Article

Influence of Parameters in SDMs Application to Citrus Presence in Mediterranean Area

Giuseppe Antonio Catalano [1], Provvidenza Rita D'Urso [1]*, Federico Maci [1] and Claudia Arcidiacono [1]

[1] University of Catania, Department of Agriculture, Food and Environment, Via S. Sofia n.100, 95123 Catania
* Correspondence: provvidenza.durso@phd.unict.it, phone: +39 0957147576, fax: +39 0957147605

Abstract: Application of innovative approaches of Precision agriculture requires assessment of the information derived from the georeferenced data elaborated. In this field, analysis of the outcome of models' application requires corroboration of accuracy measures output through map inspection as well as specific sensitivity analyses. In Suitability Distribution Models (SDM) the application of the algorithms to specific species and conditions requires an in-deep investigation of the outcomes in relation to the models' parameters, the species occurrence, and the input layers.
In this study, the application of the main algorithms used for SDM, namely Boosted Regression Tree (BRT), Generalized Linear Model (GLM), Multivariate Adaptive Regression Splines (MARS), Maximum Entropy (MAXENT) and Random Forest (RF), were considered in order to simulate the citrus probability of distribution in a Mediterranean area, based on presence data and a random background sample, and in relation to several predictors. The predictors were grouped under 19 bioclimatic variables, elevation described through a Digital terrain model, soil physical properties, and irrigation. Sensitivity analysis was carried out by: modifying the values of the main models' parameters; and reducing the input presence points.
Fine-tuning the parameters for each model according to the literature in the field produced variations in the predictors' selection and, as a consequence, probability changes in the maps, and modified values of the accuracy measures. In detail, results with modified parameters showed: a reduced overfitting for BRT yet associated to a decrease of the AUC value from 0.91 to 0.81; slight changes in AUC for GLM (equal to about 0.85) and MARS (about 0.83), and MaxEnt (from 0.86 to 0.85); a slight increase of AUC for RF (from 0.88 to 0.89).
The reduction of presence points produced a decrease of the surface area for citrus probability of presence in all the models. Therefore, for the case study analysed, it is suggested to keep input presence points above 250. In these simulations, it was also analysed which covariates and related ranges contributed most to the predicted value of citrus presence, for this case study, at different amounts of input presence points. In RF simulations, for 250 points, isothermality among others was the major predictor of citrus probability of presence (up to 0.8), while at increasing of the input points the contribution of the covariates was more uniform (0.4-0.6) in their range of variation.

Keywords: Vistrails-SAHM software; citrus; spatial distribution; probability of presence; Mediterranean climate; predictor layers

1. Introduction

In the age of agriculture 4.0, new innovative approaches are needed for sustainable process management of cultivations in order to fulfil the requirements as well as reduce environmental impact of productions.

In the Mediterranean area, the effect of the environmental pollution has produced an average annual temperature increase by about 1.4 °C [1]. A reduction of freshwater quality and availability is also expected due to saltwater intrusion and increased extraction. Therefore, resource management is object of interest for research in this field.

In this context, predictive Species Distribution Modeling (SDM) has become an essential tool in a number of environmental issues for agriculture, such as species occurrence under climate change.

 doi:10.20944/preprints202302.0379.v1

The SDMs analyse the links between species location and environmental conditions so as to identify areas with the greatest propensity to accommodate the plant [2].

Recent advances in species distribution modelling have concentrated on novel methods based on presence/absence and/or presence-only data and machine-learning algorithms to predict the probability of species occurrence [3].

A key obstacle preventing the use of SDM is creating reliable and repeatable models, thus dependable processes should be suggested and easily repeatable outcomes such as response curves for expert review should be taken into account.

Few research studies have been oriented to the agricultural sector, whereas most of the SDMs applications are in the biological one, such as for invasive species [4; 5], medicinal plants [6; 7], and species occurrence under climate change, such as *Lobaria polmonare (L.)* [8], plankton [9], and birds [10].

Examples of research studies in the agricultural sector encompass rice production in two West Africa countries [2], and some literature studies specifically aimed at studying the predicted distribution of cash crops. In this regard, Zouabi [11] investigated the direct and indirect effect of precipitations and temperatures on citrus cultivation in Tunisia. In olive grove cultivations, Ashraf et al. [12] carried out a prediction of potential distribution of *Olea ferruginea* in Pakistan. Previous studies of the authors [13] applied MaxEnt to estimate cactus pear biomass.

Since different algorithms frequently provide different results for the same modelling problem [14], thus, the choice of model selection and parameters specification are important to build a model [15].

Moreover, most of the algorithms are computationally-intensive, therefore it is of utmost importance to investigate algorithm suitability for the specific problem and fine-tuning the related parameters in order to save computational time.

Research attempts have been carried out to analyse a number of factors that may affect input data, such as the choice of resolution of environmental layers used in modelling [16]. Further research is needed in this area of interest in order to analyse other factors that may affect predictions.

Therefore, the main objectives of this paper included: the comparison among the SDM algorithms when parameters are modified in respect to the default ones in order to fine-tuning models' parameters for the specific application; and the assessment of models' sensitivity to the number of input presence data.

The analyses were applied to the case study of the citrus crop in a territorial area located in Sicily (Italy), since citrus is one of the main cultivations that contributes to the economic development of the region.

Sicilian agriculture contributes 46.5% to national production with a value of 600 million euros. The province of Syracuse, which produced about 501 million tons of product in the period 2011/2014, was the object of the case study [17; 18].

In Sicily, where precipitations are scarce, the water resource is one of the most central concern for cultivations. Therefore, focusing on the probability of presence in relation to irrigation is of utmost importance [19], and it is one of the innovative aspects of this study.

Based on the studies in the literature, this kind of research works could be valuable in land planning and resource conservation in order to build decision support systems for agriculture [20]. Due to the prediction capacity of SDM algorithms, land use policy plans in specific climatic conditions or for climate change, and improved use of sustainable resources, especially in those regions where there is resource scarcity such as in Mediterranean areas, could be supported.

2. Materials and Methods

The statistical modelling algorithms were executed in the software SAHM coupled with the visual interface VisTrails (VisTrails v.2.2.3 and SAHM v.1.2.1), which has been widely utilised in environmental niche modelling [21; 5; 22; 23; 24]. The algorithms considered were the Generalized Linear Model (GLM), Multivariate Adaptive Regression

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2023 doi:10.20944/preprints202302.0379.v1

Splines (MARS), Boosted Regression Tree (BRT), Random Forest (RF), and Maximum Entropy (MAXENT).

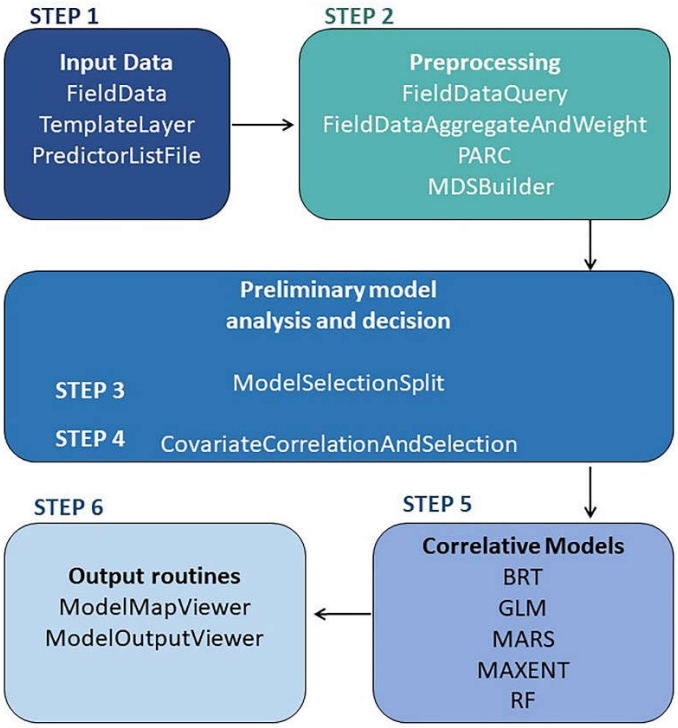


Figure 1. Pipeline of the model in VisTrails:SAHM.

