Transferability of species distribution models. A case study of the fungus *Phytophthora cinnamomi* in Andalusia and Southwest Australia

JOAQUIN DUQUE LAZO
MAY, 2013

Course Title:	Geo-Information Science and Earth Observation for Environmental Modelling and Management
Level:	Master of Science (MSc)
Course Duration:	August 2011 - June 2013
Consortium partners:	Lund University (Sweden) University of Sydney (Australia, associate partner) University of Twente, ITC (Netherlands) University of Warsaw (Poland) University of Iceland (Iceland) University of Southampton (UK)

Transferability of species distribution models. A case study of the fungus *Phytophthora cinnamomi* in Andalusia and Southwest Australia

by

JOAQUIN DUQUE LAZO

Thesis submitted to the, Faculty of Geo-Information Science and Earth Observation (ITC) of the University of Twente, in partial fulfilment of the requirements for the degree of Master of Science in Geo-information Science and Earth Observation, Specialisation: Environmental Modelling and Management.

Thesis Assessment Board

Dr. Y.A. Hussin (Chair)
Dr. R.H.G. Jongman (External examiner, Alterra, Wageningen University)

Dr. H. A. M. J. van Gils (First Supervisor)
Dr. ir. T.A. Groen (Second Supervisor)

Contribution

Dr. Rafael María Navarro Cerrillo (Cordoba University)
Dr. Eleanor Bruce (Sydney University)

Disclaimer

This document describes work undertaken as part of a programme of study at the University of Twente, Faculty of Geo-Information Science and Earth Observation (ITC), All views and opinions expressed therein remain the sole responsibility of the author, and do not necessarily represent those of the institute.

Abstract

Species distribution models (SDM) predict species occurrence based on statistical relationships with environmental conditions. Many studies have compared SDM accuracies but only a few have compared SDM transferability between two regions as distant as Andalusia and SW Australia. The R-package biomod2 which includes 10 different SDM techniques was used in this study. *Phytophthora cinnamomi* Rands is a plant pathogen oomycete which is the main factor of the Oak Decline in Andalusia and the Jarrah Dieback in SW Australia. *P. cinnamomi* location data was used to test SDMs transferability and, simultaneously, to assess the environmental response of *P. cinnamomi* in both regions. It was found that *P. cinnamomi* risk of invasion was predicted accurately with all model techniques tested, except SRE, and that different environmental conditions explained the risk of fungal invasion in each study area. Moreover, machine-learning methods had a high predictive power in the training area but low transferability, while linear based models gave a reasonable accuracy within the training area and better transferability performance. A desirable combination of good model performance and good transferability was manifested by GAM and GLM.

Key-words: Species distribution models, transferability, biomod, *Phytophthora cinnamomi,* Oak Decline, Jarrah Dieback.

Acknowledgements

It is a pleasure to thank many people who made this thesis possible.

First of all, I would like to express my deep gratitude to my supervisors, Dr. H.A.M.J. van Gils and Dr. T.A. Groen. They have been very patient, have supported me in my work and lead my study; I have learnt a lot from them. It has also been a great pleasure to know them personally.

I also thank Dr. R.M. Navarro Cerrillo, University of Cordoba, for his wise advice concerning the study, his continuous support and for providing *Phytophthora cinnamomi* data from Andalusia. His remarks have been very useful and instructive.

Thanks are due to Mike Stukely, Vegetation Health Service, Science Division, Department of Environment and Conservation, Western Australia for providing *Phytophthora cinnamomi* data from SW Australia.

I place on record, my sincere gratitude to Dr. Eleanor Bruce from the University of Sydney and Dr. Russell Turner for their welcome and support during my Australian internship.

The research group (ERSAF), University of Cordoba, have provided supplementary data of *Phytophthora cinnamomi* from Andalusia and helped me to understand *Phytophthora cinnamomi* behaviour I thank them for it.

I acknowledge the Erasmus mundus program for funding my Msc studies.

I wish to acknowledge Twente University for giving us the opportunity to realize this work and Lund University for receiving me in the first year of my Msc.

Finally but not less important, I cannot forget about the "GEM people 2011 – 2013" and all my ITC friends who made these two years an unforgettable experience and without their friendship, support and encourage this Msc thesis would not have been possible.

Table of Contents

Transferability of species distribution models. A case study of the fungus *Phytophthora cinnamomi* in Andalusia and Southwest Australia i
Abstract v
Acknowledgements vi
List of Figures ix
List of Tables xi

Chapter 1: Introduction 1
1.1 The Fungus 1
1.2 Problem Statement 2
1.3 Species Distribution Models 4
1.4 Research Questions 6

Chapter 2: Study Area, Data and Methodology 7
2.1 Study Area 7
2.1.1 Andalusia (Spain) 8
2.1.2 Southwest Australia. 10
2.2 Data 11
2.2.1 *Phytophthora cinnamomi* Location 11
2.2.1.1 Andalusia (Spain) 11
2.2.1.2 Southwest Australia 12
2.2.2 Environmental Data 13
2.3 Methodology 14
2.3.1 Models 14
2.3.1.1 Species Distribution Models 14
2.3.1.2 Statistical Models 15
2.3.1.3 Model Selection 17
2.3.1.4 Model Evaluation 18
2.3.1.5 Model Transferability 19
2.3.1.6 Model Comparison 19
2.3.2 Research Procedure 20
2.3.2.1 Variable Pre-Processing 20
2.3.2.2 Study Area Comparison 21
2.3.2.3 Variable Selection 22
2.3.2.4 Analysis of Multicollinearity 23
2.3.2.5 Variable Importance (Model Selection) Analysis ... 24
2.3.2.6 Model Transferability 30

Chapter 3: Results 31
3.1 Study Areas Comparison 31
3.1.1 Multivariate Outlier Detection 31
3.1.2 Boxplots 33
3.2 Variable Selection 34
3.2.1 Analysis of Collinearity 34
3.2.2 Model Selection 34
3.2.3 Variable Importance Analysis 40
3.2.4 Model Accuracy 42
3.2.4.1 Individual Variable Datasets 42
3.2.4.2 Common Variable dataset 45
3.2.5 Response Curve 46
3.2.6 Model Transferability 47
3.2.6.1 Common Variables Boxplots 47
3.2.6.2 Transferability Andalusia Models 50
3.2.6.3 Transferability SW Australia Models 52

Chapter 4: Discussion 55
4.1 Phytophthora cinnamomi Location 55
4.2 Environmental Variables 55
4.3 Study Area Comparison 56
4.4 Variable Importance 56
4.4.1 Number of Variables 56
4.4.2 Variable Importance Comparison 57
4.4.3 Ecological Explanation of Important Variables 59
4.5 Model Evaluation 61
4.6 Model Transferability 63

Chapter 5: Conclusion 65

Chapter 6: References 67

Appendix I 79

List of Figures

Figure 1. Worldwide Distribution of Mediterranean type ecosystems 7
Figure 2. Holm and Cork oak distribution in Andalusia (MAGRAMA, 2007) 8
Figure 3. Vegetation cover on the SW Australia study area (ABARES, 2011) 10
Figure 4. Study Areas comparison flowchart 21
Figure 5. Variable selection general process flowchart 22
Figure 6. Diagram of collinearity analysis 23
Figure 7. Flowchart summarizing the variables selection procedure 24
Figure 8. Flowchart summarizing the first 4 runs of variables selection analysis.....25
Figure 9. Methodology followed in fifth variable selection test....26
Figure 10. Flowchart summarizing the final variable selection step.. 27
Figure 11. Description of the methodology followed to the common relevant variables dataset.....28
Figure 12. Independent test procedure diagram. (SW Aus: SW Australia) 29
Figure 13. Transferability flowchart description. (SW Aus: SW Australia) 30
Figure 14. Mahalanobis Distance analysis plots of all environmental variables. (SW AUS: Southwest Australia; AND: Andalusia; sun_aut: sun shine in autumn) 32
Figure 15. Examples of variable boxplots. (AND: Andalusia; SW AUS: Southwest Australia).....33
Figure 16. Maximum and minimum AUC variability across the number of variables and model techniques in Andalusia 35
Figure 17. Maximum and minimum AUC variability across the number of variables and model techniques in SW Australia 36
Figure 18. Model accuracy variability across the number of variables and model techniques measured as AUC standard deviation. 37
Figure 19. Variables importance by model technique independent dataset. Top: Andalusia; bottom: SW Australia.....38
Figure 20. Variables importance by model technique common dataset. Top: Andalusia; bottom: SW Australia.....39
Figure 21. Variable importance average rank with all model techniques and the individual variable dataset. a) Andalusia; b) SW Australia 41

Figure 22. Variables importance average rank among all models techniques considering the common variables dataset. a) Andalusia; b) SW Australia 42
Figure 23. Example of *P. cinnamomi* risk of invasion in Andalusia (BRT) 44
Figure 24. Example of *P. cinnamomi* risk of invasion in SW Australia (FDA) 44
Figure 25. a) Elevation, b) mean temperature in summer and c) distance to water response curves in Andalusia 46
Figure 26. a) Elevation, b) maximum temperature in summer, c) distance to water and d) actual evapotranspiration in autumn response curves in SW Australia 47
Figure 27. Right, elevation (m) and left, slope (%) boxplots. (AND: Andalusia; SW AUS: Southwest Australia).....48
Figure 28. Right, distance to Water (m) and left, NDVI standard deviation in summer boxplots. (AND: Andalusia; SW AUS: Southwest Australia) 48
Figure 29. Right wet days frequency in autumn (nº of days) and left, mean summer temperature (ºC) boxplots. (AND: Andalusia; SW AUS: Southwest Australia).....49
Figure 30. Example of *P.cinnamomi* risk of invasion prediction by Andalusia GLM model 51
Figure 31. Example of *P.cinnamomi* risk of invasion in SW Australia by Andalusia by GLM model transferred.....51
Figure 32. Example of *P.cinnamomi* risk of invasion in SW Australia (RF) 53
Figure 33. Example of *P.cinnamomi* risk of invasion in Andalusia by SW Australia transfer RF model 53
Figure 34. Temperature variable importance comparison (right) and correlation between actual evapotranspiration and elevation (left) 59
Figure 35. NDVI standard deviation in summer. SW Australia.....61
Figure 36. Elevation response curve in Andalusia (right) and SW Australia (left) MaxEnt.....63
Figure 37. Elevation response curve RF model. Right, Andalusia; left, SW Australia.....64

List of Tables

Table 1. Environmental variables sources 13
Table 2. Model Techniques ensemble in biomod2 (Thuiller et al., 2013) 15
Table 3. Variables 20
Table 4. Final variables dataset 40
Table 5. Andalusian models accuracy 42
Table 6. SW Australian models accuracy (AUC and Kappa values) 43
Table 7. Andalusian models accuracy (AUC values) 45
Table 8. SW Australian models accuracy (AUC values).....45
Table 9. Andalusian models transferability result 50
Table 10. SW Australian models transferability results 52
Table 11. Variable importance summary.....58

Appendix

Table 12. Multivariate outlier detection.....79
Table 13. Boxplot variable Classification 79
Table 14. Analysis of collinearity results in Andalusia. 80
Table 15. Analysis of collinearity results in Southwest Australia.. 80
Table 16. Remnant common variables after collinearity analysis. 81
Table 17. Final variable datasets in the first test.....84
Table 18. Final variable datasets in the second test. 84
Table 19. Final variable datasets in the third test. 84
Table 20. Final variable datasets in the forth test 85
Table 21. Initial variables dataset in the fifth test 85
Table 22. Final variable datasets in the fifth test 86
Table 23. Final variable datasets in the final test 86
Table 24. Final 10 variable datasets in the final test.....86
Table 25. Andalusia models. Sensitivity and specificity 87
Table 26. SW Australian models. Sensitivity and specificity.....87

Chapter 1: Introduction

1.1 The Fungus

Phytophthora sp. is a fungal plant pathogen that affects a wide range of communities from crops to forest. The word *Phytophthora* comes from Greek and means "plant killer". One example of its relevance is that was the responsible of the Irish hunger in 1845. This genus contains numerous species one of which is *Phytophthora cinnamomi* (Zentmyer, 1988).

Phytophthora cinnamomi Rands is a soil-borne oomycete with a worldwide distribution. *P. cinnamomi* causes root rot, dieback and cankers in >3000 woody plants species, including eucalyptus, avocado, pine and oak (Hardham, 2005). Although its identification requires expert knowledge and it is costly, *P. cinnamomi* has been continental identified and isolated in the United States, Australia, South Africa and Europe. (Zentmyer, 1988).

Rands (1922) described *P. cinnamomi* from cinnamon trees (*Cinnamomum burmanni*) in Sumatra and suggested Asia as the origin of the fungus. Later, Zentmyer (1988) studied the fungi genetic variability and host plants resistance and considered South Asia, in particular Sumatra, as its centres of origin. From South Asia, the pathogen has spread worldwide.

Caprifoliaceae, Ericaceae, Rhamnaceae, Labiateae, Lauraceae, Cistaceae and Leguminosae have been identified as susceptible host families (Meentemeyer *et al.*, 2004; Moreira & Martins, 2005). The pathogen infection appears in plants weakened by drought or diseases. In addition, *P. cinnamomi* requires soil moisture and warm climate conditions to sporulate. Finally, it spreads by water, wind or infected tools, soils and root material, so an appropriate management may reduce and retard its expansion (Global Invasive Species Database, 2005; Reuter, 2005).

1.2 Problem Statement

European oak ecosystems have been affected by a severe decline and mortality during the last century, mainly caused by abiotic factors (frequent forest fires, severe drought, prolonged flooding, rapid fluctuation of soil water levels, cold winters, pollution and land use changes (Brasier, 1996; Gil Pelegrín *et al.*, 2008), but also by biotic factors (insects and fungus, (Carrasco *et al.*, 2009)).

The Oak Decline can be defined as a complex disorder or syndrome in which a wide variety of environmental factors and parasitic agents that are not capable to destroy the tree separately, interact in different combination in time and space, but produce similar symptoms, ending with the death of the tree (Brasier, 1996; Tuset *et al.*, 1996; Sánchez *et al.*, 2002; Gil Pelegrín *et al.*, 2008; Carrasco *et al.*, 2009). Two diverse symptoms could be observed: Sudden Death, expressed by a rapid drought of the crown, and Dieback characterized by a continuous decay and foliage loss (Gallego *et al.*, 1999).

In the Iberian Peninsula a general oak decline was detected at the beginning of the 1980s, mainly affecting Holm (*Quercus ilex*) and Cork oak (*Quercus suber*) (Brasier *et al.*, 1993; Brasier, 1996). However, the pathogen, *P. cinnamomi,* was not isolated from root or soil until 1991 (Tuset *et al.*, 1996). Besides, other studies related severe drought and *P. cinnamomi* as the main factors for the Oak Decline in Andalusia (Sánchez *et al.*, 2002; Gil Pelegrín *et al.*, 2008; Carrasco *et al.*, 2009). The effect of this syndrome was evaluated by Romero de los Reyes *et al.* (2007) using aerial photography of 67.292 hectares in Huelva, Andalusia. The decrease was estimated as 7.2% of canopy cover equivalent to 93.600 dead trees between 1997 and 2002. Therefore, Mediterranean oaks forests in Andalusia are currently at high risk of Oak Decline syndrome due to severe drought and *P. cinnamomi* affection (Camarero *et al.*, 2009). This is in addition to land use changes, fragmentation, overgrazing of livestock, abandoned land, increasing of wild fires, and other pest and diseases (Vogiatzakis *et al.*, 2006; Carrasco *et al.*, 2009).

Andalusian climate change scenarios forecast an average maximum and minimum temperature increase between about 0.3 – 1.5 °C per decade, an average annual precipitation increase of between 3 – 20% in the first third of the century and a decrease of between 7 – 20% for the rest of the term, and an increase of drought frequency, duration and intensity (Moreira, 2008). Furthermore, *P. cinnamomi* grows under a wide variety of hydrological circumstances and

temperature ranges, although its optimum is estimated at 30 °C and its spread through the root tissues is faster between 25 and 30 °C (Weste & Marks, 1987; Sánchez *et al.*, 2002). Consequently, *P. cinnamomi* activity may be enhancement by climate change conditions (Carrasco *et al.*, 2009).

On the other hand, plant communities in Southwest Australia have been affected by intense decline since 1921 when first *P. cinnamomi* symptoms were observed, although it was not associated with the fungus until 1964 (Weste & Marks, 1987; Hardham, 2005). The decline affected >75% of the flora in Jarrah forest (*Eucalyptus marginata*) which presented similar symptoms as Oak Decline in Europe, Sudden Death and Dieback (Weste & Marks, 1987; Dell *et al.*, 2005).

Jarrah Dieback as is *P. cinnamomi* affection known in SW Australia, is the main flora diversity threat in SW Australia (Shearer *et al.*, 2004). Its infection has been evaluated assessing direct and indirect impacts by Shearer *et al.* (2007) who pointed out the pathogen impacts such as, the weakness of endangered flora, the changes in canopy and ground cover and also the decrease in plants biodiversity in old infected areas. Furthermore, Shearer *et al.* (2012) estimated the impact on vegetation cover caused by the diseases and concluded that the loss of canopy and understory cover would affect temperature ranges and water budgets which may have a negative effect on endemic plant species.

The threat to natural ecosystems in Australia by *P. cinnamomi* invasion has been emphasized by its inclusion as a "Key Threatening Process" in the Commonwealth Environmental Protection & Biodiversity Conservation.

1.3 Species Distribution Models

Species Distribution Modelling (SDM) has acquired importance due to its faculty to forecast species occurrences from climate data, and its ability to predict species distributions in new areas or times (Elith *et al.*, 2006). Therefore SDM has become a relevant tool for biodiversity conservation and management (Guisan & Thuiller, 2005).

However, selecting a SDM approach is challenging, not only for the numerous techniques available but also for the different results yielded (Thuiller, 2004; Elith *et al.*, 2006). Furthermore, studies suggest combining several model techniques or ensemble models (Thuiller, 2003; Araújo & New, 2007).

SDMs have been widely used in ecological studies. However, biodiversity together with species spatial distribution could be the more common topic (Franklin, 2009; van Gils *et al.*, 2012). SDMs have estimate the distribution of a broad variety of organism as: butterflies (Beaumont & Hughes, 2002), fungus (Wollan *et al.*, 2008), plants (Thuiller, 2003; Benito Garzón *et al.*, 2008), mammals (Elith *et al.*, 2006), birds (Elith *et al.*, 2006), amphibious (Allouche *et al.*, 2006; Ficetola *et al.*, 2007), reptiles (Allouche *et al.*, 2006) and fishes (Buisson *et al.*, 2010). Moreover SDMs have been also used to predict potential species distribution (Benito Garzón *et al.*, 2003), or predict the impact of climate change in populations (Benito Garzón *et al.*, 2008), estimate species distribution in the past (Benito Garzón *et al.*, 2007), protection of endangered species (Benito De Pando *et al.*, 2007), predict endemic species occurrence (van Gils *et al.*, 2012), among other fields.

SDMs have been also used to predict the potential pattern of biological invasion, identify suitable areas and estimate the risk of invasion (Ficetola *et al.*, 2007; Kelly *et al.*, 2007). The estimation of invasive species distribution in a new ecosystem can be based on: suitable environmental conditions from its native area or by true locations in its new range (Franklin, 2009). Predicting the distribution of invasive species challenges SDMs. SDMs assume that the species are in equilibrium with their environment in comparison alien species could occupy a different environmental range. In addition, there are undefined suitable areas where the invader is likely to appear and are under high risk of invasion (Václavík & Meentemeyer, 2009).

Furthermore, the difference between potential and actual alien species distribution might be clarified. Potential spatial distribution agrees with the location where the invader is likely to appear

according to its environmental suitability. While actual distribution indicates areas where the invader is present in a certain time, limited under dispersal and environmental constrictions (Václavík & Meentemeyer, 2009). Normally, it is supposed that SDM predicts the potential distribution while it has been discussed that the models refers to actual distribution according to how models fits the calibration data (Franklin, 2009).

Alien species are organisms introduced in a new ecosystem which have succeeded in establishing, reproducing and expanding. Introduced species can break the ecological equilibrium causing a great economical and ecological impact in the native ecosystem (Colautti & MacIsaac, 2004). Lowe *et al.* (2000) registered and described 100 of the worst invasive species in the world, where *P. cinnamomi* was included. In addition, Sánchez *et al.* (2003) pointed out the problem of *P. cinnamomi* affection in Andalusian Oaks forest and Shearer *et al.* (2007) highlighted the incidence of *P. cinnamomi* invasion in Australian flora biodiversity.

SDM have mapped the "Sudden Oak Death" risk of invasion in California (Václavík *et al.*, 2010). This research used Multi-criteria evaluation and MaxEnt to predict the risk of infection and took host species index, precipitation, maxima and minimum temperature as predictive variables. The independent variables where weighted according to Meentemeyer *et al.* (2004) suggestions who predicted the "Sudden Oak Death" spatial distribution with GIS.

Phytophthora dieback distribution in SW Australia have been mapped using true locations and assessing infected areas (Weste & Marks, 1987; Shearer *et al.*, 2007) and similar studies have been done in Andalusia (Consejería de Medio Ambiente, 2010). However, to the best of our knowledge, no studies have mapped the risk of infection in Andalusia or SW Australia and compared both.

The spatial distribution models of invasive species are based on estimations of the native environmental range and apply these to the introduced location. This can be done by calibrating and validating a model in one location and applying the model on a different area. Model transferability allows this operation.

The spatial distribution of invasive species can be assessed by extrapolating predictions among occurrence areas; transferring models predictions throw areas. Although, extrapolate SMDs predictions are required in ecology conservation and management, model transferability have been poorly tested (Randin *et al.*, 2006;

Heikkinen *et al.*, 2012). Hence, a better knowledge of which models techniques performs more accurate extrapolations might be an improve on SDMs application (Franklin, 2009; Syphard & Franklin, 2009).

In this research we first compared the environmental distance between Southwest Australia and Andalusia (Spain) Secondly, *P. cinnamomi* SDMs were built to determine which environmental variable determine the fungus spatial distribution in each study area. Finally, we tested the transferability of SDMs between Andalusia and SW Australia and vice-versa.

The main objectives of this research were to test if *P. cinnamomi* spatial distribution is determined by similar environmental conditions in both study areas and evaluate model transferability between both study areas.

1.4 Research Questions

Q_1) Do have Andalusia and SW Australia similar environmental conditions?

Q_2) Which environmental variables determine the spatial distribution of the fungus in Andalusia and SW Australia?

Q_3) Are the spatial models accurately transferable from Andalusia to SW Australia and vice-versa?

Chapter 2: Study Area, Data and Methodology

2.1 Study Area

The Mediterranean-type ecosystem is located in regions characterized by long, hot and dry summers (at least 2 months of summer drought) and shorts, mild cold and wet winters (precipitation between 500 and 900 mm/year) (Lindner *et al.*, 2010). Such areas are around the Mediterranean Sea and equivalent areas at the West side of the continents at mid-latitudes and support the 20% of all plant species reported in the world (Figure 1) (Olson & Dinerstein, 2002). Indeed, the Mediterranean Basin occupy the 73% of the Mediterranean-type ecosystems total surface and it houses more than 25.000 species, among which more than half are endemic (Myers *et al.*, 2000).

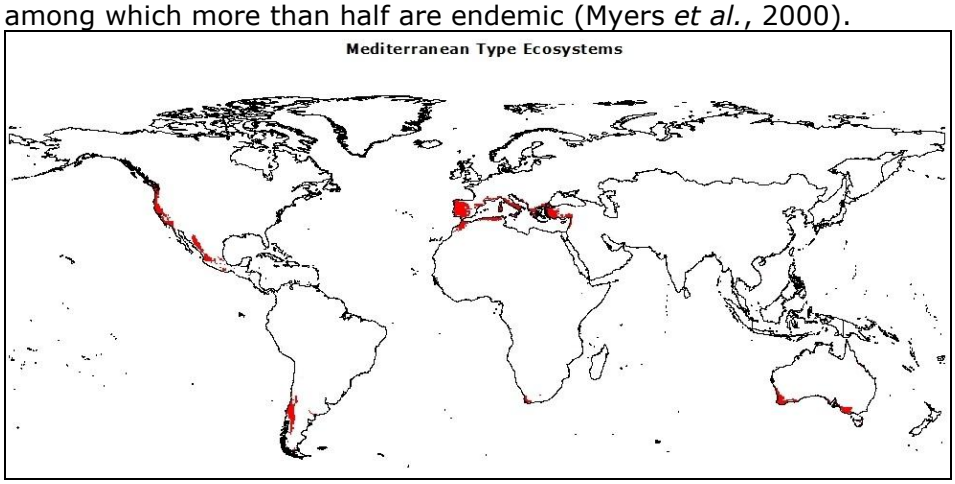

Figure 1. Worldwide Distribution of Mediterranean type ecosystems

2.1.1 Andalusia (Spain)

Andalusia is a characteristic Mediterranean Basin region of 87.268 km^2, located in South of the Iberian Peninsula, between 36° and 40° N latitude. Mediterranean oaks forests, woodlands and scrublands dominate this region with 4.6 million hectares, 51.5% of the total area, where 1.4 million hectares are populated by Oaks. Being, in this order, Holm and Cork oaks the most frequents mainly distributed in Western Andalusia (Figure 2) (Costa *et al.*, 2006).

