Using Ecological Modelling Tools to Inform Policy Makers of Potential Changes in Crop Distribution: An Example with Cacao Crops in Latin America

Juan Fernández-Manjarrés

Using ecological modelling tools to inform policy makers of potential changes in crop distribution: an example with cacao crops in Latin America

Juan Fernandez-Manjarrés

Summary

Species distribution models (SDM) is a powerful simulation tool that has become widely used in the ecological and agronomical sciences. The use of easily available presence data, global downscaled climate layers and software that can run on desktop computer has contributed to their popularity. The most used application is based on maximum entropy models that fit presence data to a series of environmental descriptors. SDM can be used to predict crop distribution under future conditions but the level of uncertainty of those models can be very high. The best use of these models is to be used as generators of hypothesis to be combined with other type of analysis.

Introduction

One consequence of climate change that is becoming increasingly clear, is the shift in species distribution of certain wild species because of climate change (Parmesan 2006). However, assigning climate effects to distributional shifts has not always been straightforward because of other factors. For instance, changes in land use can produce new empty ecological niches[1] and

[1] The definition of niche is characterized by the ecological role of a species in a natural community, but is also used in a looser form to refer to the microhabitat or the physical space occupied by a species. In this chapter, we retain the latter use.

habitats[2] that are used by local or alien species (Parmesan and Hanley 2015). Likewise, economists, agroecologists and enterprises ask themselves if the current distribution of crops would change with ongoing climate change, and if yes, to what extent. Clearly, if the climate related to crops is no longer suitable, the economic and social costs of replacing crops, or of changing cultivated areas is extremely large, so early awareness of what might happen is needed for policy makers.

To simulate the potential shifts in the distribution of species, ecologists have been using for the last 15 years or so the so called '*species distribution models*' (hereafter **SDM**) or '*niche models*'. As we will see in the following sections, SDM are statistical models that correlate the observed presence of a species (or crop for that matter) with climatic and geographic features of the zones for which occurrences of the species in question are known. They are not mechanistic models, but correlational models built upon a series of assumptions. These models have attracted the researchers in ecology, because they are less data intensive than mechanistic models (i.e., models based on photosynthesis and mineral exchanges with the air and the soil) and are easy to spatialize.

These models are extensively used not only for endangered species but for managed forests, pests and invasive species as we will see later in the text. Crops, on the other hand, have used somewhat different statistical models based mostly on matching the current requirements of a crop with its climate, but to some extent, models in ecology and agronomy may have to start to converge in the same family of modelling tools.

[2] Habitat is the locality, site and particular environment occupied by an organism and as such the definition overlaps that of niche in terms of spatial occupation. For coherence with the models, we will use only niche in this text.

The world distribution of crops has been traditionally understood as zones delimited by extremes of temperature and precipitation (Kottek, *et al.* 2006) while the changes of crop productivity and distribution has been modelled with several types of models (see Holzkämper 2017and references therein). They include empirical, suitability, biophysical, meta- and decision making models.

In this chapter, we will discuss the use of SDM in crop science, that is a type of suitability model *sensu* Holzkämper (2017). The approach might be perceived as biased, but as we will see, it may be flexible enough to forecast potential shifts not only in crops, but in their related pests and diseases as well as invasive species, all of which have economic impacts with relatively small quantities of data.

We will first review briefly the literature on SDM and crops. Second, we present the general background of the models and introduce MaxEnt (Phillips, *et al.* 2006, Phillips and Dudik 2008), that has emerged as very robust modelling platform based on maximum entropy theory models. We then present an application for crops zones in Latin America where both coffee and cacao are planted, as these zones are very likely to be affected by climate change, with impacts on two very independent value chains. We finish by discussing the limits of the approach and with a word of caution regarding the mis-use of this kind of models.

Current use of species distribution models

Overall, the use of SDM models is relatively recent. The oldest reference in our search examines the potential conflict of geese and crops (Jensen, *et al.* 2008) just about 10 years ago at the time of publishing. As said in the introduction, the question of crop distribution and climate has been treated for a long time, but it is the use of SDM that appear as a cost-efficient alternative for researchers and managers.

This first generation of use of SDM in agriculture has led to a majority of articles on staple foods likes corn, wheat and rice, but also on diseases, invasive species, pests and pollinator distribution under current and climate change conditions. A search on "SPECIES DISTRIBUTION MODELS") *AND* **TOPIC:** (CROPS) in Web of Science ® in early 2018 provided 75 records from which only four were review papers (Figure 1).