Model formulation included 6 fundamental steps (**Figure 1**). In the first step, predictor and citrus georeferenced data we considered. A *TemplateLayer* with a specific pixel size in a geographic coordinate system was defined and applied to the subsequent modelling. The second step involved synchronisation of all layers by using the *Projection, Aggregation, Resampling and Clipping* (PARC) module to match the template layer properties. The preliminary analysis carried out in the third step consisted of data splitting: 70% of the data was used for training and 30% for testing. In Step 4 uncorrelated predictors were selected by using the *CovariateCorrelationandSelection* module. Step 5 consisted of tuning the parameters for the individual algorithm. Step 6 analysed accuracy measures and provided a graphical output.

At the end, a sensitivity analysis was carried out on the number of input species presence points and on the predictors' resolution imposed by the *TemplateLayer*, which also affects the output.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2023 doi:10.20944/preprints202302.0379.v1

2.1. Study area description

Figure 2. Study area localisation within Italy and Sicily.

The territory selected to apply the methodology was the province of Syracuse, in Sicily (Italy) (**Figure 2**), since this is a widely cultivated citrus growing area, being in one of the major citrus producing regions of Italy. According to the 2014 ISTAT census, 17,000.00 ha are cultivated with citrus in the province with a production of about 350 million t [25]. The province of Syracuse has an area of approximately 2100 km² and borders with the Ionian Sea to the East and with the Catania plan to the North, whereas the south of the province is characterised by the Monti Iblei mountains (**Figure 2**).

2.2. Presence data gathering and production of predictors' maps

Predictors, in raster format, and citrus geolocation data, in shp format, were gathered and prepared for the *Templatelayer* module of the software, in the first step of the methodology. In the specific case of this study, a template layer with a pixel size of 20 m in a WGS84 geographic coordinate system was specified.

Based on data availability (time series of climate data), the simulation period was related to the year 2000. Therefore, citrus presence data were acquired for that time period by overlaying the Sicilian Technical Regional Cartography (TRC) with IT2000 orthophotos available in the Sicilian Land Information System (SITR) (https://www.sitr.regione.sicilia.it/portal/home/item.html?id=06b441f103024aa4b1b9f966b1e4e3f9) in GIS software (specifically ArcGIS® for Desktop 10.3 and QGIS 3.10.0). The resulting dataset, used as input data in VisTrails:SAHM, was composed of 10,000 citrus presence points represented as UTM WGS84 coordinates. Pseudoabsence points cannot be included, as done by Young [26], because historical data were not available, and also due to the fact that pseudoabsence points could be affected by anthropic activity, e.g., when citrus plants are eradicated due to other reasons than crop unsuitability in that area.

Linked to PARC, the *PredictorsListFile* module allowed to add predictors in the *MDSBuilder* module. In detail, the considered predictors were the following: 19 bioclimatic variables defined by WorldClim [27]; the Digital Terrain Model (DTM); soil physical properties; and irrigation.

The 19 bioclimatic variables for the three decades from 1970 to 2000 were acquired from the WorldClim database [https://www.worldclim.org/data/worldclim21.html] in .tiff format by using GIS tools. In most part of the literature, WorldClim data are utilised for this kind of studies as they are suitable to give a broad representation of monthly, seasonal and annual bioclimatic conditions.

In addition, the set of predictors was enriched by the DTM of the area, which provided valuable information on the height at which plants occurrence could be most

 doi:10.20944/preprints202302.0379.v1

probable. This layer was acquired in the Sicilian Region Land Information System website (https://www.sitr.regione.sicilia.it/portal/apps/webappviewer/index.html?id=f3f54ac44ae04a3584885eaaf0b84d70), with a resolution of 20 m, and DTM_20 was the associated predictor variable name in this study.

Soil physical properties were acquired from the European Soil Database & soil properties (available at https://esdac.jrc.ec.europa.eu/resource-type/european-soil-database-soil-properties) and entered as categorial variable in SAHM software.

The irrigation variable points were acquired from the A.C.Q.U.A project ('Agrumicultura Consapevole della Qualità e Uso dell'Acqua' – 'Awareness of quality and use of water in Citrus cultivation') [28]. These irrigation data were converted into continuous data in order to produce a raster map of the variable, named Sir_Irr (m^3 ha^{-1}) hereafter, by using the 'Kriging Ordinary' interpolation method with default settings.

All layers were transformed to match the template layer properties by using the PARC module of the SAHM. The bilinear method for resampling was utilised while the mean and majority filter methods for aggregation were selected for continuous and categorical predictors, respectively.

In fact, models implementation requires that rasters are perfectly overlapping and have exactly the same number of cells, therefore a single raster mask delimiting the study area was defined in VisTrails:SAHM to assure that all raster layers had the same dimensions, and was carried out by coupling *Templatelayer* and PARC modules.

In addition, 10,000 randomly generated background points were considered in the Merged Data Set (MDS) Builder module.

2.3. Fine-tuning models' parameters

The Software for Assisted Habitat Modelling (SAHM) uses 5 models with various default parameters. In this study, the values of the main parameters were modified by using data available in the literature in order to assess whether model performance improved.

In the following, the relevant specific settings of the models (MaxEnt, Boosted Regression Tree, Mars, Generalized Linear Model, and Random Forest) used in Vistrails:SAHM are reported in order to define the parameters and the values considered.

The three main parameters in BRT model are the Learning Rate (LR), the Tree Complexity (TC), and the Number of Trees (NT).

When default setting are applied, the BTR model adjusts the parameter values based on the input data by autoregulation [29; 6]. BRT generally suffers from overfitting [30], which takes place when the fit between predicted values and actual data in models with a large number of predictors is misleadingly good [31]. Overfitting often occurs when a great number of predictors is selected in the Pearson-Spearman-Kendall matrix and may cause random errors in the results. Therefore, although more complicated models may seem more suitable, the predictions they produce may be poorer [32].

GLM is a linear regression method based on a predefined Information Criterion to minimise overfitting, namely Akaike information criterion (AIC) or Bayesian information criterion (BIC).

In Multivariate Adaptive Regression Splines (MARS) model overfitting is controlled by a penalty term (MarsPenalty) that can optionally be set by the user [33]; default parameter is set at 2.

In MaxEnt model within SAHM software, one of the most important parameters is the betamultiplier (named regularization multiplier in MaxEnt software) [34] with a default value of 1. Other MaxEnt parameters were Replicates=1 and Maximum iterations = 5000.

The RF model has three main parameters [35]: NTREES (number of trees), MTRY (the number of possible directions for splitting at each node of each tree), and NODESIZE (the number of observations in each cell below which the cell is not split).

The value of NTREES produced by the algorithm autoregulation in this study resulted equal to 1000. The default value for MTRY, is 1 according to Biau [35], who

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2023 doi:10.20944/preprints202302.0379.v1

demonstrated that this parameter exerts a minor impact on the model performances, and in some cases [36] high values of MTRY were found to be associated to a reduction of the predictive performance. Finally, the NODSIZE value can be set to 1 for classification or to 5 for regression. In this study, the influence of the choice between two values was assessed.

The main models' parameters were refined according to the findings of some authors [29; 37; 38; 39;40; 41], and by taking into account the sensitivity analyses carried out by modifying the default values of the parameters one at a time.

2.4. Sensitivity of the model for number of presence data and raster resolution

Number of presence data was reduced to find out the models' sensitivity to this input, from 10,000 down to 250 points.

In previous research [42] the modification of the rate between the dataset for training and testing highlighted a good robustness of the models; therefore, the percentage was set to 70% for training and 30% for testing.

