Pan-american Association of Computational Interdisciplinary Sciences

Journal of Computational Interdisciplinary Sciences (2011) 2(2): 131-137

Printed version ISSN 1983-8409 / Online version ISSN 2177-8833
http://epacis.net
doi: 10.6062/jcis.2011.02.02.0040

Minimum description length principle to select environmental layers in modeling of species geographical distribution

Elisângela S.C. Rodrigues[1,2], Fabrício A. Rodrigues[2],
Ricardo L.A. da Rocha[1] and Pedro L.P. Corrêa[2]

Manuscript received on October 20, 2010 / accepted on August 30, 2011

ABSTRACT

Environmental issues are calling the attention of people all over the world, mainly in Brazil, which has one of the richest fauna and flora on Earth. Modeling of species geographical distribution is a technique that has been applied in many tasks related to biodiversity conservation. One of the problems of modeling species geographical distribution is to select an adequate set of environmental layers. A frequency distribution of each environmental layer can be represented by a histogram and the cut points of the histograms can be viewed as models. One of the classical problems in selecting a model is overfitting, that is, the super adjustment of the model to the observed data. The Minimum Description Length (MDL) principle has the property of avoiding overfitting when learning the parameters of the model. Thus, this is a promising strategy to be applied in the selection of any kind of model. The MDL principle searches for a model with the shortest description based on the observed data. This is done by finding regularities in data that are used to compress them. This principle was already successfully applied to probability density estimation by regular histograms. Nevertheless, there is a waste in the model representation when the data is non-uniformly distributed because of the high bin count needed to represent the details of high density data. Thus, the aim is to present how the MDL principle with irregular histograms can be used to select a good set of environmental layers. This strategy prevents the waste when representing parts of the data with low density.

Keywords: applied computing in space and environmental sciences, scientific computing in multidisciplinary topic, Niche-based modeling, Minimum Description Length principle.

1 INTRODUCTION

Niche-based models are the result of using an important technique in biology, known as modeling of species geographical distribution. This technique is based on localities where a species was observed and on the environmental conditions of these sites. Modeling of species geographical distribution is applied to several research fields, such as biodiversity loss prevention, identification of suitable areas to reintroduce a given species, assisting in determining priority areas for conservation, management of invasive-species spread, among others [1].

There are several Artificial Intelligence techniques that were already used to construct algorithms for modeling of species geographical distribution, such as Neural Networks [2], Support Vector Machines [3], Genetic Algorithms [4] and Maximum Entropy [5]. Basically, these algorithms search for a model which represents suitable environmental conditions for the species in a given geographic area.

Correspondence to: Elisângela S.C. Rodrigues
[1]Laboratory of Languages and Adaptive Techniques.
[2]Laboratory of Agricultural Automation, School of Engineering of the University of São Paulo, São Paulo, SP, Brazil
E-mails: elisangela.poli.usp@gmail.com / fabricio.poli.usp@gmail.com / rlarocha@usp.br / pedro.correa@usp.br

The input data for the algorithms are of two kinds: georeferenced points indicating where the species were observed and environmental layers. However, these layers are empirically selected and an excessive number of layers can lead to overfitting or underfitting. These are classical problems of machine learning. The first one happens when the model very well describes the examples used to construct it, but has poor performance with examples that were never seen before. The second one happens when the model poorly describes the examples, with low performance in both training and testing.

The Minimum Description Length (MDL) principle [6] has some interesting properties that could be useful in selecting layers for modeling species geographical distribution. Among them is the property of automatically avoiding overfitting when learning the parameters of a model. The main idea behind the MDL principle is to search for a model with the shortest description based on the observed data. This is done by finding regularities in data that are used to compress them.

Thus, the aim is to present a study of environmental layers selection using the MDL principle by histograms. These strategy was already successfully applied in probability density estimation [7], [8]. Therefore, the MDL principle is used to reduce the number of layers without decreasing the generalization ability and the predictive performance of the models.

This paper is structured as follows. In Section 2, there is an overview about the modeling of species geographical distribution. Section 3 gives the main concepts concerning the Minimum Description Length used herein. The methodology adopted to use MDL in environmental layers selection is shown in Section 4. Finally, some discussion about the studied issue is provided in Section 5.

