Species distribution modeling applied to MILPA agroecological consortium in Brazil

Matheus Gustavo Alves Sasso, matheus.sasso17@gmail.com
Paula Dornhofer Paro Costa, paulad@unicamp.br

Departamento de Engenharia de Computação e Automação (DCA)
Faculdade de Engenharia Elétrica e de Computação (FEEC)
Universidade Estadual de Campinas (Unicamp)
Campinas, SP, Brasil

Abstract – Despite playing an essential role in the Brazilian economy, traditional agriculture usually leads to several environmental issues. As an alternative to traditional agriculture, agroecology studies how poly-culture creates beneficial interactions with the environment that can reduce the ecological impact of agriculture. In this paper we propose the Species Distribution Modeling (SDM) algorithms to identify regions in which species from the MILPA agroecological consortium can grow in Brazil. We use a pipeline composed by a One Class Support Vector Machine (OCSVM) combined with multiple ensemble classifiers and neural networks to generate adaptability maps. We also propose the creation of a common database of environmental variables regarding the Brazilian territory which we call *br-env*.

Keywords– Species distribution modeling, Agroecology, MILPA, br-env, Adaptability maps.

1. Introdução

Technological advances aimed at the monocultures cultivation have fulfilled the social role of meeting the food sufficiency of the world population through the development of more efficient irrigation techniques, the mechanization of processes, the use of fertilizers, the application of pesticides, and the creation of transgenics. (more productive hybrid varieties) [1]. However, the high density of a single species (low genetic variability) promotes unstable systems to climate changes, requires constant management due to the depletion of soil nutrients (generating the need to use fertilizers), and provides a high density of pests (requiring the use of pesticides) [2] . In other words, monocultures can be held responsible for important environmental impacts such as contamination of soils, rivers, and air [2].

In the Brazilian context, other problems directly related to monoculture are also identified, such as the deforestation of Brazilian biomes (Amazon, Cerrado, Pantanal), often resulting from the land grabbing process. Also, it promotes the rural exodus and the increase in social inequalities since small farmers do not have the power to compete in the market with large farmers [3].

To face problems related to monoculture, sustainable strategies gained strength, such as agroecological crops, polycultures, and Agroforestry Systems (SAF). Those techniques use the understanding of ecological processes in favor of agricultural production, intending to mitigate the socio-environmental impacts related to monocultures and also increase output by raising the Land Equivalent Ratio Land Equivalent Ratio (LER) [4]–[6]. In these alternatives, each plant plays one or more roles within the system, such as, for example, the functions of retaining nitrogen, producing litter, protecting the system as a live fence for animals or as a barrier against winds, and acting as a ground cover, among others [2]. Past experiences of farmers and indigenous peoples discovered species that can grow together since they play complementary roles, which is defined by the theory of agroecological consortiums [7], [8].

However, agroecology faces the challenge of scale. Replacing the intensive use of agricultural techniques with agroecological consortiums required the understating of how well they can suit to environmental conditions of a region as granular as possible. In this context, the ecological niche theory establishes the relevance of studying relationships between species and the specific requirements of an area as crucial mechanisms for the survival of species and the maintenance of biodiversity [9], [10].

Therefore, the present work aims to explore a machine learning approach called Species Distribution Modeling (SDM) as a tool for identifying ecological niches of agroecological consortium. SDM techniques seek to model their suitability to a proposed region based on environmental data in coordinates of species occurrences. The model predicts how the same species could develop without major environmental interventions in inference time [11]. In this way, SDM techniques can help identify territory regions where an agroecological consortium shows potential for agricultural production.

In particular, this work focuses on the Brazilian territory. It will have as the object of study thw species of the MILPA consortium, an agroecological consortium initially cultivated by the Mayans, which integrates variations of species of corn, beans, squash, and pepper [12]. We also highlight that MILPA species are relevant to the food security objective from Food and Agriculture Organization of the United Nations [13].

2. Proposal

The general objective of this work is to study, apply and evaluate computational techniques for modeling the distribution of species from the MILPA consortium to identify regions from Brazil suitable for the establishment of agroecological crops and understand, in more detail, the potentials and limitations of SDMs. We describe the criteria to decided which species from MILPA consortium we decided to study, resulting in the following species: Zea mays, Cucurbita pepo, Cajanus cajan and Capsicum annuum.