Figure 1. Proportion of articles that use species distribution models for crop science studies. See text for details.

As we will see in the next section, the power (and weaknesses) of SDM resides in the use of geo-localized data to infer current suitable habitat that is easily transposable to future conditions if climate change projections are available. The fact that known localities are used as the main input, makes SDM highly applicable to different types of organisms (vertebrates, insects,

nematodes, etc.) and for crops that are thought to be cultivated within their normal biological niche.

Species distribution models: Maximum Entropy models

An intuitive relation between climate, soil, altitude and the distribution of animals and plants is probably one of the oldest ecological observations that human kind has made. However, what appears so self-evident and intuitive has proved enormously difficult to formalize correctly in statistical terms. As it is well known, correlational methods can adequately model and predict on models calibrated *on what is seen* (observed localities) but cannot make inferences *about what is not seen* (unknown parts of the distribution) without making assumptions and simplifications.

Earlier SDM models used an empirical approach for calculating a climatic 'envelope' for a given species based on the known occurrences. But this kind of model very soon attained their limits because a) it is frequently unknown if the current distribution of a species represent the complete physical and climatic space that a species can survive and reproduce; and b) the more variables used, the more difficult to generalize the distribution to unknown parts as almost everywhere the variable combinations are unique, so they are not transposable in space and time. Hence, several statistical methods emerged to allow for a probabilistic approach to the problem. From about a dozen that appeared in the early 2000's, the maximum entropy approach (hereafter MaxEnt) by Phillips and collaborators (Phillips, *et al.* 2006, Phillips and Dudik 2008, Steven J. Phillips, *et al.* 2018) appeared particularly robust but not completely exempt of controversy about the assumptions and meaning of the output. MaxEnt methods have been used

in more than 7000 peer-reviewed studies at the time of writing and its popularity seems to continue. Interestingly, recent generalizations of the species distribution problem is showing that many different competing methods can be related through an alternative approach of not modelling the probability of presence but by modelling the probability of a point observation in a given space (Renner, *et al.* 2015).

Before explaining the procedure, let us first formalize the input data and the goal of the simulation. Frequently, when discussing SDM, two families of models are are mentioned, those based on *presence/absence* data, and those *presence onl*y data. MaxEnt belongs to the second category but the notion of absence is necessary in the formalization of the model. The general idea of using maximum entropy methods is explained by Phillips et al. (2006) in his original paper: "*The idea of MaxEnt is to estimate a target probability distribution by finding the probability distribution of maximum entropy (i.e., that is most spread out, or closest to uniform), subject to a set of constraints that represent our incomplete information about the target distribution.*"

The idea of maximum entropy approaches imply that the goal is to find the *most spread out* distribution based on what is known from the data, i.e., a maximum entropy distribution. Next, I summarize the statistical description given by (Elith, *et al.* 2011) skipping many details for the sake of brevity. The approach assumes that the data available are a set of locations within a landscape of interest L. Next, the presence of the focus species needs to be coded in binary form: $y = 1$ denote presence, $y = 0$ denote absence. Associated to the presence points, there is a need to define a vector of environmental covariates (mean annual temperature, summer

precipitation, drought index, altitude, soil type…) which is called ***z***. Finally, there is a need to define a '*background'* in which the ***z*** vectors occur, that is defined as a random sample of locations within the landscape (Elith, *et al.* 2011). The environmental covariates *z* are available for the whole landscape as is the case for example with climate or elevation layers from geographic information systems that are found in pixel form.

The next step is to define independent probability distribution related to the covariates in the landscape, for the occurrences and for the absences. Hence, $f(z)$ defines the probability density of covariates across the landscape, $f_1(z)$ the probability density of covariates for the locations where the species is present, and $f_0(z)$ where the species is absent. It follows then that the quantity to be estimated when presence-absence data is available, is the *probability of presence of the species*, conditioned on the environment:

$$\Pr(y = 1|z).$$

Presence-only data only allow only to estimate $f_1(z)$, which cannot be used to estimate the probability of presence, because it is assumed that not *all* localities are known for the focus species. However, presence/background data allow to model both $f_1(z)$ and $f(z)$ if we knew how the two relate them through a constant *C* using Bayes' rule:

$$\Pr(y = 1| \boldsymbol{z}) = f_1(z)*C/ f(z)$$

It turns out that the needed constant $C = \Pr(y = 1)$, corresponds to the 'prevalence of the species' (or the proportion of occupied sites) in the landscape. So the challenge is to estimate $\Pr(y = 1)$, of course. In entropy terms, the probability of the distribution of the covariates across the

