Plant & Food RESEARCH
RANGAHAU AHUMĀRA KAI

www.plantandfood.com

Habitat suitability predictions for selected glasshouse biological control agents using Maxent and Multi Modelling

Logan DP, Senay SD, Narouei Khandan HA

March 2013

A confidential report prepared for

Tomatoes NZ

Logan DP
Plant & Food Research, Te Puke

Senay SD, Narouei Khandan HA
Lincoln University, Lincoln

SPTS No.8061

PUBLICATION DATA

Logan DP, Senay SD, Narouei Khandan HA. March 2013. Habitat suitability predictions for selected glasshouse biological control agents using Maxent and Multi Modelling. A report prepared for: Tomatoes NZ. Plant & Food Research Data: 51721. Contract No. 29082. Job code: P/334023/01. SPTS No. 8061

This report has been prepared by The New Zealand Institute for Plant & Food Research Limited (Plant & Food Research), which has its Head Office at 120 Mt Albert Rd, Mt Albert, Auckland.

This report has been approved by:

David Logan
Scientist, Applied Entomology
Date: 20 March 2013

Louise Malone
Science Group Leader, Bioprotection
Date:20 March 2013

Contents

Executive summary

Maxent and Multi Modelling for selected glasshouse glasshouse biological control agents

Logan DP, Senay SD, Narouei Khandan HA, March 2013, PFR SPTS No. 8061

Three exotic biological control agents (BCAs) (*Delphastus catalinae* (Horn), *Macrolophus melanotoma* (Costa)/ *M. pygmaeus* (Rambur) and *Nesidiocoris tenuis* (Reuter)) are considered to have potential for control of pests in New Zealand glasshouse production and are the subject of an application to the Environmental Protection Agency (EPA). CLIMEX modelling indicated that *D. catalinae* may persist outside glasshouses only in the warmest localities while *Macrolophus melanotoma* / *M. pygmaeus* and *N. tenuis* had potential to persist outside glasshouses in Northland, Auckland and on the east coast of the North Island. Further modelling using other methods was suggested by EPA to clarify some of the uncertainty surrounding the CLIMEX projections. The objective of this study was to use correlative-type habitat distribution models to infer environmental requirements for the three BCAs based on geographical collection records and to generate maps of suitable habitat within New Zealand

The models or algorithms were Maxent, Logistic regression, Classification and Regression Trees, Conditional trees, Naive Bayes, K-nearest neighbour, Support Vector Machines, and Artificial Neural Networks. All algorithms except Maxent were implemented using the Multi Model program developed by the Ecological Informatics group, Bio-Protection Research Centre, Lincoln University. Environmental data were gridded data with a spatial resolution of 30 arcseconds (Ca. 309 m at the equator) available from WorldClim (www.worldclim.org/). The data consist of altitude (m.a.s.l.) and 19 biologically relevant variables derived from temperature and precipitation data (BIOCLIM variables). Models were trained on presence and absence points with New Zealand excluded. Since there were no data for absences, they were generated within a limited geographic range of the known distribution by a one-class support vector machine algorithm.

Results of Maxent modelling are presented separately from those of the other seven algorithms. In the case of Maxent, the sensitivity score (proportion of true positives that were predicted correctly) for each model was used to weight the model's results for a combined Multi Model or consensus map. Sensitivity was chosen as we considered that characterizing potentially suitable habitat was more important than characterizing unsuitable areas. In other words, the Multi Model is conservative, as false positives are more acceptable than false negatives when considering the risk that any of the three BCAs may establish outside glasshouses

A summary of the model projections for New Zealand areas is as follows:

Delphastus catalinae

- The Maxent model indicated that climate suitability is generally poor for *D. catalinae* in New Zealand (most values<0.5), with coastal areas particularly in Northland slightly more favourable than elsewhere.
- The consensus Multi Model predicts that Northland is relatively well suited climatically for *D. catalinae* (scores >0.7).

- In summary, the Maxent model indicates low likelihood that preferred climate for *D. catalinae* exists in New Zealand, consistent with results from CLIMEX modelling. The CLIMEX model indicated that only small areas of Northland are suitable for *D. catalinae.* The consensus Multi Model indicates that Northland may be suitable for *D. catalinae*. Part of the difference may merely reflect the relatively small number of weather stations used by CLIMEX compared with the high resolution gridded data used by the Multi Model and Maxent.

Macrolophus melanotoma / *M. pygmaeus*

- The Maxent model indicated that climate suitability is poor for *M. melanotoma* / *M. pygmaeus* in New Zealand (values<0.5).
- The consensus Multi Model indicated that only a small area of north of Kaitaia in Northland has suitable climate for *M. melanotoma* / *M. pygmaeus.*
- In summary, the Maxent and consensus Multi Model indicate a low likelihood that suitable climate for *Macrolophus melanotoma* / *M. pygmaeus* exists in New Zealand. In contrast, the CLIMEX model indicated that Northland and the east coast of the North Island contain suitable habitat for *M. melanotoma* / *M. pygmaeus.* Other less suitable areas occur on the west coast of the North Island, the upper South Island and parts of Banks Peninsula.

Nesidiocoris tenuis

- The Maxent model indicated that suitability of climate for *N. tenuis* is poor for all New Zealand (values <0.5).
- The consensus Multi Model indicated that large areas in the northern half of the South island and many areas of the North Island have relatively suitable climate conditions (scores >0.6) for *N. tenuis.*
- In summary, the consensus Multi Model and the CLIMEX model indicated that some areas of New Zealand have relatively suitable climate for *N. tenuis,* although these areas differed. In the consensus Multi Model case, suitable climate (areas with scores between 0.5 and 0.6) was predicted to exist in many areas of the North Island and the coastal areas of the Buller, Nelson, Kaikoura, and Canterbury regions. In the CLIMEX model case, suitable climate was predicted to exist in Northland and some coastal areas of the North Island. In contrast, Maxent modelling indicates that there is likely to be no suitable climate in New Zealand for *N. tenuis*.

There is disagreement between the projections of two of the three modelling approaches for *D. catalinae* and *M. melanotoma* / *M. pygmaeus* and among all three modelling approaches for *N. tenuis.*

The Multi Model method and Maxent rely on the geographical locations from which a species was collected to identify the most important features to model its preferred environment. As the collection data for each of the three BCAs were limited (n = 14 locations for *D. catalinae*, n = 23 for *Macrolophus* spp., and n = 30 for *N. tenuis*), the models may be inaccurate and caution is advised in interpreting the results. Correlative models work best when there are many collection data. Based on other studies, 50-100 points representative of the species range may significantly reduce the uncertainty associated with habitat distribution predictions. Resolution of the environmental data is also critical.

Limited collection data can also influence the development of a CLIMEX model and the interpretation of its output. However in CLIMEX, the model developer can combine collection

data with physiological data from laboratory experiments and sometimes qualitative data on distribution and abundance to select the most important environmental features and set their critical values. In the case of the three BCAs where some physiological data exist, CLIMEX results may be more reliable than Multi Model and Maxent model results. Note that all models are subject to continuing research to refine their performance, for example through more rigorous validation and sensitivity analyses.

