Comparison of deductive classification techniques for predicting potential spatial distribution of quarantine insects

HEIT, Guillermo1, SIONE, Walter2,3, CLAPS, Lucía E.4,* & ACEÑOLAZA, Pablo5

1 Cátedra de Zoología Agrícola, Facultad de Agronomía, Universidad de Buenos Aires. Ciudad Autónoma de Buenos Aires, Argentina.
2 CEREGEO, Universidad Autónoma de Entre Ríos. Diamante, Argentina.
3 Universidad Nacional de Luján. Luján, Argentina.
4 Instituto Superior de Entomología “Dr. Abraham Willink” (INSUE), Facultad de Ciencias Naturales e Instituto Miguel Lillo, UNT. San Miguel de Tucumán, Argentina. * E-mail: luciACLAPS@gmail.com
5 CiCyTTP-CONICET & FCA-UNER. Diamante, Argentina.

RESUMEN. El objetivo de este trabajo fue comparar el desempeño de los criterios de clasificación nítidos y difusos en la construcción de modelos deductivos de la distribución potencial de insectos exóticos. Considerando criterios de clasificación binaria nítida y difusa, de capas ráster de temperatura máxima, media y mínima diaria, se generó un índice de riesgo bioclimático relativo, considerando el número de días con condiciones óptimas para el desarrollo de Bactrocera oleae (Gmelin) (Diptera: Tephritidae) y Cerotoma arcuatus (Olivier) (Coleoptera: Chrysomelidae). Se realizaron análisis de sensibilidad de los modelos. Los modelos deductivos de distribución potencial de especies realizados mediante clasificación difusa, serían más robustos y menos restrictivos en la determinación de áreas de riesgo fitosanitario potencial que aquellos realizados con criterios de clasificación nítidos. Estos últimos serían más sensibles y tendrían mayor capacidad de discriminar áreas con diferentes perfiles de riesgo ambiental.

PALABRAS CLAVE. Distribución potencial. Insectos exóticos. Lógica difusa.

ABSTRACT. The objective of this paper was to evaluate the performance of crisp and fuzzy classification criteria in the construction of deductive potential distribution models of exotic insects. As case studies, Bactrocera oleae (Gmelin) (Diptera: Tephritidae) and Cerotoma arcuatus (Olivier) (Coleoptera: Chrysomelidae) were selected. Considering crisp and fuzzy classification for raster layers of maximum, average and minimum daily temperature, a relative bioclimatic risk index was generated. The number of days with optimal conditions for pests’ development was considered. Sensitivity analyses of both models were performed. Considering each case evaluated and the variables used, deductive pest distribution models made by fuzzy classification was more robust and less conservative in the determination of potential phytosanitary risk areas than those made with crisp classification criteria. This last case was more sensitive and would have a greater capacity to discriminate areas with different environmental risk profiles.

INTRODUCTION

Agriculture, forestry, trade and other human activities have a determining role in the voluntary or accidental dispersion of species towards areas that they could not have reached without human assistance (Hlasny & Livingston, 2008). Strong increase in international trade in last decades has generated an increase in the probability that many species can be established outside their natural range. For that reason, countries worldwide make great efforts to avoid dispersion of invasive species, regulating the phytosanitary condition of plant products on international trade (Levine & D’Antonio, 2003; Brenton-Rule et al., 2016).

Although qualitative risk assessments often provide sufficient technical solutions to perform pest risk analysis for a particular route of import, there are situations in which the use of species distribution models can help in decision-making, in order to identify areas at risk of invasion or to design monitoring protocols in the field in support of eradication programs for a new quarantine pest (Baker, 2012; Cardador et al., 2016).

The challenge of estimating spatial distribution patterns of species has been addressed through different methodological approaches, which have been described and compared in several bibliographic reviews (Venette et al., 2010; Zimmermann et al., 2010; Mateo et al., 2011). Venette et al. (2010) grouped the sets of techniques into two differential methodological approaches, called "deductive" or "inductive" approaches, the latter require information on the presence sites of the species.

Deductive approaches use detailed knowledge about the biology of a species, to infer the areas of environmental suitability. These models, due to inherent characteristics of their design, may incorporate biases to the classification output. One is the selection criteria of input variables, mainly because of the variability of published information on the biology of studied species. Another dilemma is the difficulty of defining adjusted cut-off criteria for each factor that influence pest populations’ development. In this sense, it is possible that the variability of these factors and the uncertainty of the classification criteria used can be solved through the application of classification tools from fuzzy logic theory (Siler & Buckley, 2005).