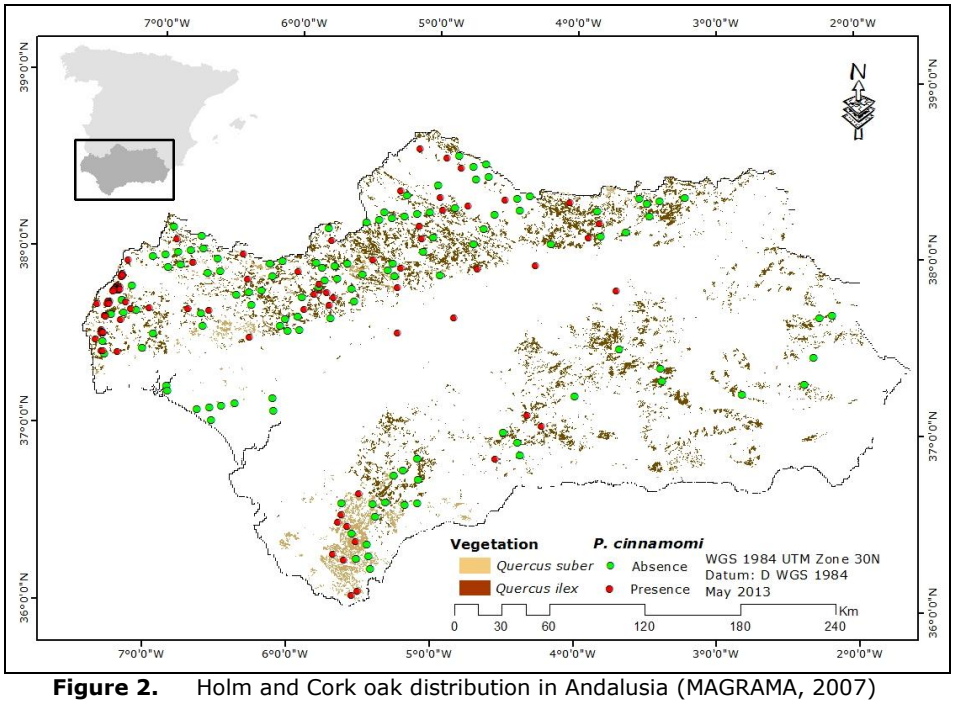

Figure 2. Holm and Cork oak distribution in Andalusia (MAGRAMA, 2007)

The value of this ecosystem has economic, (as community resources production), and ecological (rich in biodiversity) aspects, (Maranon *et al.*, 1999; Costa *et al.*, 2006). Mediterranean oaks forests contribute to the region economy with naturals and unique products such as cork, extensive livestock production, acorns, herbal and medicinal plants, and wildlife shelter (Olea & San Miguel-Ayanz, 2006). Additionally it provides social benefit, as carbon sink, ecotourism, prevention of soil erosion and desertification and creation of job opportunities and wealth (APCOR, 2010).

The ecological relevance of Mediterranean forests is protected in the 92/43/EEC Habitat Directive, included in the Nature 2000 Network (http://www.natura.org) and managed by the LIFE BioDehesa project

(Red Natura 2000, 2011). In addition, three more LIFE project are carried out to conserve endangered animal species on the IUCN list as the Spanish Imperial Eagle (*Aquila adalberti,* Vulnerable), the Iberian Lynx (*Lynx pardinus,* Critically Endangered), the Cinereous Vulture (*Aegypius monachus,* Near Threatened), and the Black Stork (*Ciconia nigra,* Least Concern), (IUCN, 2012). WWF (2012) mention that Cork oak forest contain 135 species of vascular plants per square meter, hold more than 30 different brackens and a rich diversity of fauna in land and numerous bird species nesting in trees. Moreover, Mediterranean biodiversity increases with human management (Díaz *et al.*, 2003; da Silva *et al.*, 2009; Bugalho *et al.*, 2011)

Andalusia meets the common aspect from a Mediterranean region combining (1) a marked geographical and topographical variability, (2) biseasonality climate, (3) high diversity of plant and animals (4) high degree of natural disturbances, including wild fires and (5) system exploitation often managed in a non-sustainable way (Scarascia-Mugnozza *et al.*, 2000; Vogiatzakis *et al.*, 2006). Because of the climate changes as well as the environmental and land use issues, Andalusia is a good area of study.

2.1.2 Southwest Australia.

The Australian study area was limited to Perth, Southwest Australia and the Western area of West South Costal, inside the Southwest Botanical region (Figure 3). The selection was done considering ecological value (high endemic biodiversity), climatic characteristics and area extent similar to Andalusia.

SW Australia is located in Mediterranean climate between 27º and 35º S latitude within a surface of 87.307 km^2 bounded by the Indian ocean on the west and south. The elevation increases from the coastal area to the highest point in the west limit at 782 masl.

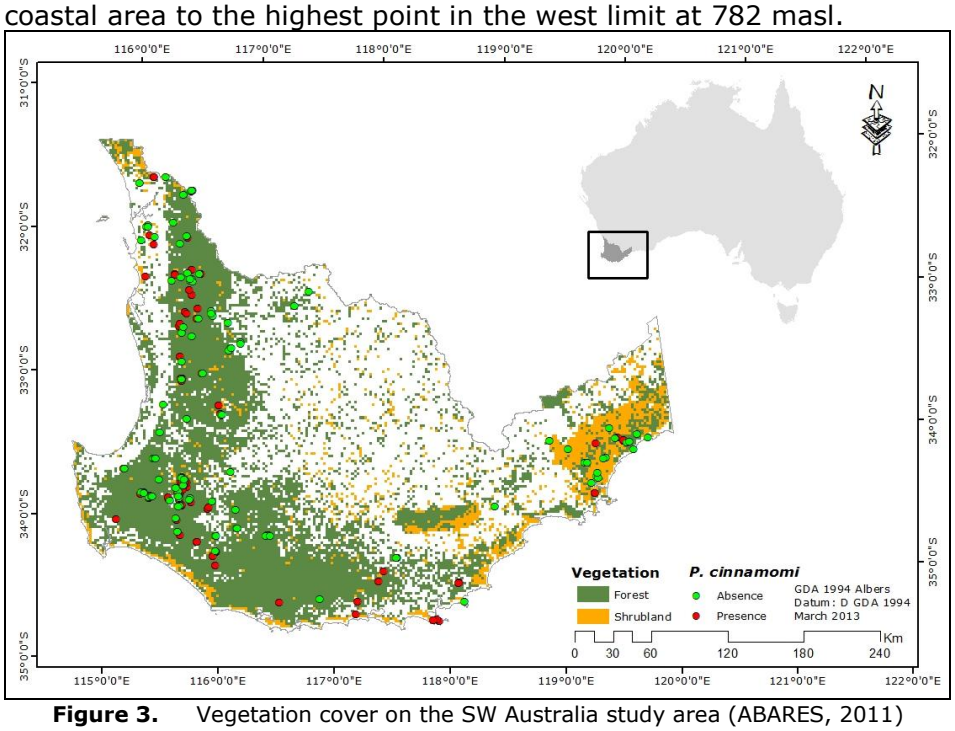

Figure 3. Vegetation cover on the SW Australia study area (ABARES, 2011)

Jarrah forest (*Eucalyptus marginata*) dominates the area (3.8 million hectares, 45% of the total area) together with woodlands, scrublands, heaths, and kwongan. (Western-Australian scrublands similar to Mediterranean maquis, Californian chaparral or South African fynbos) (Initiative & Gole, 2006).

This area host a spectacular biodiversity as a result of millions years of isolation. Three endemic Eucalyptus are the main tree species in the region, Jarrah (*Eucalyptus marginata*), Marri (*E. calophylla*), and Karri (*E. diversicolor*). In addition to hundred of endemic vertebrates, some on the IUCN list as: Carnaby's black-cockatoo (*Calyptorhynchus*

latirostris, Endangered), the Quokka (*Setonix brachyurus*, Vulnerable), the Gilbert's potoroo (*Potorous gilberti*, Critical Endangered) and the Western Swamp turtle (*Pseudemydura umbrina*, Critical Endangered), among others (IUCN, 2012).

Although, 11% of SW Australia region is protected, this area is severely at risk due to clearings of native vegetation for agriculture, bushes fires, introducing non-native species, mining and the root disease "Jarrah Dieback", causes by the fungus *Phytophthora cinnamomi* which affects >50% of rare flora (Initiative & Gole, 2006).

The ecological and economical value of this area lies in its biodiversity, where more than 3.500 species are endemic and on forest and agriculture uses with long anthropogenic influence (Initiative & Gole, 2006).

2.2 Data

2.2.1 *Phytophthora cinnamomi* Location

2.2.1.1 Andalusia (Spain)

P. cinnamomi presence-absence data was provided by Dr. Rafael María Navarro Cerrillo, Córdoba University. The data was obtained from the Andalusian Forest Monitoring Network and The Forest Phytosanitary Alert Network (Consejería de Medio Ambiente, 2010).

Sample points were located in 8x8 km vertexes network's, built according to the "Internacional Co-operative Programme on Assessment and Monitoring of Air Pollution Effects on Forests" (CEE-ICP Forest). Quercus' root and soil sample points were analysed in 133 locations from the Andalusian Forest Ecosystem Damage Monitoring Network and 34 locations from the Forestry Phytosanitary Alert Network. The Fungus identification and isolation procedure was done according to Jeffers and Martin (1986).

The Andalusian *P. cinnamomi* dataset consisted on 167 points with 47 presences and 120 absences. Presence data was increased by 48 presence point locations by literature review (Gomez-Aparicio *et al.*, 2012) and fieldwork based on previous studies and visual inspection. The final dataset was in World Geodetic System 1984 Universal Transverse Mercator Zone 30N Projected Coordinate System.

2.2.1.2 Southwest Australia

The Australian *P. cinnamomi* dataset included all samples from point locations processed in Vegetation Health Service`s laboratory between 1 July 2011 and 30 June 2012. Most samples were taken in natural ecosystems and a few were collected in nurseries and gardens. The data was Mike Stukely courtesy, Vegetation Health Service, Science Division, Department of Environment and Conservation, Western Australia (Stukely *et al.*, 2012).

This dataset contained 1.625 points with 552 presences and 1.073 absences and used Geodetic Coordinate System - Geocentric Datum of Australia 1994 Universal Transverse Mercator Zone 50 Projected Coordinate System. The original *P. cinnamomi* dataset was subsample to agree with Andalusian dataset on area extent and number of points.

The isolation and identification followed the process described in Burgess *et al.* (2009).

The initial difference in the number of *P. cinnamomi* point locations between Andalusia (167) and SW Australia (1.625) is due to their historical management of the problem with 30 years range (Weste & Marks, 1987; Tuset *et al.*, 1996; Dell *et al.*, 2005).

2.2.2 Environmental Data

Table 1 summarizes the sources where the environmental variables were downloaded.

Table 1. Environmental variables sources

Area	Source	Time Range
SW Australia	Bureau of Meteorology, 2013 ABARES, 2011	1976 - 2006
Andalusia	Junta de Andalucía, 2007 Universidad de Extremadura, 2012 MAGRAMA, 2013	
Worldwide	University of East Anglia Climatic Research Unit (CRU), 2008(1) FAO & IIASA, 2000(2)	
	Tucker *et al.*, 2004(3)	1981 - 2000

The Andalusian environmental dataset was pre-processed to 2 km spatial resolution and projected in World Geodetic System 1984 Universal Transverse Mercator Zone 30N Projected Coordinate System.

The dataset SW Australian dataset was pre-processed to agree in variables and resolution with Andalusian dataset. The Projected Coordinate System used was Geocentric Datum of Australia (GDA) Map Grid of Australia (MGA) 1994 Zone 50S.

The worldwide variables were clouds cover percentage and wet day frequency(1), Length growing period(2) and NDVI(3). Those variables were pre-processed to each study area extent (Table 1).

2.3 Methodology

2.3.1 Models

2.3.1.1 Species Distribution Models

The increasing concern over the effects of climate change to ecological conservation and biodiversity has led to the development of bioclimatic models which combine actual species occurrences with digital layers of environmental information and allow extrapolation in time and space (Elith *et al.*, 2006).

The variety of techniques accessible to model species distribution can be classified in three groups; (1) Profile techniques, which require presence-only data, environmental hype-space inhabited by a species methods as BIOCLIM, Surface Range Envelope (SRE), distance based methods as DOMAIN, Ecological Niche Factor Analysis (ENFA). (2) Discriminative techniques, which require presence-absence data, General Linear Model (GLM), General Additive Models (GAM), Multivariate Adaptive Regression Splines (MARS), Classification and Regression Tree Analysis (CTA), Boosted Regression Trees (BRT), Flexible Discriminant Analysis (FDA), Artificial Neural Network (ANN), Maximum Entropy (MaxEnt), Random Forest (RF), and (3) mix modelling approach which uses both techniques, Biomod, Generalized Regression Analysis and Spatial Prediction (GRASP), OpenModeller. Moreover, SDM can also be classified by their algorithms as: Regression methods as GAM, GLM and MARS; Machine-learning methods as ANN, BRT, MAXENT and RF; Classification methods as CTA and FDA; and Enveloping methods as SRE and BIOCLIM (Guisan & Thuiller, 2005; Elith *et al.*, 2006; Elith & Leathwick, 2009; Franklin, 2009).

SDM have described the spatial distribution of a broadly types of organism (Elith *et al.*, 2006; Václavík *et al.*, 2010). Likewise, multiple studies have compared models accuracy and performance (Benito Garzón *et al.*, 2006; Elith *et al.*, 2006; Mateo *et al.*, 2010; Gaston & Garcia-Vinas, 2011), though no superiority of any single one has been proved (Araújo & New, 2007).

To deal with model technique election, biomod2 R-package (Thuiller *et al.*, 2013), which include ten SDMs techniques, was used in this research (Thuiller, 2003; Phillips *et al.*, 2006; Phillips & Dudík, 2008; Thuiller *et al.*, 2009). Default settings of biomod2 (version 2.1.15) were used.

2.3.1.2 Statistical Models

The biomod2 R-package is a computer platform for ensemble SDMs, which works with presence-absence data and includes ANN, BRT, GLM, GAM, CTA, FDA, MARS, SRE, RF and also let to run MaxEnt (Table 2). The outputs are assessed by the goodness-of-fit, ANOVA and Akaike Information Criterion (AIC) are available for GLM and GAM, while rate of misclassification is used for CTA and RF. Model accuracy is measured by Area Under the Curve (AUC), Cohen's Kappa (κ) and True Skills Statistics (TSS) among others. Biomod2 also tests the influence of each variable in the model by a randomize procedure and displays a variable classification table. (Thuiller *et al.*, 2013).

Table 2. Model Techniques ensemble in biomod2 (Thuiller et al., 2013)

Model	**Data**	**Reference**
biomod2	S	Thuiller *et al.*, 2013
Artificial Neural Networks (ANN)	A	Lek & Guegan, 1999
Surface Range Envelope (SRE)	B	Busby, 1991
Boosting Regression Trees (BRT)	A	Elith *et al.*, 2008
Classification and Regression Trees (CTA)	A	Vayssieres *et al.*, 2000
Generalized Additive Models (GAM)	A	Guisan *et al.*, 2002
Generalized Linear Models (GLM)	A	Guisan *et al.*, 2002
Multivariate Adaptive Regression Splines (MARS)	A	Friedman, 1991
Flexible Discriminant Analysis (FDA)	A	Trevor *et al.*, 1994
Random Forest (RF)	A	Breiman, 2001
Maximum Entropy (MaxEnt)	B	Phillips *et al.*, 2006

(A: Absence, B: Background, S: Absences and Pseudo-Absences)

Artificial Neural Networks (ANN) is a machine-learning approach, widely used to deal with diverse problems (Franklin, 2009). Although the most frequent in ecology is the single layer perception, also named multi-layer feed-forward neural network. This "black box" technique estimates the species occurrence by connecting the known output (response variable) with the inputs (explanatory variables) by a middle step (hidden composite variables). The model establishes linear relation between the explanatory variables and the hidden composite variable which are non-linear related with the response variables. (Lek & Guegan, 1999; Franklin, 2009).

Surface Range Envelop (SRE) is a bioclimatic model similar to BIOCLIM. This model defines the climate range under which the species occurs, set of environmental variables where the species is present, and extrapolates the results to similar areas. This is the simple SDM technique (Busby, 1991; Beaumont & Hughes, 2002).

The Boosting Regression Tree (BRT) algorithm used in biomod2 was described in Ridgeway (1999) and implemented by Friedman (2001). BRT estimates the species occurrence by fitting numerous single models whose predictions are later ensemble to build a more robust prediction. Each single model is a simple regression or regression tree, i.e. an iterative data partitioning in homogenous groups base on the response. BRT performs a recursive method to build a final model by adding trees, reclassifying the data to highlight poor results by the previous tree model (Elith *et al.*, 2008).

Classification and Regression Trees (CTA) method is based on successive data partitions according to predictions into homogeneous groups in term of the response. The tree is done by a recursive data splitting based on a single explanatory variable. Each data division reduce the variance within the subset. The best CTA model is a mid way model between the highest variance decrease and the lower number of singles model (Vayssieres *et al.*, 2000).

Generalized Additive Models (GAM) estimates the species occurrences by fitting a response curve call "smoothers" which try to adjust the data into the curve by local fitting to data subsamples. GAM estimates more accurately complex relationships between the variables than linear models. The model fits a single smooth to each variable a then the results are additively combined (Guisan *et al.*, 2002).

Generalized Linear Models (GLM) is based on fitting a linear relationship between the independent and dependent data. GLM use linear, quadratic or polynomial functions to estimate species occurrence. The model selection is done by a stepwise procedure under Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC) which delete redundant variables and decrease collinearity (Guisan *et al.*, 2002).

Multivariate Adaptive Regression Splines (MARS) is a linear type method which gives different models coefficient according to the optimal values across each level of the explanatory variables. The threshold which indicates a modification in the model coefficient are called "Knots" which are defined automatically. It presents similarities with CTA where the data partitioning is replaced by piecewise linear functions and the reduction of the final model complexity is done by deleting not relevant basic functions. Moreover, it is also close to GAM due to use piecewise splines. MARS advantages are that: considers local variables iterations, supports large number of

explanatory variables and performs faster than GAM (Friedman, 1991).

Flexible Discriminant Analysis (FDA) use the MARS function to perform a flexible discriminant analysis for the regression part of the model. FDA is a supervise classification method which combines different models (Trevor *et al.*, 1994).

Random Forest (RF) builds many de-correlated classification trees and averages them. It constructs the same number of classification trees as data in the training set are, randomly with replacements, with a subset of explanatory variables. Each individual tree is validated with the non-used subset of the data and returns the averages predictions of all trees. Variable selection is done by rate of misclassification, for categorical outcome or mean squared error. The difference in errors, between the prediction and the values calculated by procedure of variable randomization is reflected in the weight of the predictive variable (Breiman, 2001).

MaxEnt is a machine-learning method that estimates the species distribution probability by assessing the maximum entropy distribution, so that the most spread-out, or closest to uniform. MaxEnt is performed with presence only data, though requires background points. It also gives variable comparison and test model accuracy by AUC (Phillips *et al.*, 2006; Phillips & Dudík, 2008).

2.3.1.3 Model Selection

Model selection is finding a single model with most influential predictive variables among the available (Johnson & Omland, 2004). This process is also called variable importance. This operation aims to indentify which variables are related with the prediction and identify the smaller number of variables to perform a good prediction (Gromping, 2009). Moreover, it can be done manually or systematically by backward elimination, forward selection or stepwise procedures (Franklin, 2009).

Some SDMs use Akaike Information criteria (AIC) or Bayesian Information criteria (BIC) as is implemented in GLM or GAM. AIC measures the goodness of fit of the model to the data, lower AIC better model, and reduces model complexity (Elith & Leathwick, 2009). In addition, the SDMs techniques ensemble in biomod2 had their own model selection criteria (Thuiller *et al.*, 2013).

2.3.1.4 Model Evaluation

Model evaluation calculates the model's predictive performance, based on the error evaluation and quantifying the ones classified incorrectly. There are two types of errors: commission error, which classifies an absence as present and omission error which defines presences as absence. The final model evaluation and the comparison between techniques are done by statistics which assess the discrimination capacity of the model. The optimal evaluation is done with an independent dataset to calibrate the model. However, this rarely occurs and researches applied other techniques as data partitioning or cross – validation (Franklin, 2009).

Data partitioning consist of dividing the dependent data in two sets, one to calibrate the model, training data, and other to evaluate the model, testing/validation data. The optimal data partitioning range depend on the number of predictions (Franklin, 2009); in this case 70/30 was selected to train and test the models respectivility. This range was also proposed in other studies (Thuiller, 2003; Thuiller *et al.*, 2003; Thuiller, 2004; Heikkinen *et al.*, 2012).

Data partitioning is a simple case of cross-validation where the dependent data is divided in two sets. Cross – validation consists in divide the dependent data in multiple sub-sets each one with the same number of cases. Later the model runs the same number of times as sub-sets are. Each time a different sub-set is used to test the model performance while the rest are used as training data.

Model evaluation tests the model predictive performance. The Area Under Curve of the Receiver Operating Characteristic plot (AUC) statistic was used to estimate model accuracy with presence-absence data. Moreover, Cohen's Kappa (κ) was calculated to estimate the map veracity.

AUC is a threshold independent statistic measure which represents the model's goodness of fit to the data. AUC represents graphically the model discriminative capacity. AUC plots the commission error (1 – sensitivity) in the horizontal axis, vs. omission error (sensibility) in the vertical axis. It ranges between 0.5 – 1, where 1 represents a prefect classification and 0.5 a random classification. Using AUC as evaluation metric has several advantages: It is possible to compare all SDMs, it is prevalence and threshold independent (Franklin, 2009). In contrast, requires a minimum number of presences, it does not differentiate between omission and commission errors and gives the same importance to all points across the region while the interesting area is the right top corner of the plot (Lobo *et al.*, 2008).

Kappa is statistical measures of qualitative agreement between categorical predictions. K is a threshold dependent classifier which describes the difference between observation agreement and random agreement. Kappa is calculated from the confusion matrix where both the omission and confusion errors are considered (Franklin, 2009). K was calculated using the maximum threshold for all different model techniques.

2.3.1.5 Model Transferability

SDMs use point locations in a specific region to estimate the probable species spatial distribution across the region. In these situation areas without information about the species occurrence are estimated from the known location (Wenger & Olden, 2012). In the same way, models could be used to predict the species occurrence in regions geographically apart, what is called transferability. Model transferability aims to accurately predict the species occurrence in one region from model developed in a different region. It is similar to forecast species occurrences in time but in space (Randin *et al.*, 2006).

In this research, cross-validation was used to train and test the models in each region separately and later model transferability was tested across regions.

2.3.1.6 Model Comparison

Upon the different models techniques used, the more accurate ones were elected by AUC values comparison. It was estimated that models with AUC >0.85 had a strong predictive performance.

2.3.2 Research Procedure

2.3.2.1 Variable Pre-Processing

Two equal environmental datasets in raster format each with 93 variables were prepared (Table 3). The environmental variables were selected according to *P. cinnamomi* requirements in addition to descriptive variables that could suggest the spreading pattern (Weste & Marks, 1987; Václavík *et al.*, 2010). The monthly climatic variables were composited to seasonal and annual averages. Moreover, a visual inspection of variables was done to verify their credibility. Pre-processing of environmental layers was carried out in ArcGIS 10.1.

Table 3. Variables

Variables	Description	Unit
aspect	Orientation	Degrees
asrad	Area solar radiation	Radiation/m^2
cld_xxx	Cloud cover	%
dist_road	Distance to main roads	m
dist_water	Distance to permanent rivers	m
elevation	Altitude from sea level	m
etp_xxx	Potential evapotranspiration	m
etr_xxx	Actual evapotranspiration	m
farmlands	Cropland areas	Dummy
flowdir	Flow direction	
forest	Forest areas	Dummy
frs_xxx	Frost day frequency	Days
grasslands	Grasslands areas	Dummy
lgp	Length growing period	Month
maxt_xxx	Mean maximum temperature	ºC
meant_xxx	Mean temperature	ºC
mint_xxx	Mean minimum temperature	ºC
mr_vbf	Landform valley bottom flatness	Dummy
nd_rain_xxx	Number of rainy days	Days
ndvi_av_xxx	Average NDVI	
ndvi_sd_xxx	SD NDVI	
rain_xxx	Mean precipitation	mm
ro_xxx	Run off	mm
scrublands	Scrublands areas	Dummy
sink	Sink areas	Dummy
slope	Slope	%

Variables	Description	Unit
slope_length	Slope length	m
sm_xxx	Soil moisture	m
sparse	Sparse areas	Dummy
substr	Substrate (0 = acid; 1 = basic)	Dummy
sun_xxx	Mean daily sun shine	hx10
urban	Urban areas	Dummy
wet_xxx	Wet day frequency (>1mm)	Days

(The postfix "xxx" in the climatic variables refers to annual (ann) and seasonal composite (aut, spr, sum and win)

2.3.2.2 Study Area Comparison

The study area comparison was performed by an analysis of multivariate outlier detection and visual inspection of boxplots (Figure 4). Squared Mahalanobis Distance (SMD) was calculated between both environmental variables dataset in order to assess 1) if both areas had similar environmental conditions and indentified outliers variables (De Maesschalck *et al.*, 2000).