For further information please contact:

David Logan
The New Zealand Institute for Plant & Food Research Ltd
Plant & Food Research Te Puke
412 No 1 Road
RD 2
Te Puke 3182
NEW ZEALAND
Tel: +64-7-928 9794
Fax: +64-7-928 9801
Email: david.logan@plantandfood.co.nz

1 Introduction

Three exotic biological control agents (BCAs) (a coccinellid beetle *Delphastus catalinae* (Horn), and two mirid bugs *Macrolophus melanotoma* (Costa)/ *M. pygmaeus* (Rambur) and *Nesidiocoris tenuis* (Reuter)) are considered to have potential for control of pests in New Zealand glasshouse production and are the subject of an application to the Environmental Protection Agency (EPA). *Macrolophus caliginosus* (Wagner) was named in the request for climate-matching as it is a glasshouse BCA in Europe. However, the taxon is a junior synonym of *M. melanotoma*. Furthermore there is uncertainty that *M. melanotoma* is the correct identity of the glasshouse BCA. Instead mtDNA analysis and some biological observation suggest that the correct taxon is *M. pygmaeus* (Logan 2012).

Incidence of the three BCAs outside glasshouses may have adverse effects on native fauna, as they are generalist predators. In this report the suitability of the New Zealand environment for the survival and persistence of the three BCAs was estimated using habitat distribution models. This report follows CLIMEX modelling reported by Logan (2012), which indicated that coastal areas of the North Island are potentially suitable for persistence of *M. melanotoma* / *M. pygmaeus* and *N. Tenuis,* while only limited areas in Northland were suitable for *D. catalinae*. Further modelling using other methods was requested by EPA because of some uncertainty surrounding CLIMEX projections. The objective of this study was to produce projections of the distribution of suitable areas in New Zealand for the three species using modelling methods other than CLIMEX.

2 Methods

2.1 General description

We used correlative-type habitat distribution models to infer environmental requirements for the three BCAs based on the known geographical collection records (Appendix 1, Tables A1-A3). There were 15 localities available for *D. catalinae*, 30 for *N. tenuis* and 23 for *M. melanotoma* / *M. pygmaeus*. The models or algorithms were Maxent (Phillips et al. 2006) and seven methods (Logistic regression, Classification and Regression Trees, Conditional trees, Naive Bayes, K-nearest neighbour, Support Vector Machines, and Artificial Neural Networks) implemented using a Multi Modelling framework developed by the Ecological Informatics group, Bio-Protection Research Centre, Lincoln University (Worner et al. 2010). Distribution data for each species were from the Global Biodiversity Information Facility (http://www.gbif.org/) and published literature as reported in Logan (2012). Environmental data were gridded data with a spatial resolution of 10 arcseconds (Ca. 309 m at the equator) available from WorldClim (www.worldclim.org/) (Hijmans et al. 2005). The data consist of altitude (m.a.s.l.) and 19 biologically relevant variables derived from temperature and precipitation data (BIOCLIM variables, Table 1).

Table 1. BIOCLIM variables.

Code	Variable
ALT	Altitude (m.a.s.l.)
BIO1	Annual Mean Temperature
BIO2	Mean Diurnal Range (Mean of monthly (max temp - min temp))
BIO3	Isothermality (BIO2/BIO7) (* 100)
BIO4	Temperature Seasonality (standard deviation *100)
BIO5	Max Temperature of Warmest Month
BIO6	Min Temperature of Coldest Month
BIO7	Temperature Annual Range (BIO5-BIO6)
BIO8	Mean Temperature of Wettest Quarter
BIO9	Mean Temperature of Driest Quarter
BIO10	Mean Temperature of Warmest Quarter
BIO11	Mean Temperature of Coldest Quarter
BIO12	Annual Precipitation
BIO13	Precipitation of Wettest Month
BIO14	Precipitation of Driest Month
BIO15	Precipitation Seasonality (Coefficient of Variation)
BIO16	Precipitation of Wettest Quarter
BIO17	Precipitation of Driest Quarter
BIO18	Precipitation of Warmest Quarter
BIO19	Precipitation of Coldest Quarter

2.2 Maxent

Maxent (Phillips et al. 2004; http://www.cs.princeton.edu/~schapire/maxent/) is a machine learning method that estimates a species' distribution by finding the distribution of maximum entropy, subject to some constraints based on existing knowledge. Entropy is a measure of dispersedness (Elith et al. 2011). Maxent starts with a uniform grid of probability values and then iteratively fits a model to maximise the probability of presences in relation to background data. It uses different types of features (climatic or other predictor variables) and a 'regularization' (smoothing) parameter for each of these features. There is an option to generate subsamples of background data to act as pseudo-absences, rather than using all background data and this achieves a shorter runtime than otherwise without reducing predictive ability (Phillips & Dudik 2008).. The model is validated using an independent test data set (Roura-Pascual et al. 2009). Cumulative values, as percentages, are used to show the predictions for each analysed cell (i.e. a probability value). The cell with a value of 100 is the most suitable, while cells close to 0 are the least suitable within the study area (Phillips et al. 2006).

In this study geographic co-ordinates for collection localities for each of the three BCAs were prepared as .csv files. All bioclimatic layers were used in the format of ESRI ASCII grids. We used version 3.3.3k from the command line. The main default settings were applied except that replicate number was fixed at 20 (instead of one) and maximum iterations were fixed at 1000. Replicated run type was used as the cross-validation option. We used the prepared New Zealand projection layer and clamping to omit grid cells where variables exceed the range represented in the training data.

Model performance was evaluated by the Area under the Curve (AUC) of the Receiver Operative Characteristic (ROC). AUC is calculated by plotting sensitivity against 1-specificity across the range of possible thresholds. Sensitivity is the proportion of true positives that were predicted correctly (i.e. TP/(TP+FN))(Table 2). Specificity is the proportion of true negatives that were predicted correctly (i.e. TN/(TN+FP).

Table 2. Confusion matrix.

		Predicted	
		Positive	Negative
Actual	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

Maxent can produce maps, bar charts of jack-knife results, response curves and tables of the relative contribution of the environmental variables to the final model as output. Here we have reported the ranked relative contribution tables and mapped the grid cell values in arcmap 10. Maxent calculates the contribution of each variable in the following way. To determine the first estimate, for each iteration of the training algorithm, the increase in regularised gain is added to the contribution of the corresponding variable, or subtracted from it if the change to the absolute value of lambda is negative. For the second estimate, for each environmental variable in turn, the values of that variable on training presence and background data are randomly permuted. The model is re-evaluated on the permuted data, and the resulting drop in training AUC (Area Under the ROC Curve) is shown in the table, normalised to percentage. Variable contributions

should be interpreted with caution when the predictor variables are correlated. Values shown are averages over replicate runs.

2.3 Multi Model

2.3.1 General method

The Multi Model program models species presence/absence data using seven different species distribution models (Logistic Regression (LOG), Naıve Bayes (NB), Decision Trees (CART), k-Nearest Neighbours (KNN), Support Vector Machines (SVM), and Artificial Neural Networks (ANN)) (Figure 1). All seven models were trained (or fitted) and tested using selected variables. In order to obtain the best set of parameters for some models such as KNN, SVM and ANN, initial parameterization was carried out followed by optimisation.

