In the 'discrete' form existing sites are both potential supply points and demand points, so the distance matrix is symmetric and includes all pairwise distances. In the 'continuous' form an artificial grid of demand points is created to uniformly span the ordination space, and the resulting distance matrix is most likely asymmetric unless the number of demand points is specifically set to be the same as the number of sites. The purpose of the artificial grid of demand points is to encourage the p-median procedure to select sites that are evenly spread across the ordination space.</p>	<p>Faith and Walker (1996), elaborated in Faith et al. (2004)</p>
Heuristics	<p>The goal of optimization is to find the best possible solution to a problem from a number of alternatives. In p-median for example this is the solution that best satisfies the objective of minimising travel distances. Optimization problems can be computationally demanding if a large number of alternatives exist. Heuristic procedures use a number of approaches to simplify the computational complexity so that finding a solution is feasible, but may sacrifice perfect optimality of the final solution to do so. A good heuristic will find the optimal solution, or a near optimal solution (called a feasible solution), with reasonable computation time.</p>	<p>Pearl (1984), Hromkovič (2010)</p>

Generalized dissimilarity modeling (GDM)	A regression based methodology for estimating community turnover as a function of environmental conditions and other measures of dissimilarity between sites, such as geographic distance. The method specifically caters for non-linearity of community turnover with increasing environmental distance, as well as non-linearity of turnover rates along environmental gradients. Resulting models can be used to predict community dissimilarity for pairs of sites and predictions can be used for mapping spatial patterns in community composition, environmental classification, survey site selection, conservation prioritization and predicting impacts of climate change.	(Ferrier 2002, Ferrier et al. 2002, 2007)
Species Accumulation Index (SAI)	A method for assessing the performance of surrogates for representing other aspects of biodiversity in sets of sites selected for conservation. The SAI uses the area under the curve fitted to the number of species recovered by a site selection method and the number of sites selected, and comparing this against curves for random site selection and 'optimal' site selection using the species distribution data. The index is calculated using the formula $(S-R)/(O-R)$ where S is the area under the curve for the method of interest, O is the area under the optimal curve, and R is the area under the curve obtained by random selection. Positive values indicate that the surrogate recovers	Ferrier (2002), Rodrigues and Brooks (2007)

greater levels of species diversity than random selections. Values closer to one approach optimal selections. Negative values indicate lower species recovery than random selections. Curves do not represent species accumulation curves in the true sense because each point in the dataset is the result of a unique site selection using the method being evaluated, but plots do resemble species accumulation curves in form and standard approaches for fitting functions describing those curves can be used.

Table 2: Descriptions of the heuristic procedures applied in this assessment of Environmental Diversity (ED) as a biodiversity surrogacy strategy.

Abbreviations presented here are used in Figs. 1 and 3.

Heuristic	Abbreviations	Description	References	Implementation
Interchange	swap	A simultaneous heuristic that calculates exact changes in the objective function as a result of interchanging selected and unselected sites. The interchange that will produce the largest improvement in the objective function is made at each iteration until no profitable interchange exists. The algorithm of Resende and Werneck (2007) re-uses information from one iteration to the next to improve computational performance.	Resende and Werneck (2007)	R (R Core Team 2013) and C++ using Rcpp (Eddelbuettel and Francois 2011)

Greedy algorithm	greedy	An iterative heuristic that first selects a site at random and then finds the next site that will produce the greatest decrease in the objective function. This is repeated until the desired number of sites is obtained.	Reese (2006), Mladenović et al. (2007)	R (R Core Team 2013)
Simulated annealing	SA	A simultaneous heuristic that simulates the cooling process of metallic materials to escape from local optima while searching for a globally optimal solution. A temperature parameter determines whether sites can be included in the solution, even if they will produce an increase in the objective function, and is gradually reduced as the algorithm proceeds. The algorithm converges on a solution by remembering the best solution found as it proceeds and exploring poorer solutions that may ultimately lead to better solutions. The process continues for a set number of iterations or until another stopping rule is satisfied. We used a starting temperature parameter of 1000, and decreased this by 25% every 20 iterations. We found that 500 iterations were generally sufficient to find a stable solution.	Murray and Church (1996)	Java

Table 3: Median Species Accumulation Index (SAI) values for site selections using modified maximum dispersion with various distance measures. Site selections using ED were based on a subset of six environmental variables from the Climond dataset, the total set of Climond variables together with elevation and geographic coordinates, predicted community dissimilarity derived from Generalized Dissimilarity Models (GDM) and the same predicted dissimilarity values weighted with species richness. Values closer to 1 indicate species recovery similar to complementarity based selections using raw distribution data, while values closer to zero indicate species recovery similar to random site selection.

Taxon	6 variables	All variables	GDM	GDM + richness
Artificial species	0.60	0.76	0.76	0.41
Birds	0.42	0.61	0.57	0.60
Modeled butterflies	0.41	0.69	0.49	0.25
Butterflies	0.46	0.65	0.53	0.50
Frogs	0.51	0.81	0.71	0.61
Reptiles	0.26	0.44	0.47	0.49