The first step to applying SDMs focus in Brazil was having data representing the country's environment. For this reason we propose *br-env*, a standardized ensemble of environmental information from multiple sources framed in the Brazilian territory. Next, we explore SDMs as a binary classification problem from a two steps algorithm. First, we generate pseudo-absence species using a OCSVM once this data is not naturally available. Next, we evaluated different classifiers that calculate the probability of a coordinate being present or absent. In addition, once *br-env* is composed of 99 environmental variables, we assessed the dimension reduction algorithm VIF to verify if a reduced environment space could be sufficient to classify the species well.

The output of our experiments consists of 32 distribution maps for each considered species, classifier, and environment size. We evaluated experiments according to the performance metrics AUC and TSS and qualitative aspects of the generated maps. The results of this project are accessible in https://github.com/AI-Unicamp/easy-sdm

2.1 *br-env*

A crucial part of this work consisted of preparing the data before applying machine learning models. The data engineering phase started by downloading the raw environmental databases and species occurrence to create a 3D array with 99 environment variables composed from Bioclim, Envirem, and Soilgrids databases for the Brazilian territory coordinate limits. This step is species independent and resulted in br-env, a database of aggregated environmental data for Brazil that is used in this study and can be used for further studies on any specie. From a practical point of view, the br-env consists of a Numpy array with metadata information to reference each column to an environment variable.

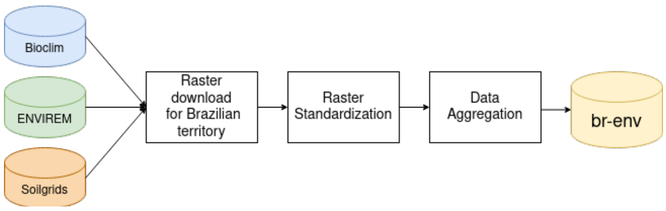


Figure 1. The figure represents the br-env creation process. First we download data from Bioclim, Envirem and Soilgrids. Next we standardize the data from each source to common specifications. Finally we describe the aggregation process to a unique array, which we call br-env.

2.2 Pseudo-absences generation

We built dataset rows corresponding to presences according to the species GBIF occurrence (OCC) coordinates. However, absent species data are not available. For this reason, we generated pseudo-absence (PSA) coordinates representing regions where the plant species could not grow.

Among the techniques mapped from the literature in we applied Random Selection with Environmental Profiling (RSEP) considering the best trade-off between simplicity and performance. To extract environmental profiling, we applied OCSVM, once we can use a semi-supervised approach, in which the algorithm is trained only with presence data and the resulting model, at inference time, labels coordinates as presence or absence [14].

We randomly generated the same number of pseudo-absences as the number of available occurrences to keep a balanced dataset and to attenuate the effects of prevalence in the classification models performance [15].

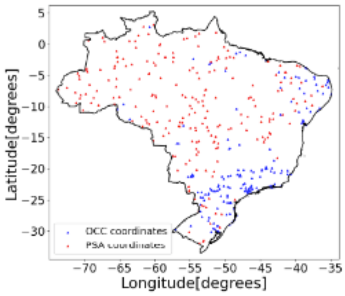


Figure 2. Example result of pseudo species generation for Zea mays.

2.3 The Variance Inflation Factor (VIF)

In the dataset construction, 99 features were present. Many of the environment variables have colinear dependencies with each other. Therefore, to evaluate behave of a reduced number of components and keep the performance metrics, we applied the VIF algorithm in parallel for the whole data before the cross-validation.

VIF is calculated by the VIF = 1/(1 − R2). This formula measures if one variable can be predicted based on the features. If R2 is small, a specific variable can not be predicted from the information from other variables, showing a small colinearity with them, offering evidence that an algorithm could use this variable as an information source to determine the target label.

We decided to keep a group of features with the constraint VIF<10 for each part compared to the others as explained in [16] when VIF>10, two variables are highly correlated, and can one of them can be discarded. Hence, we developed an iterative process that calculates the VIF factor for each feature and removed the biggest one until all of them respect the constrain.

2.4 The classifiers

In this section, we describe the proposed classifiers. We focused on comparing the performance of popular ensemble techniques (GB, XGB and RF) with a MLP. The tree ensemble algorithms are built with multiple decision trees, but they differ on how to integrate them. Decision trees learn data through an architecture represented by branch nodes and leaf nodes, the former contains a condition to split the data, and the latter helps to decide a class to a new data point [17].

GB and XGB represent the boosting ensemble technique that combines many weak decision tree classifiers to build a robust classifier. The observation weights are adjusted based on the previous classification, which replaces the approach of creating only one predictive solid model. Like ANN, boosting techniques are nonparametric machine learning, so the models can be adjusted according to the observed data, being adequate for the domain-specific tasks, as SDMs. The main difference between GB and XGB is that the last one used advanced regularization techniques like L1 and L2, which are expected to improve model generalization capabilities [18].