Fuzzy logic theory provides a method to reduce the complexity of a system through a compatibility calculation. In traditional sets, an element of the population may or may not belong to a class, and therefore the membership of an element in a set is clear, "crisp". The fuzzy set theory recognizes that certain sets have imprecise limits, in which the transition of belonging to the same is not abrupt but gradual (Castillo & Melin, 2014). These classes can overlap and this overlap corresponds to the transition from one state to the next, such as fuzzy boundaries in natural spaces (Adriaenssens et al., 2004). Velásquez & Hester (2013) argue that fuzzy logic theory provides a mathematical basis to construct decision rules for invasive species risk assessment.

The application of tools derived from fuzzy logic to natural resources is relatively new. The general areas of application include the classification of remote sensors images (Shabnam & Zhang, 2013), analysis of environmental risks (Camastra et al., 2015), floristic diversity (Ibrahim et al., 2015) and ecosystem analysis (Olivero et al., 2013). The aim of this study was to compare the performance of the classification criteria crisp and fuzzy, in the construction of deductive models for potential distribution of exotic insects in Argentina.

Two quarantine insect species absent from Argentina were selected as a case study to perform a deductive analysis of potential geographic distribution: Bactrocera oleae (Gmelin) (Diptera: Tephritidae) and Cerotoma arcuatus (Olivier) (Coleoptera: Chrysomelidae).

Bactrocera oleae is the most important olive pest worldwide. Its geographic distribution includes the Mediterranean Region, North America, Western Asia and Eastern and Southeastern Africa. It is a multivoltine species, whose larvae feed on the fruit pulp. Usually it overwinters as pupa buried a few centimeters from the soil surface. Considering a base temperature of 8 °C, a generation of B. oleae is completed at 491 degrees day (Hanife & James, 2008; Gutiérrez et al., 2009). The optimum temperature of development is between 23 and 29 °C; while the maximum development temperature is 35 °C (Ordano et al., 2015).

Cerotoma arcuatus is considered one of the major chrysomelid pests of legumes, mainly soybeans and beans. Considering a base temperature of 10 °C, a generation of C. arcuatus is completed at 489 degrees day. Optimal development temperature is between 22 and 27, while the maximum development temperature is 32 °C (Nava & Postali Parra, 2003).

MATERIAL AND METHODS

Input data

Mean daily records of maximum and minimum air temperature (°C), from the period 1991-2014, were interpolated at spatial resolution of 2 km, according to the methodology proposed by Heit et al. (2013). Records from 124 weather stations of the National Weather Service (SMN) and the National Institute of Agricultural Technology (INTA) were used. Digital terrain model of the Shuttle Radar Topography Mission (SRTM) was used as external drift variable for Kriging algorithm (Aalto et al., 2013). These raster layers were validated by generalized cross validation (Haylock et al., 2008).
A widespread problem in the construction of meteorological data interpolation algorithms in a mountainous area is the lack of meteorological stations at sites of high altitude above sea level. Some regional studies have supplemented missing observations at high altitudes with information from remote sensors (Stahl et al., 2006). This study supplemented the lack of data from meteorological stations located in high altitude sites, based on surface temperature values of 30 randomly selected points along the Andean mountain range, monthly average 2008-2012, from the information generated by the National Oceanic and Atmospheric Administration’s (NOAA) meteorological satellite.

**Bioclimatic risk indexes**

In order to establish comparisons of bioclimatic risk between the different Argentine regions, a relative risk index was generated based on the comparison of the score obtained by each pixels of the Argentine territory, in relation to those that obtained the highest absolute value. Considering criteria of binary (Crisp) and fuzzy (Fuzzy) classification, the number of days with optimum conditions of development of each one of the species was calculated, between September 1st and March 31st. Thus, the relative bioclimatic risk index (IRB) has values that range from 0: no risk, to 1: maximum risk.

Binary classification (Crisp) was based on the methodology proposed by Heit et al. (2013). Pixels were classified as follows: EITHER 1 IF (Tn> Th) OR 0 OTHERWISE; AND EITHER 1 IF (Tx< txd) OR 0 OTHERWISE; AND EITHER 1 IF (Toi>= (Tm) =<Tos) OR 0.5 IF (Tb> (Tm) <Toi) OR 0.5 IF (tos> (Tm) <Txd) OR 0 OTHERWISE.