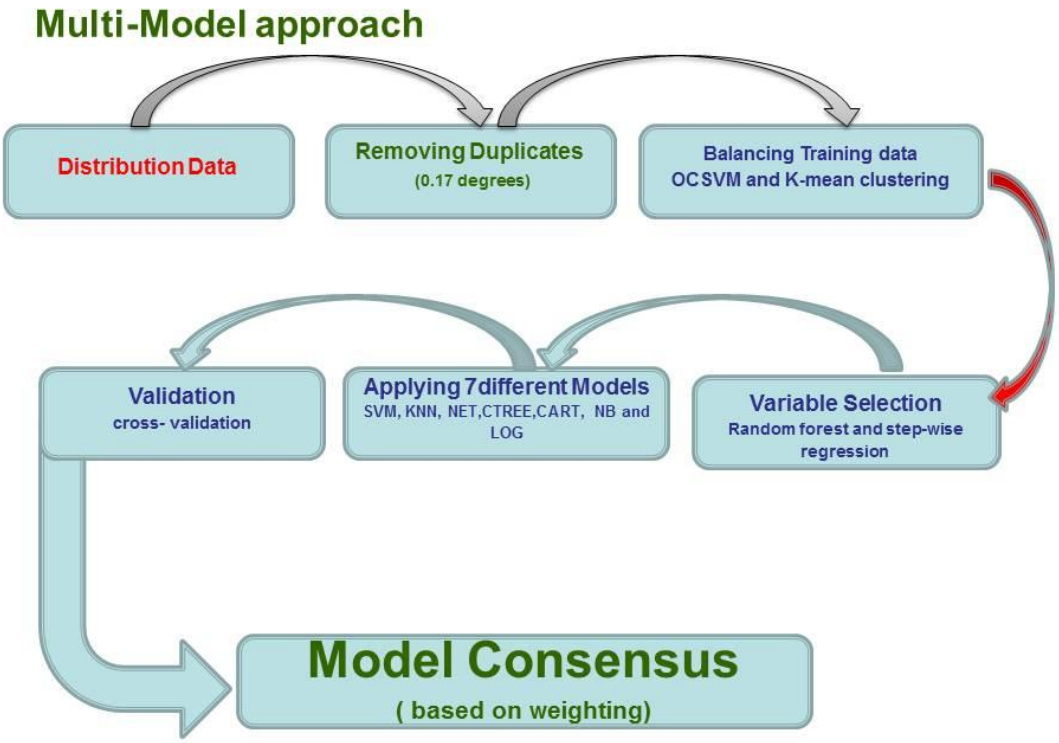


Figure 1: Flow diagram with main steps to reach a consensus model projection using the Multi Model program developed by the Ecological Informatics group, Bio-Protection Research Centre, Lincoln University.

2.3.2 Brief descriptions of algorithms

Logistic regression (LOG) is a generalisation of linear regression and used primarily for binary variables. Advantages include no distributional assumptions on the predictor variables, ease of implementation and interpretation, and ability to handle interactions between predictors, non-linear relationships and mixed continuous and discrete predictors.

Naïve Bayes (NB) is probably the most common Bayesian network model used in machine learning. In this model, the class variable Y is the root and the predictors X are leaves. The model is based on Bayes rule and it is naïve because it assumes that the attributes ($X_1 \ldots X_n$) are all conditionally independent one from another. Given the class (Y), a deterministic prediction

can be obtained by choosing the most likely class. It is easy to implement and can handle noisy data.

Classification and Regression Trees (CART) and Conditional Trees (CTREE) are types of decision trees. CARTs are built by an iterative process of splitting the data into two parts such that the sum of squared deviations from the mean is minimised at each split. CARTs are capable of handling a wide range of response variables (Speybroeck 2012). Small-sized trees are easy to interpret but may not have much predictive power. Judicious pruning is needed to limit over-fitting.

K-nearest neighbour (KNN) models assume that observations closest in the space of predictor variables will be close to each other in the space of the response variable. This method can be applied to both regression and classification problems. The advantages include simplicity and use of local information, resulting in highly adaptive behaviour.

Support Vector Machine (SVM) is a machine learning algorithm (Boser et al. 1992) used for binary classification problems. It treats objects to be classified as points in multi-dimensional space and selects a hyperplane that has a maximum margin (Noble 2006). The points on the margin of the hyperplane are called “Support Vectors” (Mahadevan & Shah 2009). Kernel functions are used to add additional dimensions to data to optimise classification. The SVM algorithm allows for some control of misclassification (i.e. setting parameters to describe a soft margin) with no effect on the final result.

Artificial Neural Networks (NNET) imitate the learning process of animal brains. A mesh of artificial neurons are organised in several linked layers, with the output of one layer used as the input of the next layer. Each layer filters information by amplifying or by reducing it. NNET has the ability to model complex non-linear relationships and detect interactions between predictor variables and so is suited to ecological systems (e.g. Watts & Worner 2008).

2.3.3 Balancing the data

One important step is to generate absence locations. Absences are virtually never recorded. Even if they are, it is difficult to be sure those absence locations are real, because in some cases there may have not been enough searching, or the area is suitable but not occupied because of historical accident or dispersal limitations (Phillips et al. 2009; Vaclavik & Meentemeyer 2009; Van der Wal et al. 2009). The problem becomes more complicated when one works with global data (in this case Worldclim data that consist of over 580,000 points) because there are a large number of absence locations that can cause class imbalance. Some researchers use a random selection of these absence locations to overcome the problem of class imbalance. The random selection of absence data, depending on the selected locations, can have a large effect on model output and consequently may result in inaccurate interpretation. To solve this problem, Multi Model uses one-class support vector machines (OCSVMs) to select appropriate absence points out of large datasets. This method has some benefits, such as model creation in a short computational time, accuracy, and ability to handle large datasets. There are few studies regarding application of OCSVMs in ecological studies (Guo et al. 2005; Zuo et al. 2008). Instead of selecting a single best-performing OCSVM model that can result in over-fitting the data in the model, a set (ensemble) of 100 models fitted to a different sample of the data, which had the lowest prediction errors, are created. The absence locations are those where the probability of their environmental suitability was 0 in all 100 models. As there are still many possible absence locations after this analysis, these points are

reduced by clustering absence locations that have similar environmental variables by k clusters, to balance the number of presence locations.

2.3.4 Variable selection

After selecting the balanced absence data, significant BIOCLIM variables were verified and selected using random forest (Breiman 2001) and stepwise regression analysis (Thompson 1995). A random forest is a classification method that uses many (≥1000) decision trees generated by a random selection of variables or features. Removing insignificant variables results in the improvement of the model fit, validity and computation time.

2.3.5 Model consensus

The goal of ensembling species distribution models is to find some consensus distribution from different algorithms. A major criticism is that it is not possible to average results obtained from models that have completely different algorithms and assumptions. At the same time, some decisions are too important or critical to be based on a single model. All models have their advantages and disadvantages. A method that can maximise prediction power by combining results of multiple models is desirable. Model consensus uses some prior criteria to weight individual model outputs for each location or grid cell and combine them into a single score to produce a final map (Marmion et al. 2009; Araujo & Peterson 2012). In our case we chose to use the sensitivity score of the Multi Models to weight the results of each model for the final consensus map. A weighted sum of the seven multiple model sensitivity scores was used to generate a probability layer of habitat suitability for each of the three species. Sensitivity (proportion of true positives that were predicted correctly in the training data) was chosen, as we considered that characterizing potentially suitable habitat was more important than characterizing unsuitable areas. In other words, the model is conservative, as false positives are more acceptable than false negatives when considering the risk that any of the three BCAs may establish outside glasshouses. Higher sensitivity can also be a result of overtraining (an inability to generalise or predict new data) of the models. Overtraining or over-fitting was avoided by checking AUC, Kappa statistic, and model uncertainty scores before generating weighting matrices.

3 Results

3.1 *Delphastus catalinae*

3.1.1 Maxent

The average test AUC for the replicate runs was 0.894±0.216 (±standard deviation). The variables BIO19 (Precipitation of Coldest Quarter), BIO03 (Isothermality), BIO18 (Precipitation of Warmest Quarter), altitude and BIO07 (Temperature Annual Range) were relatively influential (Table 3). The Maxent model for New Zealand has most habitat suitability scores of <0.5 with small areas with scores >0.6 (Figure 2).