RF is also an ensemble technique that produces each decision tree independently and combines the results at the end of the process by the majority votes of each tree. Algorithm that follow this procedure are classified as a bagging algorithms. The literature shows that this approach has good performances in problems which the number of variables are much larger than the number of observations, which is also true for SDMs [19].

MLP is a conventional neural network approach in which neurons with nonlinear activation functions are structured as a network with multiple layers. The first one has the number of neurons equal to the number of features. The last one has the number of neurons dependent on the loss function and the problem objective. ANN learns through the back-propagation process in which the neuron's weights are updated with the error between a predicted target data and the labeled one [20], [21]. For our SDM approach, the number of neurons in the first layer is the number of environment variables. As we modeled it as a binary classification problem, we used a sigmoid function in the last neuron that calculates the prediction errors through a log-loss function [16].

3. Results

The experiments results we present in this chapter correspond to distribution models that are obtained varying three different aspects of the modeling process:

- **species modeled:** Zea mays, Cucurbita pepo, Cajanus cajan or Capsicum annuum;
- **training data:** with complete set of environmental variables available or VIF-reduced number of variables;
- **binary classifier:** Random Forest, Gradient Boosting, XGBoost or Multilayer Perceptron).

In summary, we conducted 32 experiments (4 species × 2 dataset configurations × 4 binary classifiers). For each experiment, we evaluated the measured performance of the classifier (AUC and TSS metrics), and the resulting distribution map characteristics.

In Figure 3 we show an example distribution map. In Tables 1-4 we expose the results of the proposed experiments for each of the studies species.

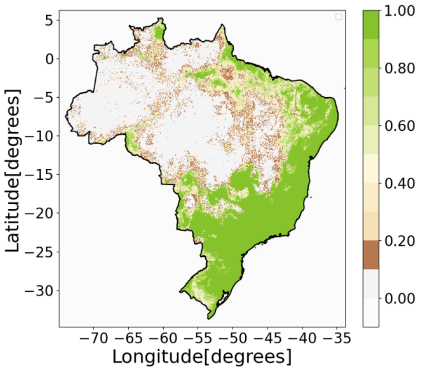


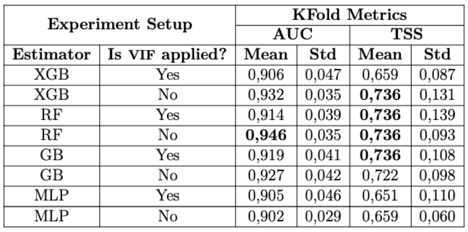
Figure 3. Example distribution for Zea mays using XGB as classifier.

Experiment Setup		KFold Metrics			
		AUC		TSS	
Estimator	**Is VIF applied?**	**Mean**	**Std**	**Mean**	**Std**
XGB	Yes	0,932	0,007	0,690	0,037
XGB	No	0,942	0,010	0,726	0,051
RF	Yes	0,943	0,014	0,735	0,050
RF	No	**0,949**	0,004	**0,749**	0,037
GB	Yes	0,924	0,018	0,672	0,038
GB	No	0,938	0,005	0,717	0,064
MLP	Yes	0,918	0,013	0,650	0,019
MLP	No	0,927	0,021	0,700	0,076

Table1. Zea mays experiments results

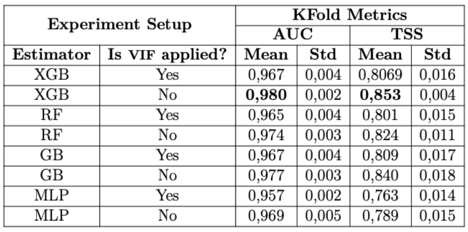
Experiment Setup		KFold Metrics			
		AUC		TSS	
Estimator	**Is VIF applied?**	**Mean**	**Std**	**Mean**	**Std**
XGB	Yes	0,835	0,077	0,565	0,118
XGB	No	0,908	0,041	**0,614**	0,1290
RF	Yes	0,881	0,060	0,565	0,057
RF	No	**0,929**	0,032	0,551	0,125
GB	Yes	0,860	0,121	0,488	0,176
GB	No	0,905	0,054	**0,614**	0,129
MLP	Yes	0,828	0,058	0,483	0,085
MLP	No	0,873	0,049	0,597	0,085