landscape $f_1(z)$, can be found through the Gibbs distribution exponential form (Elith, *et al.* 2011):

$$f_1(\boldsymbol{z}) = f(\boldsymbol{z})e^{\eta(\boldsymbol{z})}$$

where $\eta(z) = \alpha + \beta * h(\boldsymbol{z})$ and α is a normalizing constant that ensures that $f_1(z)$ sums to 1, β is vector of coefficients applied to the different terms of model, and $h(\boldsymbol{z})$ is the vector of constrained features. Hence, the target of a MaxEnt model is the exponential term that estimates the ratio $f_1(z)/f(z)$. As there is no analytical solution, the parameters are estimated by regression methods and machine learning techniques. The lack of explicit absence observations (museum samples only record presence, for example) is worked around by using random-pseudo absences during the regression iterations. Typically of machine learning techniques, MaxEnt sets aside a portion of the data to train the model and the rest to test the model.

MaxEnt transforms the original covariates (environmental information) in polynomials and splines, including piecewise linear functions data that are termed ‘features’ to allow for the complex response of organisms to climate and other biotic data. Restrictions to the features are needed to avoid the overfitting of the models. For a detailed description of the analytical development, the reader is directed to the work of Phillips and colleagues (Phillips, et al. 2006, Phillips and Dudik 2008, Steven J. Phillips, et al. 2018, Elith, et al. 2011, Elith, et al. 2006).

In its current version[3] 3.41, MaxEnt produces several types of outputs, including tests to evaluate the overall robustness of the model and for identifying which variables are more important. The raw output of MaxEnt represents a probability of *suitable* conditions issued directly from the exponential model above that are extremely low. However, the recommended output is a complementary log-log (*cloglog*) transform that is most appropriate for estimating a measure of abundance – the number of presence records per unit area (Steven J. Phillips, *et al.* 2018, Renner, *et al.* 2015, Fithian, *et al.* 2015) than probability of presence that is riddled with several theoretical and practical issues.

Potential changes of cocoa and coffee plantation zones in Latin America

We will show briefly an example of the application of MaxEnt procedures to a crop distribution that can be of great interest to economists and policy makers. Cocoa (*Theobroma* spp.) and coffee (*Coffea arabica, C. robusta*) are two staple products in Latin America. In general, they do not occupy the same ecological zones. These localities where found from an internet and library search of co-occurrences of coffee and cacao plantations using the Spanish and Portuguese languages (C. Castañeda, unpublished report). We focused on these transition areas (Figure 2) as climate change is opening new areas for the culture of cacao, that are often in zones where coffee plantations are decreasing their productivity because of warmer climates, droughts and emerging pests (Quiroga, *et al.* 2015).

[3] https://biodiversityinformatics.amnh.org/open_source/maxent/

Figure 2. Transition zones between coffee and cocoa plantation localities in Latin America use to run the species distribution model (*n* = 199).

The input data for MaxEnt is then composed of the latitude/longitude of the localities plus the different environmental variables (Table 1) chosen to explain the distribution of the species (crops in this case).

Table 1. First ten records of the MaxEnt input file. The first column is the name of the species in question, the second and third column are the geographic coordinates and the fourth column is the altitude in meters. The columns marked as *bio_* represent bioclimatic variables, and the last column is soil data[4] See text for additional details.

The most common source of downscaled environmental and climatic files is the Worldclim organization[5]. The variables typically used in the ecological field are the so called bio-climatic variables[6] because they have been shown to represent mean conditions and more importantly, limiting factors for many plant and animals. In our case we selected the following variables:

[4] http://www.fao.org/soils-portal/soil-survey/soil-maps-and-databases/harmonized-world-soil-database-v12/en/

[5] http://www.worldclim.org/

[6] http://www.worldclim.org/bioclim

BIO1 = Annual Mean Temperature, BIO4 = Temperature Seasonality (standard deviation *100 for temperatures and BIO12 = Annual Precipitation, BIO15 = Precipitation Seasonality (Coefficient of Variation) for precipitation. Typically, about a dozen or so climatic variables are used, but here we chose just to use only mean annual values and intra-annual measures of variability for illustration purposes.

Particular efforts must be done to ensure that the localities are not auto-correlated or artificially clustered around roads or research centres, which is often the case. Also, the use of several climate variables may be unnecessary and even counterproductive because of over-fit of the models, but more importantly because of correlation between variables. A first screening of pairwise variable correlation is common practice to avoid duplicate entries that can overfit models and that also obscure the interpretation of results. However, some correlation in the climate data is always present and there are no current recommendations of how to deal with this.