Table 3. Relative contribution of each predictor variable for the Maxent model of *Delphastus catalinae*

Variable	Percent contribution	Permutation importance
BIO19	25.6	26
BIO03	13.7	22.9
BIO18	12.2	6.3
alt	11.5	19.2
BIO07	10	1.5
BIO09	5.6	0.7
BIO06	5.5	3.4
BIO05	2.7	2.9
BIO15	2.7	4.9
BIO04	2.5	1.2
BIO14	2.3	7.2
BIO17	2.1	2.7
BIO11	1.5	0
BIO16	1	0.9
BIO08	0.7	0
BIO02	0.3	0
BIO13	0.1	0.1
BIO01	0	0
BIO10	0	0
BIO12	0	0

3.1.2 Multi Model consensus

The variables Bio06 (Min Temperature of Coldest Month), BIO08 (Mean Temperature of Wettest Quarter) and BIO11 (Mean Temperature of Coldest Quarter) were selected as most influential by random forests and stepwise regression analysis. Decision trees (CTREE, CART) were the best-performed individual algorithms (Table 4). The consensus Multi Model predicted that much of the North Island has a habitat suitability in the range 0.6-0.8 (Figure 3).

Table 4. Multi Model performance results for *Delphastus catalinae.*

Model/Classifier (abbreviation)	Sensitivity	Kappa	Uncertainty	AUC
Logistic regression (LOG)	0.4467	0.3830	0.0667	0.76
Naïve Bayes (NB)	0.5867	0.5333	0.0667	0.7911
Classification and Regression Tree (CART)	0.96	0.9143	0.1	0.7156
Conditional Tree (CTREE)	1.0	1.0	0.1	0.7333
K nearest neighbour (KNN)	0.95	0.8667	0.1333	0.7400
Support Vector Machine (SVM)	0.8433	0.6475	0.2667	0.8667
Artificial Neural Nets (NNET)	0.6700	0.4762	0.1667	0.8089

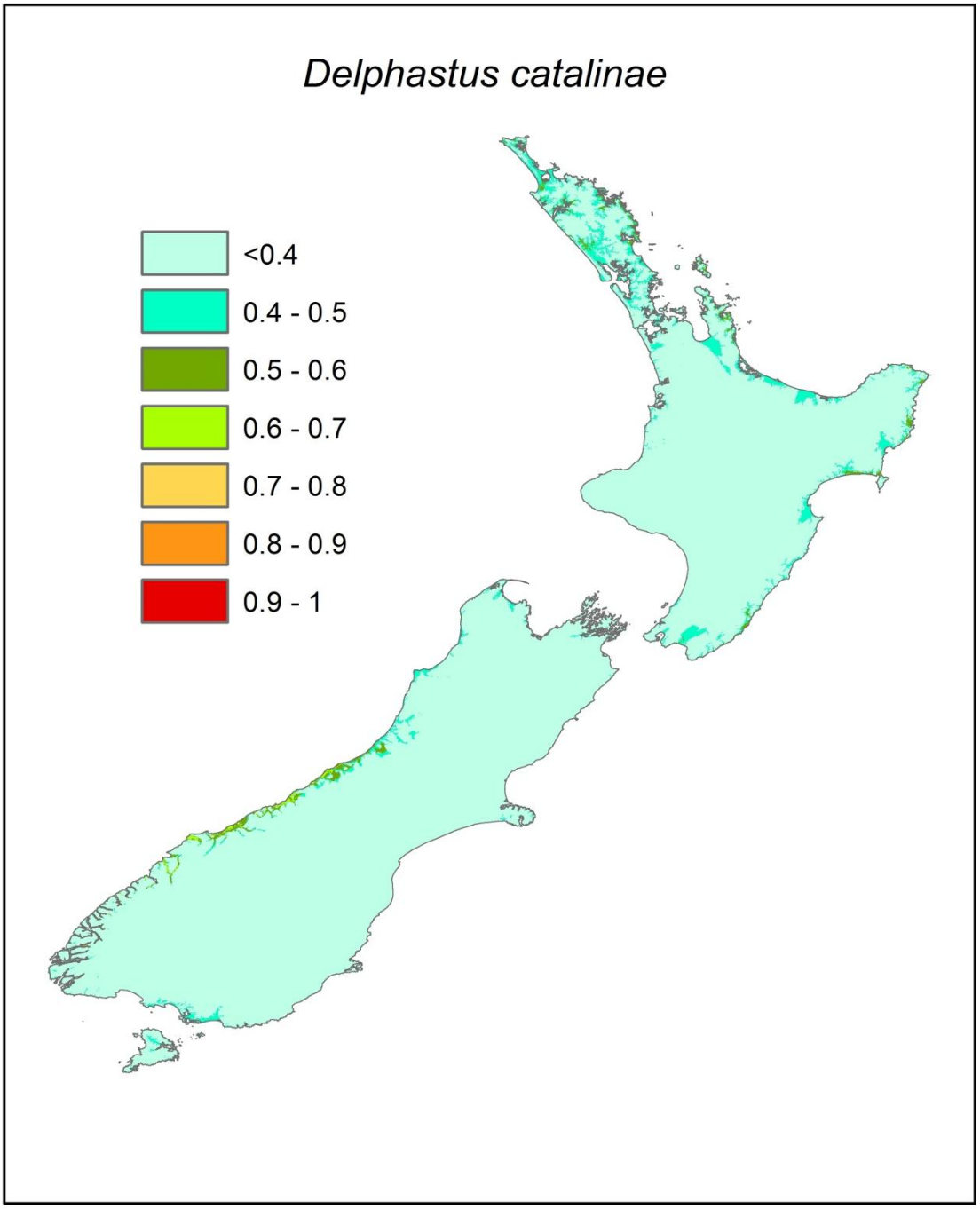


Figure 2. Maxent map of habitat suitability in New Zealand for *Delphastus catalinae.*