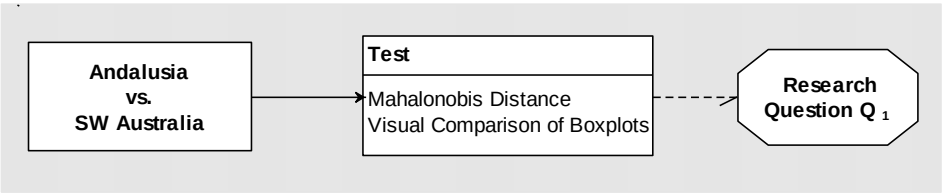

Figure 4. Study Areas comparison flowchart

Mahalanobis Distance (MD) measures the similarity between two variables based on the differences between a single variable value and the variable mean.

MD calculates the distance between the centroids of multivariable dataset and considers the correlation within the data by including the covariance matrix of the target dataset. Therefore, similar variables should have a MD close to the centroid value and lie on a ellipse or ellipsoid distribution (Farber & Kadmon, 2003). However, outliers can be confused with values in the extreme of the distribution. In order to recognise outliers variables, SMD was compared with the chi-squared distribution (χ_p^2) (Filzmoser, 2004). Moreover, in multivariable outlier detection data standardization was required to bring all variables to the same spatial scale.

For A and B two p-dimensional samples A_i ($i = a_{i1}, a_{i2}, \ldots, a_{in}$) and B_i ($i = b_{i1}, b_{i2}, \ldots, b_{in}$) where $\bar{u}_{Ai}$ ($i= \mu_{ai1}, \mu_{ai2}, \ldots, \mu_{ain}$) and $\bar{u}_{Bi}$ ($i= \mu_{bi1}, \mu_{bi2}, \ldots, \mu_{bin}$) are the standardize means. The SMD is mathematically defined as:

$$D^2 = (\bar{u}_{Ai} - \mu)^T C^{-1} (\bar{u}_{Bi} - \mu) \qquad \text{for } i = 1, 2, \ldots, n$$

Where D^2 is the SMD, μ is the arithmetic mean and C is the covariance matrix. Variables with large SMD values could be considered as outliers (Farber & Kadmon, 2003; Filzmoser, 2004). The outlaying variables were indentified and deleted in a step-by–step procedure until the (χ^2_p) distribution requirements were fulfilled ($SMD<P(\chi^2_2)$) (Filzmoser, 2004). Confidence level of 95% and p-values significance with p=0.05 were selected for statistical analysis for the degree of freedom (See Appendix I, Table 12).

Moreover, variables were compared by boxplots and identified those with anomalous ranges (See Appendix I, Table 13). Boxplots defines variable location by the median and the spread by the distance between the edges of the box, hinges. The variability of the variable is determined by the distance between the whiskers and outliers are identified by points outside the whiskers. The hinges are at 25% quartiles from the mean while the whiskers are at 50% quartiles from the mean. Outliers are outside this range (Quinn & Keough, 2002).

2.3.2.3 Variable Selection

The original 93 variables (Table 3) were reduced until the final dataset by analysis of collinearity (Guisan & Thuiller, 2005; Franklin, 2009), variable importance (model selection) stepwise procedure inherent to each model technique (Elith & Leathwick, 2009; Franklin, 2009; Thuiller *et al.*, 2013), and backwards elimination according to biomod2 outputs (Figure 5) (Thuiller *et al.*, 2013).

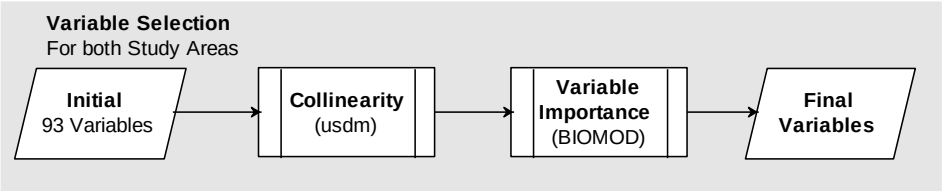

Figure 5. Variable selection general process flowchart

2.3.2.4 Analysis of Multicollinearity

Collinearity (Multicollinearity) is a statistical issue which indicate the strong correlation between two or more descriptive variables and induces uncertainty in regression models predictions. Collinearity refers to a linear relationship between two predicts variables, while multicollinearity refers to collinearity between two or more predicts variables. Collinearity affects the estimation of the regression coefficients and induces bias responses between outputs and explanatory variables. Collinearity can be detected by 1) analysis of correlation matrix and 2) Variance Inflation Factor (VIF) which is calculated as:

$$VIF = \frac{1}{1 - R_j^2}$$ Where R_j^2; it is the coefficient of determination.

The uncertainty analysis was performed in R (R Core Team, 2012) using the "usdm" R-package (Naimi, 2013). We calculated the correlation coefficient and the Variance Inflation Factor (VIF). The analysis of collinearity was done within the 93 original variables in each study area (Figure 6). Variables with R^2>0.90 and VIF>10 performed a poorly estimation of the correlation coefficient due to collinearity and were deleted (Graham, 2003; Heikkinen *et al.*, 2006). We found 46 common non collinear variables (See Appendix I, Tables 14, 15 & 16).

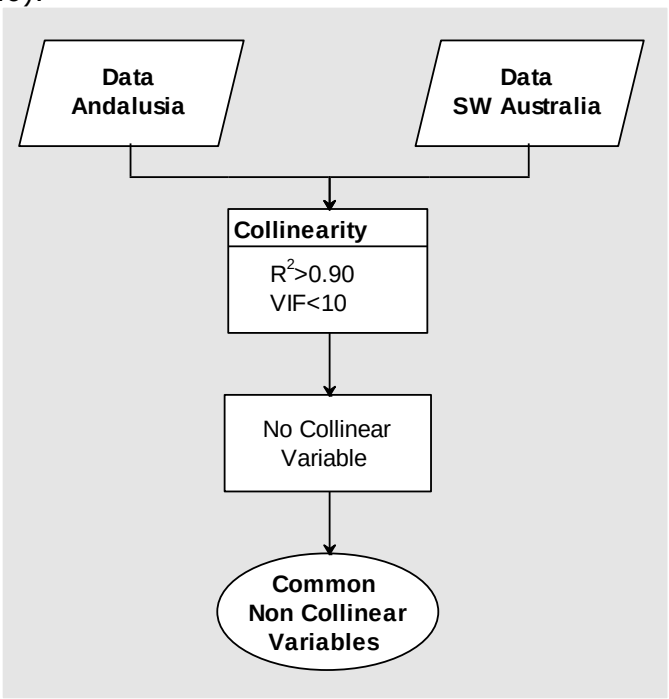

Figure 6. Diagram of collinearity analysis

2.3.2.5 Variable Importance (Model Selection) Analysis

A backwards elimination procedure according to the variable importance function in biomod2 R-package (Thuiller *et al.*, 2013) was done. Moreover, each model techniques performed itself a stepwise variable selection from the available variable dataset. The process was tedious and time consuming due to the large dataset and the numerous test performed (Figure 7). We did 4 replicates per each study area running the 10 model techniques 50 times each deleting variables one by one until there were 10 variables remaining (Figure 8). In addition, we performed a fifth test where we took as initial dataset all the variables resultant from the 4 previous replicates and followed the same procedure until there were 10 variables remaining (Figure 9). We decided to stop at 10 variables randomly in order to ensure a final number of common variables between both areas.
Finally, we did two final tests; one with an independent variables dataset, called final test, (Figure 10) and the other with a common variable dataset, called transfer test (Figure 11).

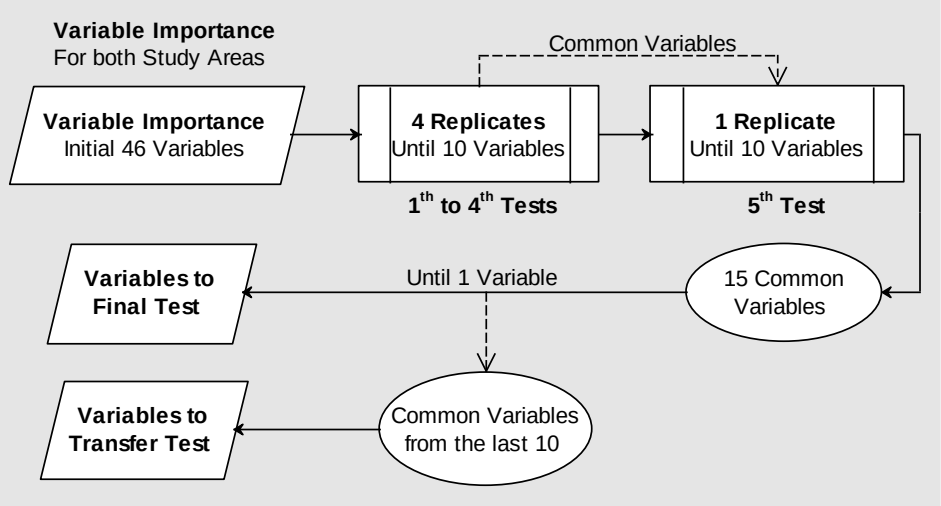

Figure 7. Flowchart summarizing the variables selection procedure

First Test

We performed a multi-model variable selection, from the initial 46 common non collinear variables, deleting in each run the less relevant variable, that with lower correlation coefficient in the variable importance table of biomod2 outputs, from the more accurate model, the one with higher AUC (Figure 7 & 8).

The first test highlighted three models techniques as the more accurate (highest AUC). Therefore, we based our next variables selection procedure according to the variable importance output from one single model each time. The selected models were BRT, MaxEnt and RF (See Appendix I, Table 17).

Second, Third and Fourth Test

The next tests were performed following the same methodology as in the first one but following the variable importance classification from a single statistical model, BRT, MaxEnt and RF, each time (Figure 8) (See Appendix I, Tables 18, 19 & 20).

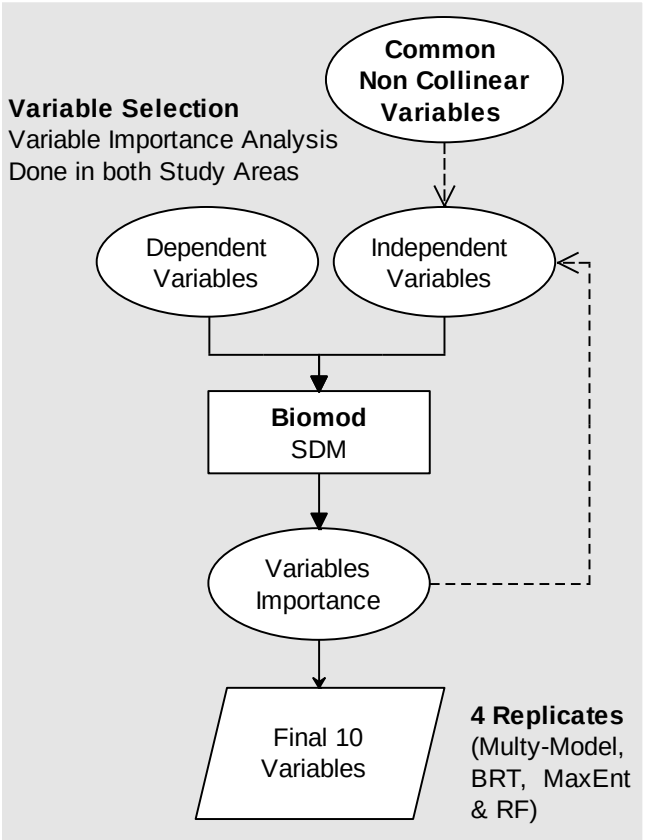

Figure 8. Flowchart summarizing the first 4 runs of variables selection analysis

Fifth Test

The fifth test was done using as initial variable dataset those variables in common founded in the 1st to 4th tests. Variable selection was done according to MaxEnt variable importance classification (Figure 9) (See Appendix I, Table 21 & 22).

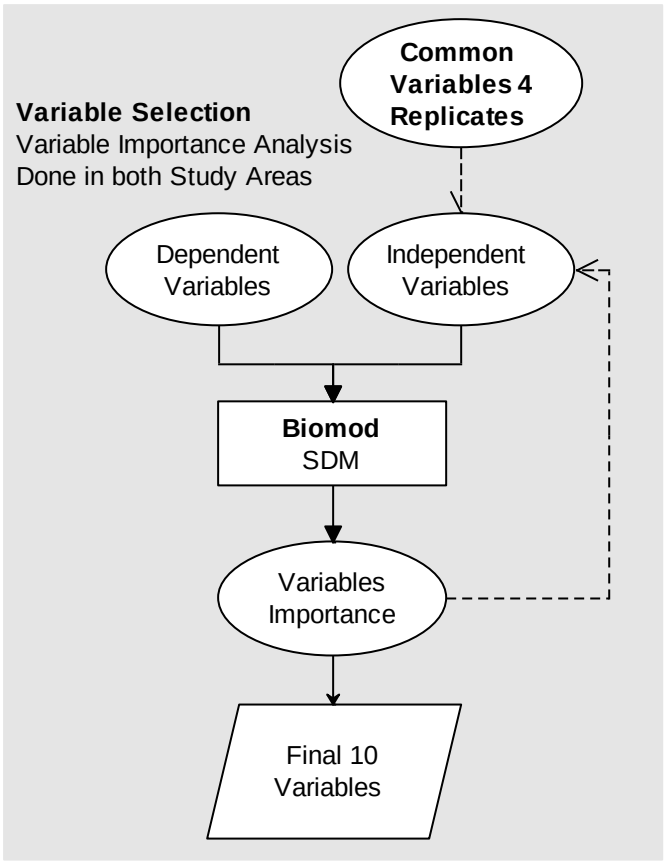

Figure 9. Methodology followed in fifth variable selection test

Final Test

The final variable selection test was based on the common final variables resulted from the fifth test, 15 environmental variables (See Appendix I, Tables 23 & 24). During the step-by-step variable deleting process, standard deviation, maximum and minimum AUC values were calculated in order to assess how the number of variables influenced the model accuracy. This analysis was repeated until one variable remained (Figure 10).

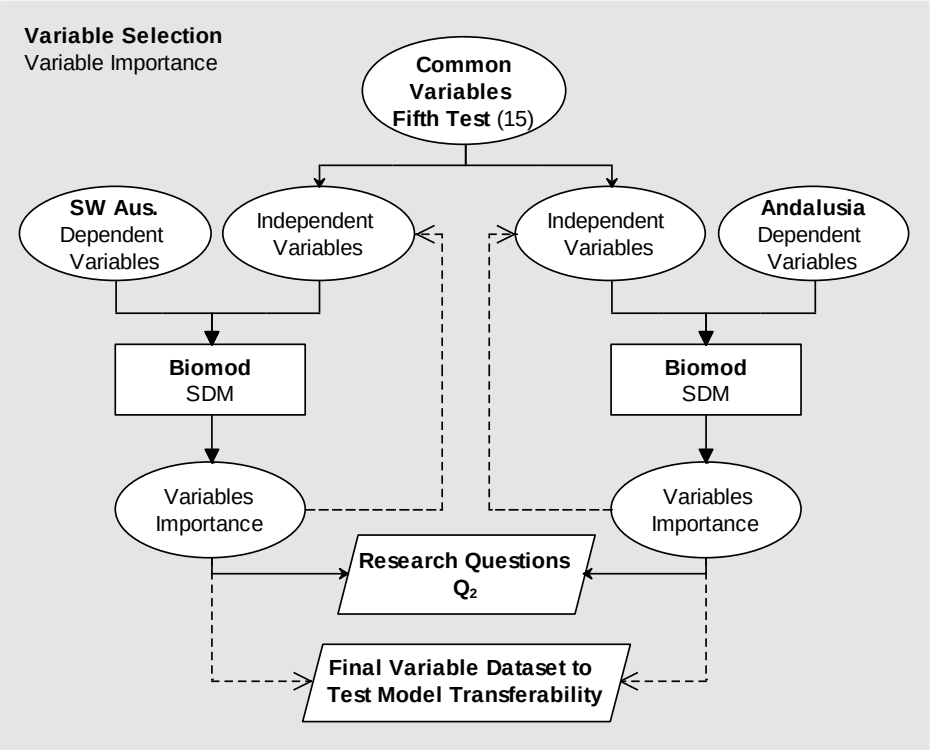

Figure 10. Flowchart summarizing the final variable selection step

Moreover, we did an intermediate variable selection which included the 10 most relevant variables in each study areas (10 variables each). Both dataset were compared and the common ones were elected to perform the transfer test (Figure 11) (See Appendix I, Table 24).

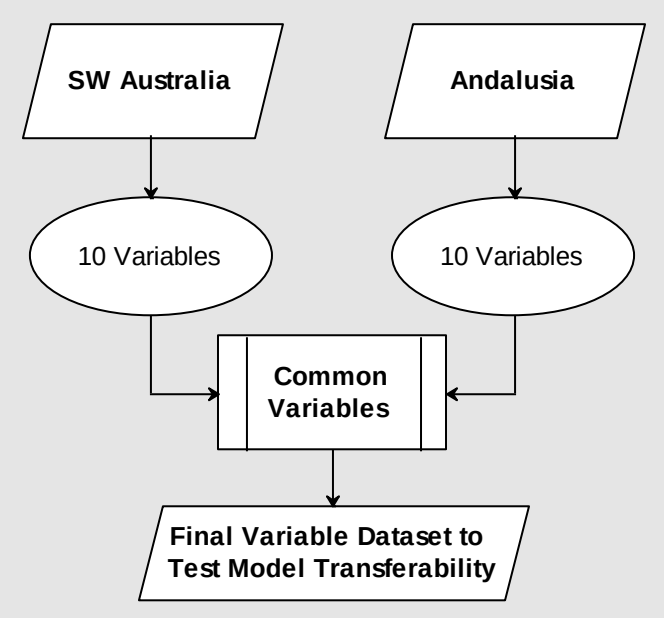

Figure 11. Description of the methodology followed to the common relevant variables dataset

Independent Test

The independent test was done with the more relevant variables in each study area, 5 for Andalusia and 7 for SW Australia. This time the model runs 300 with 10 different techniques with each variable set. In this analysis it was calculated the standard deviation, maximum, mean and minimum AUC values for each model technique (Figure 12).

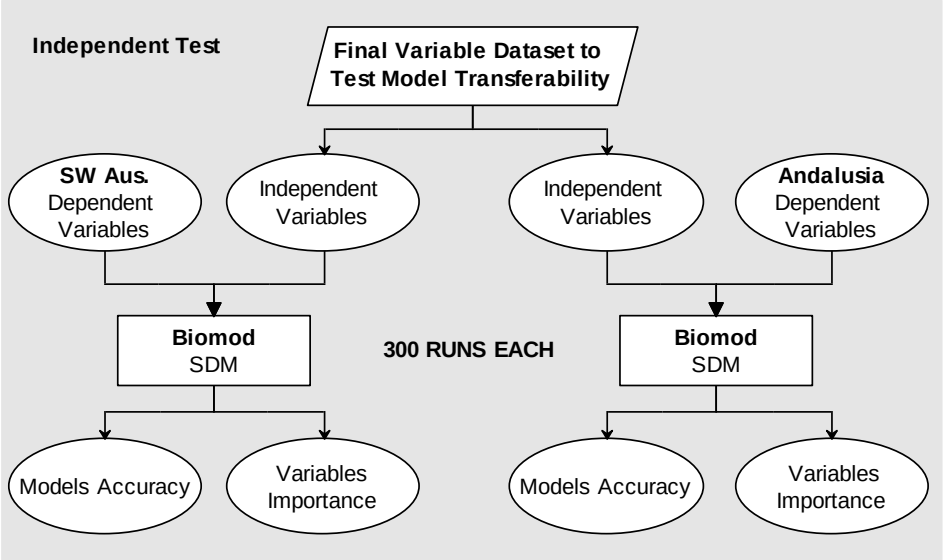

Figure 12. Independent test procedure diagram. (SW Aus: SW Australia)

2.3.2.5.1 Ranking Variable Importance

The variable importance function in biomod2 described in section 2.3.1.2 give a good overview of the variable influence in the model but meaningless to compare between models. This lack of similarity of the variable importance classification led us to develop a ranking system so that we could compare importance variables between the different models.

For all models we ranked the environmental variables from 1 to "n". Being "n" the number of variables presented in the analysis. The rank 1 was given to the most important variable. If there were two variables with the same importance we ranked both of them with the same number. Similar approach was used by Syphard and Franklin (2009).

Later, to avoid model influence in the variable importance ranking we sorted the ranked variable importance into one single classification by mean and mode. The variable importance ranking was done considering those models with the highest AUC values per each model technique.

2.3.2.6 Model Transferability

The models with highest AUCs values from each technique were transferred. Model transferability was done according to the common environmental dataset selected for the transfer test. Transferability was carried out in biomod2 (Thuiller *et al.*, 2013) by projecting models in the "new environment". Therefore, model trained and validated in SW Australia were projected in Andalusia and vice versa. Transfer predictions were validated using the complete set of point locations available in the new environment (Figure 13). In SW Australia model transferability was validated with same subsample that calibrated and validated the models initially. Model transferability projections were validated by the "PresenceAbsence" R-package (Freeman & Moisen, 2008).

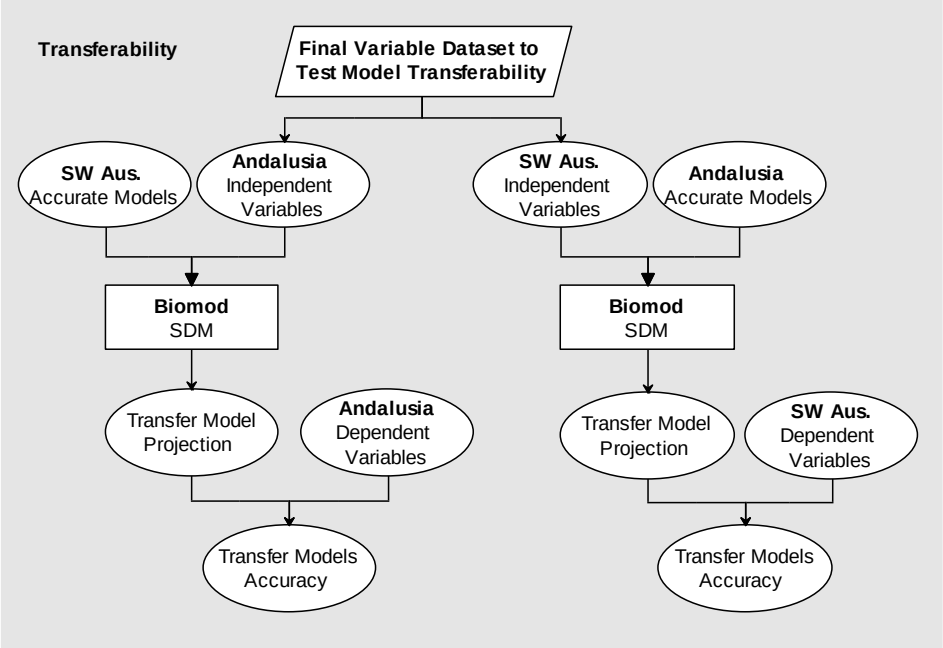

Figure 13. Transferability flowchart description. (SW Aus: SW Australia)

Chapter 3: Results

3.1 Study Areas Comparison

3.1.1 Multivariate Outlier Detection

The analysis of multivariate outlier detection revealed that considering the complete environmental variables datasets, both study areas did not present similar environmental conditions under a confidence level of 95.00% according to the chi-squared distribution test. After the outlier variables were indentified and deleted, 29 (See Appendix I, Table 12). The remaining dataset, without the 29 outlier variables, fulfil the SDM<P((χ_2^2)) requirements, therefore both areas had similar environmental conditions (Figure 14).

In plot (14a) points closer to the centroid represent variables with similar means. Points inside the ellipses are at 1 and 2 MD time radius from the centroid respectively. Points outside the ellipses could be considered as outliers, due to large difference in their means.

Plot (14b) highlights variables that don't follow a chi-squared distribution according to the elected confident level, those with larger SMD. Points above the red line represent variables which follows a chi-squared distribution, those with lower SMD.

Plot (14c) presents the variables distribution after deleting outliers, points above the line follows a chi-squared distribution. Point apart from the line could be considered outliers. The point surrounding by the red circle is the one with highest SMD, although according with the selected confidence level was rejected as an outliers. The point represented the variable sun shine in autumn.

Plot (14d) shows the final SMD density distribution, without outliers, where the red line indicates the variable with highest SMD value (SMD = 5.49), the one in the edge. The final variable distribution without outliers had a bell-shape normal distributed.

Any variable was deleted in this step due to the principal aim was to verified if both areas were environmentally similar and detected those that could be problematic to perform transferability.