2 OVERVIEW ABOUT MODELING OF SPECIES GEOGRAPHICAL DISTRIBUTION

Modeling tools for species geographical distribution estimate models based on recorded occurrences of a given species and environmental variables collected in the study area. The model is an approximation to the species ecological niche, which represents the set of resources and ecological conditions essential for species to sustain a population for a long period of time [9]. A niche-based model is a probability function that represents the suitability of the environment for the species [5]; in other words, this is a species potential distribution.

Independently of the modeling technique chosen, the input data are the same: occurrence records, also known as occurrence points, and environmental layers, also known as environmental variables. Occurrence data are georeferenced points, that is, latitude and longitude, representing sites where the species were observed. If the species absence is also recorded, the localities are distinguished as presence and absence points, indicating the existence or the non-existence of the species, respectively.

Environmental layers represent the ecological niche of the species [10], that is, the environmental conditions needed for the survival of the species, such as temperature and precipitation. All environmental layers used in modeling of species geographical distribution are also georeferenced and they must belong to the same study region [5]. A potential distribution map can represent current, previous or future scenarios, depending on if the environmental layers are from the present, the past or future estimations.

The modeling process of species geographical distribution can be seen as a method for assisting decision making. Thus, before beginning the modeling process, it is necessary to identify the problem to be solved. In the problem specification, the queries to be answered by the model are formulated and the data to be used are defined [11]. Computationally, the modeling stages of the species geographical distribution can be viewed as illustrated in Figure 1.

After the problem delineation, the species occurrence data and the environmental layers are selected and treated. The data treatment includes, but is not limited to, removal of digitization errors and transformation to some geographically coordinated system. Afterwards, an algorithm should be chosen and its parameters set. A reasonable understanding of the algorithms is recommended for the definition of its parameters because it can affect the algorithm behavior in the model estimation.

Running the selected algorithm generates a model which is projected into a geographic region of interest. The projection results in a georeferenced map that contains the suitability of the species. Finally, the researcher validates the model based on the specialist's previous knowledge. The validation can be assisted by statistics generated by the modeling tool and some data which were not used by the algorithm during the model generation can be used to test the predictive ability of the model. At this level, the model can be accepted and the modeling process finishes, or one may decide to return to any of the previous levels.

The amount of available data as well as the quality of the selected data can affect the modeling results. Hence, methods of pre-analysis can be used to improve the performance of the estimator. The main idea of this work is to use the MDL principle to assist the choice of a suitable set of environmental layers in a

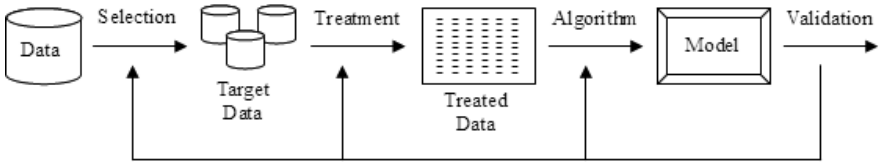

Figure 1 – Computational view point of the Modeling of Species Geographical Distribution.

model estimation for a given species. This is an important task in modeling the species geographical distribution but not much explored so far.

3 THE MINIMUM DESCRIPTION LENGTH PRINCIPLE

Model is a representation of the behavior or characteristics of a process. It is not necessarily a representation of a species distribution. Thus, the aim of this section is to show how the MDL principle can be used in the selection of environmental layers. The MDL principle with histograms has been successfully used in probability density estimation [7], [8]. The histogram is a conceptually simple strategy but capable of creating models with complex properties [7].

Generally, the histograms with bins of equal width, also known as regular histograms, are used for density estimation and the aim of methods for constructing them is to determine the optimal number of bins [8]. Nevertheless, there is a waste in the model representation when the data is non-uniformly distributed because of the high bin count needed to represent the details of high density data [7]. In the same way, there will be a high bin count to represent low density data, unnecessarily.

The irregular histograms, that is, histograms with bins of variable width can be used to avoid this problem. In this case, besides the bin counts, it is necessary to find the best width of each bin, i.e., the best set of cut points. It makes the problem naturally difficult. However, the sets of cut points can be seen as models and the MDL principle can be useful to select the best model according to some complexity measure.

The MDL principle, an approach rooted in the Kolmogorov Complexity [12], uses the regularities found in the observed data as a learning process. One of the advantages of this principle is that it does not need any prior distribution to estimate the probability density of a given data set [6], unlike other methods. The general idea is to calculate the Normalized Maximum Likelihood (NML) distribution of each environmental layer for using the Stochastic Complexity (SC) as a metric of classification.