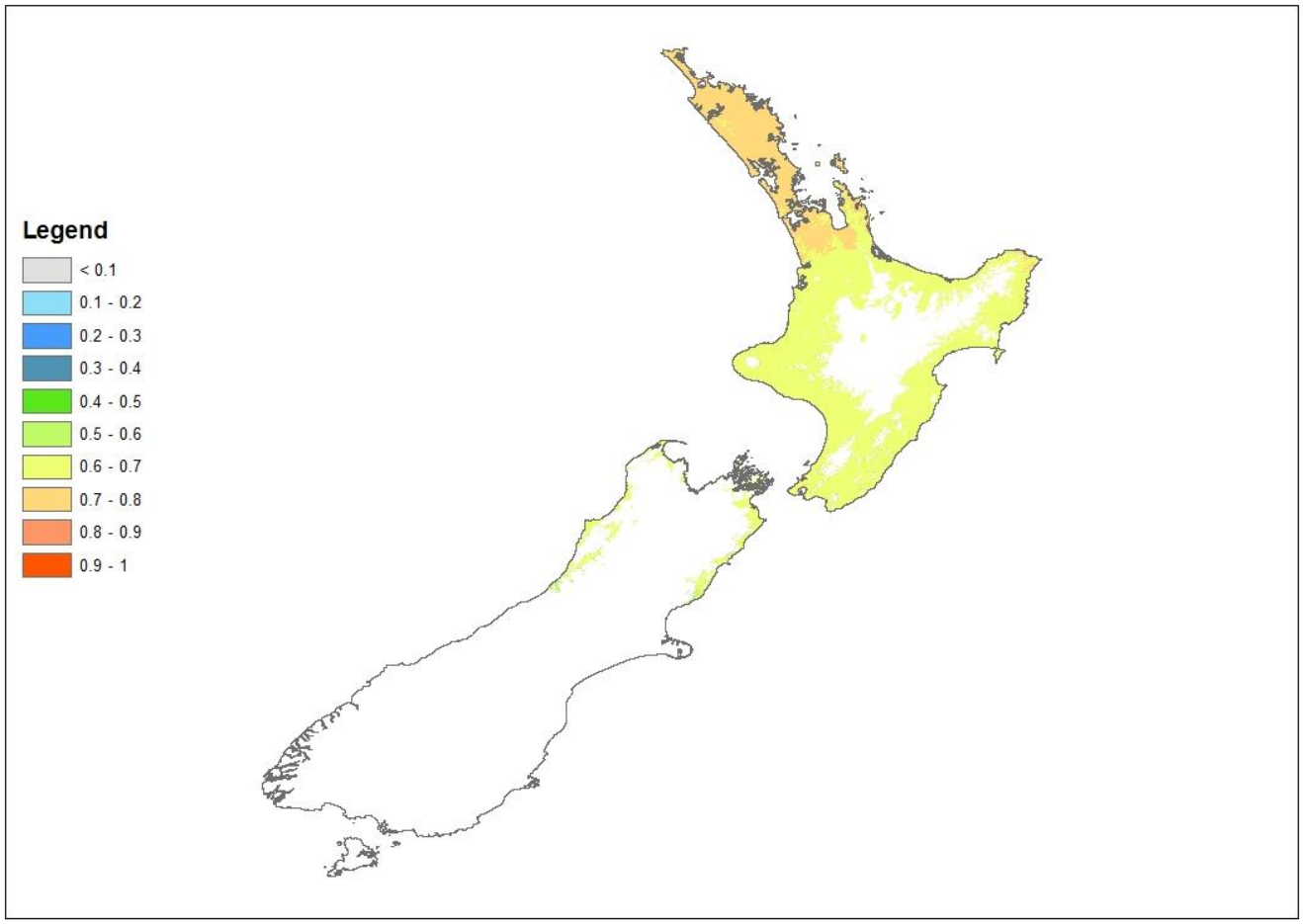


Figure 3. Consensus map of habitat suitability in New Zealand for *Delphastus catalinae*. The map is a consensus of seven different algorithms (Logistic Regression, Naive Bayes, Classification and Regression Trees, Conditional Trees, K-Nearest Neighbours, Support Vector Machines, and Artificial Neural Networks) weighted by their sensitivity scores. White areas are where no predictions were made because of dissimilarities to training data.

3.2 *Macrolophus melanotoma* / *M. pygmaeus*

3.2.1 Maxent

The average test AUC for the replicate runs was 0.995±0.003 (±standard deviation). The variables BIO09 (Mean Temperature of Driest Quarter), BIO06 (Min Temperature of Coldest Month), and BIO08 (Mean Temperature of Wettest Quarter) were relatively influential (Table 5). All grid cells in New Zealand have habitat suitability scores less than 0.5 (Figure 4).

Table 5. Relative contribution of each predictor variable for the Maxent model of *Macrolophus melanotoma* / *M. Pygmaeus*.

Variable	Percent contribution	Permutation importance
BIO09	26.9	47.7
BIO06	25.7	2.8
BIO08	14.5	0
BIO02	8.3	0.1
BIO14	6.2	1.4
BIO15	5.5	2.8
BIO04	5.3	28.2
alt	2.1	0.8
BIO01	1.4	12
BIO07	1.4	0
BIO03	1.3	3.9
BIO18	1.1	0
BIO13	0.1	0
BIO05	0.1	0.1
BIO10	0.1	0
BIO12	0.1	0
BIO17	0	0
BIO11	0	0
BIO16	0	0
BIO19	0	0

3.2.2 Multi Model consensus

The variables BIO01 (Annual Mean Temperature), BIO09 (Mean Temperature of Driest Quarter) and BIO11 (Mean Temperature of Coldest Quarter) were selected as most influential by random forests and stepwise regression analysis. Artificial Neural Nets (NNET) and Naïve Bayes (NB) were the best-performed individual algorithms (Table 6). The consensus Multi Model predicted that small areas in Northland are highly suitable (>0.8) for *M. melanotoma* / *M. pygmaeus* (Figure 5).

Table 6. Multi Model performance results for *Macrolophus melanotoma / M. pygmaeus*

Model/Classifier (abbreviation)	Sensitivity	Kappa	Uncertainty	AUC
Logistic regression (LOG)	0.9333	0.7059	0.1250	0.9434
Naïve Bayes (NB)	1.00	0.7487	0.0938	0.9609
Classification and Regression Tree (CART)	0.8933	0.1575	0.3125	0.5215
Conditional Tree (CTREE)	0.40	0.00	0.7188	0.1406
K nearest neighbour (KNN)	0.95	0.6618	0.2188	0.9219
Support Vector Machine (SVM)	0.8933	0.7433	0.2188	0.9219
Artificial Neural Nets (NNET)	1.00	0.8059	0.1875	0.9492

