Genetics and Molecular Research

Exhaustive search for conservation networks of populations representing genetic diversity

J.A.F. Diniz-Filho[1], J.V.B.P.L. Diniz[2] and M.P.C. Telles[3]

[1]Departamento de Ecologia, Instituto de Ciências Biológicas, Universidade Federal de Goiás, Campus II, Samambaia/Itatiaia, Goiânia, GO, Brasil
[2]Rede de Pesquisa GENPAC, Universidade Federal de Goiás, Goiânia, GO, Brasil
[3]Departamento de Genética, Instituto de Ciências Biológicas, Universidade Federal de Goiás, Goiânia, GO, Brasil

Corresponding author: J.A.F. Diniz-Filho
E-mail: diniz@ufg.br

Genet. Mol. Res. 15 (1): gmr.15017525
Received August 25, 2015
Accepted November 4, 2015
Published January 29, 2016
DOI http://dx.doi.org/10.4238/gmr.15017525

ABSTRACT. Conservation strategies routinely use optimization methods to identify the smallest number of units required to represent a set of features that need to be conserved, including biomes, species, and populations. In this study, we provide R scripts to facilitate exhaustive search for solutions that represent all of the alleles in networks with the smallest possible number of populations. The script also allows other variables to be added to describe the populations, thereby providing the basis for multi-objective optimization and the construction of Pareto curves by averaging the values in the solutions. We applied this algorithm to an empirical dataset that comprised 23 populations of *Eugenia dysenterica*, which is a tree species with a widespread distribution in the Cerrado biome. We observed that 15 populations would be necessary to represent all 249 alleles based on 11 microsatellite loci, and that the likelihood of representing all of the alleles with random networks is less than 0.0001. We selected the solution (from two with the smallest number

 ©FUNPEC-RP www.funpecrp.com.br

of populations) obtained for the populations with a higher level of climatic stability as the best strategy for *in situ* conservation of genetic diversity of *E. dysenterica*. The scripts provided in this study are a simple and efficient alternative to more complex optimization methods, especially when the number of populations is relatively small (i.e., <25 populations).

Key words: Cerrado; Computational search; Conservation planning; Genetic diversity; Optimization; R platform

INTRODUCTION

Systematic conservation planning or spatial conservation planning (SCP) aims to establish a set of localities (to form a conservation network) that are necessary to establish a conservation goal (Margules and Pressey, 2000; Sarkar and Illoldi-Rangel, 2010). Implementation of conservation actions faces constraints such as low budget or conflicts with other socio-economic interests; therefore, optimization methods need to be applied to define the minimum number of localities necessary to achieve the goals. Networks can also be designed by incorporating information regarding the constraints and conflicts as well as existing conservation areas (Cabeza and Moilanen, 2001).

In general, SCP has been applied using species (or higher-level units, such as biomes or vegetation formations) as conservation targets. However, there have also been discussions on how to conserve intraspecific genetic diversity, which began with the debate regarding the definitions of “evolutionarily significant units” and management units (Diniz-Filho and Telles, 2002, 2006). More recently, Diniz-Filho et al. (2012) applied SCP reasoning to define a set of priorities at the population level using alleles from microsatellite loci as variables, where the goal was to establish the smallest number of local populations required to represent all known genetic diversity of the species (alleles). Schlottfeldt et al. (2015a,b) developed a more complex model based on a set of multi-objective algorithms, which also allowed several properties of the populations to be optimized simultaneously.

However, for relatively small and straightforward problems (i.e., small number of populations), it is possible to use much simpler algorithms and to search for all possible solutions to the representation problem using an exhaustive search strategy. In this study, we present a set of R scripts for finding solutions to these problems and to select the best (i.e., the smallest number of populations required to represent all of the alleles from a sample), as well as facilitating the simultaneous evaluation of other variables in the solution. We also applied the proposed algorithm to a real problem by finding the set of populations of *Eugenia dysenterica*, a tree species from the Cerrado biome, that are necessary to represent all of the alleles based on 11 microsatellite loci. We then compared the smallest possible networks required to represent the known genetic diversity with populations situated in more climatically stable regions.