Table 2. Cucurbita pepo experiments results



Experiment Setup		KFold Metrics			
		AUC		TSS	
Estimator	**Is VIF applied?**	**Mean**	**Std**	**Mean**	**Std**
XGB	Yes	0,906	0,047	0,659	0,087
XGB	No	0,932	0,035	**0,736**	0,131
RF	Yes	0,914	0,039	**0,736**	0,139
RF	No	**0,946**	0,035	**0,736**	0,093
GB	Yes	0,919	0,041	**0,736**	0,108
GB	No	0,927	0,042	0,722	0,098
MLP	Yes	0,905	0,046	0,651	0,110
MLP	No	0,902	0,029	0,659	0,060

Table 3. Cajanus cajan experiments results



Experiment Setup		KFold Metrics			
		AUC		TSS	
Estimator	**Is VIF applied?**	**Mean**	**Std**	**Mean**	**Std**
XGB	Yes	0,967	0,004	0,8069	0,016
XGB	No	**0,980**	0,002	**0,853**	0,004
RF	Yes	0,965	0,004	0,801	0,015
RF	No	0,974	0,003	0,824	0,011
GB	Yes	0,967	0,004	0,809	0,017
GB	No	0,977	0,003	0,840	0,018
MLP	Yes	0,957	0,002	0,763	0,014
MLP	No	0,969	0,005	0,789	0,015

Table 4. Capsicum annuum experiments results

4. Conclusions

Despite being the Brazilian economy flagship, traditional agriculture has several problems regarding environmental degradation and efficiency in production by area. Understanding the regions in Brazil where producers can grow agroecological consortiums is a starting point to increase the adoption of sustainable crops in the place of monoculture and reduces it negative impact on natural resources. For this reason, this work aimed to explore how Species Distribution Modeling (SDM) can suggest regions for the MILPA consortium to grow in Brazil. According to ecological concepts, this consortium considers species from big groups (corn, squash, bean, and pepper) to create positive interactions between them and the environment to amplify food production without requiring pesticides and fertilizers.

We presented br-env, a standardized three-dimensional array that uses raster image data from Bioclim, Envirem and Soilgrids to represent the Brazilian environment conditions. Br env could be used in further research to create SDM independent of species and modeling choices. Besides, we detail how we used a pipeline of a OCSVM plus an ensemble classifier to generate distribution maps based on species occurrences and pseudo-absences. According to several experiments, we compare the proposed ensemble classifiers (XGB, GB, RF, and MLP) applied to the MILPA species we selected (Zea mays, Cucurbita pepo, Cajanus cajan and Capsicum annuum) to evaluate the classification algorithms according to the AUC and TSS metrics and the VIF algorithm.

Based on the performed experiments, we concluded this work by answering the research questions. The algorithms had similar performances according to the performance metrics for each species. They show that using and OCSVM to generate pseudo absences had a more significant impact independent of the classification algorithm in terms of metrics. Still, those classifiers were relevant in the style of distribution, with RF being granular in its decisions and XGB, GB and MLP being more decisive. AUC and TSS metrics should be analyzed as a pair once they evaluate the predicted habitats according to different perspectives. Also, we can notice that apply VIF variables. Besides, we identify species with fewer occurrences as Cucurbita pepo, which tend to perform worse in the metrics, and species with concentrated occurrences as Capsicum annuum tend to perform well. Finally, we can conclude that species to compose the MILPA consortium could grow together in a coastal part of the Northeast region of Brazil. We could join part of the Southeast region if we remove the optional species Capsicum annuum.

Finally, we highlight the importance of considering statistical, computational, and ecological knowledge to evaluate the generated habitat distributions. Understanding specific properties of a target species, is substantial to make a crop succeed when using the distribution's manual.

Acknowledgements

I would like to thank Paula Dornhofer Paro Costa for guiding me on the journey of being a researcher and for enabling me to turn the initial ideas I had from this project into solid work. I would also like to thank my parents, Maria Angela Alves and Antonio Carlos Sasso, for their financial, loving support and for making me believe in education since I was a kid. Finally, I would like to thank all my friends who have substantially influenced the person and professional I am.

References

[1] M. Duru, O. Therond, and M. Fares, 'Designing agroecological transitions; A review', *Agron. Sustain. Dev.*, vol. 35, no. 4, pp. 1237–1257, Oct. 2015, doi: 10.1007/s13593-015-0318-x.

[2] S. R. Gliessman, *Agroecology: The Ecology of Sustainable Food Systems, Third Edition*, 0 ed. CRC Press, 2014. doi: 10.1201/b17881.

[3] L. A. Martinelli, R. Naylor, P. M. Vitousek, and P. Moutinho, 'Agriculture in Brazil: impacts, costs, and opportunities for a sustainable future', *Curr. Opin. Environ. Sustain.*, vol. 2, no. 5–6, pp. 431–438, Dec. 2010, doi: 10.1016/j.cosust.2010.09.008.