Response curves were computed for the simulations at different amounts of input presence data. In detail, Response curves describes a measure of predictors' importance in explaining the species distribution in the territory [43] by providing general relationship between each predictor range and the suitability for the species. These curves represent a useful tool for experienced researchers for assessing the outcomes of the elaborations by the biological meaning of the species.

Input raster resolution was modified from 20 m to 1 km to verify the sensitivity of the models to a change in resolution; these analyses were carried out also in relation to the number of input presence points ranging from 250 to 10,000.

2.5. Assessment of models applications

Accuracy measures were considered to assess the models results and allowed comparisons among them. In detail, evaluation accuracy measures derived from the confusion matrix, represented by True skill statistic (TSS), and the area under the receiver operating characteristic curve (ROC-AUC), a standard statistical method widely used to evaluate the accuracy of species distribution models, were utilised to assess them.

According to D'Arrigo [44], the area under the curve shows values between 0.5 e 1.0. In detail, the greater the area under the curve (i.e., the closer the curve gets to the top of the graph), the greater the discriminating power of the test.

Prediction accuracy is considered to be similar to random for ROC/AUC values lower than 0.5; poor, for values in the range 0.5–0.7; fair in the range 0.7–0.9, and excellent for values greater than 0.9 [45]. Moreover, ΔAUC values greater than 0.05 indicate that the model is subjected to overfitting.

TSS values range between −1 (performance no better than random) to +1 (perfect agreement) [44].

3. Results

To facilitate comparison between maps, for each simulation, the threshold computed by SAHM for each model was acquired to convert the continuous probability maps into binary maps that identify suitable and unsuitable territorial areas for citrus. In detail, the threshold method of probability computes the threshold value by considering equal the probability that the model correctly classifies a suitable area and the probability that the model correctly classifies an unsuitable area (i.e., Sensitivity = Specificity) [66].

3.1. Fine-tuning models' parameters

The execution of the 5 different algorithms was carried out by modifying the default values of the main models' parameters considered. The refined values of those parameters were the following: mTry = 2, nTrees = 500, and nodesize = 1, for RF; LearningRate = 0.001, NumberOfTrees = 1000 and TreeComplexity = 3, for BRT; MarsPenality = 2.5, for MARS;

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2023 doi:10.20944/preprints202302.0379.v1

AIC SimplificationMethod, for GLM; and Replicates=15 and Maximum iterations = 5000, and betamultiplier =1, for MaxEnt. These values of the parameters were set according to the findings of some authors [29; 37; 38; 22; 39;40;41] and based on specific sensitivity analyses performed by modifying the default values of the parameters one at a time and keeping the others at their default values. This analysis, based on refined parameters, was compared with that performed by using default parameters and models' autoregulation, obtained in a previous study [19], reported in **Figure 3** and **Table 1**.

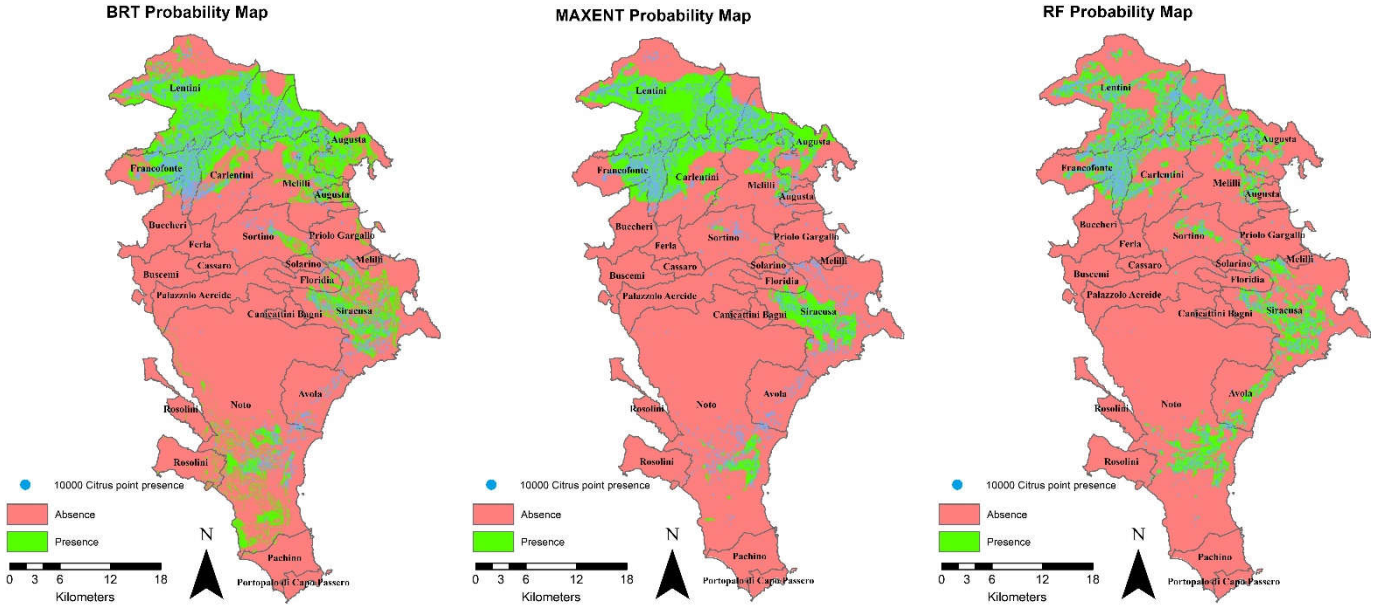


Figure 3. Probability maps of species distribution for 10,000 presence points, default values of models' parameters, and 20-m resolution, for BRT, MaxEnt, and RF models (maps related to GLM and MARS results are reported in the supplementary material **Figure S1**).

Table 1. Presence and absence surface areas for species distribution and AUC values for training, 10,000 presence points, default values of parameters, and 20-m resolution.

Surface area [Km²]	BRT	GLM	MARS	Maxent	RF
red (absence)	1589.39	1618.61	1597.80	1426.66	1701.51
green (presence)	519.59	484.35	505.17	676.30	401.45
AUC for training	0.91	0.85	0.83	0.86	0.88
ΔAUC	0.082	0.002	0.001	0.006	0.000

The analysis based on refined parameters produced the results reported in **Figure 4** and **Tables 2**.

The overall results related to BRT showed that, although the accuracy measures decreased, the values were higher than 0.80 and overfitting was greatly reduced. However, from the analysis of species distribution in the territory results incoherent with citrus actual presence have been obtained due to the absence of the species in the southern area of the province and a general increase of less detailed predicted areas for the species (i.e., more uniform areas without holes).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2023 doi:10.20944/preprints202302.0379.v1

Table 2. Surface areas of citrus probability of presence or absence for each SDMs, and accuracy measures for training and related ΔAUC, obtained by using refined models' parameters.