Stochastic Complexity [13] can be defined as the minimum description length which represents the data considering a class of models C. In the case of histograms, the class of models is the set of cut points. Thus, consider a set of n values $f_j^n = (f_{j1}, \ldots, f_{jn})$ for a given environmental layer f_j in a range $[f_{j(\min)}, f_{j(\max)}]$, in which $f_{j(\min)}$ and $f_{j(\max)}$ are the minimum and maximum values of f_j, respectively. The data are stored in ascending order with a finite precision γ. It means that each element of f_j^n pertains to the set Γ (1). According to Kontkanen and Myllymäki [7], the effect of this parameter in the SC is a constant that can be ignored in the model selection process.

$$\Gamma = \left\{ f_{j(\min)} + t\gamma \mid t = 0, \ldots, \frac{f_{j(\max)} - f_{j(\min)}}{\gamma} \right\} \quad (1)$$

The first step for the histogram construction is to choose the set of cut points that will be considered. There are several possible sets of cut points. A reasonable choice is to put two cut points between all pairs of consecutive values in the data [7]. As the extreme points are automatically included in all sets of cut points, C is

$$\begin{aligned} C = \left(\left\{ f_{ji} - \frac{\gamma}{2} \mid f_{ji} \in f_j^n \right\} \cup \left\{ f_{ji} + \frac{\gamma}{2} \mid f_{ji} \in f_j^n \right\} \right) \\ - \left\{ f_{j(\min)} - \frac{\gamma}{2}, f_{j(\max)} + \frac{\gamma}{2} \right\}. \end{aligned} \quad (2)$$

Once the set of cut points has been defined, the task is to find the best subset c in C according to the SC. The SC is instantiated for the problem of density estimation of environmental layers by histograms as in equation (3).

$$\begin{aligned} SC(f_j^n \mid c) &= \log R_n(h_K) \\ &+ \sum_{k=1}^{K} -h_k \big(\log(\gamma \cdot h_k) - \log(L_k \cdot n) \big), \end{aligned} \quad (3)$$

where K is the number of bins in the histogram, h_k is the number of sample in the k^{th} bin, L_k is the width of the k^{th} bin and $R_n(h_K)$ is the parametric complexity of the histogram with K bins, expressed by equation (4) [7].

$$R_n(h_K) = \sum_{f_{ji} \in f_j^n} P\big(f_{ji} \mid \hat{\theta}(f_{ji})\big), \quad (4)$$

where $P\big(f_{ji} \mid \hat{\theta}(f_{ji})\big)$ is the probability distribution of f_{ji} according to $\hat{\theta}(f_{ji})$, the maximum likelihood estimator.

Parametric complexity is a very difficult formula to compute. However, Kontkanen and Myllymäki [7] developed an algorithm to efficiently compute the parametric complexity in the histogram case. It makes the MDL principle usable in practice and their code was used herein. The MDL principle with irregular histograms was used as a pre-analysis step in the modeling process of species geographical distribution. The purpose was to define the best set of environmental layers for a given species. A pre-analysis over the data can be useful for improving the performance of the algorithm in the model generation. Although the MDL principle was applied to rank the environmental layers, further studies are needed to identify the best number of layers to generate a model.

4 METHODOLOGY

This section describes the occurrence data and the environmental layers used in the experiments. Besides this, the experimental description and some results are presented.

4.1 Occurrence Data

Experiments were carried out with 65 occurrence points of *Xylopia aromatica* species, which is a small tree of the *Annonaceae* family. It is popularly known as *malagueto* and is found in the Brazilian cerrado. It is commonly used for firewood. These data are derived from SinBiota – an environmental information system for the program Biota/Fapesp[1]. From the 65 occurrence points used in the experiments, 45 records were used in the training and 20 records were used to test. Figure 2 shows the training points, represented by black triangles, and the testing points, represented by circles.