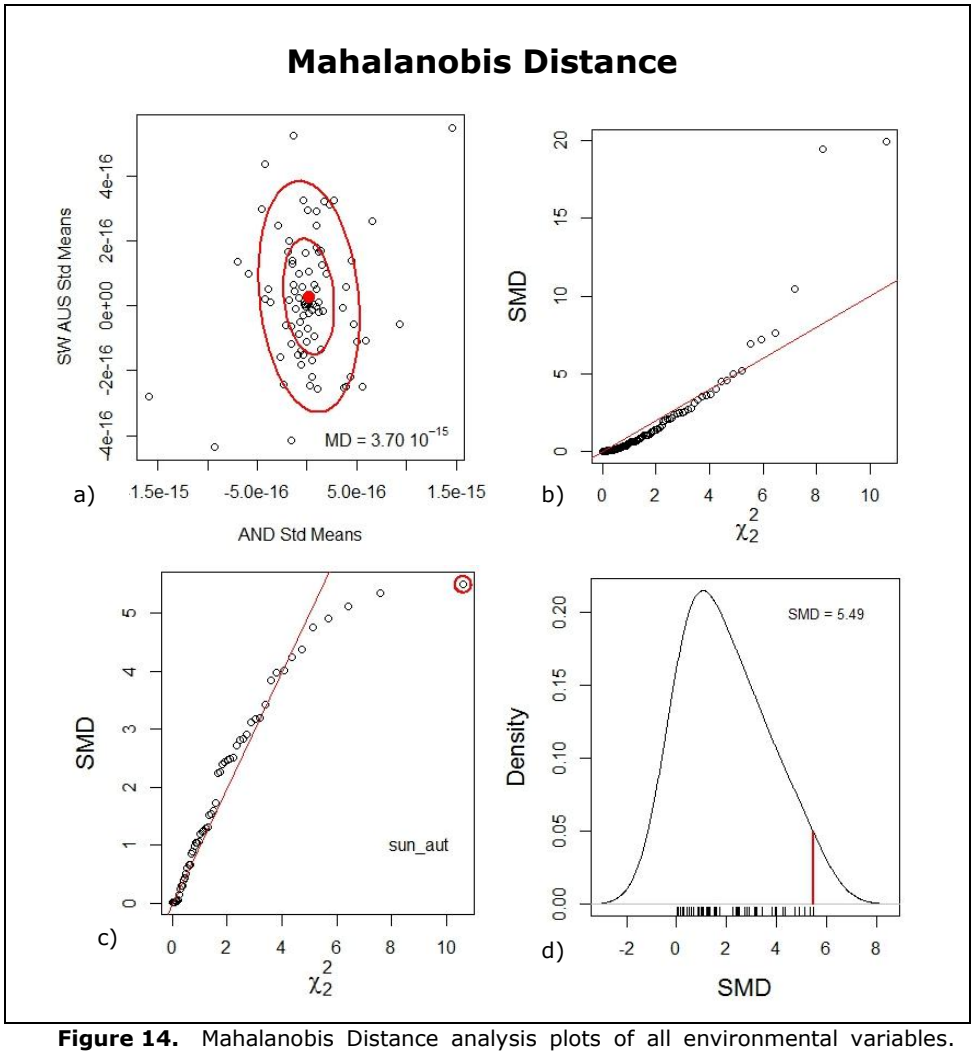

Figure 14. Mahalanobis Distance analysis plots of all environmental variables. (SW AUS: Southwest Australia; AND: Andalusia; sun_aut: sun shine in autumn)

Plot (a) displays the standardized mean of variables values location between both study areas. It represents Andalusia in the horizontal axis versus SW Australia in the vertical axis in respect to the Mahalanobis centroid position, (red point). Plot (b) shows Q-Q plot of SMD vs. quantiles of chi-squared initial situation, including outliers, where in the horizontal axis is chi-squared distribution and in the vertical axis is the SMD. Plot (c) presents Q-Q plot of SMD vs. quantiles of chi-squared without outliers. Plot (d) density plot of squared SMD where in the horizontal axis are the SMD and in the vertical axis the SMD frequency respect to the total.

3.1.2 Boxplots

Examples of boxplots are shown in Figure 15.

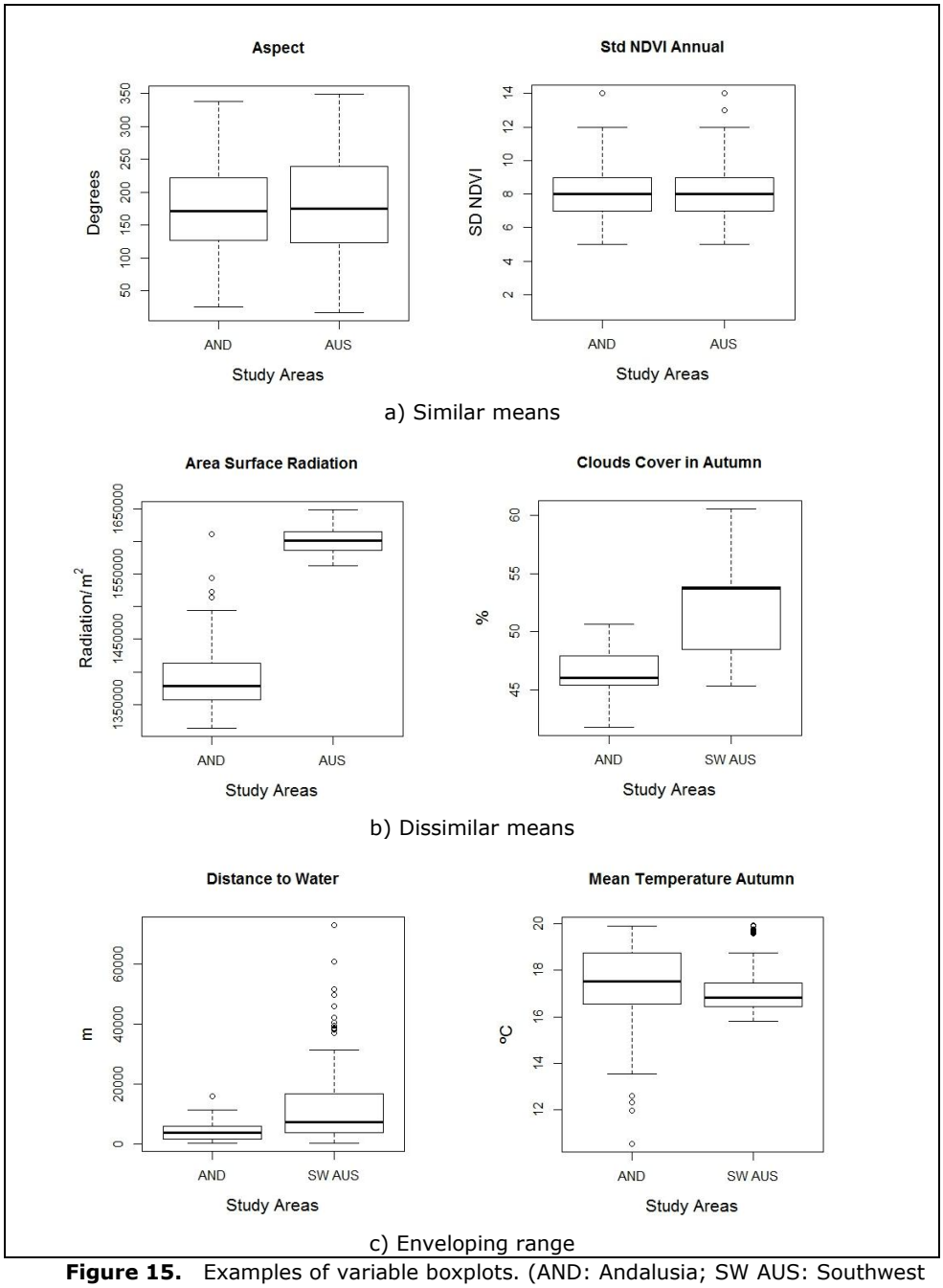

Figure 15. Examples of variable boxplots. (AND: Andalusia; SW AUS: Southwest Australia)

The boxplots analysis pointed out similarities and difference between variables. Among the numerous plots done, we recognised three mains patterns: similar means, dissimilar means and enveloping range. As similar mean were considered those variables with closer means and ranges. As dissimilar means those variables with different means and ranges; and enveloping range those variables that included the other area range in theirs. Boxplots visual inspection confirmed the results obtained in the analysis of multivariate outlier detection (Figure 15) (See Appendix I, Table 13).

3.2 Variable Selection

3.2.1 Analysis of Collinearity

The Variance Inflation Factor (VIF) and the correlation coefficient between variables were calculated separately in each study area. We found 27 and 32 variables with collinearity problems in SW Australia and Andalusia respectively. Therefore, there were 66 and 61 variables remaining of which 46 were common in both datasets. Those 46 variables were sorted in the next step. (See Appendix I, Table 16)

3.2.2 Model Selection

The model selection test showed that the number of variables did not influence in the model performance in terms of the highest AUC. The maximum and minimum AUC remained constant across number of variables, model techniques and study areas. However, in Andalusia the minimum AUC value in GAM, GLM, MaxEnt and RF decreased with <1 variables remaining (Figure 16), while in SW Australia, this fact was observed with <4 (Figure 17).

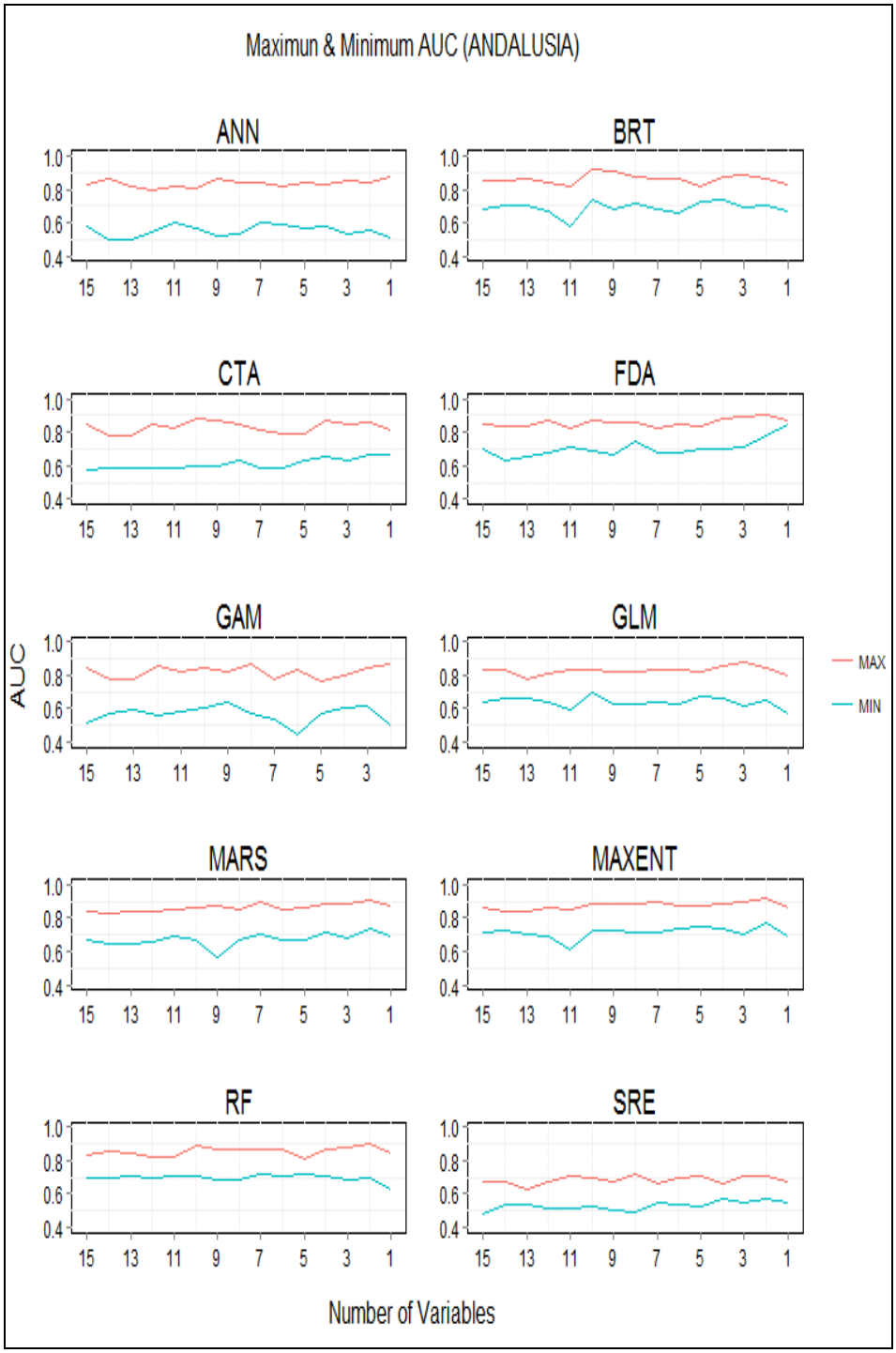

Figure 16. Maximum and minimum AUC variability across the number of variables and model techniques in Andalusia

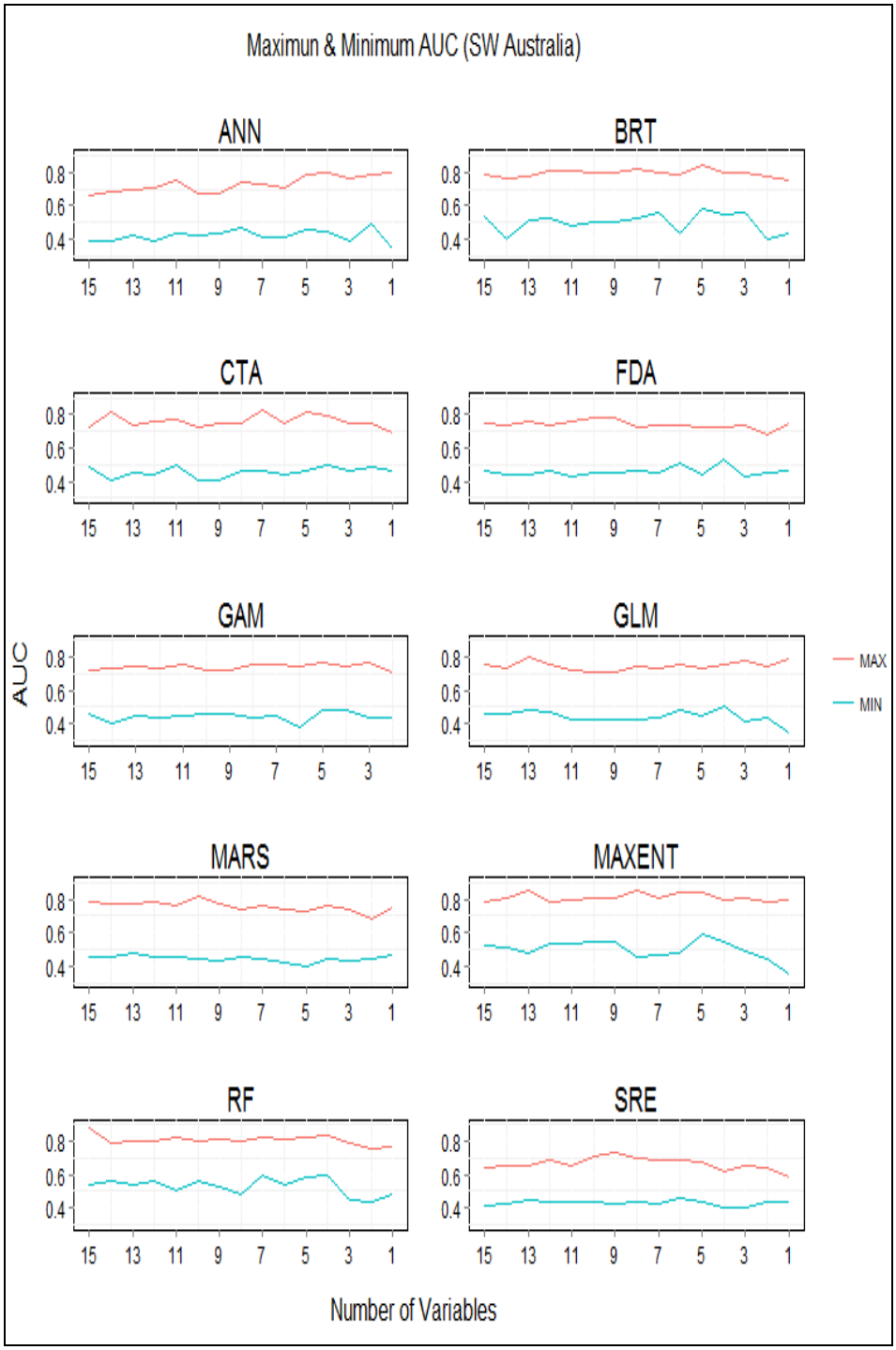

Figure 17. Maximum and minimum AUC variability across the number of variables and model techniques in SW Australia

Model variability tested by the AUC standard deviation was minimized at 5 and 7 variables in Andalusia and SW Australia respectively. At lower number of variable the standard deviation increased drastically. Therefore, this numbers of variables, 5 in Andalusia and 7 in SW Australia, were set at this level to assess *P. cinnamomi* risk of invasion (Figure 18).

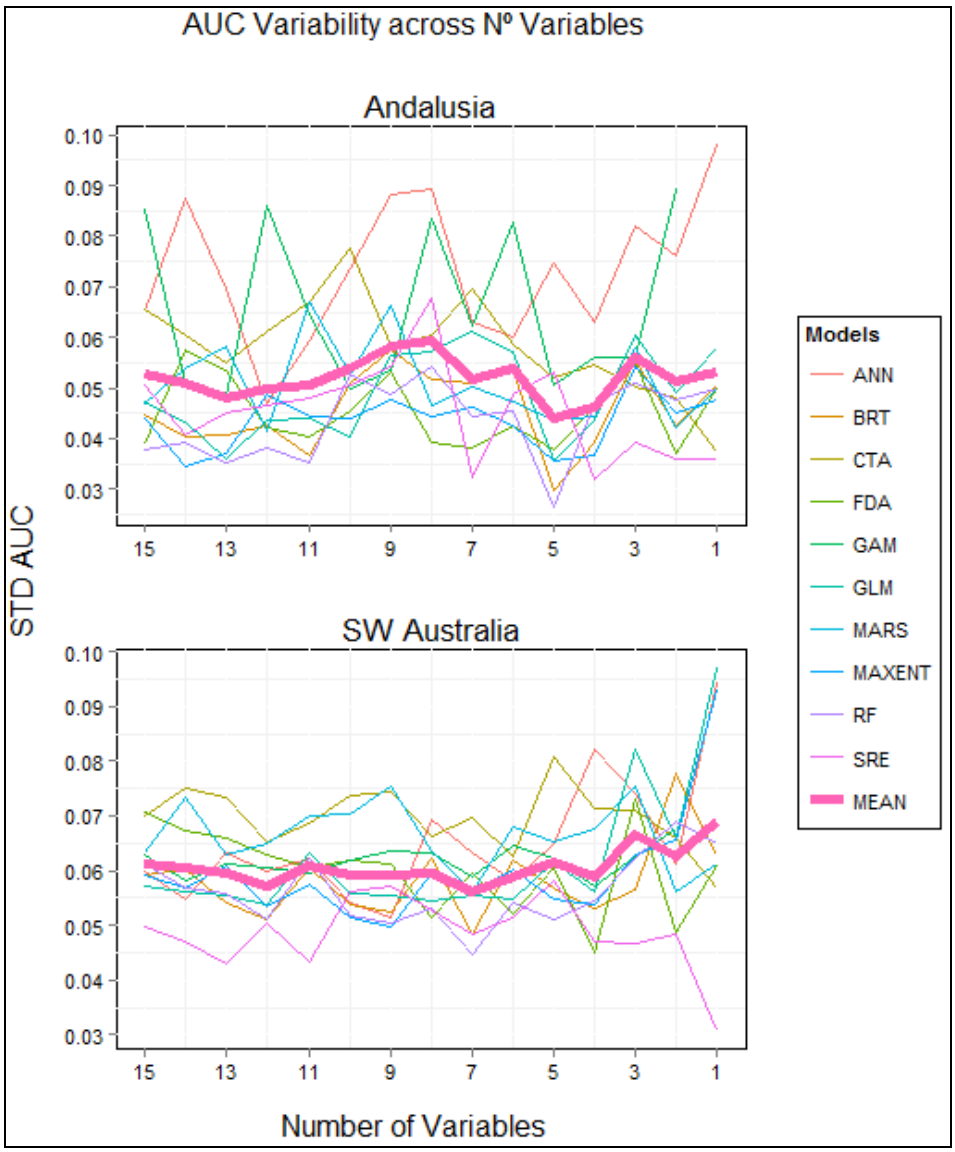

Figure 18. Model accuracy variability across the number of variables and model techniques measured as AUC standard deviation.

Finally, models techniques differed in number and type of environmental variables and their respective importance rank to return an accurate response. ANN, BRT, CTA, MaxEnt, RF and SRE required more variables than FDA, GAM, GLM and MARS in order to return an accurate response. This fact was observed in both datasets (Figures 19 & 20).

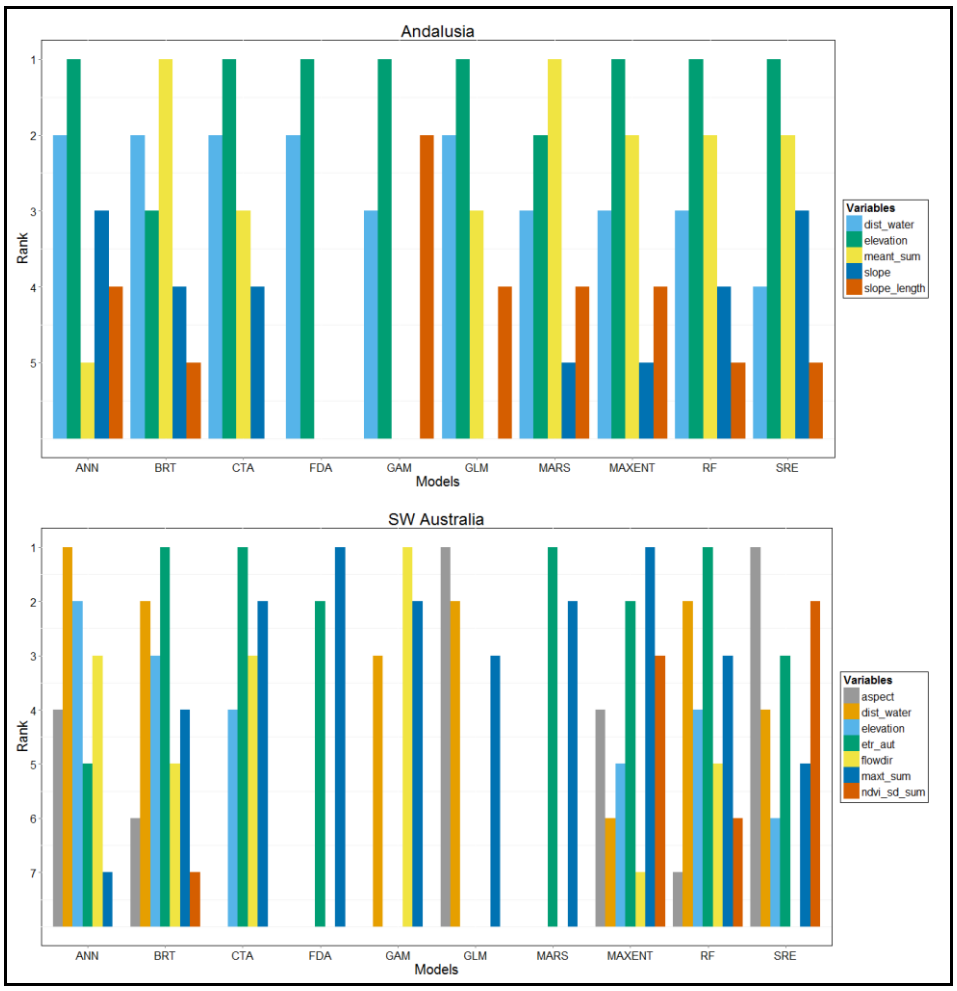

Figure 19. Variables importance by model technique independent dataset. Top: Andalusia; bottom: SW Australia

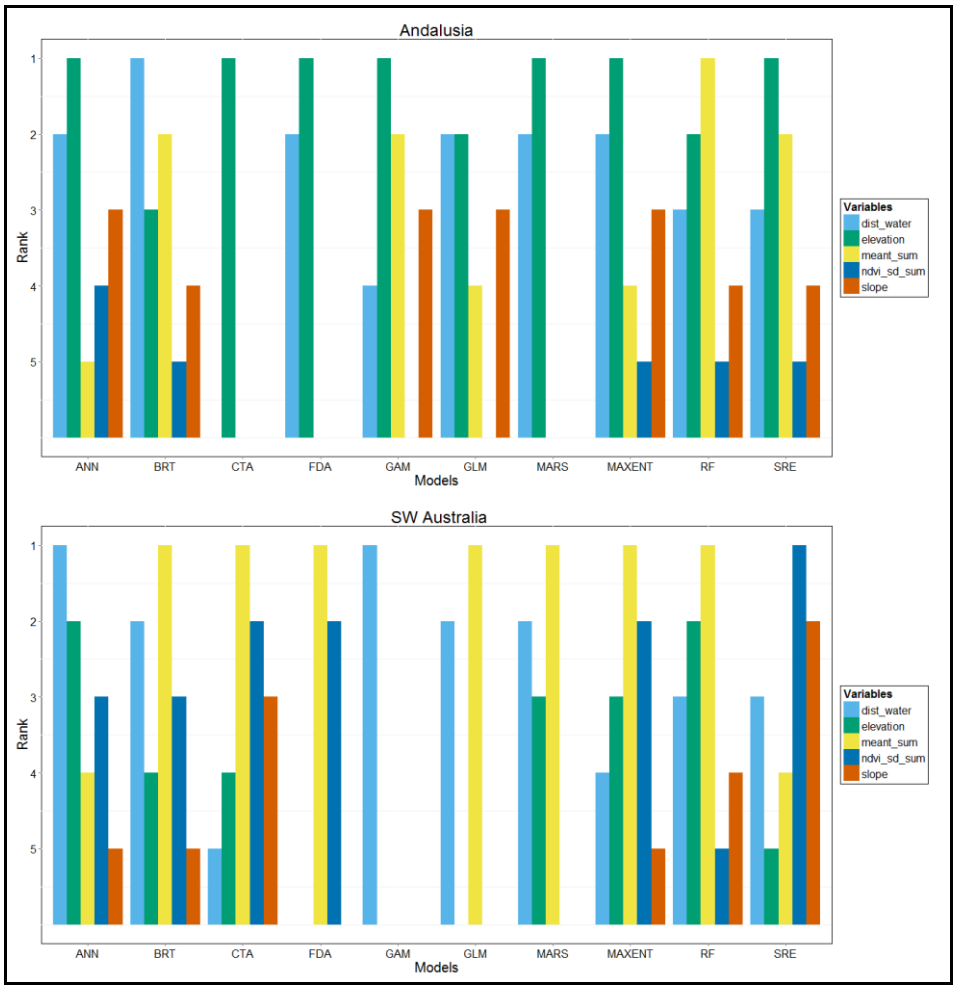

Figure 20. Variables importance by model technique common dataset. Top: Andalusia; bottom: SW Australia

3.2.3 Variable Importance Analysis

The step-by-step variable selection procedure led to several results. On the one hand, we got two sets of variables which determined the fungus risk of invasion in Andalusia and SW Australia. On the other hand, we obtained a common variable dataset that was used to transfer the models (Table 4).