Surface area [Km²]	BRT	GLM	MARS	MAXENT	RF
red (absence)	1546.14	1618.61	1597.80	1650.26	1718.07
green (presence)	562.83	484.35	505.16	452.709	384.89
AUC	0.81	0.84	0.83	0.85	0.89
TSS	0.74	0.52	0.51	0.55	0.62
ΔAUC	0.006	0.002	0.001	0.006	0.000

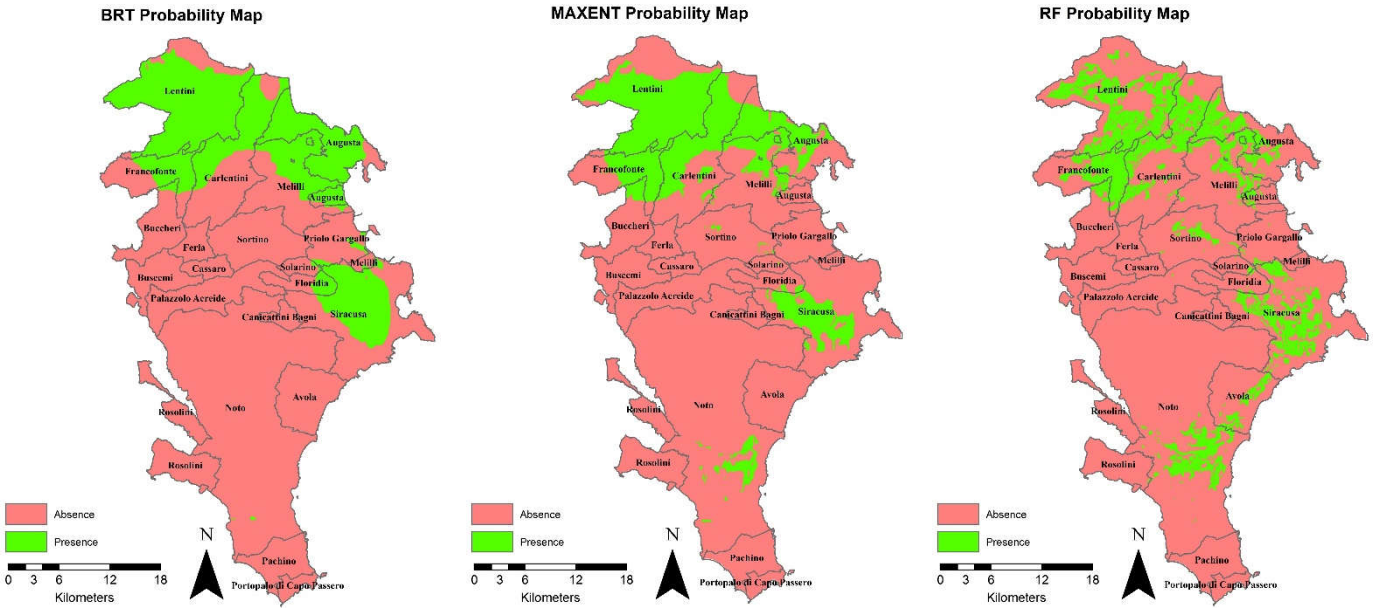


Figure 4. Probability maps of species, with 10,000 presence points, 20-m resolution, and refined values of models' parameters for BRT, MaxEnt, and RF models (maps related to GLM and MARS results are reported in the supplementary material **Figure 2S**).

Moreover, from the sensitivity analyses performed, in relation to increments of LR from 0.001 to 0.1, predicted areas for the species decreased to the point of determining no presence in some areas, and correspondingly ΔAUC values increased. Finally, at increasing of NTREE, from 500 to 5000, the predicted surface area increased, though elaboration time increased, as observed also by Elith [29].

Sensitivity on parameters for GLM and MARS produced low variations for both predicted presence of the species and accuracy measures.

For Maxent, at increasing of the betamultiplier (i.e., 0.5, 1, 1.5, 3, and 5) the AUC values progressively decreased (from 0.86 to 0.83), and ΔAUC reduced from 0.014 to 0; therefore, overfitting decreased yet model performance was reduced. The choice of a high value of this parameter could be useful in the cases when MaxEnt suffers from overfitting, with the aim of keeping it one of the choice models.

As regards the amplitude and distribution of predicted surfaces in the study area, the increase of the regulator reduces the quality of the model simulation. In fact, with a low regulator equal to 0.5 produced a smaller and more detailed predicted surface; at increasing of the regulator, citrus presence is no longer predicted in the southern zone of the province.

For the RF model, the application of the value 5000 for NTREES, compared to the value 500, produced an increase in computational times and small differences in the results, both for the accuracy measures (AUC increased from 0.88 to 0.89) and in the distribution of the predicted surfaces. MYTRY and NODESIZE variations did not

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2023 doi:10.20944/preprints202302.0379.v1

significantly affect the results; this finding could be related to the high number of input presence points [41].

In conclusion, the comparison between Table 2 and Table 1 highlights that GLM, MARS and RF models provided more stable results, with surface area variations ranging from 0 to about 17 km². The variations to parameter settings produced a slight impact on BRT model outcomes (surface area variation equal to 43.24 km²) and the highest on MAXENT model findings (surface area variation equal to 223.60 km²). For this latter model, the prediction is highly modified in the South and Nord-East areas of the province (especially within Sortino municipality), thus in the areas having a lower number of input points of citrus presence.

Table 3 shows that all models performed much better than random (AUC > 0.5) since they all exhibited AUCs >0.8. Also, all models produced TSS > 50%. The models RF, MAXENT, GLM, MARS and BRT showed, in that order, high predictive performance for training, whereas in terms of consistent evaluation accuracy measures between training and testing, RF, MARS and GLM performed better compared to MAXENT and BRT.

In summary, the use of the specific parameters suggested by the literature made it possible reducing the overfitting for the BRT but with a decrease in the AUC value, from 0.91 to 0.81, and an increase in the TSS for all models was encountered.

3.2 Sensitivity of the model for the number of presence data

The modification of the input presence points determined a variation in SDM predictions and in the accuracy measures. For instance, the higher the input presence points are reduced the higher is the reduction of predicted presence for BRT, MARS and GLM models in the East and South areas where the number of input points is lower (i.e., about the 13% of the input points). All the models showed an increase of the surface area of the predicted presence at increasing of the number of input presence points. In detail, the GLM model predicted a wider surface in the North of the study area compared to the other models, while in MARS the surface widened in the South. When the number of input points reduced, the RF model preserved the presence areas but with less detail (i.e., more uniform areas without holes) (see supplementary materials **Figure 3S**).

Furthermore, the analysis of the surface data highlights that for the GLM, MARS and RF models, the predicted surface area decreases as the input points increase, making the results more refined (**Table 3**).

Table 3. Predicted citrus surface area at different values of input presence points, for 20-m resolution, and default parameters with autoregulation.

		Input presence points			
Models		**250**	**500**	**1000**	**10000**
BRT	**absence**	1,778.6	1,749.4	1,782.8	1,589.4
	presence	330.4	359.6	326.1	519.6
GLM	**absence**	1,565.2	1,577.6	1,562.4	1,618.6
	presence	543.8	531.4	546.6	484.4
MARS	**absence**	1,591.4	1,587.9	1,579.1	1,597.8
	presence	517.6	521.1	529.9	505.2
MAXENT	**absence**	1,679.1	1,665.0	1,652.8	1,426.7
	presence	429.9	444.0	456.2	676.3
RF	**absence**	1,583.0	1,648.2	1,685.1	1,701.5
	presence	525.9	460.8	423.8	401.5

By analysing the accuracy measures, the models GLM and RF were influenced by the reduction in the points number, whereas the MARS, BRT and MAXENT models were less affected.

In detail, with regard to AUC, the BRT model showed a range between 0.91 and 0.94 for the training, yet it was influenced by the overfitting for all the hypotheses, with ΔAUC

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2023 doi:10.20944/preprints202302.0379.v1

values between 0.08 and 0.10. Conversely, the other models were less affected by overfitting with a maximum value of ΔAUC equal to 0.04, produced by MaxEnt for the 250-points simulation. The GLM model showed AUC values between 0.81 and 0.82, and reached the value of 0.85 in the simulation at 10,000 points. The MARS (AUC=0.82-0.83) and MaxEnt (AUC=0.86-0.88) models did not exhibit large variations. The RF model (AUC=0.83-0.88) was initially affected by the lower number of points and reached the maximum AUC value in the simulation at 10,000 points.

With regard to the analysis of the TSS values, the RF model produced a gradual increase of the values as the number of input points increased, from 0.49 for 250 points to 0.62 for 10,000 points, and a minimum of 500 input points was required to have TSS≥0.5 . GLM and MARS were stable on values around the threshold of 0.5 and exceeded it only in the simulation at 10,000 points; therefore, for these models it is advisable to have a large number of input presence data. TSS generally decreased as input points increased, with minimum values of 0.55 and 0.65 (for training) in the 10,000-point simulation, for MaxEnt and BRT, respectively.