4.2 Environmental Data

Since the occurrence points of the species used in the experiments were from the São Paulo state, Brazil, the selected environmental layers were from the same region. Generally, about seven layers are selected to estimate a model. However, as the aim was to show that the MDL principle can assist this choice, 55 environmental layers were selected to generate the model. The spatial resolution used was 30 arc-second and the layers were provided by WorldClim – Global Climate Data [15]. The layers used were:

- Annual mean temperature (Bio_1);
- Mean diurnal range (Bio_2);
- Isothermality (Bio_3);
- Temperature seasonality (Bio_4);
- Maximum temperature of warmest month (Bio_5);
- Minimum temperature of coldest month (Bio_6);
- Temperature annual range (Bio_7);
- Mean temperature of wettest quarter (Bio_8);
- Mean temperature of driest quarter (Bio_9);
- Mean temperature of warmest quarter (Bio_10);
- Mean temperature of coldest quarter (Bio_11);
- Annual precipitation (Bio_12);
- Precipitation of wettest month (Bio_13);
- Precipitation of driest month (Bio_14);
- Precipitation seasonality (coefficient of variation) (Bio_15);
- Precipitation of wettest quarter (Bio_16);
- Precipitation of driest quarter (Bio_17);
- Precipitation of warmest quarter (Bio_18);
- Precipitation of coldest quarter (Bio_19);
- Average monthly minimum temperature (Tmin_1 - Tmin_12);
- Average monthly maximum temperature (Tmax_1 - Tmax_12);
- Average monthly precipitation (Prec_1 - Prec_12);

4.3 Experimental Sketching

The aim of the experiments carried out was to demonstrate that the MDL principle with histograms can be used in the pre-analysis of the modeling process of species geographical distribution to select a good set of environmental layers. The first step was to calculate the Stochastic Complexity of each environmental layer. Just the values of the environmental layers corresponding to the records separated for training were used to calculate the Stochastic Complexity. Figure 3 shows an example of a histogram generated in the Stochastic Complexity calculation.

Afterwards, the layers were arranged according to the increasing order of its Stochastic Complexity and they were divided into three groups. The choice of dividing the set of variables into three groups was random. The purpose of this division was to

[1] http://sinbiota.cria.org.br/

Figure 2 – Training points, represented by black triangles, and test points, represented by circles. All the points are from the São Paulo state, Brazil.

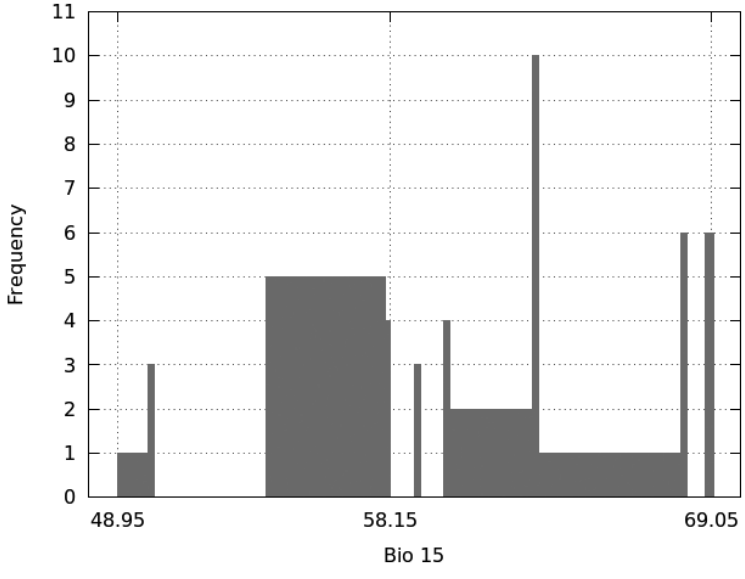

Figure 3 – Histogram generated for the Bio_15 environmental layer.

show that the Stochastic Complexity can be used as an indicator of importance of the layers. Thus, the first hypothesis was that the group with the lowest complexity could estimate better models. Therefore, more experiments are needed to define the appropriate amount of layers in the estimation of species geographical distribution.

Each subset of environmental layers was used to estimate the distribution of the species *Xylopia aromatica* with the maximum entropy algorithm available in the *openModeller* tool, which is a framework with a lot of resources developed in C++ for modeling of species geographical distributions [14]. The last step was to evaluate the results obtained. This evaluation was made using occurrence points unseen by the algorithm during the training. The measure used to evaluate the predictive performance of the estimated models was the AUC (Area Under the Curve).