Table 4. Final variables dataset

Selected Variable		
SW Australia	*Andalusia*	*Common Set*
aspect	**dist_water**	**dist_water**
dist_water	**elevation**	**elevation**
elevation	meant_sum	meant_sum
etr_aut	slope	ndvi_sd_sum
flowdir	slope_length	slope
maxt_sum		wet_aut
ndvi_sd_sum		

In bold variables in common in the three sets and highlighted variable that was selected in the common dataset but had a large SDM value. It was not included in the transfer analysis.

P. cinnamomi risk of invasion was explained accurately in Andalusia by distance to water, elevation, mean temperature in summer, slope and slope length. AUC values ranged between 0.925 (BRT) and 0.733 (SRE). MaxEnt, FDA, and RF gave AUCs>0.90 and the remaining model techniques achieved AUCs>0.80 (Table 5). Elevation, mean temperature in summer and distance to water were the variables that showed highest importance for *P. cinnamomi* prediction in Andalusia (Figure 21a).

On the other hand, aspect, distance to water, elevation, actual evapotranspiration in autumn, flow direction, maxima temperature in summer and NDVI standard deviation in summer predicted accurately *P. cinnamomi* risk of invasion in SW Australia. Model performance ranged between 0.897 (FDA) and 0.675 (SRE) AUC values. In addition to BRT, MaxEnt, RF, CTA and MARS which gave AUC>0.80 (Table 6). Actual evapotranspiration in autumn, maximum temperature in summer and distance to water were found the most important variables for *P. cinnamomi* prediction in SW Australia (Figure 21b).

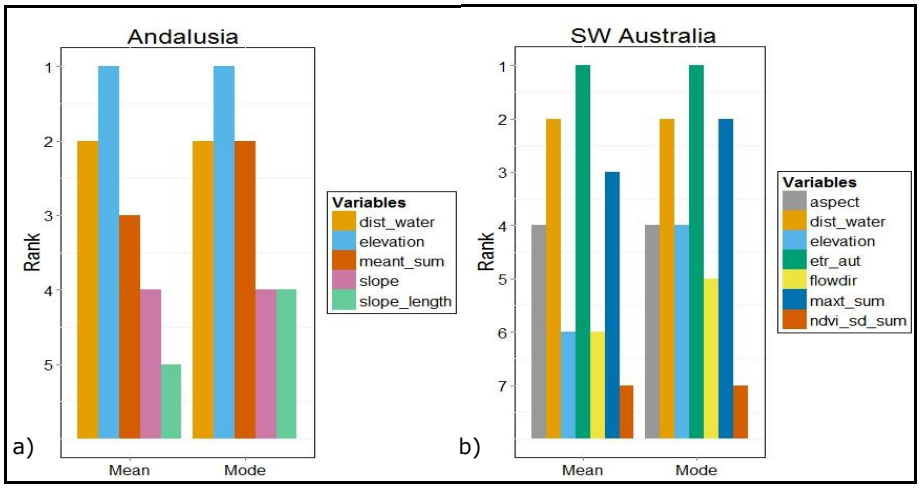

Figure 21. Variable importance average rank with all model techniques and the individual variable dataset. a) Andalusia; b) SW Australia

Furthermore, a common environmental variables dataset formed by distance to water, elevation, mean temperature in summer, NDVI standard deviation in summer and slope were elected to test model transferability. In deed frequency of wet days in autumn was also elected in the beginning, although it was removed for the transferability test because it was underlined as outliers by the analysis of multivariate outlier detection. The correlation with the prediction pointed out that in Andalusia elevation, distance to water and slope were the more important (Figure 22a). While in SW Australia NDVI standard deviation in summer, mean temperature in summer and distance to water had the highest influence in the predictions (Figure 22b). Finally, in Andalusia model accuracy ranged between 0.771 – 0.925 AUCs values in SRE and MaxEnt respectivility while in SW Australia varied from 0.643 (SRE) to 0.841 (MaxEnt) AUCs values (Tables 5 and 6).

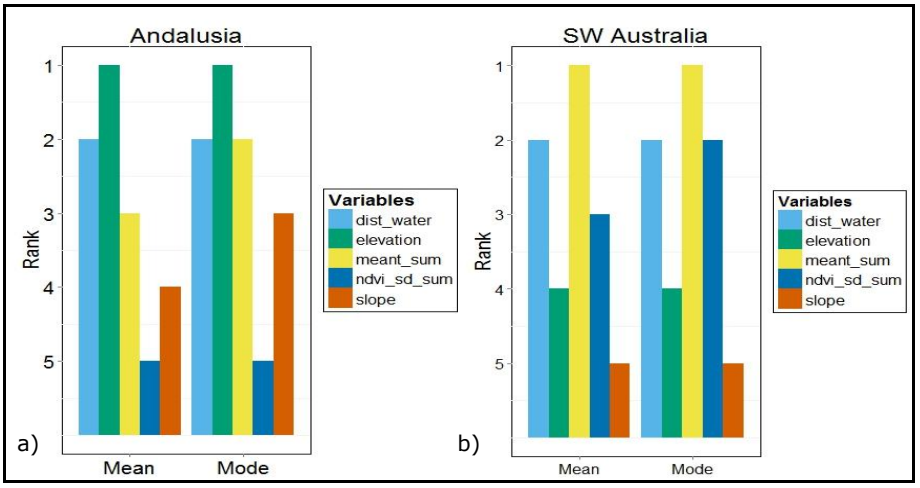

Figure 22. Variables importance average rank among all models techniques considering the common variables dataset. a) Andalusia; b) SW Australia

3.2.4 Model Accuracy

Models performance was test by AUC and the map veracity was estimated by Kappa (κ).

3.2.4.1 Individual Variable Datasets

In Andalusia all models except SRE returned accurate results, AUCs>0.8. BRT gave the highest AUC and Kappa, and MaxEnt the highest mean and minimum AUC; and minimum AUC standard deviation (Table 5). Figure 23 shows an example of the fungus spatial distribution in Andalusia.

Table 5. Andalusian models accuracy

Andalusia					
Models	**Max AUC**	**Mean AUC**	**Min AUC**	**Std AUC**	**KAPPA**
ANN	0.874	0.722	0.528	0.073	0.654
BRT	0.925	0.792	0.627	0.047	0.780
CTA	0.857	0.730	0.565	0.057	0.624
FDA	0.904	0.797	0.689	0.042	0.689
GAM	0.828	0.710	0.531	0.055	0.528
GLM	0.888	0.743	0.547	0.050	0.748
MARS	0.891	0.786	0.622	0.051	0.629
MAXENT	0.920	0.810	0.667	0.041	0.721
RF	0.909	0.797	0.646	0.044	0.661
SRE	0.733	0.626	0.501	0.046	0.599

(Underline more accurate models)

In SW Australia BRT, RF and CTA, gave reasonable maximum AUCs values >0.80. BRT presented highest minimum AUC and, BRT and RF had the highest mean AUC. However, the more accurate model was FDA which also had the best Kappa (0.745) and the highest AUC standard deviation. Further, MaxEnt returned also an accurate prediction (Table 6). Figure 24 displays an example of the fungus spatial distribution in SW Australia.

Table 6. SW Australian models accuracy (AUC and Kappa values)

SW Australia					
Models	**Max AUC**	**Mean AUC**	**Min AUC**	**Std AUC**	**KAPPA**
ANN	0.742	0.582	0.379	0.061	0.540
BRT	0.886	0.701	0.573	0.056	0.645
CTA	0.829	0.674	0.547	0.062	0.612
FDA	0.897	0.690	0.438	0.076	0.745
GAM	0.740	0.581	0.446	0.059	0.463
GLM	0.728	0.596	0.441	0.056	0.364
MARS	0.820	0.644	0.412	0.069	0.533
MAXENT	0.836	0.678	0.434	0.060	0.547
RF	0.833	0.703	0.540	0.051	0.550
SRE	0.675	0.527	0.400	0.052	0.331

(Underline more accurate models)

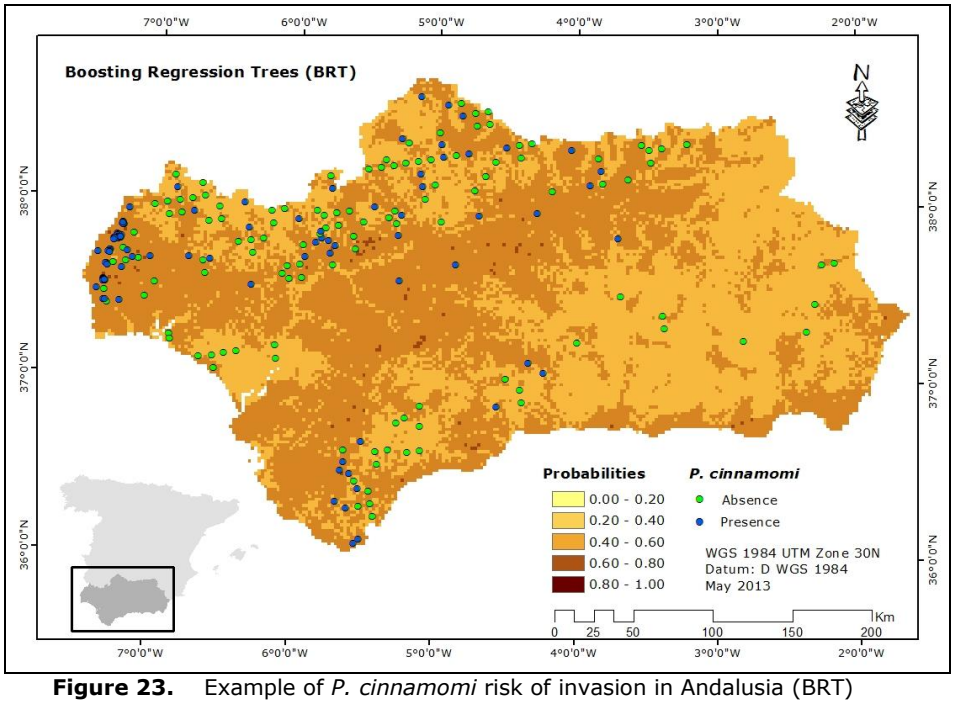

Figure 23. Example of *P. cinnamomi* risk of invasion in Andalusia (BRT)

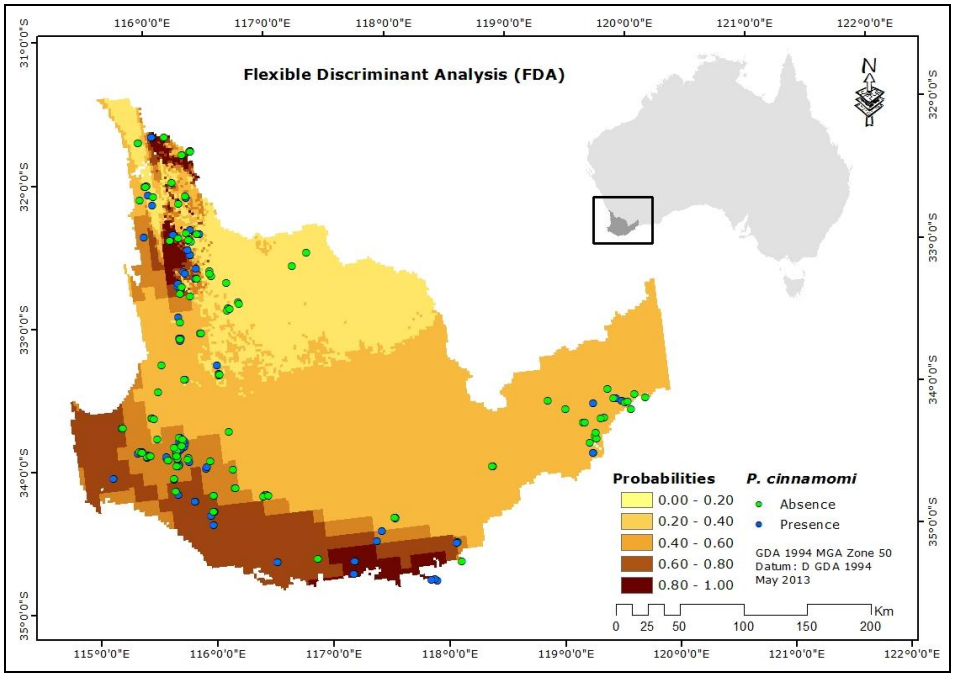

Figure 24. Example of *P. cinnamomi* risk of invasion in SW Australia (FDA)

3.2.4.2 Common Variable dataset

In Andalusia, MaxEnt, RF and BRT showed better performance. Although, MaxEnt had highest maximum, mean, minimum AUC values and minimum AUC standard deviation. Moreover, all models except SRE predicted the risk of invasion accurately, AUCs >0.86 (Table 7).

Table 7. Andalusian models accuracy (AUC values)

Andalusia				
Models	**Max AUC**	**Mean AUC**	**Min AUC**	**Std AUC**
ANN	0.915	0.720	0.528	0.072
BRT	0.923	0.789	0.572	0.055
CTA	0.868	0.730	0.530	0.062
FDA	0.911	0.795	0.667	0.047
GAM	0.879	0.713	0.400	0.067
GLM	0.876	0.740	0.579	0.054
MARS	0.909	0.782	0.565	0.059
MAXENT	0.925	0.806	0.671	0.046
RF	0.924	0.797	0.626	0.051
SRE	0.771	0.640	0.470	0.050

(Underline more accurate models)

On the other hand, in SW Australia CTA, MaxEnt and RF were the models with better performance. However, CTA was the one with lower AUC standard deviation and higher mean AUC while RF had the minimum values (Table 8).

Table 8. SW Australian models accuracy (AUC values)

SW Australia				
Models	**Max AUC**	**Mean AUC**	**Min AUC**	**Std AUC**
ANN	0.741	0.570	0.352	0.066
CTA	0.805	0.660	0.408	0.059
BRT	0.784	0.603	0.364	0.063
FDA	0.715	0.579	0.419	0.056
GAM	0.724	0.573	0.419	0.058
GLM	0.729	0.573	0.422	0.058
MARS	0.767	0.579	0.433	0.061
MAXENT	0.841	0.636	0.423	0.063
RF	0.804	0.676	0.427	0.062
SRE	0.643	0.503	0.325	0.051

(Underline more accurate models)

Comparing both situations within the common dataset, MaxEnt was the model which gave better AUCs, SRE returns the lowest AUC and ANN presented the highest AUC standard deviation.

3.2.5 Response Curve

In Andalusia high risk of *P. cinnamomi* invasion was determined by areas between 0 – 1800 masl. within the complete range of mean temperature in summer and distance to water (Figure 25)

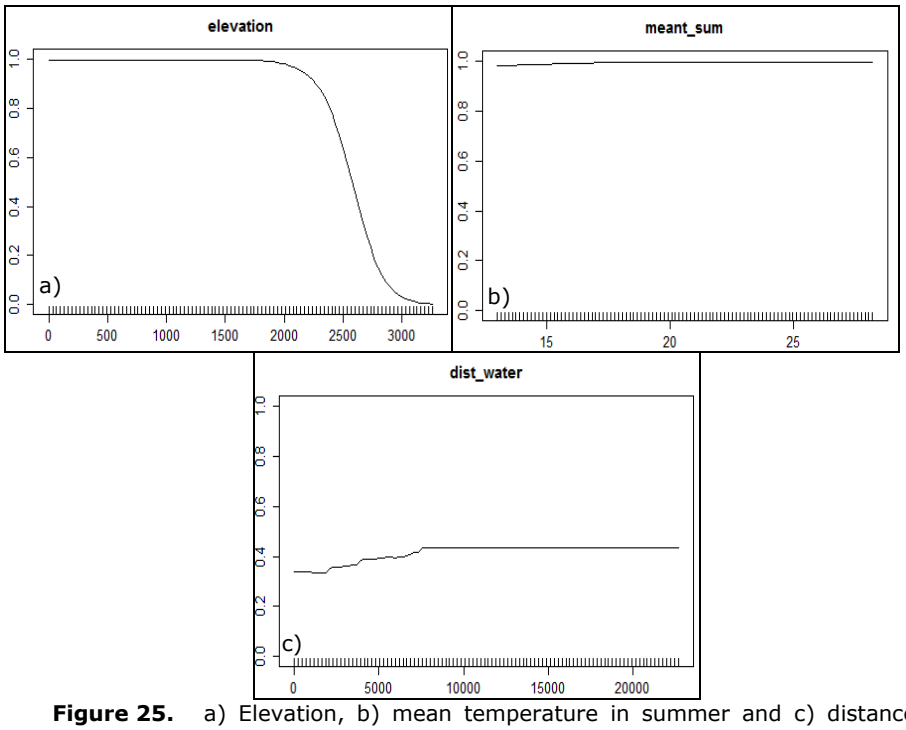

Figure 25. a) Elevation, b) mean temperature in summer and c) distance to water response curves in Andalusia

On the other hand, In SW Australia the probability occurrences increased after 300 masl with temperatures above 30 ºC and together with actual evapotranspiration. Moreover, areas close to sea level where also pointed out by elevation as areas at high risk of invasion. Maximum temperature in summer had a pick of occurrence around 23 ºC with a soft decrease until 31 ºC and a later increase. Finally, there was found a negative correlation between distance to water and probability of occurrence (Figure 26).

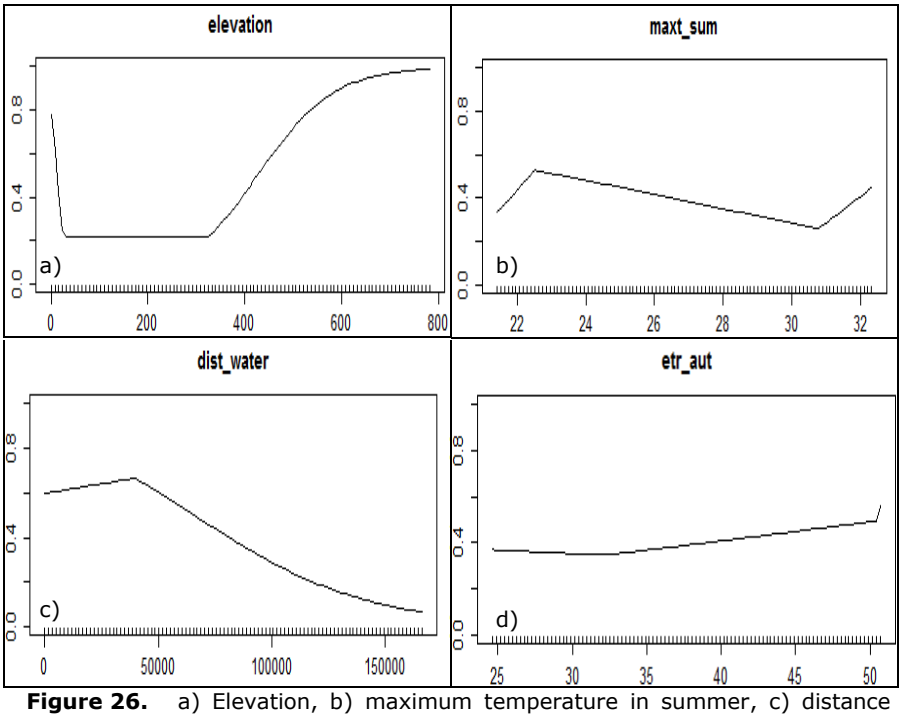

Figure 26. a) Elevation, b) maximum temperature in summer, c) distance to water and d) actual evapotranspiration in autumn response curves in SW Australia

Significantly differences were found in both situations, influenced by elevation and temperatures in summer. These differences should affect transferability.

3.2.6 Model Transferability

3.2.6.1 Common Variables Boxplots

The predictive variable initially selected to test models transferability were: distance to water, elevation, mean temperature in summer, NDVI standard deviation in summer, slope and wet frequency days in autumn.

Elevation and slope boxplots revealed large difference in means and variability between the study areas. However SW Australian ranges were included in the Andalusian ones (Figure 27).

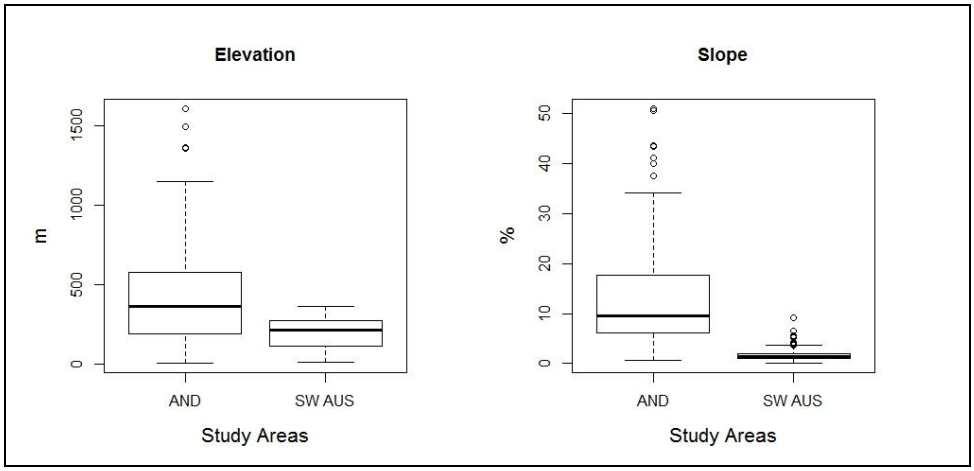

Figure 27. Right, elevation (m) and left, slope (%) boxplots. (AND: Andalusia; SW AUS: Southwest Australia)

Andalusia presented relatively low distances in comparison with SW Australia the later included the Andalusian range. NDVI standard deviation in summer showed similarities between the study areas. Means were close however they differed in the inferior quartile (Figure 28).

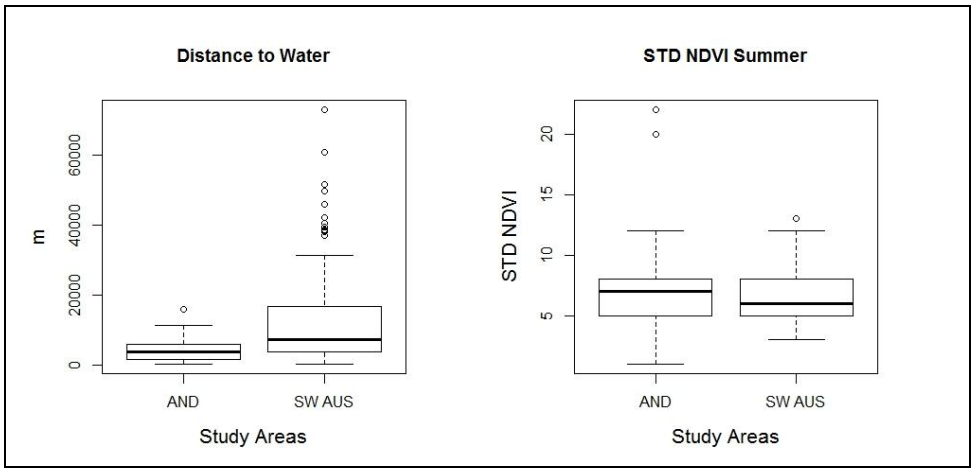

Figure 28. Right, distance to Water (m) and left, NDVI standard deviation in summer boxplots. (AND: Andalusia; SW AUS: Southwest Australia)

Wet day frequency in autumn presents a significant difference. SW Australia presented a large mean and variability while in Andalusia both were lower. Moreover, Andalusia "box" fitted below SW Australia first quartile. In addition, wet day frequency in autumn was classified as conflictive variable by SMD and mean temperature in summer had one of the highest SMD in the non-conflictive variables remaining

(See Appendix I, Table 12). Consequently, wet day frequency in autumn was not considered to perform the transferability analysis (Figure 29).

Although, mean temperature in summer also presented dissimilar means and ranges, this was not marked by the multivariate outlier detection analysis as a conflictive variable.