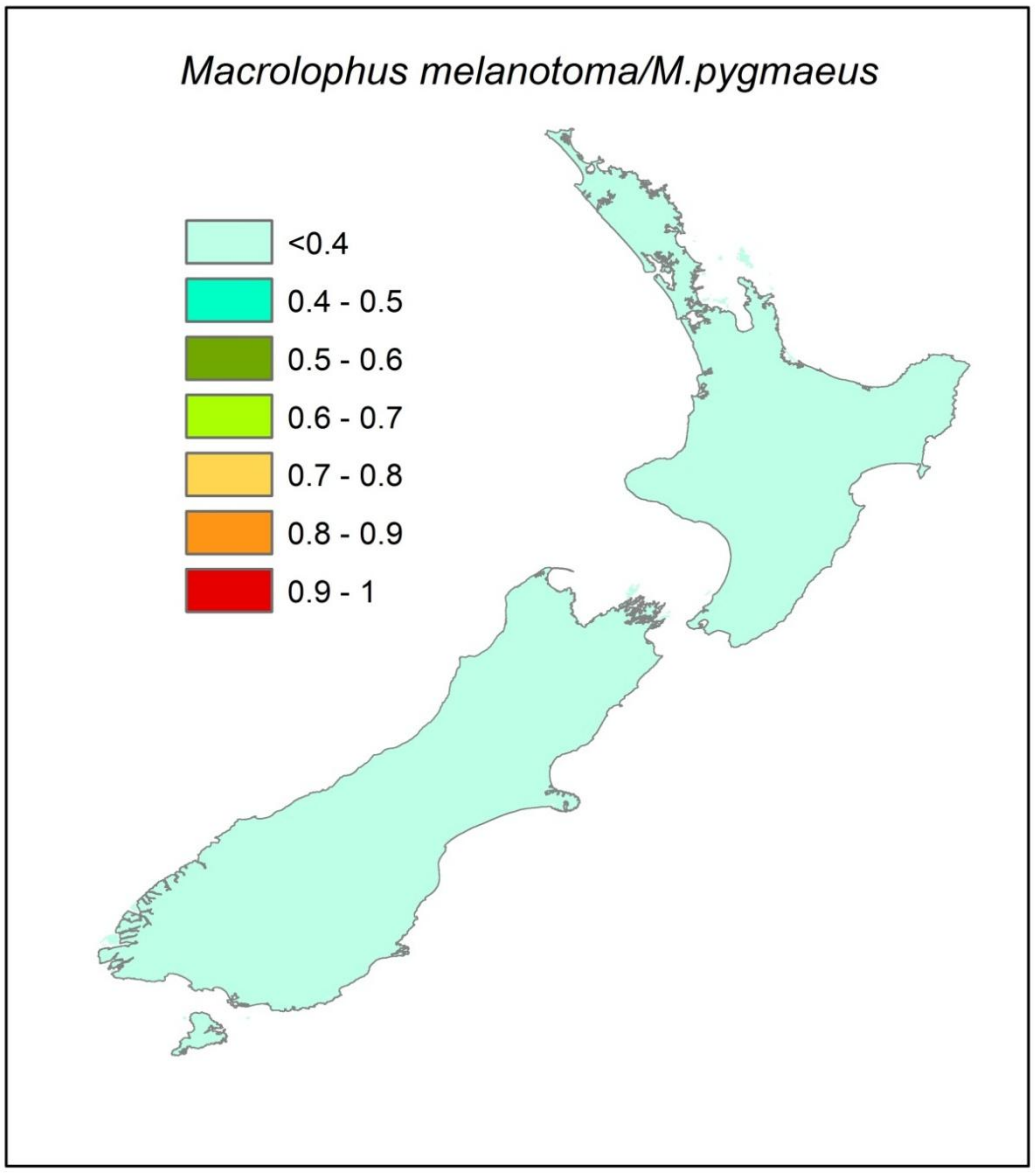


Figure 4. Maxent map of habitat suitability in New Zealand for *Macrolophus melanotoma / M. pygmaeus.*

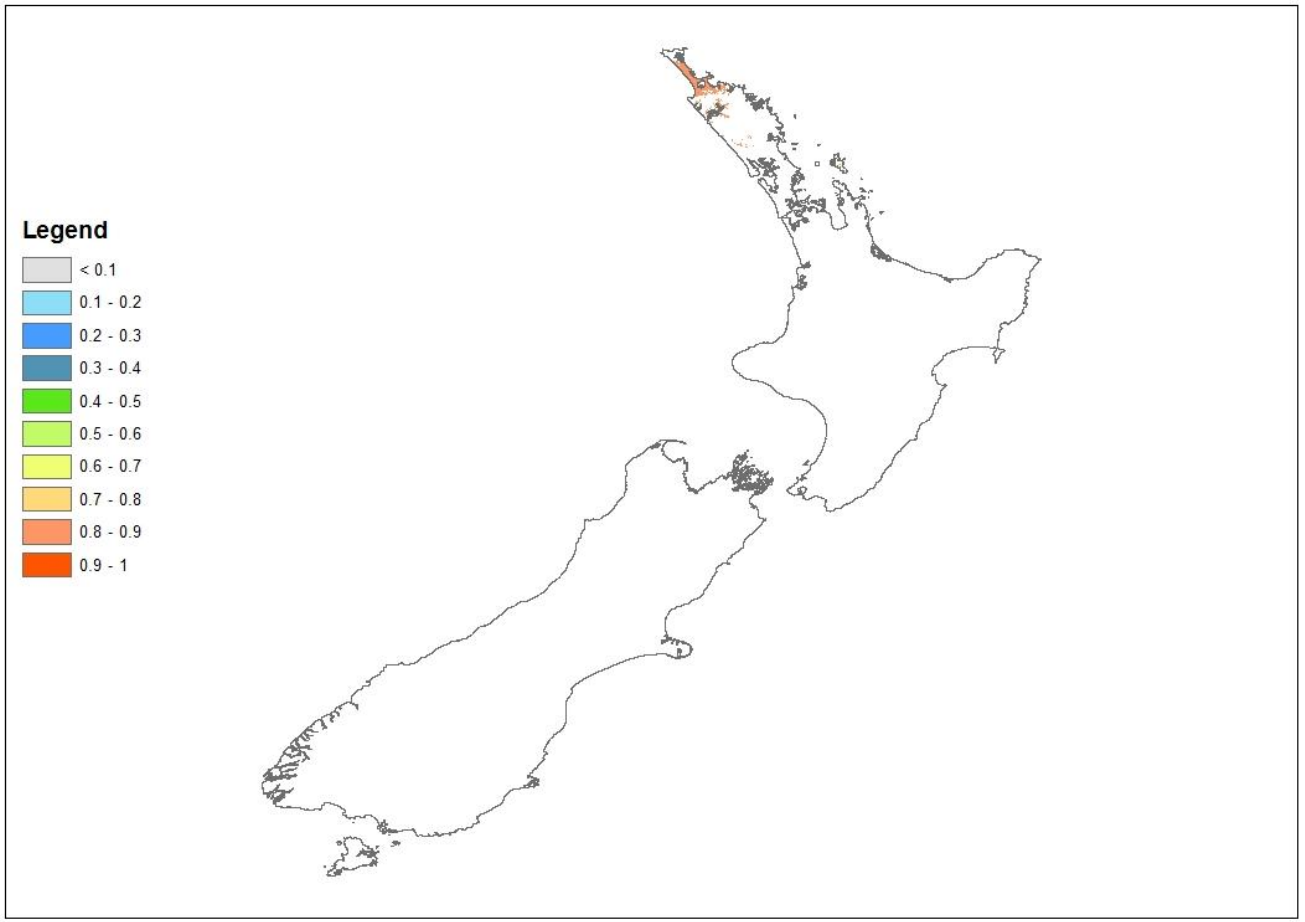


Figure 5. Consensus map of habitat suitability in New Zealand for *Macrolophus melanotoma / M. pygmaeus*. The map is a consensus of seven different algorithms (Logistic Regression, Naive Bayes, Classification and Regression Trees, Conditional Trees, K-Nearest Neighbours, Support Vector Machines, and Artificial Neural Networks) weighted by their sensitivity scores.

3.3 *Nesidiocoris tenuis*

3.3.1 Maxent

The average test AUC for the replicate runs was 0.928±0.049. The variables BIO09 (Mean Temperature of Driest Quarter), BIO02 (Mean Diurnal Range), BIO13 (Precipitation of the Wettest Month) and BIO18 (Precipitation of the Warmest Quarter) were relatively influential (Table 7). The Maxent model for *N. tenuis* predicted habitat suitability scores of less than 0.5 for New Zealand (Figure 6).

Table 7. Relative contribution of each predictor variable for the Maxent model of *Nesidiocoris tenuis.*

Variable	Percent contribution	Permutation importance
BIO09	36.9	48.7
BIO02	12.6	8.5
BIO13	12.6	10
BIO18	10.7	1.7
alt	8.2	0.6
BIO14	4.9	0.9
BIO04	4.6	24.2
BIO19	3.7	1.6
BIO07	2.2	0.2
BIO16	1.9	0
BIO12	0.9	0.8
BIO08	0.3	1.2
BIO17	0.2	0
BIO06	0.1	0.7
BIO15	0.1	0.3
BIO01	0	0.2
BIO10	0	0.3
BIO03	0	0.3
BIO11	0	0
BIO05	0	0

3.3.2 Multi Model consensus

The variables BIO04 (Temperature Seasonality), Bio06 (Min Temperature of Coldest Month), BIO07 (Temperature Annual Range) and BIO09 (Mean Temperature of Driest Quarter) were selected as most influential by random forests and stepwise regression analysis. Logistic regression (LOG) and Naïve Bayes (NB) were the best-performed individual algorithms (Table 8).The Multi Model consensus predicted much of the North Island as climatically suitable, as well as areas in the northern half of the South Island (Figure 7).