Where (Tx): maximum temperature, (Tn): minimum temperature, (Txd): maximum temperature of development of the species, Tb: base temperature of development of the species, Tni: optimum temperature of development-threshold lower, Tos: optimum development temperature-upper threshold, Th: temperature of agrometeorological frost. Agrometeorological frost is defined as any thermal decrease equal to or lower than 3 °C, measured in the meteorological shelter, which would be equivalent to 0 °C, or less, to the surface outdoors (Gusta & Wisniewski, 2013).

The classification criteria of the input variables to the model using fuzzy logic (fuzzy), are shown in Figures 1a and 1b. Membership function could take any value from the interval (0,1) representing the full or null degree to which each proposition occurs, respectively. In our case, 1 (one) represents the total fulfillment of the conditions that determine the maximum risk of pest presence. Otherwise, 0 (zero) represent the total fulfillment of the condition that minimize potential risk of pest establishment, based on knowledge available in the literature. Linear membership functions were used and monotonic linear transitions from one state to another shows continuous changes of the variable under study, from a minimum to a maximum level. The outputs from fuzzy prepositions were combined by using a set of rules, total fulfillment of the conditions was 1 IF (Toi>= (Tm) =<Tos) AND (Tn> Tb) AND (Tx<tb). Total fulfillment of the condition that minimize potential risk of pest establishment were 0 IF (Tb<Tm>Txd) AND (Tn<Th) AND (Tx>Txd).

The same biological parameters were used to classify the environmental aptitude for the development of the species for both criteria of classification. The theoretical assumptions of the models were: first, biotic interactions are not important on a regional scale or are constant in time and space. Second, genetic and phenotypic composition of the species is constant in space and time. Third, there are no limitations to the dispersal of the species (Urban et al., 2007).

Raster masks of distribution area of susceptible vegetation for each pest species were applied to bioclimatic risk indexes output. The potential distribution of susceptible crops was obtained on the basis of the 2002 National Agricultural Census (National Institute of Statistics and Census) presented by political subdivisions (Departmental). Vector layer were rasterized to a grid of 1 km wide.

In addition, United Nations Land Cover Classification System (UN-LCCS, GLOBCOVER-ESA, 300 m spatial resolution), was used to select in each Department those pixels associated with the host crop (Bicheron et al., 2006). Finally, the average monthly vegetation index MOD13Q1 of the 2009-2012 series (250 m spatial resolution) was used to discard cells with a high percentage of bare soil (NDVI<0.3).

**Sensitivity analysis**

Sensitivity analysis was performed to assess the robustness and sensitivity of the model to systematic changes of input variables to the system. The selection criteria of the input variables to the model were systematically increased or reduced to assess relative changes in the model results (Ferraro, 2010). Two
specific tests were evaluated:

**Test 1**: For each of the evaluated months, the selection ranges of climate variables input to the model were independently increased, or decreased, by 10% to reach 50% change. The results of sensitivity, both of increase or reduction of the selection ranges, were calculated using a base scenario, such as 0 (minimum risk) or 1 (maximum risk).

**Test 2**: All input variables to the system were simultaneously increased, or decreased, by 20 or 50%. Sensitivity results were calculated using the same base scenarios in test 1.

Since the delimitation of susceptible areas depended largely on the classification of qualitative variables, the estimated percentages of relative change was performed considering the following classification criteria: a) to estimate the percentage reduction in the area potentially planted by the host crop, the agricultural area of the departments was reduced by 20% and 50%, based on the INDEC statistics; and/or to estimate the percentage increase in the area potentially planted by the reference host crops, the agricultural surface of the departments was increased by 20 and 50%; considering as the maximum area, the total area of the departments with an NDVI>0.3 (in order to discard areas with a high percentage of bare soil). R software (gstat, gdal and automap libraries), QGIS 2.12 (Quantum GIS Development Team, 2015), IDRISI Selva and “Infostat estudiantil” were used for the treatment of the information.

### RESULTS AND DISCUSSION

Figure 2 shows the relative bioclimatic risk index (RBRI) for the potential establishment of *B. oleae* in Argentina, considering both classification criteria. It can be observed that the RBRI differs significantly between the crisp and fuzzy classification criteria.

The relative bioclimatic risk index for *B. oleae*, based on the binary classification (Crisp), estimates that 3.1% of the area potentially implanted with its main host (olive), would have a high establishment risk for this species; 21.1% of the area had a moderate risk and the remaining 75.8% would have a low establishment risk for *B. oleae*. Furthermore, based on fuzzy logic classification, 11% of this area could be classified as high risk of establishment; 13.6% with a moderate risk and 75.4% of the potentially implanted area with olive would have a low risk of establishment for *B. oleae*.