[4] E. Barrios *et al.*, 'The 10 Elements of Agroecology: enabling transitions towards sustainable agriculture and food systems through visual narratives', *Ecosyst. People*, vol. 16, no. 1, pp. 230–247, Jan. 2020, doi: 10.1080/26395916.2020.1808705.

[5] T. Juniper, *What has nature ever done for us? how money really does grow on trees*. Santa Fe, NM: Synergetic Press, 2013.

[6] Y. Yu, T.-J. Stomph, D. Makowski, and W. van der Werf, 'Temporal niche differentiation increases the land equivalent ratio of annual intercrops: A meta-analysis', *Field Crops Res.*, vol. 184, pp. 133–144, Dec. 2015, doi: 10.1016/j.fcr.2015.09.010.

[7] C. Kremen, A. Iles, and C. Bacon, ‘Diversified farming systems: an agroecological, systems-based alternative to modern industrial agriculture’, *Ecol. Soc.*, vol. 17, no. 4, 2012.

[8] E. A. Frison and I. P. of E. on S. F. Systems, ‘From uniformity to diversity: a paradigm shift from industrial agriculture to diversified agroecological systems’, IPES, Report, 2016. Accessed: Oct. 30, 2021. [Online]. Available: https://cgspace.cgiar.org/handle/10568/75659

[9] T. Václavík and R. K. Meentemeyer, ‘Equilibrium or not? Modelling potential distribution of invasive species in different stages of invasion: Equilibrium and invasive species distribution models’, *Divers. Distrib.*, vol. 18, no. 1, pp. 73–83, Jan. 2012, doi: 10.1111/j.1472-4642.2011.00854.x.

[10] J. L. Brown, ‘SDMtoolbox: a python-based GIS toolkit for landscape genetic, biogeographic and species distribution model analyses’, *Methods Ecol. Evol.*, vol. 5, no. 7, pp. 694–700, Jul. 2014, doi: 10.1111/2041-210X.12200.

[11] J. Elith and J. R. Leathwick, ‘Species Distribution Models: Ecological Explanation and Prediction Across Space and Time’, *Annu. Rev. Ecol. Evol. Syst.*, vol. 40, no. 1, pp. 677–697, Dec. 2009, doi: 10.1146/annurev.ecolsys.110308.120159.

[12] S. Teran and R. Heilskov, ‘Las plantas de la milpa entre los mayas: etnobotanica de las plantas cultivadas por campesinos mayas en las milpas del noreste de Yucatan’, 1998.

[13] FAO, ‘FAO Strategic framework 2022-31’. FAO, 2021. Accessed: Jun. 30, 2022. [Online]. Available: http://www.fao.org/3/ne577en/ne577en.pdf

[14] M. Goldstein and S. Uchida, ‘A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data’, *PLOS ONE*, vol. 11, no. 4, p. e0152173, Apr. 2016, doi: 10.1371/journal.pone.0152173.

[15] M. Barbet-Massin and W. Jetz, ‘A 40-year, continent-wide, multispecies assessment of relevant climate predictors for species distribution modelling’, *Divers. Distrib.*, vol. 20, no. 11, pp. 1285–1295, Nov. 2014, doi: 10.1111/ddi.12229.

[16] J. Rew, Y. Cho, and E. Hwang, ‘A Robust Prediction Model for Species Distribution Using Bagging Ensembles with Deep Neural Networks’, *Remote Sens.*, vol. 13, no. 8, p. 1495, Apr. 2021, doi: 10.3390/rs13081495.

[17] A. J. Myles, R. N. Feudale, Y. Liu, N. A. Woody, and S. D. Brown, ‘An introduction to decision tree modeling’, *J. Chemom.*, vol. 18, no. 6, pp. 275–285, Jun. 2004, doi: 10.1002/cem.873.

[18] T. Chen and C. Guestrin, ‘XGBoost: A Scalable Tree Boosting System’, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco California USA, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[19] G. Biau and E. Scornet, ‘A random forest guided tour’, *TEST*, vol. 25, no. 2, pp. 197–227, Jun. 2016, doi: 10.1007/s11749-016-0481-7.

[20] A. K. Jain, Jianchang Mao, and K. M. Mohiuddin, ‘Artificial neural networks: a tutorial’, *Computer*, vol. 29, no. 3, pp. 31–44, Mar. 1996, doi: 10.1109/2.485891.

[21] L. Noriega, ‘Multilayer perceptron tutorial’, *Sch. Comput. Staffs. Univ.*, 2005.