MATERIAL AND METHODS

Algorithm and R scripts

The procedure began by creating a small function to decode a number into a binary vector with *n* populations, using the following R script:

 ©FUNPEC-RP www.funpecrp.com.br

```
#function to decode a number into a vector
#x is the number to decode, and s is the length of the vector
decode <-function(x,n){
vec <-numeric(n)
for(j in 1:n){
vec[j] <-x %% 2
x <-floor(x/2)
}
return(vec)
}
```

The function above returns a vector comprising zeros and ones with *n* populations (a "solution", where a value of 1 indicates whether a population is present in the network and 0 otherwise), and must be loaded in advance. Next, using the following script, we found the combinations of populations that form the 2^n solutions and we accumulated some results from each of these solutions for further analysis.

```
pops <-nrow(FAbin)
nall <-ncol(FAbin)
X <-runif(pops,0,1) #defined here as a random variable for the example
nsim <-(2^pops)-1
sgrid <-matrix(0,nsim,3)
progress.bar <-txtProgressBar(min=0,max=nsim,style=3)
for(i in 1:nsim){
setTxtProgressBar(progress.bar,i)
popS <-decode(i,pops)
wpop <-which(popS>0)
BinS <-as.matrix(FAbin[wpop,])
A <-apply(BinS,2,sum)
rich <-sum(ifelse(A>1.0001,1,A))
Xsel <-as.matrix(X[wpop])
sgrid[i,1] <-sum(popS)
sgrid[i,2] <-rich
sgrid[i,3] <-mean(Xsel)
}
end(progress.bar)
```

The results in the matrix *sgrid* were obtained by creating and analyzing subsets from the original input matrix *FAbin*, which contains 1 if the allele is found in the population and 0 otherwise. The matrix *FAbin* can be obtained from the allele frequency matrix *FA* with *n* lines (populations) and *p* columns (allele frequencies) by

```
FAbin <-ifelse(FA >0, 1,0)
```

where *FA* is the matrix with allele frequencies. But we may also decide to exclude rare alleles, with low frequencies, and in this case it may be necessary to exclude some columns from

the matrix *FAbin* before the analysis (if this is not done the target above will be larger than feasible and no solutions would be found). This exclusion can be done by establishing a *threshold* (in the example below, 0.05) and using the following script

```
pa <-ifelse(FA>0,1,0)
threshold <-0.05
pres <-which(apply(FA,2,max)>=threshold,1,0)
FAbin <-pa[,pres]
```

Thus, if the threshold is equal to 0, then all of the alleles present will appear as 1 in *FAbin,* regardless of their frequencies.

The output matrix *sgrid* is set as three columns in the script given above, which in each solution comprise the number of populations in the network, the allelic richness, and the mean of a third variable (X). The variable X was defined above, for simplicity, as a random uniform variable in the populations, but must be actually another "extrinsic" variable measured in each population (see below) and it must be read as an independent object. Thus, other variables (actually, other descriptive statistics) can also be added to the algorithm (increasing the number of columns in *sgrid*), thereby allowing the description of solutions using other characteristics.

After running the script, it is possible to use the matrix *sgrid* for any further analyses of the solutions. For example, it is possible to find the solutions that contain all of the alleles (setting *propA* to a value of 1, below) with *n* number of populations in the solution, after which we can find the solutions with the smallest number of populations using the following script.

```
propA <-1 #percentage of alleles to represent
target <-round(propA*nall)
sbin <-numeric()
for(k in 2:pops){
v <-as.matrix(which(sgrid[,1]==k&sgrid[,2]>=target))
ns <-nrow(v)
sbin[k] <-ns
}
sb <-ifelse(sbin>0,1,0)
minpop<-min(which(sb==1))
best <-as.matrix(which(sgrid[,1]==minpop&sgrid[,2]>=target))
decode(best[1],pops) # or select other values, as for example best[1]
ngood<-nrow(best)
sgrid[best[1],] # look at best solutions and find the best for X (the last column of sgrid)
sgrid[best[2,]
```

The object *minpop* gives the minimum number of populations necessary to represent all alleles, whereas *ngood* gives how many solutions with this minimum number were found. If *ngood* is larger than 1, multiple solutions to the problem exists and one can select which solution to use based on extrinsic attributes of the solution, such as the mean of maximum value of the variable X measured in the populations. The mean values of X in each of the best solutions is obtained by

```
#mean X
meanX <-numeric()
for(i in 1:ngood){
meanX[i] <-t(X)%*%decode(best[i],pops)/minpop
}
```