The relative bioclimatic risk index in the potential area of susceptible hosts for the development of *C. arcuatus* populations is shown in Figure 3. Similar to what was observed for *B. oleae*, it can be observed that RBRI differs between the two classification criteria evaluated. The Crisp classification estimates that 14.8% of the area potentially planted with its main hosts (soybeans and beans) was classified as high establishment risk; 77% of moderate risk and the remaining 8.1% as areas of low risk of establishment for this species.

On the other hand, the fuzzy classification estimated that 67.7% of the area occupied with the main plant hosts for *C. arcuatus* could be considered as a high risk of establishment; 32.2% with moderate risk and 0.1% with low risk of establishment of the species.

Figure 4 shows the sensitivity analysis of test 1 for *B. oleae*, broken down by classification criteria. Therein, it shows that the crisp classification method is more sensitive to changes in input variables to the fuzzy classification system ($\chi^2(9, 0.05)$: 57.39, p<0.001). For crisp classification criteria, a negative correlation was observed between output risk areas and systematic change of climate variables input. Greater relative change in the risk areas where found in the lower range limit of systematic change of the input variables, and the smaller relative changes in risk area on the upper range limit, Spearman’s correlation coefficient ($\rho$): -0.94, p: 0.0048 for high risk class and $\rho$: -0.89, p: 0.007 for moderate risk class. In Fuzzy classification criteria, no linear correlation was observed between the change in risk area and the systematic change of variables for high risk class ($\rho$: -0.41, p: 0.2232), and moderate risk class ($\rho$: -0.14, p: 0.6758). No significant differences in sensitivity among risk categories, due to systematic changes in input variables were observed considering crisp classification criteria ($\chi^2(18, 0.05)$: 3.46, p>0.9) or fuzzy classification criteria ($\chi^2(18, 0.05)$: 12.64, p: 0.8127).

Sensitivity tests 1 for *C. arcuatus*, differentiated by classification criteria, are shown in Figure 5. The same patterns of variability are observed as for the case study of *B. oleae*, although the differences between both are significantly higher ($\chi^2(9, 0.05)$: 100.35, p<0.0001). No significant differences in test 1 sensitivity analysis among different risk categories due to systematic changes in input variables were observed considering crisp classification criteria ($\chi^2(18, 0.05)$: 21.96, p: 0.234) or fuzzy classification criteria ($\chi^2(18, 0.05)$: 5.84, p: 0.996).

A negative correlation criterion between output risk areas and systematic change of climate variables input was observed for crisp classification, from -50% systematic change to +50 systematic change of input variables, ($\rho$): -0.99, p: 0.003 for high risk class and $\rho$: -0.89, p: 0.0075 for moderate risk class. No linear correlation was observed between systematic change of variables and risk area considering Fuzzy classification criteria, for high risk class ($\rho$: -0.57, p: 0.084), and moderate risk class ($\rho$: -0.39, p: 0.266). This difference is primarily attributable to the wide geographic extent of the potential range of the main host crops (soybean, bean, etc.), relative to olive cultivation in Argentina.

Figures 6 and 7 show the sensitivity analysis of test 2, for *B. oleae* and *C. arcuatus*, for each of the classification criteria. In the sensitivity test 2, in addition to varying the climatic classification criteria, systematic changes were made in the potential surface implanted with the host crops of the pest in analysis.
Test 2 of sensitivity analysis for crisp classification criteria showed greater variability on risk area delimitation, due to systematic changes of all the classification variables, than fuzzy classification criteria. Crisp classification method is more sensitive to changes in input variables to the fuzzy classification criteria in test 2 sensitivity analysis for *B*. *oleae* ($\chi^2(3, 0.05) = 55.71$, $p<0.001$). When comparing the sensitivity analysis of test 2 against test 1 in *B*. *oleae* according to crisp classification criteria, variation of estimated area classified as high and moderate bioclimatic risk differed between 20 and 50%, when decreasing systematic variations of the input variables were made. While there were decreasing variations, from 1 to 15% of the risk area, when incremental systematic variations in the criteria for classifying the input variables to the system were made.

Area classified as “low risk” showed the lowest relative variations between both tests. Sensitivity analyzes of classifications based on fuzzy logic showed similar behavior in both tests, but different magnitude. Being the variations of the surface estimated as high or moderate risk, 30% higher than those observed in test 1, regardless of the percentage of systematic change made.

Significant differences in test 2 sensitivity analysis between classification criteria were found ($\chi^2(3, 0.05) = 71.84$, $p<0.001$). When comparing the sensitivity analysis of test 2 against test 1 in *C*. *arcuatus*, using the crisp classification criterion, surface variation estimated as high and moderate risk bioclimatic differed between 60 and 110%, when decreasing systematic variations of the input variables were made. While there were decreasing variations of 10 to 30% due to incremental variations in the criteria for classifying the input variables to the system.