4.4 Results

After the calculation of the Stochastic Complexity, the environmental layers were arranged by increasing order of the calculated value. Thus, the three groups of layers were:

1. Bio_3, Bio_15, Prec_7, Prec_5, Prec_4, Prec_9, Bio_14, Prec_8, Bio_2, Tmin_1, Tmin_3, Tmin_2, Tmin_4, Tmin_12, Bio_8, Bio_10, Prec_6 e Bio_12;
2. Tmin_5, Tmin_6, Prec_10, Tmax_2, Bio_6, Tmin_7, Tmax_12, Tmin_11, Tmin_10, Tmax_1, Bio_7, Tmax_3, Prec_3, Bio_1, Tmin_9, Bio_11, Prec_12 e Tmin_8;
3. Tmax_4, Prec_2, Bio_9, Tmax_5, Prec_11, Tmax_6, Tmax_11, Tmax_10, Tmax_7, Bio_13, Prec_1, Bio_19, Tmax_8, Tmax_9, Bio_17, Bio_4, Bio_5, Bio_18 e Bio_16.

Table 1 shows the values of AUC collected in the training and in the testing with each one of the three subsets of environmental layers presented above.

Table 1 – Results of the modeling using the three subsets of environmental layers.

Group	AUC – training	AUC – testing
1	0.91	0.92
2	0.80	0.90
3	0.28	0.27

As can be observed in Table 1, the best results were obtained with the first subset of environmental layers. The second group also reached good results, but declined in relation to the first one. The algorithm was not able to associate the last subset of environmental layers to the presented occurrence points, that is, it was not able to learn and estimate a model. This is the reason why the model was not able to generalize, either.

Figure 4 presents the models generated in the training using the environmental layer groups 1, 2 and 3, respectively. Hot colors indicate areas with more suitable environmental conditions to the species and cold colors represent regions less suitable for the species survival. Although the occurrence points were from the São Paulo state, the model was projected for Brazil because the Brazilian cerrado covers a region larger than the São Paulo state and it allows a broader view of the suitable areas for the *Xylopia aromatica* species.

As the aim of this work was to show that the MDL principle can be used to determine the importance of the environmental layers, assisting their choice, it is possible to conclude that the results were satisfactory. However, new analysis to define the best number of environmental layers for a given species are needed, besides the measure of importance of each layer.

5 FINAL DISCUSSION

As can be observed in the previous section, the best results were with the first set of layers. The second group was a little bit worse than the first and the third group was the worst of all. For modeling of species geographical distribution, the test performance is very important because models are generally used to make predictions over unseen data. Thus, it was possible to verify that the MDL principle with irregular histograms can indicate the importance of a layer for a given species.

These results motivate the research and give indications that the MDL principle can be used in other parts of the modeling process, such as to select the niche-based models, for example. Nevertheless, for selecting niche-based models, the instantiation of the MDL principle should be modified, since the histograms are a graphical representation of the frequency distribution of a single variable.

One of the methods available in *openModeller* to determine the importance of a layer is the *Jackknife* [16]. In this method, it is necessary to generate the same number of models as the number of layers, since the estimators are calculated based on

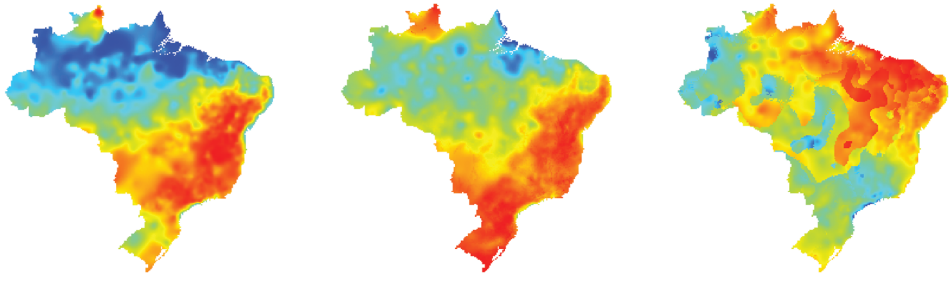

Figure 4 – Estimated models with the maximum entropy algorithm available in *openModeller* using different groups of environmental layers, projected in Brazil.

the generated model. An advantage of the MDL principle in relation to *Jackknife* is that it does not need to create models to give a measure of importance to each layer. Therefore, the computational performance is much better with the MDL principle than with *Jackknife*.