It is possible to choose among the alternative solutions based on these mean values (see below, in the empirical example). Alternatively, it is also common to calculate a frequency in which populations appear in all solutions, sometimes called “irreplaceability”. The vector of irrepleceability *IRR* can be obtained by

```
#If ngood>2...
irr <-matrix(0,ngood,pops)
for(i in 1:ngood){
irr[i,] <-decode(best[i],pops)
}
IRR <-apply(irr,2,mean)
```

Thus, using these scripts it is possible to find all the possible solutions (i.e., combinations of populations) and to evaluate them, e.g., by selecting the solutions that represent a given proportion of alleles with the smallest number of populations and those with the maximum/minimum value for the extrinsic variable X (if this maximum is of interest). Other scripts for dealing with the *sgrid* matrix are available from J.A.F. Diniz-Filho upon request.

Empirical application

The data used for our empirical example were described by Barbosa et al. (2015; also see Telles et al., 2013). In summary, 736 individuals of *Eugenia dysenterica* (Myrtaceae), a tree species that is distributed widely in the Cerrado region of Central Brazil, were sampled from 23 localities (i.e., populations) throughout most of the species range (Figure 1). Eleven microsatellite loci were genotyped, generating 249 alleles. The presence or absence of these alleles in each population (i.e., without excluding rare alleles) provided the basic matrix used by our algorithm (Diniz-Filho et al., 2012).

We also used Ecological Niche Models (ENMs) based on occurrence records for the species (see Terribile et al., 2012) to estimate the climatic stability in each of the 23 populations by comparing the current climatic suitability with the suitability predicted for the year 2080, employing several algorithms for the ENM and climatic models (Diniz-Filho et al., 2015, 2016).

RESULTS

We observed north-south gradients in the genetic diversity (H_E) and in the principal coordinate of the pairwise F_{ST} matrix (Figure 2). These gradients coincided with the climatic stability patterns obtained using the ENMs, which showed that there will be a future shift in climatic conditions in the northern/northwestern part of the species range.

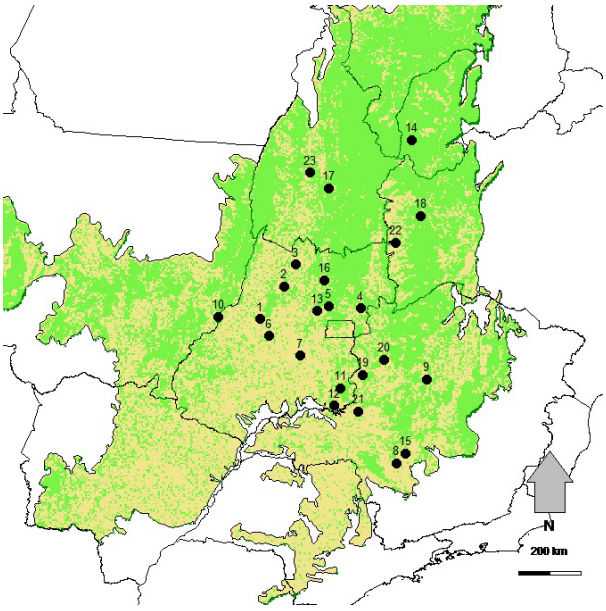

Figure 1. Geographic distribution of 23 populations of *Eugenia dysenterica* analyzed here (the shadow represents the distribution of Cerrado biome).