No significant differences in test 2 sensitivity analysis among different risk categories due to systematic changes in input variables were observed for both classification criteria and species ($p>0.05$). No linear correlation was observed between systematic change of variables and risk area considering Fuzzy classification criteria, for *B*. *oleae* and *C*. *arcuatus* ($p>0.05$). However, a negative correlation criterion between output risk areas and systematic change of climate variables input was observed. Based on crisp classification, Spearman’s correlation coefficient, from $-50\%$ systematic change to $+50\%$ systematic change of input variables, for high risk class in *B*. *oleae* was $-0.98$ ($p: 0.0194$), and for *C*. *arcuatus* $p: -0.99$ ($p: 0.0133$).

The area considered as low bioclimatic risk did not show relative variations between the two evaluated tests.
When comparing the sensitivity analyzes of both types of tests in the fuzzy classification criterion, it can be noticed that with decreasing systematic variations in the input variables, the estimated area classified as high or moderate bioclimatic risk in test 2, varied from 15 to 50% in relation to the surface variations observed in test 1. Furthermore, due to incremental changes to the criteria for classification of the input variables to the system, variations of the surface considered high or moderate bioclimatic risk is about half the estimated before decreasing variations of such variables. For this reason, it can be considered that most of the variability observed in test 2, in relation to that observed with test 1, it is based on the estimation of the area implanted with susceptible crop/host to the pest.
An important goal of pest risk assessment is to provide assistance to policy makers in environmental protection. Areas of higher agro-ecological vulnerability should be protected over all others, and appropriate trade regulations and surveillance systems should be developed regarding environmental protection (Franklin, 2009). Although there is a growing number of scientific studies comparing different methodological approaches for species distribution models (Heikkinen et al., 2006; Jescheck & Strayer, 2008), it is not possible to generate a unique protocol analysis to estimate the potential quarantine pests distribution for the majority of cases addressed by a National Plant Protection Organization. Comparison of species distribution models in terms of their assumptions, approaches and results provides a perspective on the uncertainty of the prediction and ultimately allows policymakers make better decisions (Schneiderman et al., 2015).
It should be noted that the accuracy of the risk mapping was dependent on different factors, like data quality, data processing criteria and selection of threshold for classification of factors. Uncertainties regarding data sources may introduce even larger uncertainties within environmental evaluations (Peche & Rodríguez, 2009).

Deductive models are highly dependent on the quality of the scientific literature on a pest. If literature does not indicate which factors are most influential in the distribution of the species, policymakers must choose the variables that, based on their own experience, have a significant effect in restrict the potential geographical distribution of the pest evaluated, at the cartographic scale study selected. This also represents a critical point for inductive models due that the modeler must select and prioritize system input variables, since they are the most relevant covariates for creating potential distribution maps (Dupin et al., 2011).

Sensitivity analyses indicated that results of the proposed deductive method approach are highly sensitive to the selection of threshold for classification of factors. In that sense, results provide evidences that the choice of the classification criteria can have statistically significant effects on spatial patterns of deductive species distribution model predictions.

Many authors have highlighted and compared the predictive power of correlative species distribution models (SDM) (Zimmermann et al., 2010; Mateo et al., 2011; Kehlenbeck et al., 2012); while others have argued that this is unimportant, since many assumptions of the species distribution models are not reasonable and therefore their results lack scientific validity (Rose & Burton, 2010; Sinclair et al., 2010). It is therefore necessary to judge the merits of the SDMs according to the objectives for which they have been developed. In the case of using these models in the framework of pest risk analysis or phytosanitary surveillance tasks, they have a strategic role in the alerting to decision-makers about the scenarios that National Plant Protection Organization must face in case of an exotic pest invasion.

Deductive species distribution model carried out considering Fuzzy classification criteria, would be more robust and less restrictive in identifying areas of potential phytosanitary risk, than those made using crisp classification criteria, which would be more sensitive and would have a greater capacity to discriminate areas with different environmental risk profiles.

Regarding the inherent characteristics to the variables used as input in the models, it was observed, in both classification criteria, a lower sensitivity in the results of the classifications area when performing incremental systematic variations of the selected classification ranges, than when decreasing them. Most of the variability observed in the areas, identified by both classification criteria, was due to the estimation of potential area occupied by the susceptible crops rather than the method used.

### LITERATURE CITED