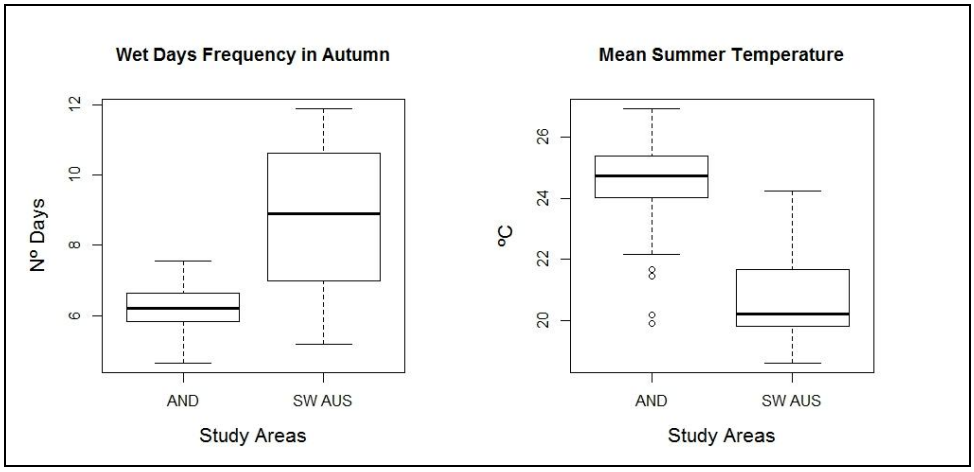

Figure 29. Right wet days frequency in autumn (nº of days) and left, mean summer temperature (ºC) boxplots. (AND: Andalusia; SW AUS: Southwest Australia)

The chosen environmental variables to assess model transferability differ in means, but the range involves both situations except mean temperature in summer.

3.2.6.2 Transferability Andalusia Models

Models trained and calibrated in Andalusia and transferred to SW Australia presented a percentage AUC loss that ranges between 27.89 – 40.84% in GLM and MaxEnt respectively (Table 9). Andalusian model performance was better in Andalusia than in SW Australia. However, SW Australian models showed a better transferability (Table 9 & 10). Figure 30 shows an example of the fungus spatial distribution in Andalusia and Figure 31 presents an example of the fungus spatial distribution in SW Australia performed with the model trained in Andalusia.

Table 9. Andalusian models transferability result

Andalusia			
Models	**Max. AUC**	**Transfer AUC**	**%AUC Loss**
ANN	0.915	0.567	37.99
BRT	0.923	0.637	31.01
CTA	0.868	0.530	38.92
FDA	0.911	0.607	33.42
GAM	0.879	0.625	28.87
GLM	0.876	0.632	27.89
MARS	0.909	0.565	37.89
MAXENT	0.925	0.547	40.84
RF	0.924	0.550	40.43
SRE	0.771	0.500	35.15

(Underline models with lower %AUC loss)

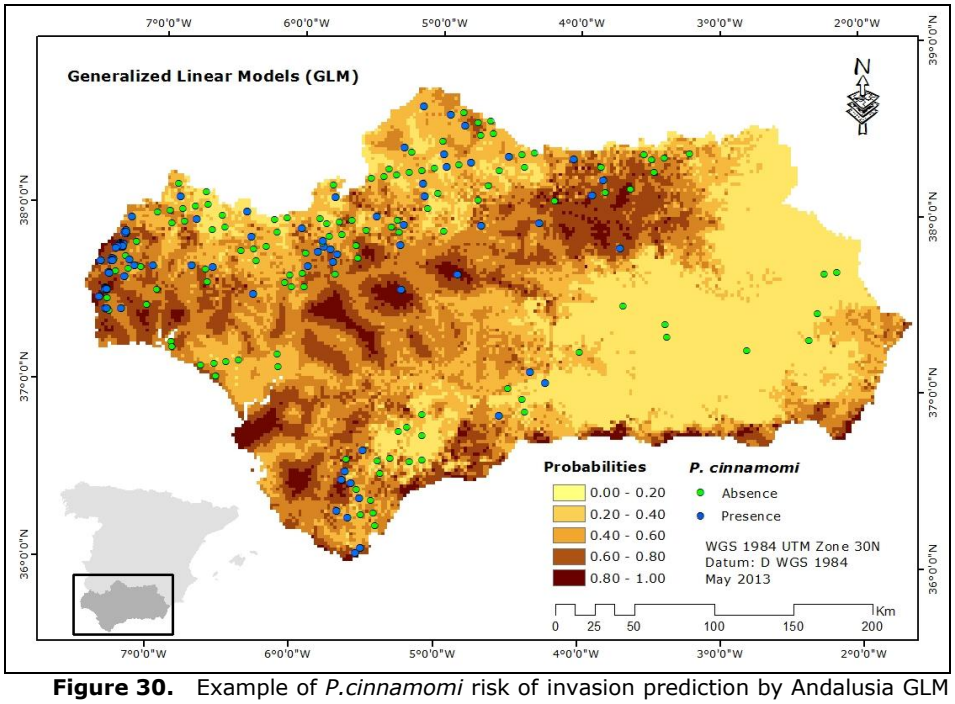

Figure 30. Example of *P.cinnamomi* risk of invasion prediction by Andalusia GLM model

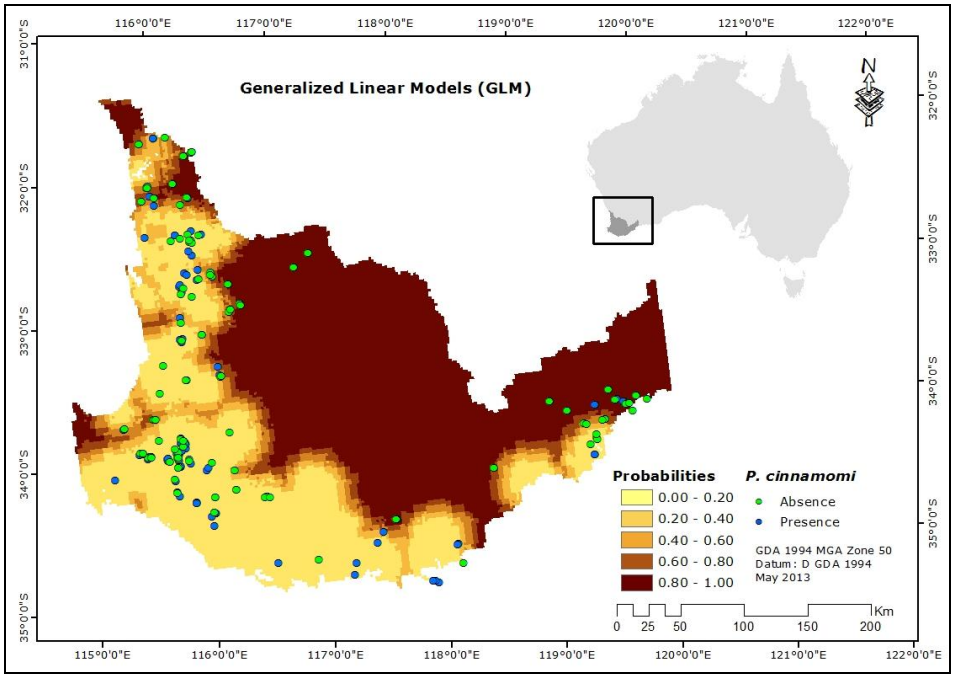

Figure 31. Example of *P.cinnamomi* risk of invasion in SW Australia by Andalusia by GLM model transferred

3.2.6.3 Transferability SW Australia Models

Models trained and calibrated in SW Australia and transferred to Andalusia presented a percentage AUC loss that ranges between 3.18 – 33.73% in GAM and MARS respectively. Moreover, RF (3.61%) and GLM (5.39%) presented low loss in AUC (Table10). Figure 32 shows an example of the fungus spatial distribution in SW Australia and Figure 33 presents an example of the fungus spatial distribution in Andalusia performed with the model trained in SW Australia.

Table 10. SW Australian models transferability results

SW Australia			
Models	**Max. AUC**	**Transfer AUC**	**%AUC Loss**
ANN	0.741	0.667	10.00
BRT	0.805	0.620	22.94
CTA	0.784	0.705	10.13
FDA	0.715	0.631	11.69
GAM	0.724	0.701	3.18
GLM	0.729	0.690	5.39
MARS	0.767	0.508	33.73
MAXENT	0.841	0.678	19.44
RF	0.804	0.775	3.61
SRE	0.643	0.500	22.24

(Underline models with lower %AUC loss)

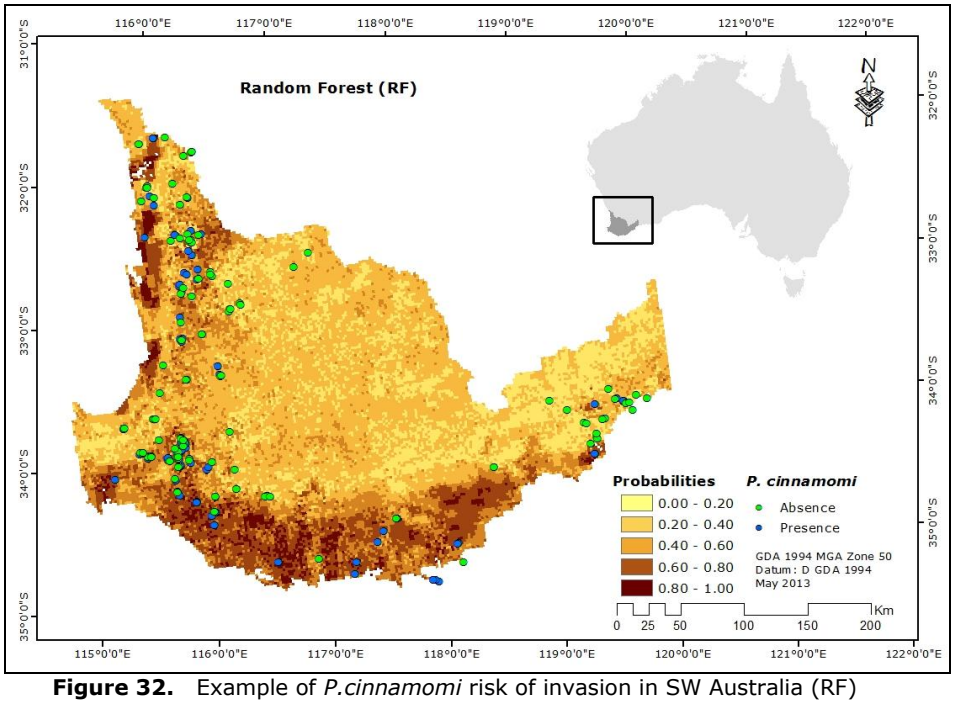

Figure 32. Example of *P.cinnamomi* risk of invasion in SW Australia (RF)

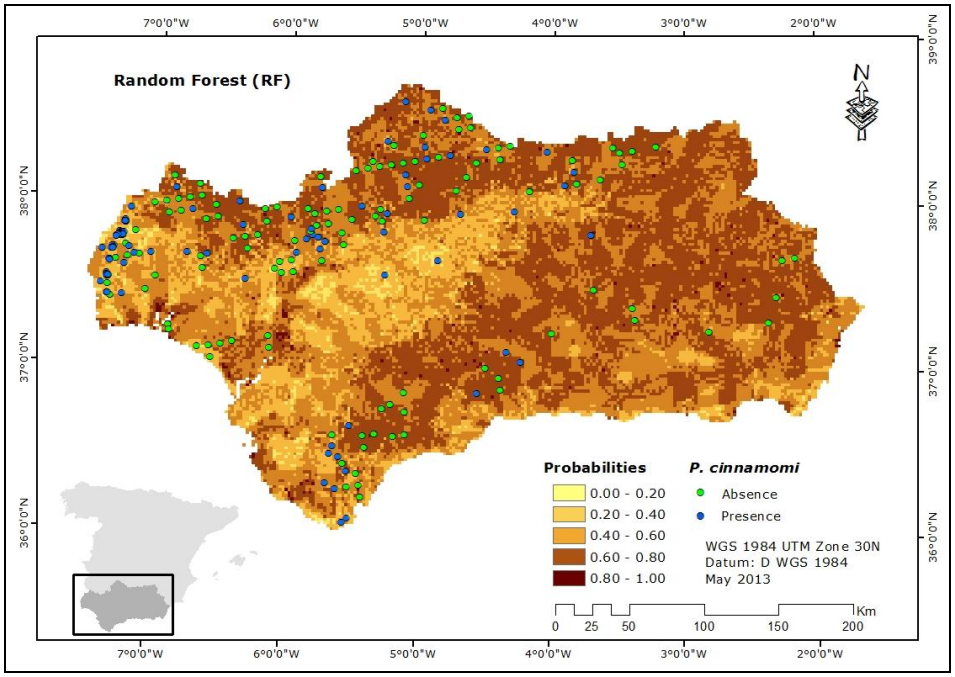

Figure 33. Example of *P.cinnamomi* risk of invasion in Andalusia by SW Australia transfer RF model

Chapter 4: Discussion

4.1 Phytophthora cinnamomi Location

The number of data point used in this research, (95 presence and 120 absence, prevalence 44 – 56%), differed significantly with other studies that predicted the risk of invasion of Oak Decline (Meentemeyer *et al.*, 2004; Kelly *et al.*, 2007; Václavík *et al.*, 2010). Nevertheless, Franklin (2009) stated that 100 – 500 data points with a prevalence of (50 – 50%) should be enough to obtain accurate predictions. On the other hand, using true absences it is an improvement in comparison with similar studies (Kelly *et al.*, 2007; Václavík *et al.*, 2010) because it have been found that true absences increase model accuracy (Václavík & Meentemeyer, 2009).

The predictive variables differed with other studies that modelled *Phytophthora sp*. risk of invasion. While (Meentemeyer *et al.*, 2004; Václavík *et al.*, 2010) used pre-classified predictive variables, we use preconceived establishment conditions. However, we used environmental continuous variable as has already been done for macrofungi (Wollan *et al.*, 2008), invasive species (Ficetola *et al.*, 2007) or *Phytophthora ramorum* (Kelly *et al.*, 2007) with accurate results. We use preconceived establishment conditions because we wanted to determine which environmental variables determined the risk of invasion and compare predictions between Andalusia and SW Australia.

4.2 Environmental Variables

"The Niche Theory" considers that species occurrence is determined by environmental, dispersal and biotic interaction factors (Soberón & Nakamura, 2009). Our dataset included environmental information, variables that could suggest fungus dispersal (i.e. flow direction, distance to water and roads) and preferences areas (forest, scrublands and flat valley bottom areas). However, we missed relations with other plants or hosts considered in other studies (Kelly *et al.*, 2007; Václavík *et al.*, 2010) and highlighted in some biological descriptions of the fungus (Hardham, 2005). Moreover, some studies concluded that different soils types enhance fungus distribution (Weste & Marks, 1987) but we did not consider soil information. These assumptions were considered during the research because of the flora and soil types dissimilarity between both study areas (FAO & IIASA, 2000). A factor considered to influence model transferability.

4.3 Study Area Comparison

Though SW Australia and Andalusia are both Mediterranean regions (Peel *et al.*, 2007), the multivariate area comparison carried out revealed difference in climate variables such as maximum temperature, mean temperature and potential evapotranspiration. However, both areas had similar environmental conditions after deleting outliers. Moreover, visual inspection of boxplots confirmed those differences that we suggest could be related to topography.

SW Australia is a relatively flat area, without mountain ranges. The altitude increases inland up to 700 masl. Andalusia is characterized by a central valley surrounded by mountains. Andalusia´s elevation range changes rapidly from sea level up to 3.800 masl. The difference in topography explains the variability in temperature and precipitation between the study areas. Moreover, Andalusia elevation variability is associated to climatic variability including sub-climates that range from desert to alpine (Junta de Andalucía, 2007). This is not the case in SW Australia which is characterized by a uniform climate.

Our finding suggests that model transferability should be performed carefully between Andalusia and SW Australia. The limitations were highlighted by the analysis of multivariate outlier detection flagging dissimilar variables. Model transferability can be performed confidently between both regions excluding dissimilar variables.

4.4 Variable Importance

4.4.1 Number of Variables

The number of variables to explain the species occurrence varies with the organism studied. Franklin (2009) included a revision of recent studies with SDMs and the number of explanatory variables used in each, varied from 3 - 4 up to 40. Other researchers found that 1, 2 and rarely 3 explanatory variables were sufficient (van Gils *et al.*, 2012).

The result for the optimal number of variables is inconclusive. We found that Maximum and minimum AUC values were not influenced by the number of variables used to make the predictions (Figures 16 & 17). Moreover, model performance variability measured by AUC standard deviation fluctuated across the number of variables without any pattern. However, we decided to choose the number of variables that minimized the AUC standard deviation in the awareness that this result could be due to chance. Minimum AUC standard deviation measures the model accuracy variability across the number of

variables. So, more robust model can be identified by lower AUC standard deviation. Furthermore, we had to decide on a final number of variables and neither in the literature nor our test we found a conclusive approach.

The fact that model performances were independent of the number of variables emphasized model ability to find relationships between occurrences and explanatory variables and disagree with other studies which suggest a decreases of model accuracy with the increase of number of variables, in some cases (Barry & Elith, 2006; Zhang & Zhang, 2012). Moreover, this issue might be influenced by the fact that we compared model performance according to the maximum and minimum AUCs values instead of the mean AUC as other studies did (Elith *et al.*, 2006; Syphard & Franklin, 2009; Heikkinen *et al.*, 2012; Zhang & Zhang, 2012).

4.4.2 Variable Importance Comparison

The process to assess variable importance could be unbiased due to the fact that we only considered the more accurate model from each single technique. The more accurate model was found among 300 "runs". Choosing the models that returned the highest AUC can be misleading. This is because maximum values do not represent model prediction variability and could in some cases be outliers. Therefore, a good model performance could be due to chance. A proper analysis could have been done considering all models and using the mean AUC instead of maximum AUC and an average of variable importance across models as preferred by Syphard and Franklin (2009). Mean AUC values and average variable importance would have taken into account all the models performance.

We found that model techniques differed in number and type of environmental variables and in their respective importance rank to return an accurate response. So, model techniques could mimic complex relations between variables which enhance model performance but challenged the ecological relationship between predictive and response variables (Barry & Elith, 2006). Machines learning algorithms found complex relationships, with numerous variables involved, while linear models gave a smoother relation between variables with lower number of variables. Additionally, the lack of a single important variables have been considered as a source of uncertainty in model prediction (Barry & Elith, 2006), a general issue in our study (Figure 19 & 20).

We suggest that our statistical models gave a reasonable approach of *P. cinnamomi* environmental response in consonance with the

variables selected but with unclear ecological information due to the variable importance misclassification. We found accurate model results using different variable sets to predict *P. cinnamomni* occurrence in each study area.

Likewise, we found that *P. cinnamomi* risk of invasion is predicted by different environmental situations in each study area. In the individual datasets there were two common variables. Distance to water and summer temperature, were classified in 2nd and 3rd order in the more accurate model. However, the variables more related to the predictions (elevation and actual evapotranspiration) differed (Figure 21, Table 11).

Table 11. Variable importance summary

High Correlated Variable		
Rank	***Andalusia***	***SW Australia***
	Individual set	*Individual set*
1	elevation	etr_aut
2	dist_water	maxt_sum
3	meant_sum	dist_water

According to our finding we can suggest that the risk of invasion of *P. cinnamomi* in Andalusia and SW Australia was determined by different environmental variables where distance to water was common. Moreover, temperature in summer seemed to be also a relevant variable although they differed in their description. A boxplot visual inspection comparing both variables showed difference in their means and range (Figure 34a). Furthermore, the boxplot shows that the temperature range defined by both temperature variables agree with the optimal growing range of *P. cinnamomi* defined by Weste and Marks (1987) in their biological description of the fungus. Finally, an inspection of correlation between evapotranspiration in autumn and elevation showed that there was a low relationship between both. This finding confirms that the fungus distribution is determined by different environmental conditions in each study area (Figure 34b).

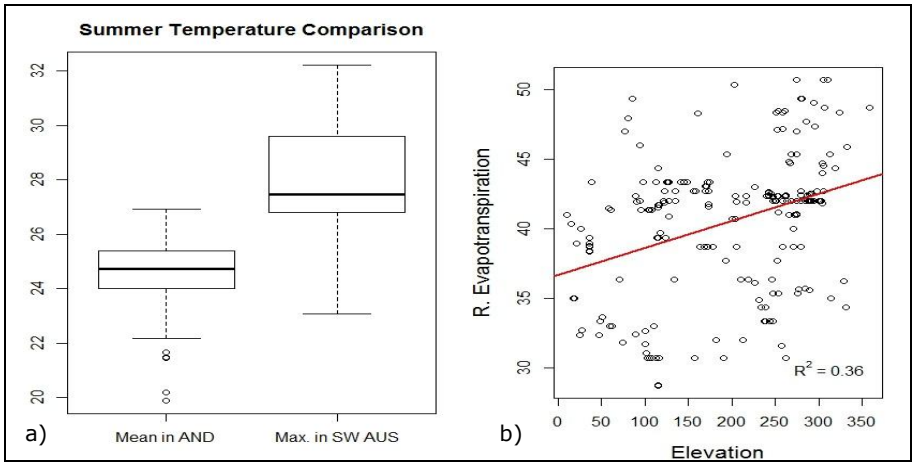

Figure 34. Temperature variable importance comparison (right) and correlation between actual evapotranspiration and elevation (left)

The difference in the methodological process to compare variable importance in this study with Syphard and Franklin (2009) suggest that our result could be doubtful because it just considered the average influence of the "best" single models between all techniques.

4.4.3 Ecological Explanation of Important Variables

The found predictive variables can be categorised by four classes: Climatic, topographical, land cover and dispersal. Temperature has been assessed as an important factor to *P. cinnamomi* growth and distribution (Weste & Marks, 1987; Sánchez *et al.*, 2003). Mean temperature in summer and maximum temperature in summer were found as relevant variables. The fact that summer is the most influential season suggest that temperature in summer influences the survival of the fungus by forming Chlamydospores, a life-stage capable to resist unfavourable conditions (Weste & Marks, 1987). Visual comparison of summer temperatures response curves pointed out difference between both regions (Figures 25b & 26b). The high probability of occurrence in Andalusia for the complete temperature range (12 - 30ºC) coincided with fungus growing interval. However, this did not occur in SW Australia (Weste & Marks, 1987).

Moreover, actual evapotranspiration in autumn is directly related to soil moisture (Droogers, 2000). Soil moisture has also been pointed out as an important factor of distribution and growth of *P. cinnamomi* (Weste & Marks, 1987). Warm temperatures and free water in the soils enhance dispersal, production and growth of *P. cinnamomi* (Weste & Marks, 1987). Our findings agree with some biological

studies which suggested that in warm climates the fungus distribution was controlled by soil moisture. Moreover, host resistance may vary with temperature (Weste & Marks, 1987). The response curve presented a positive relationship. The habitat suitability increased with the soil moisture which agrees with (Figure 26d) (Hardham, 2005).

Elevation might be related to the risk of invasion according to their relation with the main host species, *Quercus* in Andalusia and Jarrah in SW Australia. The vegetations cover maps (Figures 2 & 3) revealed that presence points were mainly located in Jarrah and *Quercus* forest which were in altitude range between 200 and 500 masl. However, this result could be biased due to sample strategy followed in each area. Sample were collected from *Quercus* and Jarrah forest while the fungus may occurs in a wider range (Dell *et al.*, 2005). The visual comparison between response curves revealed opposite pattern. In Andalusia the altitude range might be related to *Quercus* forest which grows from sea level up to 2.000 masl in some cases (Aronson *et al.*, 2009). While in SW Australia, the pattern is incongruent because according to Weste and Marks (1987) the fungus avoid high altitudes. Moreover, in SW Australia temperature and dryness increased inland and these are unsuitable conditions for the fungus (Figure 25a & 26a).

NDVI standard deviation in summer explained the distribution of *P.cinnamomi*. This could be related to land cover and climate. NDVI might be linked to land cover in the same way that elevation is with *Quercus* and Jarrah forest. The SW Australian response curve shows that with lower NDVI standard deviation the probability of invasion was higher (Figure 35a). Forest NDVI standard deviation is often low through time. This means that the risk of invasion is associated with a low variability of NDVI as Jarrah forest cover. On the other hand, NDVI might be related to climate. The underline season was summer. In the Mediterranean, during long warm summers herbaceous vegetation dries out, so the NDVI low variability may describe forest and scrubland covers. The fungus affects woody species.

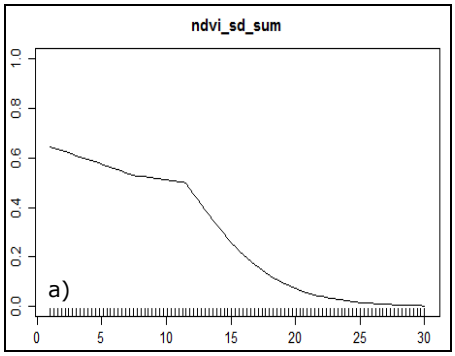

Figure 35. NDVI standard deviation in summer. SW Australia

Distance to water, flow direction, slope and slope-length simulates the fungus natural dispersion by water flow (Shearer *et al.*, 2007). Therefore, downhill and ponding areas have been defined at high risk of infection. The distance to water response curve for SW Australia presented a negative correlation with the probability of occurrence which might indicate the importance of this variable in the fungus dispersal, while in Andalusia remains almost constant possibly because of its dense waterway (Figure 25c & 26c).

The visual comparison of the response curve revealed meaningful ecological relationship especially in SW Australia. We found positive correlation with actual evapotranspiration (soil moisture), negative correlation with distance to water (dispersing Agent) in addition to summer temperatures.