With regard to Response curves, the simulations of RT model for 250 and 10,000 input points are reported in **Figure 5**. In detail, **Figure 5** graphically depicts the shape and the magnitude of the covariates, displaying the link between the values of the covariates and the citrus suitability according to the predictions of the RF algorithm.

At increasing of the input points, the number of covariates increased from 7 to 8 by including the DTM, and some of them changed. Consequently, the contribute of the biovariables reduced, markedly in some cases such as those of Bio_3 and Bio_9. The shape of the curves generally changed, especially for Bio_16 that increased its contribute for values above 245 Millimeters, and for the maximum contribute of Bio_17 that shifted from 20 to 24 mm in the x axis. Elevation (Dtm_20) contributed up to about 400 m, as previously found [19], and irrigation (Sir_irr) gave a constant contribute in its range of variation.

Isothermality (BIO_3) showed a left-skewed response curve and a maximum suitability between 35 % and 38 %; this range would indicate that high suitability is connected to lower variability of daily and nightly temperatures within a month compared to the year.

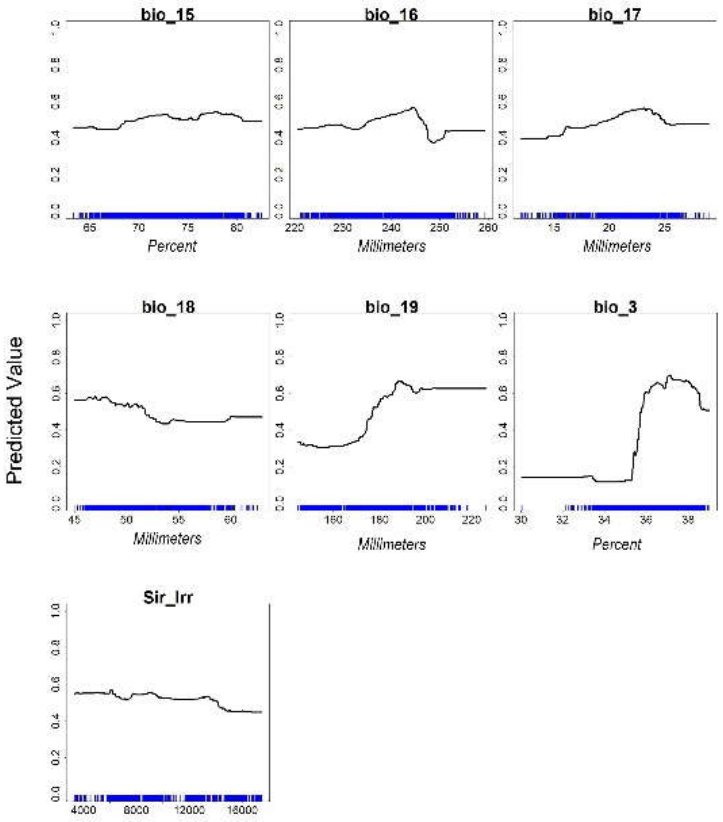
Precipitation of wettest quarter (BIO_16) showed the highest suitability between 220 and 245 mm for the 250-points simulation, whereas the suitability was high still above 245 mm in the 10,000-points simulation.

Bio_15 exhibited a constant curve with the highest values ranging between 68 % and 80 %. This variable describes the variability of the precipitations in the year; the higher is the index value the higher is the variability of the precipitations. According to the Intergovernmental Panel on Climate Change 2012 report, high variability indicates a concentration of precipitation in a short period of time, such as in Mediterranean regions [47].

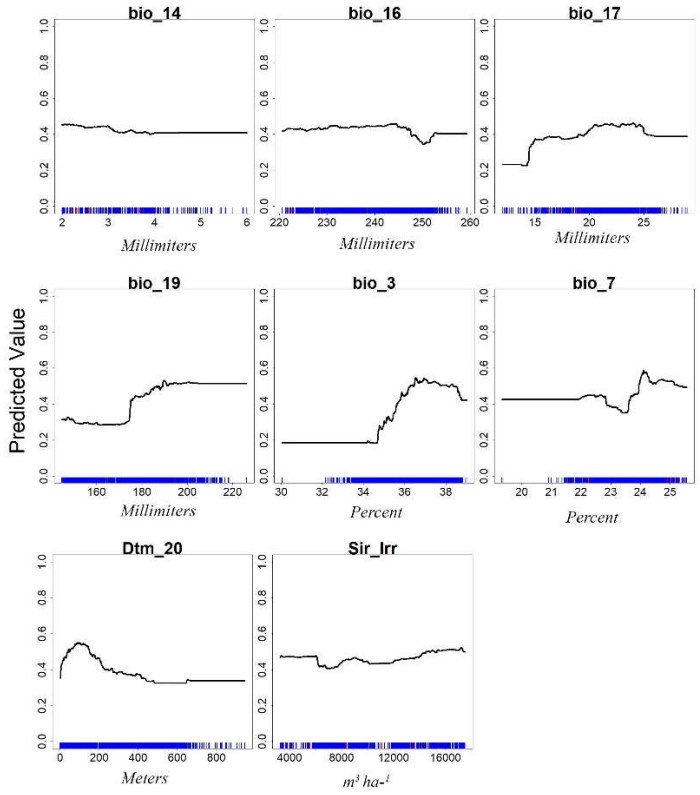
Also Bio_19 exhibited a sigmoid response curve with the highest values above 190 mm of precipitation in the coldest quarter.

The Bio_7 temperature annual range, between about 20° and 25°C, was in the interval considered for citrus species in Spain, i.e., 21°-34° [48].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2023 doi:10.20944/preprints202302.0379.v1



a)



b)

Figure 5. Response Curves for RF model at 250 pt (a) and 10,000 pt (b) of input citrus presence points.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2023 doi:10.20944/preprints202302.0379.v1

3.3. Resolution

Comparison between a 20-m resolution and a 1-Km resolution simulations, keeping the number of input presence points equal to 10000 or reducing it to 250, allowed analysing whether the models were affected by resolution and to what extent when input presence data were modified.

In **Fig. 6 (and Figure 4S)**, the maps of the 1-Km simulations carried out by the different models are reported for the 10,000-points simulation. In green colour the areas of predicted presence generally encompass the input presence points (in blue) except for GLM that most failed to simulate correctly in the south-eastern coastal area of the province (mainly in Avola and Noto municipalities) and also in the central one (Sortino municipality). The comparison of these maps (**Figure. 6 and Figure 4S**) with those at a 20-m resolution (**Figure 3**) confirm the failure of GLM, and of MARS to some extent, to predict the citrus presence in those areas and in the south of the province. Overall, the lower the resolution the higher the surface areas of predicted presence; in detail, the difference between the values of surface areas for the two resolutions ($S_{20\ m} - S_{1\ Km}$) ranged between 182.2 Km^2 of BRT and 411,7 Km^2 of RF, except for MaxEnt that reduced by 47.1 Km^2.