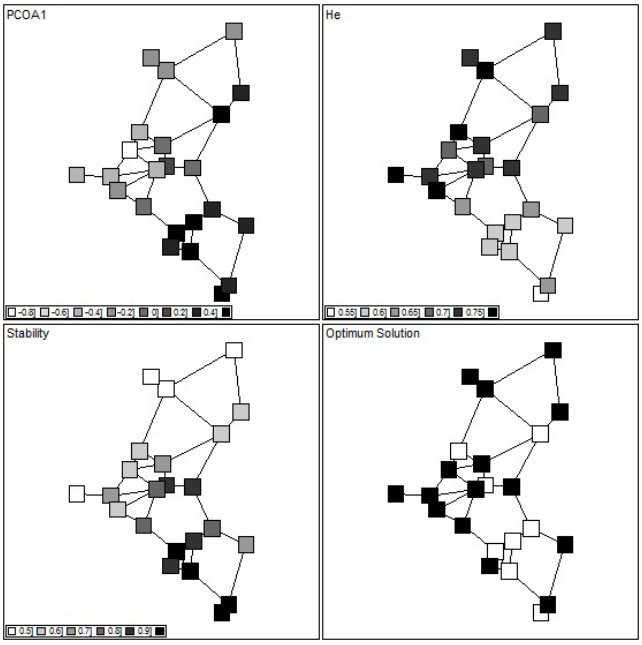

Figure 2. Geographical patterns obtained for the analyses of 249 alleles from 11 microsatellite loci, including the first principal coordinate of pairwise F_{ST} matrix, the expected heterozygosity (H_E), the climatic stability comparing ENM suitabilities from present to 2080, and the optimum solution in SCP, necessary to represent all alleles with the smallest number of populations (15) situated in localities with maximum climatic stability.

For a set of 23 populations, there were 8,388,608 possible solutions, where the 23 populations were combined to form networks, with sizes that varied from one to 23. The SCP's goal is to find the solutions that represent all 249 alleles with the smallest number of populations. Thus, after applying the scripts described above, we found that two slightly different networks satisfied this requirement, each with 15 populations. The relatively high number of populations required to represent all of the alleles was due to the fairly large number of rare alleles in some (or one) populations. In addition, after running 10,000,000 simulations to create random networks with N = 15 populations, we found that on an average only 235 ± 5 alleles were represented by chance alone and that the likelihood of retaining all of the alleles in a random solution was less than 1/10,000 (Figure 3).

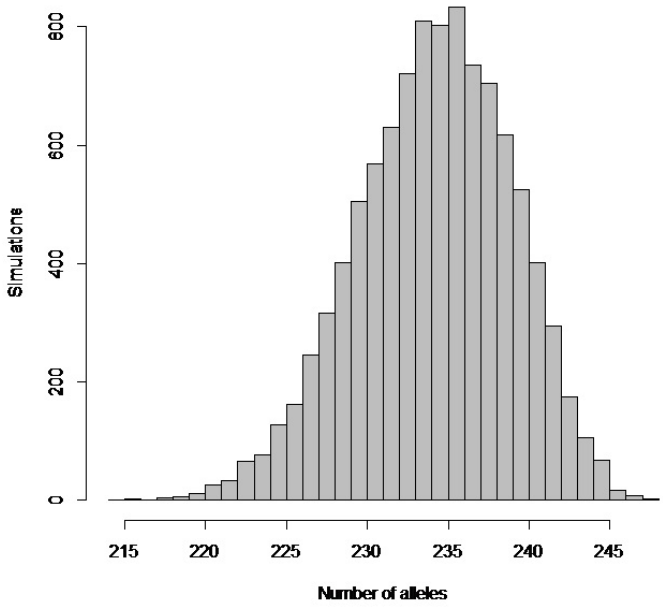

Figure 3. Statistical distribution of the number of alleles represented in 10,000 stochastic combination of populations, showing that none of them was able to represent all 249 alleles.

The two networks with 15 populations differed in terms of their climatic suitability. One of them retained an average of 64.2% of its climatic suitability (Figure 2, lower left) whereas the other retained 61.7% (these solutions were only in one population). The descriptions of all the possible solutions are stored in the *sgrid* object, so it is possible to search for other solutions of interest even if they are not at the minimum. For example, it would be possible to select all of the solutions with k < 16 populations and to compare the climatic stability for all of them. Ultimately, it is more interesting from a conservation perspective, to use a solution with k =16 if the average climatic stability in the populations is much higher than that in the smallest possible set with k = 15. We can also evaluate how increasing the number of populations in the network is related to the climatic stability by plotting the highest stability in the populations with k = 15, 16, 17, etc. (Figure 4), thereby forming a Pareto curve. In our example, the maximum stability was achieved with 20 populations, and if this factor is actually important, we could adopt this solution, although the number of populations is much higher than the minimum required. This is actually the basis of the multi-objective optimization process, as described by Schlottfeldt et al. (2015a,b).