To produce an output in graphic form, MaxEnt requires that the user provides a directory with each one of the layers included in the input data table, i.e., from altitude, bioclimatic variables and soil types for current conditions. If the distribution model is intended also for a simulation under climate change for example, the same equivalent files area needed for the period in question as we will see later in the text. MaxEnt will produce maps for current and expected distributions that can be used as input for other analyses. The command line to run the model in our case was:

"*java density.MaxEnt nowarnings noprefixes -E "" -E cacao_coffee responsecurves jackknife "outputdirectory=F:\Maxent Cacao\current_out" "samplesfile=F:\Maxent Cacao\maxent_cacaocoffee.csv" "environmentallayers=F:\Maxent Cacao\current" replicates=10 -t soil*"

but the program has an interface that does not require command line commands. Note that the program is instructed to use the specific sample file "*maxent_cacaocoffee.csv"* (Table 1), produce response curves for the different variables and to jackknife the data to have replicates of training/test runs. In the case of a climate change simulation, an additional "projection layers" instruction is needed as well as an additional output directory for the projection. Finally, it is specified that 'soil' is not a continuous variable but a discrete one indicated by '*-t*'.

Figure 3. Output for the probability of abundance from the maximum entropy model for a reference climate derived from 1970-2000 observations (see text for details).

Typically, more than one climate change model will be used, not only for each frame time, but from each representative concentration pathways (*rcp*) as currently defined in the fifth IPCC report[7]. Here we present only one simulation (**Figure 4**) for the rcp 6.0 and the IPSL global circulation model[8]

[7] http://www.ipcc-data.org/guidelines/pages/glossary/glossary_r.html
[8] http://ocmip5.ipsl.fr/models_description/ipsl_ipsl-cm4.html

Figure 4. Output for the probability of abundance from the maximum entropy model for the projection in 2070 of the model fitted on 1970-2000 climate (Figure 3).

The most popular statistic for examining the robustness of a SDM is the AUC or area under the receiver operating characteristic (ROC) curve that remains controversial (Jiménez-Valverde 2012). This index depends on the use of thresholds that remain themselves a matter of research (Liu, *et al.* 2016). . In general the presence threshold is varied from 0 to 1 to be able to compute how many false positives and false negatives you get at each level. Figure 5 shows the MaxEnt output

Figure 5. Average sensitivity (true positive rate of simulated locations) against false positive rate (1 – Specificity). Sensitivity is the probability that a model correctly classifies a presence. Specificity is the probability a model correctly classify an absence. The average AUC for this model is 0.897.

The cocoa/coffee was split into two partitions, 75% for training and 25% for testing during 10 different runs (cross-validation). The red (training) line shows the “fit” of the model to the training data while the blue (testing) line indicates the fit of the model to the testing data,

corresponding to the predictive power of the model. The black diagonal line depicts a model no better than random (Figure 5).

For what kind of application can these distributions be used in the economic sciences? In general terms, the SDM predict suitable areas for a species or for a crop if the known distribution covers enough of the climate niche of the species or crop. Hence, one straightforward analysis would be to subtract the future suitability from the current suitability to examine which areas will loose suitability and which will gain. Such information could be easily translated into economic models (but see last section) and help managers and decision makers (Figure 6). Clearly, the lowlands show a deterioration of climatic conditions and the Andean region seems to be the most suitable for cacao and coffee plantations for the conditions predicted with this model for 2070. Likewise, one could do the same exercise regarding coffee/cocoa pests and diseases to further refine this analysis if the data were available.

Figure 6. Differences in expected abundance between a baseline scenario (1970-2000) and 2070 (Figures 3 and 4). Light green-blue areas represent zones for which the two cultures will potentially increase their abundance as a result of shifting climatic conditions; yellow and the different shades of brown represent zones where potentially the conditions for 2070 will reduce the abundance of the two crops.