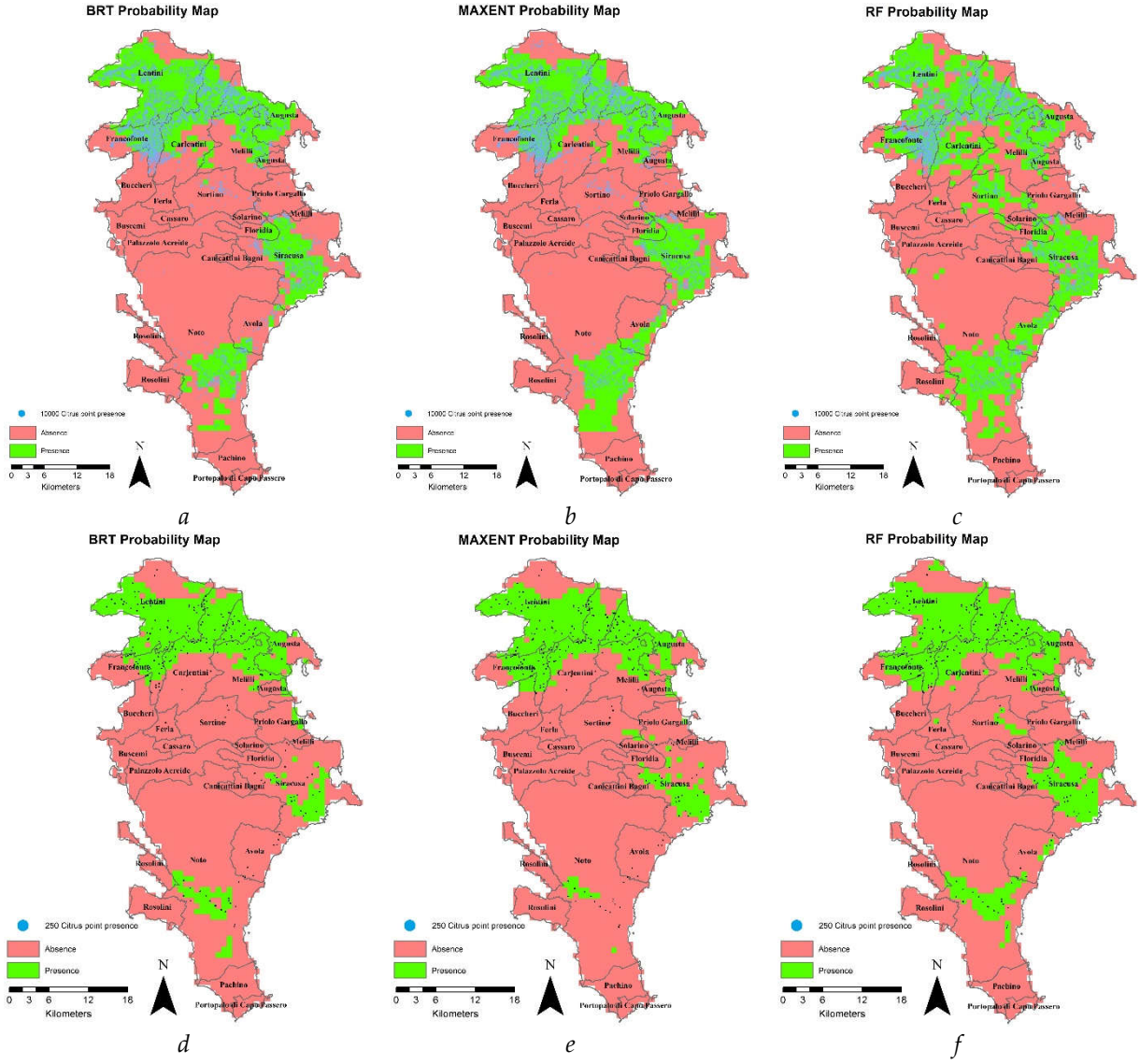


Figure 6. Maps of predicted citrus presence (green) or absence (red) for the different models, at a 1-Km resolution and 10,000 (a,b,c) input presence points (blue) and 1-Km resolution and 250 input presence points (d,e,f) for BRT, MaxEnt , and RF models (maps related to GLM and MARS results are reported in the supplementary material).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 February 2023 doi:10.20944/preprints202302.0379.v1

Table 4. Surface area [Km²] of predicted citrus presence or absence for the different models, at a 1-Km resolution and 10,000 or 250 input presence points.

		Surface areas [Km²]				
	Input points	**BRT**	**GLM**	**MARS**	**MAXENT**	**RF**
Absence (red)	10,000	1,482	1,357	1,436	1,453	1,302
	250	1,633	1,509	1,556	1,641	1,515
Presence (green)	10,000	627	752	673	656	807
	250	476	600	553	468	594

Table 5. Accuracy measures for the different models, at a 1-Km resolution and 10,000 or 250 input presence points.

	BRT		**GLM**		**MARS**		**MAXENT**		**RF**	
	Training	**Testing**	**Training**	**Testing**	**Training**	**Testing**	**Training**	**Testing**	**Training**	**Testing**
AUC_{250}	0.95	0.83	0.79	0.83	0.80	0,84	0.85	0.84	0.75	0.81
AUC_{10000}	0.77	0.72	0.70	0.72	0.73	0.75	0.75	0.75	0.53	0.57
TSS_{250}	0.55	0.53	0.44	0.55	0.45	0.55	0.54	0.57	0.39	0.82
TSS_{10000}	0.39	0.33	0.29	0.29	0.33	0.38	0.36	0.35	0.05	0.12

The reduction of spatial resolution to 1 Km produced a general reduction of the models' performance in terms of accuracy measures (**Table 5**). In detail, the values of TSS indicated a low accuracy for the models, and the AUC values dropped drastically by about 0.2. Conversely, *overfitting* was not encountered in the 10000-points simulations, as the values related to ΔAUC_{10000} of the models did not reach the threshold of 0.05. At 250 input presence points, the values of AUC were high (>0.75) but with a great *overfitting* for BRT and RF, whereas the TSS decreases under 0.5 for GLM, MARS and RF.

4. Discussion

The analyses carried out in this study on models' parameters allowed investigating the performance of the models in relation to the specific case study.

Based on refined parameters, the analysis of the maps produced by the SDMs showed significant changes in the prediction of the BRT model. In detail, a reduction of overfitting was obtained but associated with the prediction of more uniform areas and a lower precision in the southern area compared to the use of the default parameters.

Similarly, the MaxEnt model was affected by the changes of the settings in the East and South areas, where there were a lower number of input presence points.

GLM, MARS, RF and Maxent were the best performing models, while BRT underperformed regardless of the changed parameters. In fact, the self-tuning capability of the BRT can be reduced by the settings of the model parameters [6]. Differences in model performance are often associated with model complexity; models with longer running times appear to produce better accuracy measures [2].

With regard to RF, as confirmed by Diaz [41], the Mtry parameter can lead to higher error rates when few input presence points are used. This confirms the importance of working on large amounts of input data and that, in this case study, no significant effects were observed due to the 10,000 input presence points. In fact, models with a large number of occurrences in the training set performed, on average, better and had smaller variances than models built with few occurrences. These results indicated that poorly performing models are less likely to be fitted for species with a sufficient number of occurrences. Therefore, the use of a high number of background points (i.e., 10,000 in this case study) and a good number of presence points increase the performance and prediction of SDMs. This is in line with the study of Barbet-Massin [49].

Since MARS, BRT and MAXENT were found to be less affected, in terms of accuracy measures, by the number of input presence data, they could be utilised in simulations where multiple species are considered, when characterised by a different amount of input

 doi:10.20944/preprints202302.0379.v1

presence data. However, *overfitting* should be duly considered in the choice of the model to apply; BRT, for instance, showed this drawback in this study, as observed in other studies [30].

In this study, elevation was among the main predictors, when a high number of presence points were considered, increasing suitability up to a maximum predictor value of 400 m. This could be explained by a reduced temperature with an increase in elevation, following the gradient south-north, so that areas located above 400 m presented marginal climatic conditions for citrus cultivation.

However, model performance was found to be greatly affected by the resolution. The spatial scale of the study was affected by the study extent and the resolutions of the available input rasters that ranged from a 20-m DTM, and a high presence data density, to the 1-km resolution of WorldClim biovariables. This suggested to consider that DTM could have a higher effect on the probability of presence compared to bioclimatic variables that were less detailed.

The MaxEnt model is one of the most applied models and the impact of beta-regulations on performance and final results has been investigated. Based on the results obtained, the regulator penalised the areas with fewer input presence points. Therefore, it should be investigated whether the regulator value can be related to the number of input presence points.

According to Guisan [16], it is necessary to carry out more tests to determine the resolution ratio, the number of points and the extension of the study area for the good result of the prediction since the significant influence of the grain size could be due to multiple factors. For example, the use of a low resolution encloses different conditions in one pixel and consequently would lead to the selection of unsuitable habitats for the plant.