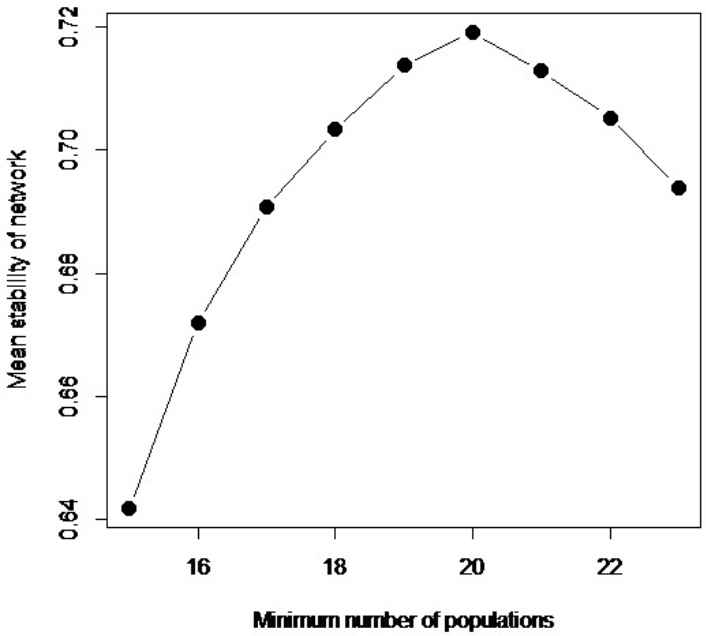

Figure 4. Highest mean climatic stability in networks of *Eugenia dysenterica* with number of populations ranging from 15 to 23.

DISCUSSION

In this study, we presented simple R scripts to search for solutions that represent all of the alleles in networks with the smallest possible number of populations, and we also employed these scripts to analyze a dataset that comprised 23 populations of *E. dysenterica.* Because there are a relatively large number of alleles found in a unique population, this required number is relatively large. This opens the possibility of reducing the number of variables by assuming that very rare alleles would quickly disappear due to effects of genetic drift, and it would not be worthwhile to increase the size of the solution because of such rare alleles. For example, if rare alleles (i.e., with maximum frequencies lower than 0.05) are deleted from populations, about 16% of the alleles are deleted and the minimum solution reduces from 15 populations to 12 populations. Of course, excluding such alleles is only valid under the perspective that these alleles are neutral or quasi-neutral and are being used just to assess overall genetic diversity in the species.

Moreover, all of the possible combinations of populations that can form networks are stored as binary numbers, so it is simple to evaluate these solutions using any other variables of interest (in our example, we evaluated climatic stability), as well as incorporating these variables when deciding between one solution or another. In our study case, selecting the climatic stability as an extrinsic variable to choose among solutions may be interesting because our previous analyses showed a correlation between climatic shifts since the last glacial maximum and current genetic diversity (Diniz-Filho et al. 2015), so selecting populations in more stable regions may provide a more resilient solution to ongoing climate changes.

The scripts presented above are limited to working within the number of distinct solutions

that the R platform is able to represent, i.e., the number of populations cannot be larger than the number of bits used to represent an integer by R. However, this is an exhaustive search process and the complexity of the problem increases as 2^n, so it may be difficult to use this exhaustive search method with more than 25 or 30 populations (see below). For example, running the above algorithm with 23 populations (the empirical example described below) on a Dell desktop using an Intel i-5-3330S with a CPU running at 2.7 GHz and 6 GB of RAM memory required ca. 2.15 h. The complexity of the problem grows at 2^n; therefore, increasing to 25 populations (for example) would require 9.8 h (2.15 x 2 x 2). Expanding the reasoning, approximately 11.5 days would be required to run the algorithm for 30 populations. In addition, the *sgrid* output matrix rapidly becomes difficult to work with, because a set of 25 populations generates a matrix with 33,554,432 lines and 3 columns. For dealing with larger number of populations, it may be necessary to modify the script to avoid accumulating the matrix *sgrid* and to retain only minimum solutions (loosing flexibility in the analyses and not allowing building directly, for example, Pareto curve). Thus, if the problem involves more than 25-30 populations, we consider that other optimization algorithms should be used instead of our exhaustive search script (Diniz-Filho et al., 2012; Schlottfeldt et al., 2015a,b).

More complex algorithms and computer programs for SCP are available (e.g., Diniz-Filho et al., 2012; Schlottfeldt et al., 2015a,b), but they are not available as R scripts or packages at present, while the scripts presented in this study can be easily integrated with other population genetic analyses. In addition, our proposed algorithm may be computationally feasible when dealing with a relatively small number of populations (i.e., <25 populations). Due to constraints in field sampling and laboratory work, most studies on population genetics deal with a small number of populations. Thus, the algorithms and scripts presented in this study will be of interest to a large number of researchers in this field.