4.5 Model Evaluation

We compared model accuracy within and between study areas. We found that Andalusian models had better performance with higher AUC's compared to Australian models. We suggest that this difference is due to the sampling strategy. Andalusia point location were collected by a systematic stratified sampling design in *Quercus* forest areas while the SW Australia points were taken by a purposive method within the areas where decline symptoms in trees and scrublands were present. Moreover, Edwards *et al.* (2006) found similar effects of sampling design although with (15-45%) prevalence ratio and Hirzel and Guisan (2002) demonstrated that regular and equal stratified sampling were the more accurate strategies. In addition, Andalusia point location were distributed throughout the entire study area while in SW Australia the points were not that well distributed (Figure 2 & 3), which may have influenced negatively the SDM extrapolation (Franklin, 2009).

AUC values with the independent dataset were higher in both study areas than with the common dataset which supported the variable selection process and verified that different variables predict accurately *P. cinnamomi* risk of invasion in Andalusia and SW Australia.

In Andalusia BRT with an AUC = 0.925, high performance, and K = 0.78, very good, was the "best" model in the individual test while MaxEnt with AUC = 0.925 had a high performance with the common dataset.

In SW Australia FDA resulted as more accurate model with AUC = 0.897 and K = 0.745 in the individual test which meant a high model performance and very good level of agreement between data and predictions. MaxEnt returned the highest AUC = 0.841 which was adequate result with the common dataset.

A Comparison of the complete set of models across study areas and variable sets highlighted that machine-learning methods ANN, BRT, MaxEnt and RF had the overall higher accuracy. However, machine-learning method tend to over-fit predictions even though this have been considered a desirable property to model invasive species (Jiménez-Valverde *et al.*, 2011). Regression methods GAM, GLM and MARS together with classification methods CTA and FDA returned an acceptable to high accurate response. These techniques have been suggested to predict the fundamental niche more efficiently (Jiménez-Valverde *et al.*, 2011). On the other hand, the low predictive power of SRE has been also pointed out. SRE belong to bioclimatic envelop techniques which are the simple and "older" species distribution method. Therefore, the improvements on SDM techniques come also to light.

Finally, the ability of AUC statistic to assess model performance individually have been criticized by several authors (Lobo *et al.*, 2008). They say that sensitivity and specificity should also be reported in model comparison. However, some drawbacks are not applicable to this research because same species are compared within the same extent (See Appendix I, Tables 25 & 26).

4.6 Model Transferability

The transferability test highlighted the poor ability of machine-learning methods (ANN, BRT, MaxEnt and RF) to extrapolate across space. Our results disagree with Heikkinen *et al.* (2012) who found good transferability in ANN, BRT and MaxEnt. However, we both agree about GAM and GLM good performance and transferability. Wenger and Olden (2012) also marked the transferability of GAM and GLM. We suggest that the transferability strengths of GLM and GAM versus machine-learning methods is linked to the common issue of feeding SDMs with lower number of variables to avoid complexity or over-fitting (Wiens *et al.*, 2009). Complexity may come up due to fitting non-linear relationships between species and environment, as did BRT and MaxEnt (Elith *et al.*, 2006). In addition, model complexity might be caused by the inclusion of too many descriptive variables (Thuiller *et al.*, 2008). According to our results the latter might be the case why machine-learning methods have lower transferability in comparison with regression methods (Figures 19 & 20, Tables 9 & 10).

Moreover, the lack of transferability in machine-learning methods could be due to the inclusion of elevation as a predictive variable while GAM and GLM did not in SW Australia (Figure 20). The elevation response curve in MAxEnt highlighted a mismatch in ranges between high probability of occurrence and elevation in both areas (Figure 36).

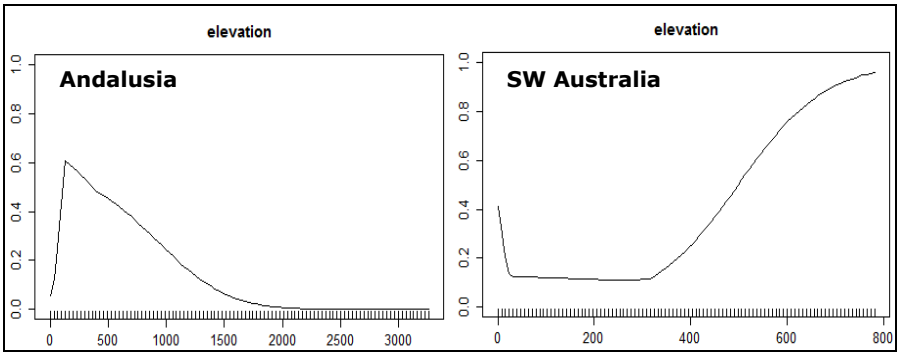

Figure 36. Elevation response curve in Andalusia (right) and SW Australia (left) MaxEnt.

This also explains the high transferability power presented by RF between SW Australia and Andalusia (Figure 37). High probability of occurrence was not related to elevation in this RF model. On the contrary, Heikkinen *et al.* (2012) and Wenger and Olden (2012)

found poor transferability ability using RF in comparison with the other methods.

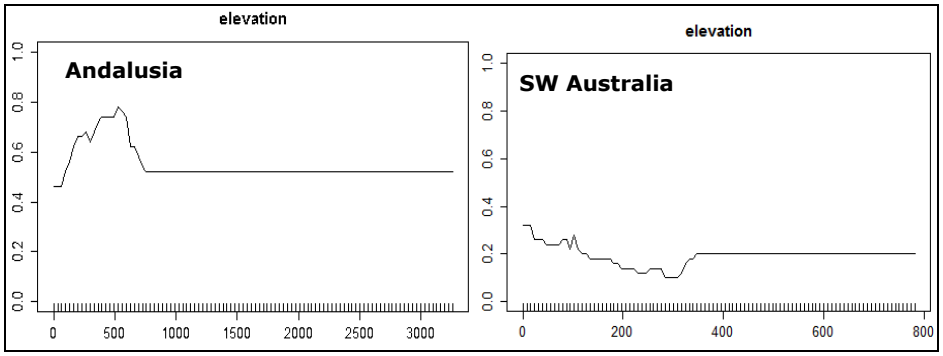

Figure 37. Elevation response curve RF model. Right, Andalusia; left, SW Australia

The ability of MARS to produce realistic extrapolations is doubtful. MARS presented low results in both areas and agreed with Heikkinen *et al.* (2012). Additionally Prasad *et al.* (2006) pointed out MARS low extrapolation capacity to perform future projections. In the same line, the tendency to over-fit of CTA that was mentioned by Thuiller (2003) might explains its poor transferability results in addition to complexity.

Finally, according to our results GAM and GLM are the "best" methods for extrapolation. Transferability together with their acceptable performance and their continuous response curves (Austin, 2007), makes these model techniques a suitable tool to; (1) predict species across regions, (2) indentify species occurrence in restrings areas, (3) predicts the influence of climate change on biodiversity and (4) predict the potential invasion of alien species.

Chapter 5: Conclusion

Model transferability extrapolates predictions across regions. It requires similar environmental conditions between the regions. In some studies the environmental similarities between regions are assumed (Randin *et al.*, 2006; Heikkinen *et al.*, 2012). Our finding suggests that an analysis of dissimilarity between environmental conditions may exclude dissimilar environmental variables from modelling and improve model transferability.

The risk of invasion by *P. cinnamomi* is predicted by elevation, distance to water and mean temperature in summer in Andalusia. While in SW Australia evapotranspiration in autumn, maximum temperature in summer and distance to water are found as the most important variables. The environmental conditions differ between both areas. The visual comparison between response curves result ambiguous, although ecologically meaningful. Finally, our research emphasize the importance of the variable selection process that should be done carefully and requires expert ecological knowledge of the species modelled (Barry & Elith, 2006).

Species distribution models predicted *Phytophthora cinnamomi* risk of invasion accurately in Andalusia (AUCs>0.85) and SW Australia (AUCs>0.72) with nine out of ten model techniques. The predictive power of machine-learning methods as BRT, RF and MaxEnt (AUCs> 0.90 in Andalusia and >0.83 in SW Australia) and the classification method FDA (AUCs>0.89 in both) were superior to the others. The sample strategy design may have caused the lower model performance in SW Australia (Edwards *et al.*, 2006).

We found that machine-learning methods ANN, BRT, MaxEnt and RF give an accurate response in the training area while having a low transferability. On the contrary, regression methods as GAM and GLM show lower AUCs in the training areas but have best transferability. MARS, CTA and FDA show a similar predictive power as regression methods, though with lower transferability. SRE predictive and transferability ability are the lowest. A lower number of explanatory variables might increase model transferability although further research should be done in this area. In conclusion, GAM and GLM are the models that provide good performance combined with transferability results.

Our results suggested to consider carefully model predictions outside their training data range, so that in extrapolation through time or space. In order to achieve more confident results in SDM

extrapolations we suggest to tests model transferability and incorporate its results to model accuracy analysis. In addition, it is suggested to use different SDM techniques depending on the aim of the study. In studies where the aim is to predict species occurrence in inaccessible area, predict the distribution of rare species or predict the potential distribution of an invasive species we recommend to use GAM or GLM.

Chapter 6: References

ABARES (2011) *The 2011 Australian national map layers*. Available at: www.daff.gov.au/abares/mcass (accessed 15 October 2012).

Allouche, O., Tsoar, A. & Kadmon, R. (2006) Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology*, **43**, 1223-1232.

APCOR (2010) Cork Nature, Future, Culture. Cork Environmental Importance. In: *Cork Information Bureau 2010* (ed. P.C. Association). Portuguese Cork Association, Portugal.

Araújo, M.B. & New, M. (2007) Ensemble forecasting of species distributions. *Trends in Ecology & Evolution*, **22**, 42-47.

Aronson, J., Pereira, J.S., Pausas, J.G., Society for Ecological Restoration International. & ebrary Inc. (2009) Cork oak woodlands on the edge ecology, adaptive management, and restoration. In: *The science and practice of ecological restoration*, pp. xvii, 315 p. Island Press, Washington, D.C.

Austin, M. (2007) Species distribution models and ecological theory: A critical assessment and some possible new approaches. *Ecological Modelling*, **200**, 1-19.

Barry, S. & Elith, J. (2006) Error and uncertainty in habitat models. *Journal of Applied Ecology*, **43**, 413-423.

Beaumont, L.J. & Hughes, L. (2002) Potential changes in the distributions of latitudinally restricted Australian butterfly species in response to climate change. *Global Change Biology*, **8**, 954-971.

Benito De Pando, B., Peñas de Giles, J. & e-libro Corp (2007) Aplicación de modelos de distribución de especies a la conservación de la biodiversidad en el sureste de la Península Ibérica. In: *E-Libro*, pp. 100-119 p. Asociación de Geógrafos Españoles,, Madrid.

Benito Garzón, M., Sánchez de Dios, R. & Sáinz Ollero, H. (2007) Predictive modelling of tree species distributions on the Iberian Peninsula during the Last Glacial Maximum and Mid-Holocene. *Ecography*, **30**, 120-134.

Benito Garzón, M., Sánchez de Dios, R. & Sainz Ollero, H. (2008) Effects of climate change on the distribution of Iberian tree species. *Applied Vegetation Science*, **11**, 169-178.

Benito Garzón, M., Maldonado Ruiz, J., Sánchez de Dios, R. & Sainz Ollero, H. (2003) Predicción de la potencialidad de los bosques esclerófilos españoles mediante redes neuronales artificiales. In: *Graellsia*, pp. 345–358. CSIC

Benito Garzón, M., Blazek, R., Neteler, M., Dios, R.S.d., Ollero, H.S. & Furlanello, C. (2006) Predicting habitat suitability with machine learning models: The potential area of Pinus sylvestris L. in the Iberian Peninsula. *Ecological Modelling*, **197**, 383-393.

Brasier, C.M. (1996) Phytophthora cinnamomi and oak decline in southern Europe. Environmental constraints including climate change. *Annales Des Sciences Forestieres*, **53**, 347-358.

Brasier, C.M., Robredo, F. & Ferraz, J.F.P. (1993) Evidence for Phytophthora cinnamomi involvement in Iberian oak decline. *Plant Pathology*, **42**, 140-145.

Breiman, L. (2001) Random forests. *Machine Learning*, **45**, 5-32.

Bugalho, M.N., Caldeira, M.C., Pereira, J.S., Aronson, J. & Pausas, J.G. (2011) Mediterranean cork oak savannas require human use to sustain biodiversity and ecosystem services. *Frontiers in Ecology and the Environment*, **9**, 278-286.

Buisson, L., Thuiller, W., Casajus, N., Lek, S. & Grenouillet, G. (2010) Uncertainty in ensemble forecasting of species distribution. *Global Change Biology*, **16**, 1145-1157.

Bureau of Meteorology (2013) *Climate and past weather*. Available at: http://www.bom.gov.au/climate/ (accessed 13 October 2012).

Burgess, T.I., Webster, J.L., Ciampini, J.A., White, D.W., Hardy, G.E.S.J. & Stuckley, M.J.C. (2009) Re-evaluation of Phytophthora species isolated during 30 years of vegetation health surveys in Western Australia using molecular techniques. *Plant Disease*, **3**, 215-223.

Busby, J.R. (1991) BIOCLIM - a bioclimatic analysis and prediction system. *Nature Conservation: cost effective biological surveys and data analysis*, pp. 64-68.

Camarero, J.J., Lloret, F., Corcuera, L., Peñuelas, J. & Gil-Pelegrín, E. (2009) Cambio global y decaimiento del bosque. *Ecología del bosque mediterráneo en un mundo cambiante.* (ed. by F. Valladares Ros), p. 27. Organismo Autónomo Parques Nacionales, Madrid.

Carrasco, A., Fernández Cancio, A., Trapero Casas, A., López Pantoja, G., Sánchez Osorio, I., Ruiz Navarro, J., Jiménez Molina, J., Domínguez Nevado, L., Romero Martín, M.A., Carbonero Muñoz, M.D., Sánchez Hernández, M.E., Lucas Caetano, P.A., Gil Hernández, P., Fernández Rebollo, P., Navarro Cerrillo, R.M., Sánchez de la Cuesta, R., Raposo Llobet, R. & Rodríguez Reviriego, S. (2009) *Procesos de Decaimiento Forestal (la Seca). Situación del Conocimiento.* Consejería de Medio Ambiente, Junta de Andalucía, Córdoba.

Colautti, R.I. & MacIsaac, H.J. (2004) A neutral terminology to define 'invasive' species. *Diversity and Distributions*, **10**, 135-141.

Consejería de Medio Ambiente (2010) Estudio de procesos de mortalidad de pinares (sierra de Filabres y Baza - Almería) y especies del genero *Quercus* en Andalucía. In: (ed. Red Andaluza De Seguimiento De Daños Y Red De Alerta Fitosanitaria Forestal. Presencia De Phytophthora Cinnamomi Y Pythium Spiculum En Puntos Con Presencia De Quercus En Andalucía), p. 19. Consejería de Medio Ambiente de la Junta de Andalucía y Universidad de Córdoba (UCO), Córdoba.

Costa, J.C., Martín, A., Fernández, R. & Estirado, M. (2006) *Dehesas de Andalucía. Caracterización ambiental.* Consejería de Medio Ambiente, Junta de Andalucía, Sevilla, Spain.

da Silva, P.M., Aguiar, C.A.S., Niemela, J., Sousa, J.P. & Serrano, A.R.M. (2009) Cork-oak woodlands as key-habitats for biodiversity conservation in Mediterranean landscapes: a case study using rove and ground beetles (Coleoptera: Staphylinidae, Carabidae). *Biodiversity and Conservation*, **18**, 605-619.

De Maesschalck, R., Jouan-Rimbaud, D. & Massart, D.L. (2000) The Mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems*, **50**, 1-18.

Dell, B., Hardy, G.E.S.J. & Vear, K. (2005) History of Phytophthora cinnamomi management in Western Australia. *A Forest Conscienceness: Proceedings 6th National Conference of the Australian Forest History Society* (ed. by M.S. Publishers), pp. 391-406. Millpress Science Publishers, Rotterdam

Díaz, M., Pulido, F.J. & Marañón, T. (2003) Diversidad biológica y sostenibilidad ecológica y económica de los sistemas adehesados. In: *Revista Ecosistemas*. Sociedad Española de Ecología Terrestre, 2003/3.

Droogers, P. (2000) Estimating actual evapotranspiration using a detailed agro-hydrological model. *Journal of Hydrology*, **229**, 50-58.

Edwards, T.C., Cutler, D.R., Zimmermann, N.E., Geiser, L. & Moisen, G.G. (2006) Effects of sample survey design on the accuracy of classification tree models in species distribution models. *Ecological Modelling*, **199**, 132-141.

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.

Elith, J., Leathwick, J.R. & Hastie, T. (2008) A working guide to boosted regression trees. *Journal of Animal Ecology*, **77**, 802-813.

Elith, J., H. Graham, C., P. Anderson, R., Dudík, M., Ferrier, S., Guisan, A., J. Hijmans, R., Huettmann, F., R. Leathwick, J., Lehmann, A., Li, J., G. Lohmann, L., A. Loiselle, B., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., McC. M. Overton, J., Townsend Peterson, A., J. Phillips, S., Richardson, K., Scachetti-Pereira, R., E. Schapire, R., Soberón, J., Williams, S., S. Wisz, M. & E. Zimmermann, N. (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, **29**, 129-151.

FAO & IIASA (2000) Global agro-ecological zones. In: (ed. Fao and Iiasa 2007), Mapping biophysical factors that influence agricultural production and rural vulnerability, by H. von Velthuizen et al.

Farber, O. & Kadmon, R. (2003) Assessment of alternative approaches for bioclimatic modeling with special emphasis on the Mahalanobis distance. *Ecological Modelling*, **160**, 115-130.

Ficetola, G.F., Thuiller, W. & Miaud, C. (2007) Prediction and validation of the potential global distribution of a problematic alien invasive species—the American bullfrog. *Diversity and Distributions*, **13**, 476-485.

Filzmoser, P. (2004) A multivariate outlier detection method. *Proceedings of the Seventh International Conference on Computer Data Analysis and Modeling* (ed by, pp. 18-22.

Franklin, J. (2009) *Mapping species distributions: spatial inference and prediction*. Cambridge University Press, New York (U.S.A).

Freeman, E.A. & Moisen, G. (2008) *PresenceAbsence: An R Package for Presence Absence Analysis*.

Friedman, J.H. (1991) Multivariate Adaptive Regression Splines. *Annals of Statistics*, **19**, 1-67.

Friedman, J.H. (2001) Greedy function approximation: a gradient boosting machine.(English summary). *Ann. Statist*, **29**, 1189-1232.

Gallego, F.J., de Algaba, A.P. & Fernandez-Escobar, R. (1999) Etiology of oak decline in Spain. *European Journal of Forest Pathology*, **29**, 17-27.

Gaston, A. & Garcia-Vinas, J.I. (2011) Modelling species distributions with penalised logistic regressions: A comparison with maximum entropy models. *Ecological Modelling*, **222**, 2037-2041.

Gil Pelegrín, E., Peguero Pina, J.J., Camarero, J.J., Fernández Cancio, A. & Navarro Cerrillo, R.M. (2008) Drought and Forest Decline in the Iberian Peninsula: A Simple Explanation for a Complex Phenomenon? *Droughts: Causes, Effects and Predictions* (ed. by E.Y.P.P. Gil Pelegrín, J.J. Y Camarero, J.J. Y Fernández

Cancio, Angel Y Navarro Cerrillo, R.). Nova Science Publishers Inc, Madrid.

Global Invasive Species Database (2005) *Phytophthora cinnamomi Rands*. Available at: http://www.issg.org/database/species/ecology.asp?si=143&fr=1&sts=&lang=EN (accessed 6th February 2013).

Gomez-Aparicio, L., Ibanez, B., Serrano, M.S., De Vita, P., Avila, J.M., Perez-Ramos, I.M., Garcia, L.V., Sanchez, M.E. & Maranon, T. (2012) Spatial patterns of soil pathogens in declining Mediterranean forests: implications for tree species regeneration. *New Phytologist*, **194**, 1014-1024.

Graham, M.H. (2003) Confronting multicollinearity in ecological multiple regression. *Ecology*, **84**, 2809-2815.

Gromping, U. (2009) Variable Importance Assessment in Regression: Linear Regression versus Random Forest. *American Statistician*, **63**, 308-319.

Guisan, A. & Thuiller, W. (2005) Predicting species distribution: offering more than simple habitat models. *Ecology Letters*, **8**, 993-1009.

Guisan, A., Edwards Jr, T.C. & Hastie, T. (2002) Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecological Modelling*, **157**, 89-100.

Hardham, A.R. (2005) Phytophthora cinnamomi. *Molecular Plant Pathology*, **6**, 589-604.

Heikkinen, R.K., Marmion, M. & Luoto, M. (2012) Does the interpolation accuracy of species distribution models come at the expense of transferability? *Ecography*, **35**, 276-288.

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**, 751-777.

Hirzel, A. & Guisan, A. (2002) Which is the optimal sampling strategy for habitat suitability modelling. *Ecological Modelling*, **157**, 331-341.

Initiative, S.A.E. & Gole, C. (2006) *The Southwest Australia Ecoregion: Jewel of the Australian Continent*. Southwest Australia Ecoregion Initiative.

IUCN (2012) *IUCN Red List of Threatened Species Version 2012.1*. Available at: www.iucnredlist.org (accessed 17 July 2012).

Jeffers, S.N. & Martin, S.B. (1986) Comparison of two media selective for Phytophthora and Pythium species. *Plant Disease*, **70**, 1038-1043.

Jiménez-Valverde, A., Peterson, A., Soberón, J., Overton, J., Aragón, P. & Lobo, J. (2011) Use of niche models in invasive species risk assessments. *Biological Invasions*, **13**, 2785-2797.

Johnson, J.B. & Omland, K.S. (2004) Model selection in ecology and evolution. *Trends in ecology & evolution (Personal edition)*, **19**, 101-108.

Junta de Andalucía (2007) *Red de Información Ambiental de Andalucía. (REDIAM)*. Available at: http://www.juntadeandalucia.es/medioambiente/site/web/rediam (accessed 17 July 2012).

Kelly, M., Guo, Q., Liu, D. & Shaari, D. (2007) Modeling the risk for a new invasive forest disease in the United States: An evaluation of five environmental niche models. *Computers, Environment and Urban Systems*, **31**, 689-710.

Lek, S. & Guegan, J.F. (1999) Artificial neural networks as a tool in ecological modelling, an introduction. *Ecological Modelling*, **120**, 65-73.

Lindner, M., Maroschek, M., Netherer, S., Kremer, A., Barbati, A., Garcia-Gonzalo, J., Seidl, R., Delzon, S., Corona, P., Kolström, M., Lexer, M.J. & Marchetti, M. (2010) Climate change impacts, adaptive capacity, and vulnerability of European forest ecosystems. *Forest Ecology and Management*, **259**, 698-709.

Lobo, J.M., Jiménez-Valverde, A. & Real, R. (2008) AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, **17**, 145-151.

Lowe, S., Browne, M., Boudjelas, S. & De Poorter, M. (2000) *100 of the world's worst invasive alien species: a selection from the global invasive species database*. Invasive Species Specialist Group Auckland, New Zealand.

MAGRAMA (2007) *Tercer Inventario Forestal Nacional (IFN_3)*. Available at: http://www.magrama.gob.es/es/biodiversidad/servicios/banco-datos-naturaleza/informacion-disponible/ifn3.aspx (accessed 20 July 2012).

MAGRAMA (2013) *Raster*. Available at: http://servicios2.magrama.es/sia/visualizacion/descargas/capas.jsp (accessed 02 January 2013).

Maranon, T., Ajbilou, R., Ojeda, F. & Arroyo, J. (1999) Biodiversity of woody species in oak woodlands of southern Spain and northern Morocco. *Forest Ecology and Management*, **115**, 147-156.

Mateo, R.G., Croat, T.B., Felicísimo, Á.M. & Muñoz, J. (2010) Profile or group discriminative techniques? Generating reliable species distribution models using pseudo-absences and target-

group absences from natural history collections. *Diversity and Distributions*, **16**, 84-94.

Meentemeyer, R., Rizzo, D., Mark, W. & Lotz, E. (2004) Mapping the risk of establishment and spread of sudden oak death in California. *Forest Ecology and Management*, **200**, 195-214.

Moreira, A.C. & Martins, J.M.S. (2005) Influence of site factors on the impact of Phytophthora cinnamomi in cork oak stands in Portugal. *Forest Pathology*, **35**, 145-162.

Moreira, J.M. (2008) El Cambio Climático en Andalucía. Escenarios actuales y futuros del clima. In: *Medioambiente*, pp. 33 - 39. Consejería Medio Ambiente. Junta de Andalucía, Sevilla.

Myers, N., Mittermeier, R.A., Mittermeier, C.G., da Fonseca, G.A. & Kent, J. (2000) Biodiversity hotspots for conservation priorities. *Nature*, **403**, 853-8.

Naimi, B. (2013) *usdm: Uncertainty analysis for species distribution models*. R package version 1.1-8.

Olea, L. & San Miguel-Ayanz, A. (2006) The Spanish dehesa, a traditional Mediterranean silvopastoral system. In: *21st General Meeting of the European Grassland Federation*, Badajoz (Spain).