Conversely, the use of a high resolution can lead to forced resampling and, consequently, the result could not provide consistent information with real conditions.

5. Conclusions

Since fitting an SDM involves a series of steps, each requiring a number of choices and well-justified decisions, this study have investigated on what would be the effect of changing algorithms parameters, and data width on SDMs performance.

The results of this study demonstrated that the number of the presence points have a significant impact on the expected presence of the species, therefore it is crucial to consider an adequate number of input presence points in relation to the SDM sensitivity to this parameter. Furthermore, the resolution chosen for the input levels must be proportional to the study area and the number of input presence points in order to maximise model performance and obtain reliable predictions.

Although this modelling application was wide in terms of covariates and presence data width, as well as models' parameter tuning in order to improve the outcomes, further research is needed to explore other potential important predictors and their quality. In fact, in the context of crop suitability mapping, uncertainties may arise by a number of other circumstances such as the adoption of novel techniques, new crop varieties, specific economic drivers, and trade that could influence crop production. Thus, the effort to introduce new spatial explicit predictors' data related those drivers of change in the species distribution could significantly improve the connected models' predictive capability.

The outcomes of this study have broaden the information basis thus contributing to support aware utilisation of SDMs, coupled with GIS tools, in the studies related to the environmental sector.

Author Contributions: "Conceptualization, C.A. and G.A.C.; methodology, C.A. and G.A.C.; software, G.A.C.; validation, P.R.D. and C.A.; formal analysis, P.R.D. and F.M.; investigation, P.R.D. and C.A.; resources, C.A.; data curation, G.A.C. and F.M; writing—original draft preparation, C.A. and G.A.C.; writing—review and editing, C.A., F.M., and P.R.D.; visualization, P.R.D.; supervision,

C.A.; project administration, C.A.; funding acquisition, C.A. All authors have read and agreed to the published version of the manuscript.

Funding: The research study was funded and APC entirely paid by the University of Catania through the project 'PON "RICERCA E INNOVAZIONE" 2014–2020, Azione II – Obiettivo Specifico 1b – Progetto "Miglioramento delle produzioni agroalimentari mediterranee in condizioni di carenza di risorse idriche—WATER4AGRIFOOD", Cod. progetto: ARS01_00825, CUP: B64I20000160005. This study was also carried out within the Agritech National Research Center and received funding from the European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.4 – D.D. 1032 17/06/2022, CN00000022). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

Acknowledgments: The authors wish to thank the Sicilian Region for SITR data (https://www.sitr.regione.sicilia.it/).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Cramer, Wolfgang, et al. "Climate change and interconnected risks to sustainable development in the Mediterranean." Nature Climate Change 8.11 (2018): 972-980.
2. Akpoti, K., Kabo-Bah, A. T., Dossou-Yovo, E. R., Groen, T. A., & Zwart, S. J. (2020). Mapping suitability for rice production in inland valley landscapes in Benin and Togo using environmental niche modeling. Science of the total environment, 709, 136165. https://doi.org/10.1016/j.scitotenv.2019.136165.
3. Elith, J., & Leathwick, J. R. (2009). Species distribution models: ecological explanation and prediction across space and time. Annual review of ecology, evolution, and systematics, 40, 677-697, doi:10.1146/annurev.ecolsys.110308.120159.
4. Baer, K. C.; Gray, A. N. Biotic predictors improve species distribution models for invasive plants in Western US Forests at high but not low spatial resolutions. Forest Ecology and Management, 518, 120249. 2022. https://doi.org/10.1016/j.foreco.2022.120249.
5. West, A. M., Jarnevich, C. S., Young, N. E., & Fuller, P. L. (2019). Evaluating potential distribution of high-risk aquatic invasive species in the water garden and aquarium trade at a global scale based on current established populations. Risk Analysis, 39(5) 1169-1191. doi.org/10.1111/risa.13230. B.
6. Yang, X. Q.; Kushwaha, S. P. S.; Saran, S.; Xu, J., Roy, P. S. Maxent modeling for predicting the potential distribution of medicinal plant, Justicia adhatoda L. in Lesser Himalayan foothills. Ecological engineering, 51, 83-87. 2013. https://doi.org/10.1016/j.ecoleng.2012.12.004.
7. Yi, Y. J.; Cheng, X.; Yang, Z. F.; Zhang, S. H. *Maxent modeling for predicting the potential distribution of endangered medicinal plant (H. riparia Lour) in Yunnan, China*. Ecological Engineering, 92, 260-269. 2016. https://doi.org/10.1016/j.ecoleng.2016.04.010.
8. Nascimbene, J.; Casazza, G.; Benesperi, R.; *et al.*; Climate change fosters the decline of epiphytic Lobaria species in Italy, Biological Conservation, Volume 201, 2016, Pages 377-384, ISSN 0006-3207, https://doi.org/10.1016/j.biocon.2016.08.003.
9. Brun, P., Vogt, M., Payne, M.R., Gruber, N., O'Brien, C.J., Buitenhuis, E.T., Le Quéré, C., Leblanc, K. and Luo, Y.-W. (2015), Ecological niches of open ocean phytoplankton taxa. Limnol. Oceanogr., 60: 1020-1038. https://doi.org/10.1002/lno.10074.
10. Diniz-Filho, JAF; Mauricio Bini, L; Fernando Rangel, T; Loyola, R.D., Hof, C., Nogués-Bravo, D. and Araújo, M.B. (2009), Partitioning and mapping uncertainties in ensembles of forecasts of species turnover under climate change. Ecography, 32: 897-906. 2009. https://doi.org/10.1111/j.1600-0587.2009.06196.x
11. Zouabi, O., Kadria, M. The direct and indirect effect of climate change on citrus production in Tunisia: a macro and micro spatial analysis. Climatic Change 139, 307–324 (2016). https://doi.org/10.1007/s10584-016-1784-0.
12. Ashraf, U.; Ali, H.; Chaudry, M.N.; Ashraf, I.; Batool, A.; Saqib, Z. Predicting the Potential Distribution of Olea fer-ruginea in Pakistan incorporating Climate Change by Using Maxent Model. Sustainability 2016, 8, 722. https://doi.org/10.3390/su8080722.
13. Leanza, P.M.; Valenti, F.; D'Urso, P. R.; Arcidiacono, C. *A combined MaxEnt and GIS-based methodology to estimate cactus pear biomass distribution: application to an area of southern Italy*. Biofuels. Bioproducts and Biorefining. 2022. 16.1: 54-67. https://doi.org/10.1002/bbb.2304.
14. Qiao, Liu, Christoph Mayer, and Shiyin Liu. "*Distribution and interannual variability of supraglacial lakes on debris-covered glaciers in the Khan Tengri-Tumor Mountains, Central Asia*." Environmental Research Letters 10.1 (2015): 014014.
15. Jarnevich, C. S., Stohlgren, T. J., Kumar, S., Morisette, J. T., & Holcombe, T. R. (2015). Caveats for correlative species distribution modeling. Ecological Informatics, 29, 6-15. https://doi.org/10.1016/j.ecoinf.2015.06.007
16. Guisan, A., Graham, C.H., Elith, J., Huettmann, F., and the NCEAS Species Distribution Modelling Group. 2007. Sensitivity of predictive species distribution models to change in grain size. Diversity and Distributions, (Diversity Distrib.) (2007) 13 (3), 332–340. DOI: 10.1111/j.1472-4642.2007.00342.x
17. Istat, Atlante dell'agricoltura in Sicilia. Una lettura guidata delle mappe tematiche. Istat, Rome (2014)
18. Del Bravo F.; Finizia A.; Lo Moriello M. S.; Ronga M. La competitività della filiera agrumicola in Italia. Rete Rurale Nazionale 2014-2020, 2020
19. Catalano, G.A.; Maci, F.; D'Urso, P.R.; Arcidiacono, C. GIS and SDM-Based Methodology for Resource Optimisation: Feasibility Study for Citrus in Mediterranean Area. Agronomy 2023, 13, 549. https://doi.org/10.3390/agronomy13020549.