A word of caution

We have seen that with relatively few records (n = 199) we are able to produce more than decent suitability maps for the crops in question as their AUC was quite five (~0.90). All this with relative low computer power and using available downscaled climate change models. However, several points must be mentioned first before it is recommended to use this kind of models:

- SDM models produce better predictions near the real observations. Predictions outside of these areas have large uncertainty. MaxEnt provides various analysis to verify this, and they should be taken seriously.
- The biggest danger is probably due to the fact that a modest number of observation produce decent models and always a very appealing map can be produced. The more data, the better (but see next)
- Data need to be trimmed as auto-correlated data, like excess of collection around research stations biased the models. Remember the models assume a random sample of the presence of the species
- The use of several auto-correlated climate layers can improve the AUC of an otherwise mediocre model
- The use of statistically downscaled climate layers for future conditions is subject to debate. The output of global circulation models is quite coarse (typically 5 degrees or so) and statistical downscaling might be considered an artefact to get high resolution maps

- For many invasive species with short life cycles, it has been shown that their original climatic niche does not correspond to their climatic niche when they invade new areas as evolutionary adaptation can occur very rapidly in some organisms. Thus, SDM may underestimate the invasive potential of a species.
- And last, but not least, SDM should be used to generate hypothesis of what might happen and not to anticipate events.

References

1. Parmesan, C., Ecological and Evolutionary Responses to Recent Climate Change. *Annual Review of Ecology, Evolution, and Systematics* **2006,** 37, (1), 637-669.
2. Parmesan, C.; Hanley, M. E., Plants and climate change: complexities and surprises. *Ann. Bot.* **2015,** 116, (6), 849-864.
3. Kottek, M.; Grieser, J.; Beck, C.; Rudolf, B.; Rubel, F., World map of the Koppen-Geiger climate classification updated. *Meteorologische Zeitschrift* **2006,** 15, (3), 259-263.
4. Holzkämper, A., Adapting Agricultural Production Systems to Climate Change—What's the Use of Models? *Agriculture* **2017,** 7, (10), 86.
5. Phillips, S. J.; Anderson, R. P.; Schapire, R. E., Maximum entropy modeling of species geographic distributions. *Ecol. Model.* **2006,** 190, (3-4), 231-259.
6. Phillips, S. J.; Dudik, M., Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography* **2008,** 31, (2), 161-175.
7. Jensen, R. A.; Wisz, M. S.; Madsen, J., Prioritizing refuge sites for migratory geese to alleviate conflicts with agriculture. *Biol. Conserv.* **2008,** 141, (7), 1806-1818.

8. Steven J. Phillips; Miroslav Dudík; Schapire, R. E. *Maxent software for modeling species niches and distributions (Version 3.4.1).* , 2018.

9. Renner, I. W.; Elith, J.; Baddeley, A.; Fithian, W.; Hastie, T.; Phillips, S. J.; Popovic, G.; Warton, D. I., Point process models for presence-only analysis. *Methods in Ecology and Evolution* **2015,** 6, (4), 366-379.

10. Elith, J.; Phillips, S. J.; Hastie, T.; Dudik, M.; Chee, Y. E.; Yates, C. J., A statistical explanation of MaxEnt for ecologists. *Divers. Distrib.* **2011,** 17, (1), 43-57.

11. Elith, J.; Graham, C. H.; Anderson, R. P.; Dudik, M.; Ferrier, S.; Guisan, A.; Hijmans, R. J.; Huettmann, F.; Leathwick, J. R.; Lehmann, A.; Li, J.; Lohmann, L. G.; Loiselle, B. A.; Manion, G.; Moritz, C.; Nakamura, M.; Nakazawa, Y.; Overton, J. M.; Peterson, A. T.; Phillips, S. J.; Richardson, K.; Scachetti-Pereira, R.; Schapire, R. E.; Soberon, J.; Williams, S.; Wisz, M. S.; Zimmermann, N. E., Novel methods improve prediction of species' distributions from occurrence data. *Ecography* **2006,** 29, (2), 129-151.

12. Fithian, W.; Elith, J.; Hastie, T.; Keith, D. A., Bias correction in species distribution models: pooling survey and collection data for multiple species. *Methods in Ecology and Evolution* **2015,** 6, (4), 424-438.

13. Quiroga, S.; Suárez, C.; Solís, J. D., Exploring coffee farmers' awareness about climate change and water needs: Smallholders' perceptions of adaptive capacity. *Environmental Science & Policy* **2015,** 45, 53-66.

14. Jiménez-Valverde, A., Insights into the area under the receiver operating characteristic curve (AUC) as a discrimination measure in species distribution modelling. *Glob. Ecol. Biogeogr.* **2012,** 21, (4), 498-507.

15. Liu, C.; Newell, G.; White, M., On the selection of thresholds for predicting species occurrence with presence-only data. *Ecol. Evol.* **2016,** 6, (1), 337-348.