Olson, D.M. & Dinerstein, E. (2002) The Global 200: Priority Ecoregions for Global Conservation. In, pp. 199-224. Missouri Botanical Garden Press, Annals of the Missouri Botanical Garden.

Peel, M.C., Finlayson, B.L. & McMahon, T.A. (2007) Updated world map of the Köppen-Geiger climate classification. *Hydrology and Earth System Sciences Discussions Discussions*, **4**, 439-473.

Phillips, S.J. & Dudík, M. (2008) Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography*, **31**, 161-175.

Phillips, S.J., Anderson, R.P. & Schapire, R.E. (2006) Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, **190**, 231-259.

Prasad, A., Iverson, L. & Liaw, A. (2006) Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction. *Ecosystems*, **9**, 181-199.

Quinn, G.G.P. & Keough, M.J. (2002) *Experimental design and data analysis for biologists*. Cambridge University Press.

R Core Team (2012) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.

Randin, C.F., Dirnböck, T., Dullinger, S., Zimmermann, N.E., Zappa, M. & Guisan, A. (2006) Are niche-based species distribution models transferable in space? *Journal of Biogeography*, **33**, 1689-1703.

Rands, R.D. (1922) *Streepkanker van kaneel, veroorzaakt door phytophthora cinnamomi n. sp. (Stripe canker of cinnamon, caused by Phytophthora cinnamomi n. sp.)*. Drukkerij Ruygrok & Co., Batavia.

Red Natura 2000 (2011) *BIODEHESA. Desarrollo de Políticas y Herramientas para la Gestión y Conservación de la Biodiversidad.* Available at: http://www.rednatura2000.info/index.php?option=com_content&view=article&id=1329:life-biodehesa-&catid=80:life (accessed 16 May 2013).

Reuter, C. (2005) *Phytophthora cinnamomi Rands* Available at: http://www.cals.ncsu.edu/course/pp728/cinnamomi/p_cinnamomi.htm (accessed 6th February 2013).

Ridgeway, G. (1999) The state of boosting. *Computing Science and Statistics*, 172-181.

Romero de los Reyes, E., Navarro Cerrillo, R.M. & García - Ferrer Porras, A. (2007) Aplicación de ortofotos para la estimación de pérdidas de individuos en dehesas de encinas (*Quercus ilex* L. subps. *ballota* (Desf.) Samp.) afectados por processos de decaimiento. *Boletín de Sanidad Vegetal - Plagas*, **33**, 121 - 134.

Sánchez, M.E., Caetano, P., Ferraz, J. & Trapero, A. (2002) Phytophthora disease of Quercus ilex in south-western Spain. *Forest Pathology*, **32**, 5-18.

Sánchez, M.E., Sánchez, J.E., Navarro Cerrillo, R.M., Fernández, P. & Trapero, A. (2003) Incidencia de la podredumbre radical causada por Phytophthora cinnamomi en masas de Quercus en Andalucía. *Boletín de Sanidad Vegetal - Plagas*, **29**, 87-108.

Scarascia-Mugnozza, G., Oswald, H., Piussi, P. & Radoglou, K. (2000) Forests of the Mediterranean region: gaps in knowledge and research needs. *Forest Ecology and Management*, **132**, 97-109.

Shearer, B., Crane, C. & Cochrane, A. (2004) Quantification of the susceptibility of the native flora of the South-West Botanical Province, Western Australia, to Phytophthora cinnamomi. *Australian Journal of Botany*, **52**, 435-443.

Shearer, B., Crane, C. & Dunne, C. (2012) Variation in vegetation cover between shrubland, woodland and forest biomes invaded by Phytophthora cinnamomi. *Australasian Plant Pathology*, **41**, 413-424.

Shearer, B.L., Crane, C.E., Barrett, S. & Cochrane, A. (2007) Phytophthora cinnamomi invasion, a major threatening process to conservation of flora diversity in the South-west Botanical Province of Western Australia. *Australian Journal of Botany*, **55**, 225-238.

Soberón, J. & Nakamura, M. (2009) Niches and distributional areas: Concepts, methods, and assumptions. *Proceedings of the National Academy of Sciences*, **106**, 19644-19650.

Stukely, M., Webster, J. & Ciampini, J. (2012) Results of Phytophthora sample testing. In, Vegetation Health Service: Annual Report 2011-2012 Phytophthora detection.

Syphard, A.D. & Franklin, J. (2009) Differences in spatial predictions among species distribution modeling methods vary with species traits and environmental predictors. *Ecography*, **32**, 907-918.

Thuiller, W. (2003) BIOMOD - optimizing predictions of species distributions and projecting potential future shifts under global change. *Global Change Biology*, **9**, 1353-1362.

Thuiller, W. (2004) Patterns and uncertainties of species' range shifts under climate change. *Global Change Biology*, **10**, 2020-2027.

Thuiller, W., Araújo, M.B. & Lavorel, S. (2003) Generalized models vs. classification tree analysis: Predicting spatial distributions of plant species at different scales. *Journal of Vegetation Science*, **14**, 669-680.

Thuiller, W., Georges, D. & Engler, R. (2013) *biomod2: Ensemble platform for species distribution modeling*. R package version 2.0.3/r539.

Thuiller, W., Lafourcade, B., Engler, R. & Araujo, M.B. (2009) BIOMOD - a platform for ensemble forecasting of species distributions. *Ecography*, **32**, 369-373.

Thuiller, W., Albert, C., Araujo, M.B., Berry, P.M., Cabeza, M., Guisan, A., Hickler, T., Midgely, G.F., Paterson, J., Schurr, F.M., Sykes, M.T. & Zimmermann, N.E. (2008) Predicting global change impacts on plant species' distributions: Future challenges. *Perspectives in Plant Ecology Evolution and Systematics*, **9**, 137-152.

Trevor, H., Robert, T. & Andreas, B. (1994) Flexible Discriminant Analysis by Optimal Scoring. *Journal of the American Statistical Association*, **89**, 1255-1270.

Tucker, C.J., Pinzon, J.E. & Brown, M.E. (2004) Global Inventory Modeling and Mapping Studies. In: eds. (Global Land Cover Facility and University of Maryland), College Park, Maryland.

Tuset, J.J., Hinarejos, C., Mira, J.L. & Cobos, J.M. (1996) Implicación de Phytophthora cinnamomi Rands en la enfermedad de la «seca» de encinas y alcornoques *Boletín de Sanidad Vegetal - Plagas,* , **22**, 491-499.

Universidad de Extremadura (2012) *Servicio de Cartografía Digital e IDE (SECAD)*. Available at: http://ide.unex.es/geonetwork/srv/en/main.home (accessed 15 July 2012).

University of East Anglia Climatic Research Unit (CRU), [Phil Jones & Ian Harris] (2008) *CRU Time Series (TS) high resolution gridded datasets,*. Available at: http://badc.nerc.ac.uk/view/badc.nerc.ac.uk_ATOM_dataent_1256223773328276 (accessed 11 March 2013).

Václavík, T. & Meentemeyer, R.K. (2009) Invasive species distribution modeling (iSDM): Are absence data and dispersal constraints needed to predict actual distributions? *Ecological Modelling*, **220**, 3248-3258.

Václavík, T., Kanaskie, A., Hansen, E.M., Ohmann, J.L. & Meentemeyer, R.K. (2010) Predicting potential and actual distribution of sudden oak death in Oregon: Prioritizing landscape contexts for early detection and eradication of disease outbreaks. *Forest Ecology and Management*, **260**, 1026-1035.

van Gils, H., Conti, F., Ciaschetti, G. & Westinga, E. (2012) Fine resolution distribution modelling of endemics in Majella National Park, Central Italy. *Plant Biosystems - An International Journal Dealing with all Aspects of Plant Biology*, 1-12.

Vayssieres, M.P., Plant, R.E. & Allen-Diaz, B.H. (2000) Classification trees: An alternative non-parametric approach for predicting species distributions. *Journal of Vegetation Science*, **11**, 679-694.

Vogiatzakis, I.N., Mannion, A.M. & Griffiths, G.H. (2006) Mediterranean ecosystems: problems and tools for conservation. *Progress in Physical Geography*, **30**, 175-200.

Wenger, S.J. & Olden, J.D. (2012) Assessing transferability of ecological models: an underappreciated aspect of statistical validation. *Methods in Ecology and Evolution*, **3**, 260-267.

Weste, G. & Marks, G.C. (1987) The biology of phytophthora cinnamomi in australasian forests. *Annual Review of Phytopathology*, **25**, 207-229.

Wiens, J.A., Stralberg, D., Jongsomjit, D., Howell, C.A. & Snyder, M.A. (2009) Niches, models, and climate change: Assessing the assumptions and uncertainties. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 19729-19736.

Wollan, A.K., Bakkestuen, V., Kauserud, H., Gulden, G. & Halvorsen, R. (2008) Modelling and predicting fungal distribution patterns using herbarium data. *Journal of Biogeography*, **35**, 2298-2310.

WWF (2012) *Cork Oak Landscapes*. Available at: http://mediterranean.panda.org/about/forests/cork/ (accessed 16 August 2012).

Zentmyer, G.A. (1988) Origin and distribution of four species of Phytophthora. *Transactions of the British Mycological Society*, **91**, 367-378.

Zhang, Q. & Zhang, X. (2012) Impacts of predictor variables and species models on simulating Tamarix ramosissima distribution in Tarim Basin, northwestern China. *Journal of Plant Ecology*,

Appendix I

Table 12. Multivariate outlier detection

Conflictive variables (SMD> χ^2_2)
asrad, cld_aut, cld_sum, cld_win, etp_sum, maxt_sum, cld_ann, etr_ann, mint_spr, sun_spr, sun_sum, cld_spr, meant_spr, mint_sum meant_aut, mint_aut, wet_win, meant_ann, nd_rain_spr, sun_win, etp_spr, nd_rain_win, wet_spr, meant_win, ndvi_sd_win, wet_aut, etp_ann, maxt_aut, etp_win
Similar Variables (SMD< χ^2_2)
aspect, dist_road, dist_water, elevation, etp_aut, etr_aut, etr_spr, etr_sum, etr_win, farmland, flowdir, forest, frs_ann, frs_aut, frs_spr, frs_sum, frs_win, grassland, lgp, maxt_ann, maxt_spr, maxt_win, meant_sum, mint_ann, mint_win, mr_vbf, nd_rain_ann, nd_rain_aut, nd_rain_sum, ndvi_av_ann, ndvi_av_aut, ndvi_av_spr, ndvi_av_sum, ndvi_av_win, ndvi_sd_aut, ndvi_sd_spr, ndvi_sd_ann, ndvi_sd_sum, rain_ann, rain_aut, rain_spr, rain_sum, rain_win, ro_ann, ro_aut, ro_spr, ro_sum, ro_win, schrubland, sink, slope, slope_length, sm_ann, sm_aut, sm_spr, sm_sum, sm_win, sparce, substr, sun_ann, sun_aut, urban, wet_ann, wet_sum

Table 13. Boxplot variable Classification

Similar means
Aspect, lgp, mr_vbf, ndvi_sd_ann, ndvi_sd_sum, ndvi_sd_ann, substr, forest, sparce, grassland, urban, farmland, schrubland,
Dissimilar means
asrad, cld_aut, cld_sum, cld_win, etp_sum, maxt_sum, cld_ann, etr_ann, mint_spr, sun_spr, sun_sum, cld_spr, meant_spr, mint_sum meant_aut, mint_aut, wet_win, meant_ann, nd_rain_spr, sun_win, etp_spr, nd_rain_win, wet_spr, meant_win, ndvi_sd_win, wet_aut, etp_ann, maxt_aut, etp_win, sm_ann, sm_aut, sm_spr, sm_sum, sm_win, sun_ann, sun_aut
Enveloping range
dist_road, dist_water, elevation, etp_aut, etr_aut, etr_spr, etr_sum, etr_winflowdir, forest, frs_ann, frs_aut, frs_spr, frs_sum, frs_win, maxt_ann, maxt_spr, maxt_win, meant_sum, mint_ann, mint_win, nd_rain_ann, nd_rain_aut, nd_rain_sum, ndvi_av_ann, ndvi_av_aut, ndvi_av_spr, ndvi_av_sum, ndvi_av_win, ndvi_sd_aut, ndvi_sd_spr, rain_ann, rain_aut, rain_spr, rain_sum, rain_win, ro_ann, ro_aut, ro_spr, ro_sum, ro_win, sink, slope, slope_length, , wet_ann, wet_sum

Table 14. Analysis of collinearity results in Andalusia.

Variables	VIF	Variables	VIF
aspect	1.016584	ndvi_av_spr	NaN
asrad	2.80241	ndvi_av_sum	NaN
cld_aut	4.416347	ndvi_av_win	NaN
cld_spr	3.075054	ndvi_sd_aut	1.473013
cld_sum	6.880399	ndvi_sd_spr	1.401879
cld_win	4.605479	ndvi_sd_sum	1.488448
dist_road	1.314359	ndvi_sd_win	1.446093
dist_water	1.110111	rain_spr	2.603423
elevation	4.58486	rain_sum	3.475938
etp_aut	5.028966	ro_aut	2.961508
etr_aut	3.189802	ro_spr	4.706019
etr_spr	2.464264	ro_sum	2.425971
etr_sum	2.047741	shrubland	1.244987
etr_win	8.021377	sink	4.530363
farmland	1.393689	slope	1.246836
flowdir	1.063668	slope_length	1.173909
frs_aut	5.977137	sm_aut	5.207821
frs_spr	5.509032	sm_spr	3.965157
frs_sum	2.437951	sm_sum	1.419309
grassland	1.056727	sm_win	2.149801
lgp	3.45744	sparce	1.017311
maxt_ann	7.668116	substr	2.004189
maxt_sum	3.972388	sun_sum	5.16054
meant_sum	5.138166	sun_win	5.415015
mint_sum	5.565177	urban	1.030116
mint_win	5.131128	wet_ann	4.787577
mr_vbf	1.077351	wet_aut	5.933211
nd_rain_sum	1.762096	wet_spr	3.698527
nd_rain_win	1.647782	wet_sum	2.156721
ndvi_av_ann	NaN	wet_win	4.214086
ndvi_av_aut	NaN		

Table 15. Analysis of collinearity results in Southwest Australia.

Variables	VIF	Variables	VIF
aspect	1.56265	nd_rain_spr	5.383356
asrad	2.208482	nd_rain_sum	6.583005
cld_win	7.408131	ndvi_av_ann	NaN
dist_road	2.433441	ndvi_av_aut	NaN
dist_water	4.920254	ndvi_av_spr	NaN
elevation	3.723748	ndvi_av_sum	NaN
etp_ann	4.389227	ndvi_av_win	NaN
etp_aut	4.613975	ndvi_sd_ann	1.632248

Variables	VIF	Variables	VIF
etp_spr	5.825692	ndvi_sd_aut	7.792278
etr_aut	5.781966	ndvi_sd_sum	7.193667
etr_win	8.060153	ndvi_sd_win	5.089221
flowdir	1.638937	rain_aut	3.157844
forest	4.872873	rain_spr	8.834589
frs_sum	3.727892	rain_sum	6.932753
frs_win	4.669365	ro_aut	NaN
grassland	3.007996	ro_spr	NaN
lgp	2.471522	ro_sum	NaN
maxt_ann	5.89776	ro_win	1.508466
maxt_aut	8.115661	shrubland	4.679568
maxt_spr	6.535298	sink	2.622465
maxt_sum	4.529628	slope	1.939261
meant_ann	7.524935	slope_length	1.589986
meant_aut	4.497451	sm_aut	2.048551
meant_spr	6.336262	sm_spr	9.94033
meant_sum	4.407796	sm_win	8.115136
meant_win	7.817776	sparce	1.988436
mint_ann	5.498913	substr	1.758227
mint_aut	6.149601	sun_aut	6.844044
mint_spr	6.8774	sun_spr	8.185003
mint_sum	8.199376	sun_win	5.746109
mint_win	5.433885	urban	2.868092
mr_vbf	1.574735	wet_aut	3.033147
nd_rain_aut	4.046758	wet_sum	6.146046

Table 16. Remnant common variables after collinearity analysis

N	Variables	Variables SW Aus.	Variables And.
1	aspect	aspect	aspect
2	asrad	asrad	asrad
3	cld_ann		
4	cld_aut		cld_aut
5	cld_spr		cld_spr
6	cld_sum		cld_sum
7	cld_win	cld_win	cld_win
8	dist_road	dist_road	dist_road
9	dist_water	dist_water	dist_water
10	elevation	elevation	elevation
11	etp_ann	etp_ann	
12	etp_aut	etp_aut	etp_aut
13	etp_spr	etp_spr	
14	etp_sum		
15	etp_win		
16	etr_ann		

N	Variables	Variables SW Aus.	Variables And.
17	etr_aut	etr_aut	etr_aut
18	etr_spr		etr_spr
19	etr_sum		etr_sum
20	etr_win	etr_win	etr_win
21	farmland		farmland
22	flowdir	flowdir	flowdir
23	forest	forest	
24	frs_ann		
25	frs_aut		frs_aut
26	frs_spr		frs_spr
27	frs_sum	frs_sum	frs_sum
28	frs_win	frs_win	
29	grassland	grassland	grasslands
30	lgp	lgp	lgp
31	maxt_ann	maxt_ann	maxt_ann
32	maxt_aut	maxt_aut	
33	maxt_spr	maxt_spr	
34	maxt_sum	maxt_sum	maxt_sum
35	maxt_win		
36	meant_ann	meant_ann	
37	meant_aut	meant_aut	
38	meant_spr	meant_spr	
39	meant_sum	meant_sum	meant_sum
40	meant_win	meant_win	
41	mint_ann	mint_ann	
42	mint_aut	mint_aut	
43	mint_spr	mint_spr	
44	mint_sum	mint_sum	mint_sum
45	mint_win	mint_win	mint_win
46	mr_vbf	mr_vbf	mr_vbf
47	nd_rain_ann		
48	nd_rain_aut	nd_rain_aut	
49	nd_rain_spr	nd_rain_spr	
50	nd_rain_sum	nd_rain_sum	nd_rain_sum
51	nd_rain_win		nd_rain_win
52	ndvi_av_ann	ndvi_av_ann	ndvi_av_ann
53	ndvi_av_aut	ndvi_av_aut	ndvi_av_aut
54	ndvi_av_spr	ndvi_av_spr	ndvi_av_spr
55	ndvi_av_sum	ndvi_av_sum	ndvi_av_sum
56	ndvi_av_win	ndvi_av_win	ndvi_av_win
57	ndvi_sd_ann	ndvi_sd_ann	
58	ndvi_sd_aut	ndvi_sd_aut	ndvi_sd_aut
59	ndvi_sd_spr		ndvi_sd_spr
60	ndvi_sd_sum	ndvi_sd_sum	ndvi_sd_sum

N	Variables	Variables SW Aus.	Variables And.
61	ndvi_sd_win	ndvi_sd_win	ndvi_sd_win
62	rain_ann		
63	rain_aut	rain_aut	
64	rain_spr	rain_spr	rain_spr
65	rain_sum	rain_sum	rain_sum
66	rain_win		
67	ro_ann		
68	ro_aut	ro_aut	ro_aut
69	ro_spr	ro_spr	ro_spr
70	ro_sum	ro_sum	ro_sum
71	ro_win	ro_win	
72	schrubland	schrubland	schrubland
73	sink	sink	sink
74	slope	slope	slope
75	slope_length	slope_length	slope_length
76	sm_ann		
77	sm_aut	sm_aut	sm_aut
78	sm_spr	sm_spr	sm_spr
79	sm_sum		sm_sum
80	sm_win	sm_win	sm_win
81	sparce	sparce	sparce
82	substr	substr	substr
83	sun_ann		
84	sun_aut	sun_aut	
85	sun_spr	sun_spr	
86	sun_sum		sun_sum
87	sun_win	sun_win	sun_win
88	urban	urban	urban
89	wet_ann		wet_ann
90	wet_aut	wet_aut	wet_aut
91	wet_spr		wet_spr
92	wet_sum	wet_sum	wet_sum
93	wet_win		wet_win

Table 17. Final variable datasets in the first test.

1st test							
Model	MaxEnt	**AUC**	0.747	**Model**	MaxEnt	**AUC**	0.868
Variables SW Australia				**Variables Andalusia**			
aspect				asrad			
dist_water				dist_water			
elevation				elevation			
etr_aut				etp_aut			
flowdir				frs_sum			
meant_sum				maxt_sum			
mint_win				meant_sum			
rain_spr				mint_win			
slope				sun_win			
wet_aut				wet_aut			

Table 18. Final variable datasets in the second test.

2nd test (RF)							
Model	RF	**AUC**	0.842	**Model**	MaxEnt	**AUC**	0.903
Variables SW Australia				**Variables Andalusia**			
aspect				asrad			
dist_road				dist_water			
elevation				elevation			
maxt_ann				maxt_sum			
maxt_sum				meant_sum			
meant_sum				mint_sum			
mint_sum				nd_rain_sum			
nd_rain_sum				ro_spr			
ndvi_sd_sum				ro_sum			
slope				sun_win			

Table 19. Final variable datasets in the third test.

3rd test (BRT)							
Model	BRT	**AUC**	0.852	**Model**	BRT	**AUC**	0.890
Variables SW Australia				**Variables Andalusia**			
cld_win				aspect			
dist_road				dist_road			
dist_water				dist_water			
elevation				elevation			
etr_aut				meant_sum			
meant_sum				mint_sum			
ndvi_sd_sum				ro_spr			
slope				ro_sum			
sm_aut				slope_length			
wet_aut				sun_win			

Table 20. Final variable datasets in the forth test

4th test (MaxEnt)							
Model	MaxEnt	**AUC**	0.901	**Model**	MaxEnt	**AUC**	0.913
Variables SW Australia				**Variables Andalusia**			
cld_win				dist_water			
dist_water				elevation			
elevation				meant_sum			
etr_aut				mint_win			
frs_sum				ndvi_sd_sum			
meant_sum				ro_aut			
ndvi_av_win				ro_spr			
ndvi_sd_sum				slope			
sm_aut				slope_length			
wet_aut				wet_aut			

Table 21. Initial variables dataset in the fifth test

Fifth Test's Initial Variable Dataset	
Variables SW Australia	**Variables Andalusia**
aspect	aspect
cld_win	asrad
dist_road	dist_road
dist_water	dist_water
elevation	elevation
etr_aut	etp_aut
flowdir	frs_sum
maxt_ann	maxt_sum
maxt_sum	meant_sum
meant_sum	mint_sum
mint_sum	mint_win
mint_win	nd_rain_sum
nd_rain_sum	ndvi_sd_sum
ndvi_sd_sum	ro_aut
rain_spr	ro_spr
slope	ro_sum
sm_aut	slope
wet_aut	slope_lenght
	sun_win
	wet_aut

Table 22. Final variable datasets in the fifth test

5th test (Combine)							
Model	MaxEnt	**AUC**	0.830	**Model**	MaxEnt	**AUC**	0.921
Variables SW Australia				**Variables Andalusia**			
aspect				dist_water			
dist_water				elevation			
elevation				meant_sum			
etr_aut				mint_win			
flowdir				ndvi_sd_sum			
meant_sum				ro_aut			
ndvi_sd_sum				ro_spr			
rain_spr				slope			
slope				slope_lenght			
wet_aut				wet_aut			

Table 23. Final variable datasets in the final test

Final test	
Final Variables Dataset	
aspect	ndvi_sd_sum
dist_water	rain_spr
elevation	ro_aut
etr_aut	ro_spr
flowdir	slope
maxt_sum	slope_lenght
meant_sum	wet_aut
mint_win	

Table 24. Final 10 variable datasets in the final test

6th test (FINAL 10 Variables)							
SW AUSTRALIA				**ANDALUSIA**			
Model	MARS	**AUC**	0.822	**Model**	BRT	**AUC**	0.925
Final Variables Dataset							
aspect				dist_water			
dist_water				elevation			
elevation				meant_sum			
etr_aut				mint_win			
flowdir				ndvi_sd_sum			
maxt_sum				ro_aut			
meant_sum				ro_spr			
ndvi_sd_sum				slope			
slope				slope_length			
wet_aut				wet_aut			

Table 25. Andalusia models. Sensitivity and specificity

Models	Andalusia			
	Individual dataset		**Common dataset**	
	Sensitivity	*Specificity*	*Sensitivity*	*Specificity*
ANN	75.86	91.67	86.21	91.67
BRT	82.76	94.44	68.97	100.00
CTA	75.86	86.11	79.31	83.33
FDA	82.76	86.11	82.76	91.67
GAM	68.97	83.33	89.66	83.33
GLM	79.31	94.44	86.21	83.33
MARS	86.21	80.56	79.31	88.89
MAXENT	86.21	86.11	100.00	75.00
RF	86.21	80.56	86.21	83.33
SER	82.76	63.89	93.10	61.11

Table 26. SW Australian models. Sensitivity and specificity

Models	SW Australia			
	Individual dataset		**Common dataset**	
	Sensitivity	*Specificity*	*Sensitivity*	*Specificity*
ANN	53.57	60.00	75.00	77.14
BRT	78.57	85.71	82.14	71.43
CTA	75.00	85.71	82.14	65.71
FDA	89.29	85.71	53.57	91.43
GAM	78.57	68.57	60.71	77.14
GLM	92.86	45.71	75.00	71.43
MARS	96.43	45.71	78.57	71.43
MAXENT	71.43	82.86	75.00	94.29
RF	75.00	80.00	89.29	57.14
SRE	89.29	45.71	57.14	71.43