20. Miller, J. (2010), Species Distribution Modeling. Geography Compass, 4: 490-509. https://doi.org/10.1111/j.1749-8198.2010.00351.x.
21. Jarnevich, C. S., Talbert, M., Morisette, J., Aldridge, C., Brown, C. S., Kumar, S., ... & Holcombe, T. (2017). Minimizing effects of methodological decisions on interpretation and prediction in species distribution studies: An example with background selection. Ecological Modelling, 363, 48-56.
22. West, A. M., Kumar, S., Brown, C. S., Stohlgren, T. J., & Bromberg, J. (2016). Field validation of an invasive species Maxent model. Ecological Informatics, 36, 126-134. https://doi.org/10.1016/j.ecoinf.2016.11.001.A
23. Hayes, M. A., Cryan, P. M., & Wunder, M. B. (2015). Seasonally-dynamic presence-only species distribution models for a cryptic migratory bat impacted by wind energy development. PLoS One, 10(7), e0132599, doi.org/10.1371/journal.pone.0132599.
24. Chang, T., Hansen, A. J., & Piekielek, N. (2014). Patterns and variability of projected bioclimatic habitat for Pinus albicaulis in the Greater Yellowstone Area. PLoS One, 9(11), e111669. https://doi.org/10.1371/journal.pone.0111669.
25. Valenti, Francesca, et al. A GIS-based model to estimate citrus pulp availability for biogas production: an application to a region of the Mediterranean Basin. Biofuels, Bioproducts and Biorefining, 2016, 10.6: 710-727. https://doi.org/10.1002/bbb.1707.
26. Young NE, Jarnevich CS, Sofaer HR, Pearse I, Sullivan J, Engelstad P, et al. (2020) A modeling workflow that balances automation and human intervention to inform invasive plant management decisions at multiple spatial scales. PLoS ONE 15(3): e0229253. https://doi.org/10.1371/journal.pone.0229253.
27. O'donnell, M. S., & Ignizio, D. A. (2012). Bioclimatic predictors for supporting ecological applications in the conterminous United States. US geological survey data series, 691(10), 4-9. NODOI
28. University of Catania, CREA, Distretto Agrumi Sicilia and CocaCola Foundation, A.C.Q.U.A. PROJECT RESULTS, 2020 (https://www.distrettoagrumidisicilia.it/wp-content/uploads/Dossier-Acqua5.pdf accessed on 06/09/2022).
29. Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. Journal of animal ecology, 77(4), 802-813. https://doi.org/10.1111/j.1365-2656.2008.01390.x.
30. Suleiman, A., Tight, M.R. & Quinn, A.D. Hybrid Neural Networks and Boosted Regression Tree Models for Predicting Roadside Particulate Matter. *Environ Model Assess* **21**, 731–750 (2016). https://doi.org/10.1007/s10666-016-9507-5.
31. Breiner, F.T., Guisan, A., Bergamini, A. and Nobis, M.P. (2015), Overcoming limitations of modelling rare species by using ensembles of small models. Methods Ecol Evol, 6: 1210-1218. https://doi.org/10.1111/2041-210X.12403.
32. Heikkinen, R. K., Luoto, M., Araújo, M. B., Virkkala, R., Thuiller, W., & Sykes, M. T. (2006). Methods and uncertainties in bioclimatic envelope modelling under climate change. Progress in Physical Geography, 30(6), 751-777. doi.org/10.1177/030913330607195.
33. Morisette, J.T.; Jarnevich, C.S.; Holcombe, T.R.; Talbert, C.B.; Ignizio, D.; Talbert, M.K.; Silva, C.; Koop D.; Swanson, A.; Young, N.E.; *VisTrails SAHM: visualization and workflow management for species habitat modeling*. Ecography, 36: 129-135. 2013. https://doi.org/10.1111/j.1600-0587.2012.07815.x.
34. Morales NS, Fernández IC, Baca-González V. 2017 . Configurazione dei parametri di MaxEnt e piccoli campioni: stiamo prestando attenzione alle raccomandazioni? Una revisione sistematica . PeerJ 5 : e3093 https://doi.org/10.7717/peerj.3093.
35. Biau, G., & Scornet, E. (2016). A random forest guided tour. *Test*, *25*(2), 197-227.
36. Valavi, R., Elith, J., Lahoz-Monfort, J. J., & Guillera-Arroita, G. (2021). Modelling species presence-only data with random forests. Ecography, 44(12), 1731-1742, https://doi.org/10.1111/ecog.05615.
37. Eilers, P. H. C., & Marx, B. D. (2002). Generalized linear additive smooth structures. Journal of computational and graphical statistics, 11(4), 758-783. doi.org/10.1198/106186002844.
38. Chen, X., Aravkin, A. Y., & Martin, R. D. (2018). Generalized linear model for gamma distributed variables via elastic net regularization. arXiv preprint arXiv:1804.07780. https://doi.org/10.48550/arXiv.1804.07780.
39. Phillips, S. J., & Dudík, M. (2008). Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. Ecography, 31(2), 161-175, https://doi.org/10.1111/j.0906-7590.2008.5203.x.
40. Phillips, S. J., Anderson, R. P., & Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. Ecological modelling, 190(3-4), 231-259, https://doi.org/10.1016/j.ecolmodel.2005.03.026.
41. Díaz-Uriarte, R., Alvarez de Andrés, S. Gene selection and classification of microarray data using random forest. BMC Bioinformatics 7, 3 (2006). https://doi.org/10.1186/1471-2105-7-3 .
42. Catalano G. A., Maci F., Valenti F., D'Urso P.R., Arcidiacono C. (in press) Application of geospatial models for suitability and distribution potential of citrus: a case study in eastern Sicily. 12th International AIIA Conference: September 19-22, 2022 Palermo – Italy "Biosystems Engineering towards the Green Deal", Springer.
43. Naimi, B., & Araújo, M. B. (2016). sdm: a reproducible and extensible R platform for species distribution modelling. Ecography, 39(4), 368-375, https://doi.org/10.1111/ecog.01881.
44. D'Arrigo, G., Provenzano, F., Torino, C., Zoccali, C., & Tripepi, G. (2011). I test diagnostici e l'analisi della curva ROC. G Ital Nefrol, 28(6), 642-647.
45. Di Napoli, M., Carotenuto, F., Cevasco, A. *et al.* Machine learning ensemble modelling as a tool to improve landslide susceptibility mapping reliability. *Landslides* **17**, 1897–1914 (2020). https://doi.org/10.1007/s10346-020-01392-9
46. Salas, E. A. L., Seamster, V. A., Boykin, K. G., Harings, N. M., & Dixon, K. W. (2017). Modeling the impacts of climate change on Species of Concern (birds) in South Central US based on bioclimatic variables. AIMS Environmental Science, 4(2), 358-385. doi: 10.3934/environsci.2017.2.358.
47. Saidi, H., Dresti, C., & Ciampittiello, M. Il cambiamento climatico e le piogge: analisi dell'evoluzione delle piogge stagionali e degli eventi estremi negli ultimi 50 anni nella stazione di Pallanza. Biologia Ambientale, Vol.28, n°2,2014.

48. Primo-Capella, A.; Martínez-Cuenca, M.-R.; Forner-Giner, M.Á. Cold Stress in Citrus: A Molecular, Physiological and Biochemical Perspective. Horticulturae 2021, 7, 340. https://doi.org/10.3390/horticulturae7100340.
49. Barbet-Massin, M., Jiguet, F., Albert, C. H., & Thuiller, W. (2012). Selecting pseudo-absences for species distribution models: How, where and how many?. Methods in ecology and evolution, 3(2), 327-338. https://doi.org/10.1111/j.2041-210X.2011.00172.x.