Table 8. Multi Model performance results for *Nesidiocoris tenuis.*

Model/Classifier (abbreviation)	Sensitivity	Kappa	Uncertainty	AUC
Logistic regression (LOG)	0.8798	0.7368	0.1111	0.7572
Naïve Bayes (NB)	0.8317	0.6357	0.0370	0.7599
Classification and Regression Tree (CART)	0.7111	0.3534	0.0926	0.6324
Conditional Tree (CTREE)	0.3778	0.1023	0.5185	0.6543
K nearest neighbour (KNN)	0.8603	0.5374	0.1111	0.8615
Support Vector Machine (SVM)	0.6781	0.5454	0.1852	0.8436
Artificial Neural Nets (NNET)	0.5924	0.5136	0.2593	0.9218

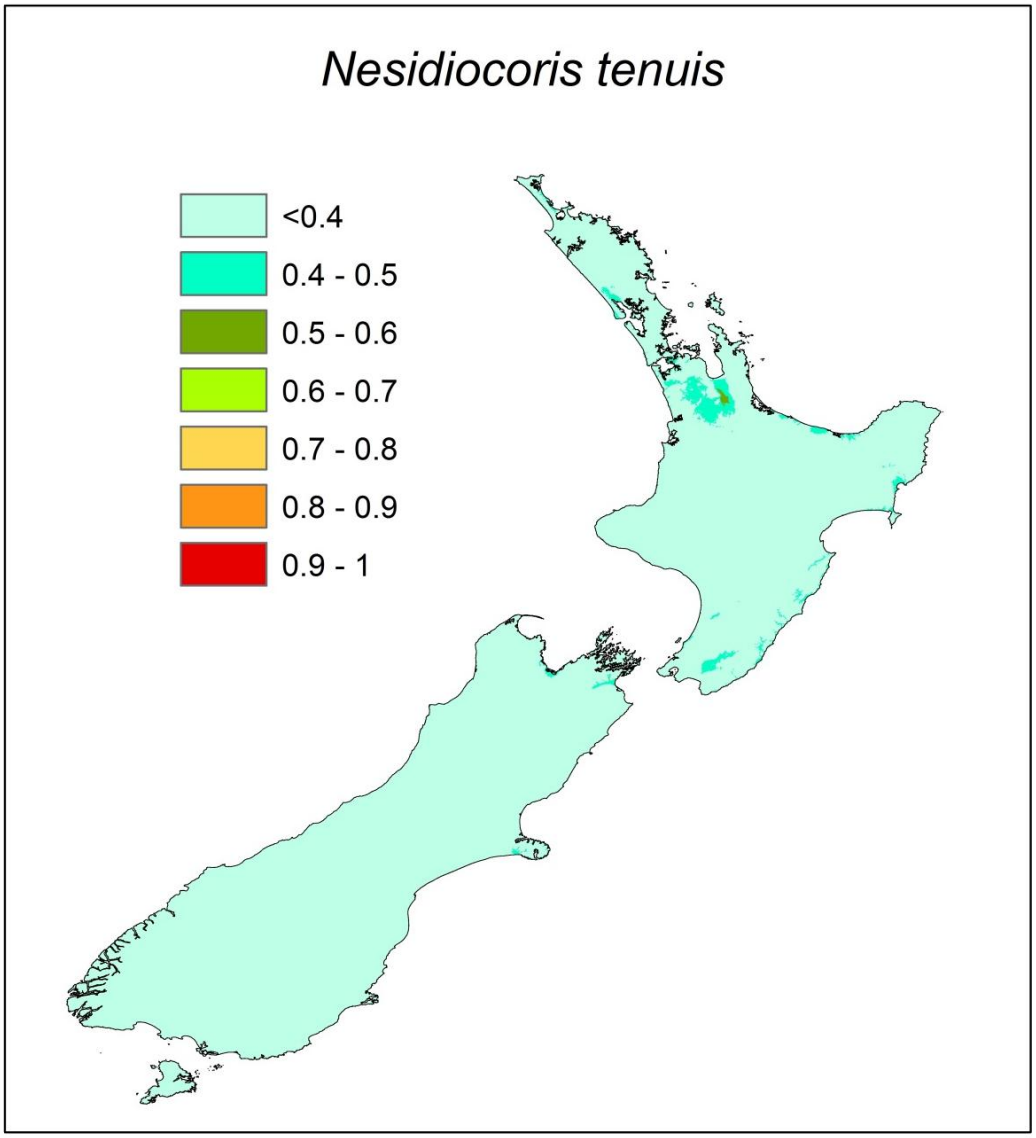


Figure 6. Maxent map of habitat suitability in New Zealand for *Nesidiocoris tenuis*.

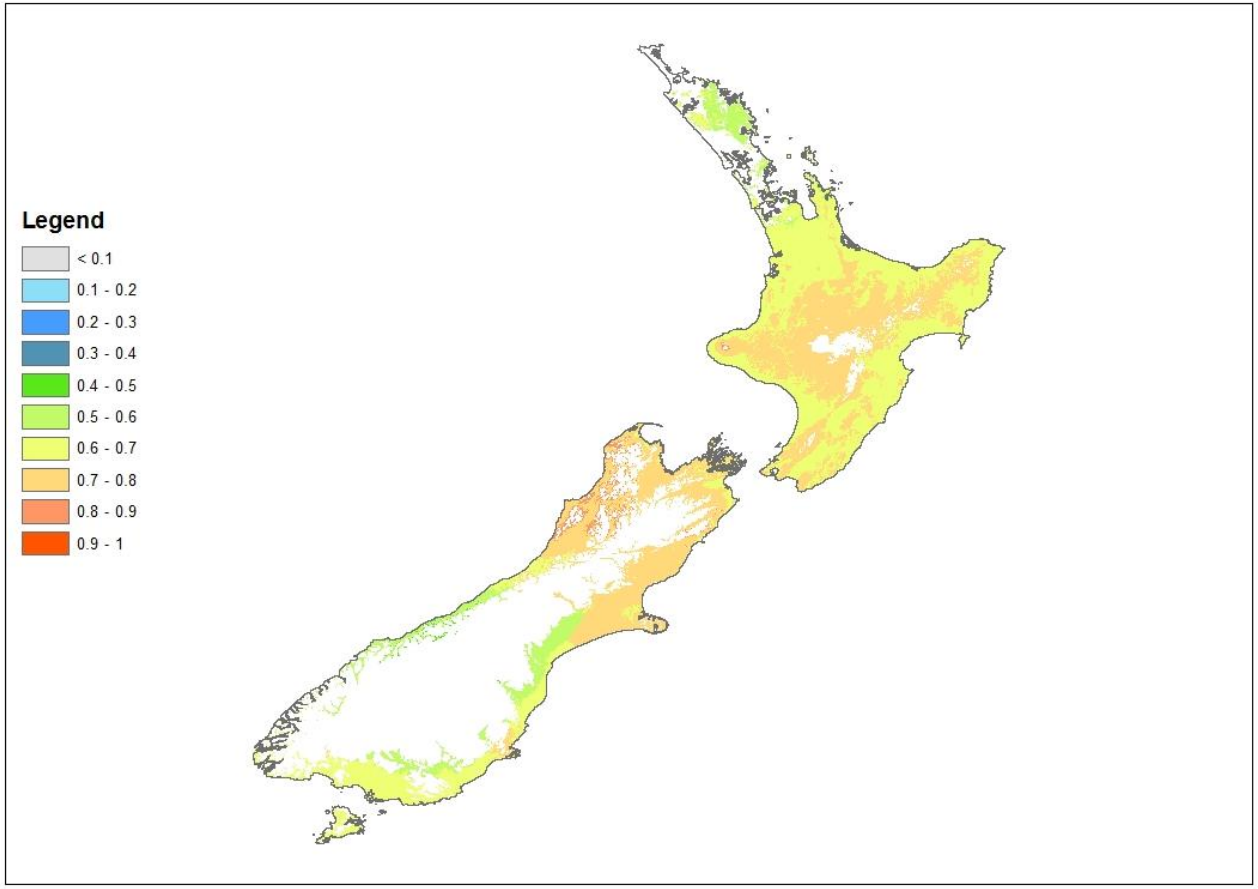


Figure 7. Consensus map of habitat suitability in New Zealand for *Nesidiocoris tenuis*. The map is a consensus of seven different algorithms (Logistic Regression, Naıve Bayes, Classification and Regression Trees, Conditional Trees, K-Nearest Neighbours, Support Vector Machines, and Artificial Neural Networks) weighted by their sensitivity scores.