Conflicts of interest

The authors declare no conflict of interest.

ACKNOWLEDGMENTS

Our research on population genetics of Cerrado plants was supported by the project “Núcleo de Excelência em Genética e Conservação de Espécies do Cerrado” - GECER (PRONEX/FAPEG/CNPq #CP07-2009), and by several grants and fellowships for the research network GENPAC (“Geographical Genetics and Regional Planning for Natural Resources in Brazilian Cerrado”) from CNPq/MCT/CAPES/FAPEG. Research by J.A.F. Diniz-Filho and M.P.C. Telles was also supported by productivity grants from CNPq.

REFERENCES

Barbosa ACOF, Collevatti RG, Chaves LJ, Guedes LBS, et al. (2015). Range-wide genetic differentiation of *E. dysenterica* (Myrtaceae) populations in Brazilian Cerrado. *Biochem. Syst. Ecol.* 59: 288-296. http://dx.doi.org/10.1016/j.bse.2015.02.004

Cabeza M and Moilanen A (2001). Design of reserve networks and the persistence of biodiversity. *Trends Ecol. Evol. (Amst.)* 16: 242-248.http://dx.doi.org/10.1016/S0169-5347(01)02125-5

Diniz-Filho JAF and Telles MPC (2002). Spatial autocorrelation analysis and the identification of operational units for conservation in continuous populations. *Conserv. Biol.* 16: 924-935. http://dx.doi.org/10.1046/j.1523-1739.2002.00295.x

Diniz-Filho JAF and Telles MPC (2006). Optimization procedures for establishing reserve networks for biodiversity conservation

taking into account population genetic structure. *Genet. Mol. Biol.* 29: 207-214. http://dx.doi.org/10.1590/S1415-47572006000200004

Diniz-Filho JAF, Melo DB, de Oliveira G, Collevatti RG, et al. (2012). Planning for optimal conservation geographical genetic variability within species. *Conserv. Genet.* 13: 1085-1093. http://dx.doi.org/10.1007/s10592-012-0356-8

Diniz-Filho JAF, Rodrigues H, Telles MPC, Oliveira GD, et al. (2015). Correlation between genetic diversity and environmental suitability: taking uncertainty from ecological niche models into account. *Mol. Ecol. Resour.* 15: 1059-1066. http://dx.doi.org/10.1111/1755-0998.12374

Diniz-Filho JAF, Barbosa ACOF, Collevatti RG, Chaves LJ, et al. (2016). Spatial autocorrelation analysis and ecological niche modelling allows inferring geographic range dynamics driving population genetic structure in a Neotropical savanna tree. *J. Biogeog.* 46: 167-177. (doi:10.1111/jbi.12622).

Margules CR and Pressey RL (2000). Systematic conservation planning. *Nature* 405: 243-253. http://dx.doi.org/10.1038/35012251

Sarkar S and Illoldi-Rangel P (2010). Systematic conservation planning: an updated protocol. *Nat. Conserv.* 8: 19-26. http://dx.doi.org/10.4322/natcon.00801003

Schlottfeldt S, Walter MEMT, Carvalho ACPLF, Soares TN, et al. (2015a). Multi-objective optimization for plant germplasm collection conservation of genetic resources based on molecular variability. *Tree Genet. Genomes* 11: 16. http://dx.doi.org/10.1007/s11295-015-0836-3

Schlottfeldt S, Walter MEMT, L F Carvalho AC, Soares TN, et al. (2015b). Multi-objective optimization in systematic conservation planning and the representation of genetic variability among populations. *Genet. Mol. Res.* 14: 6744-6761. http://dx.doi.org/10.4238/2015.June.18.18

Telles MPC, Silva JB, Resende LV, Vianello RP, et al. (2013). Development and characterization of new microsatellites for *Eugenia dysenterica* DC (Myrtaceae). *Genet. Mol. Res.* 12: 3124-3127.http://dx.doi.org/10.4238/2013.February.6.3

Terribile LC, Lima-Ribeiro MS, Araújo M, Bizao N, et al. (2012). Areas of climate stability of species ranges in the Brazilian Cerrado: disentangling uncertainties through time. *Nat. Conserv.* 10: 152-159. http://dx.doi.org/10.4322/natcon.2012.025