There are several branches of the MDL principle to be researched in modeling of species geographical distribution. Consequently, some future works proposed herein are: the comparison of the results obtained with other approaches to selecting environmental variables; the use of MDL for clustering since the data are multivariate; the study of the viability of increasing the performance regarding the running time using a cluster; the establishment of a method to define the best number of layers.

The experiments carried out and presented herein were the beginning of a promising research field. The aim was to verify the viability of using the MDL principle to select environmental layers in modeling of species geographical distribution. The purpose was successfully attained and several studies regarding this issue are already under consideration.

ACKNOWLEDGMENTS

The authors thank Ph.D. Petri Kontkanen for the MDL code cession used in the experiments. The financial support provided by CAPES (Coordination for the Improvement of Higher Education Personnel) is highly appreciated.

REFERENCES

[1] GRAHAM CH, FERRIER S, HUETTMAN F, MORITZ C & PETERSON AT. 2004. New developments in museum-based informatics and applications in biodiversity analysis, TRENDS in Ecology and Evolution, 19(9): 497–503.

[2] RODRIGUES FA, AVILLA AO, RODRIGUES ESC, CORRÊA PLP, SARAIVA AM & ROCHA, RLA, 2009. Species distribution modeling with neural networks. e-Biosphere 2009, London – UK.

[3] LORENA AC, SIQUEIRA MF, GIOVANNI R, CARVALHO ACPLF & PRATI RC. 2008. Potential Distribution Modelling Using Machine Learning Classifiers. In: The Twenty First International Conference on Industrial, Engineering & Other Applications of Applied Intelligent Systems, 2008, Wroclaw. Lecture Notes in Artificial Intelligence, 5027: 255–264.

[4] PERSONA L, CORRÊA PLP & SARAIVA AM. 2003. Environmental Niche Modeling in Biodiversity with Genetic Algorithms. In: 2nd International Information and Telecommunication Technologies Symposium, Florianópolis. Proceedings of the IEEE – I2TS'2003. (In Portuguese).

[5] PHILLIPS SJ, ANDERSON RP & SCHAPIRE RE. 2006. Maximum entropy modeling of species geographic distributions. Ecological Modelling, 190: 231–259.

[6] GRÜNWALD PD. 2005. Introducing the Minimum Description Length Principle. Advances in Minimum Description Length – Theory and Applications. The MIT Press, 2005, pp. 3–21.

[7] KONTKANEN P & MYLLYMÄKI P. 2007. MDL histogram density estimation. In: M. Meila and S. Shen, editors, Proceedings of the Eleventh International Workshop on Artificial Intelligence and Statistics.

[8] CHAPEAU-BLONDEAU F & ROUSSEAU D. 2009. The minimum description length principle for probability density estimation by regular histograms. Physica A, 388: 3969–3984.

[9] HUTCHINSON GE. 1981. Introduction to Ecology of Populations. Barcelona, Editorial Blume, 492p. (In Spanish).

[10] SIQUEIRA MF. 2005. Use of Fundamental Niche Modeling in the Pattern Evaluation of Vegetal Species Geographic Distribution. PhD Thesis. Department of Ambient Engineering of University of São Carlos. São Carlos/SP – Brazil. (In Portuguese).

[11] SANTANA F, SIQUEIRA MF, SARAIVA AM & CORRÊA PLP. 2008. A reference business process for ecological niche modelling. Ecological Informatics 3: 75–86.

[12] LI M & VITÁNYI P. 1997. An Introduction to Kolmogorov Complexity and its Applications, Springer, Berlin.

[13] RISSANEN J. 1986. Stochastic Complexity and Modeling. Annals of Statistics, 14(3): 1080–1100.

[14] MUÑOZ MES, GIOVANNI R, SIQUEIRA MF, SUTTON T, BREWER P, PEREIRA RS, CANHOS DAL & CANHOS VP. 2009. openModeller: a generic approach to species' potential distribution modeling. GeoInformatica.

[15] HIJMANS RJ, CAMERON SE, PARRA JL, JONES PG & JARVIS A. 2005. Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology, 25: 1965–1978.

[16] RODRIGUES FA, RODRIGUES ESC, SATO LM, MIDORIKAWA ET, CORRÊA PLP & SARAIVA AM. 2008. Parallelization of the Jackknife Algorithm Applied to a Biodiversity Modeling System. Proceedings of the 7th International Information and Telecommunication Technologies Symposium – I2TS'2008, pp. 58–65.