4 Interpretation of maps

4.1 *Delphastus catalinae*

- The Maxent model indicated that climate suitability is generally poor for *D. catalinae* in New Zealand (most values<0.5), with coastal areas particularly in Northland slightly more favourable than elsewhere.
- The consensus Multi Model predicts that Northland is relatively well suited climatically for *D. catalinae* (scores >0.7).
- In summary, the Maxent model indicates low likelihood that preferred climate for *D. catalinae* exists in New Zealand, consistent with CLIMEX modelling. The CLIMEX model indicated that only small areas of Northland are suitable for *D. catalinae.* The consensus Multi Model indicates that Northland may be suitable for *D. catalinae*. Part of the difference may merely reflect the relatively small number of weather stations used by CLIMEX compared with the high resolution gridded data used by the Multi Model and Maxent.

4.2 *Macrolophus melanotoma* / *M. pygmaeus*

- The Maxent model indicated that climate suitability is poor for *M. melanotoma* / *M. pygmaeus* in New Zealand (values<0.5).
- The consensus Multi Model indicated that only a small area of north of Kaitaia in Northland has suitable climate for *M. melanotoma* / *M. pygmaeus.*
- In summary, the Maxent and consensus Multi Model indicate low likelihood that preferred climate for *Macrolophus melanotoma* / *M. pygmaeus* exists in most or all of New Zealand. In contrast, the CLIMEX model indicated that Northland and the east coast of the North Island contains suitable habitat for *M. melanotoma* / *M. pygmaeus.*

4.3 *Nesidiocoris tenuis*

- The Maxent model indicated that suitability of climate for *N. tenuis* is poor for all New Zealand (values <0.5).
- The consensus Multi Model indicated that large areas in the northern half of the South island and central areas of the North Island have relatively suitable climate conditions (scores >0.7) for *N. tenuis.*
- In summary, the consensus Multi Model and the CLIMEX model indicated that some areas of New Zealand have relatively suitable climate for *N. Tenuis,* although these areas differed. In the consensus Multi Model case, suitable climate was predicted to exist in the central north island and the coastal areas of the Buller, Nelson, Kaikoura, and Canterbury regions. In the CLIMEX model case, suitable climate was predicted in Northland and some coastal areas of the North Island. Maxent modelling indicates that there is likely to be no suitable climate in New Zealand for *N. tenuis*.

4.4 Summary

There is disagreement between the projections of two of the three modelling approaches for *D. catalinae* and *M. melanotoma* / *M. pygmaeus* and among all three modelling approaches for *N. tenuis.* Model performance for all three species and particularly for *D. catalinae* is likely to be compromised significantly by small training data sets (i.e. the limited geographical collection records available: n=14 for *D. catalinae*, n=23 for *Macrolophus* spp., n=30 for *N. tenuis*). Small training sets (n<30) are likely to be subject to various errors, including incidental correlation with environmental variables and poor representation of the species range (Stockwell & Peterson 2002; Wisz et al. 2008). Caution is advised in interpreting results of the Maxent and Multi Models in particular. The limited collection data can also influence interpretation of CLIMEX output; however, these data are augmented by physiological data from laboratory experiments and by expert opinion and in this case CLIMEX results may be more reliable than Multi Model and Maxent model results.

5 Acknowledgements

Thanks to Sue Worner for advice and making resources available to complete modelling.

6 References

Araujo MB, Peterson AT 2012. Uses and misuses of bioclimatic envelope modelling. Ecology 93(7): 1527-1539.

Boser BE, Guyon IM, Vapnik VN 1992. A training algorithm for optimal margin classifiers. Paper presented at the Proceedings of the fifth annual workshop on Computational learning theory.

Breiman L 2001. Random forests. Machine Learning 45(1): 5-32.

Elith J, Phillips SJ, Hastie T, Dudík M, Chee YE, Yates CJ 2011. A statistical explanation of MaxEnt for ecologists. Diversity and Distributions 17(1): 43-57.

Guo QH, Kelly M, Graham CH 2005. Support vector machines for predicting distribution of Sudden Oak Death in California. Ecological Modelling 182(1): 75-90.

Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A 2005. Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology 25: 1965-1978.

Logan DP 2012. CLIMEX models for selected glasshouse biological control agents. A report prepared for Horticulture New Zealand. PFR SPTS No. 6938.

Mahadevan S, Shah SL 2009. Fault detection and diagnosis in process data using one-class support vector machines. Journal of Process Control 19(10): 1627-1639.

Marmion M, Miska L, Heikkinen RK, Thuiller W 2009. The performance of state-of-the-art modelling techniques depends on geographical distribution of species. Ecological Modelling 220: 3512-3520.

Noble WS 2006. What is a support vector machine? Nature Biotechnology 24 (12):1565-1567.

Phillips SJ, Anderson RP, Schapire RE 2006. Maximum entropy modeling of species geographic distributions. Ecological Modelling 190(3): 231-259.

Phillips SJ, Dudík M 2008. Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. Ecography 31: 161-175.

Phillips SJ, Dudik M, Schapire RE.2004. A maximum entropy approach to species distribution modeling. – In: Greiner R and Schuurmans D.(eds), *Proc. 21st Int. Conf. on Machine Learning*. ACM press, New York, pp.655–662.

Phillips SJ, Dudík M, Elith J, Graham CH, Lehmann A, Leathwick J, Ferrier S 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecological Applications 19(1): 181-197.

Roura-Pascual, N, Brotons L, Peterson AT, Thuiller W 2009. Consensual predictions of potential distributional areas for invasive species: a case study of Argentine ants in the Iberian Peninsula. Biological Invasions 11(4): 1017-1031.

Speybroeck N 2012. Classification and regression trees. International Journal of Public Health 57(1): 243-246.

Stockwell DRB, Peterson AT. 2002 Effects of sample size on accuracy of species distribution models. Ecological Modelling 148(1): 1-13.

Thompson B 1995. Stepwise regression and stepwise discriminant analysis need not apply here: A guidelines editorial. Educational and Psychological Measurement. 55(4): 525-534.

Vaclavik T, Meentemeyer RK 2009. Invasive species distribution modeling (iSDM): Are absence data and dispersal constraints needed to predict actual distributions? Ecological Modelling 220(23): 3248-3258.

Van Der Wal J, Shoo LP, Graham C, Williams SE 2009. Selecting pseudo-absence data for presence-only distribution modeling: How far should you stray from what you know? Ecological Modelling 220(4): 589-594.

Watts MJ, Worner SP 2008. Using artificial neural networks to determine the relative contribution of abiotic factors influencing the establishment of insect pest species. Ecological Informatics 3(1): 64-74.

Wisz MS, Hijmans RJ, Li J, Petersomn AT, Graham CH, Guisan A . 2008. Effects of sample size on the performance of species distribution models. Diversity and Distributions 14 (5): 763-773.

Worner SP, Ikeda T, Leday G, Joy M 2010. Surveillance tools for freshwater invertebrates. MAF Biosecurity New Zealand Technical Paper No: 2010/21, 112 Pp.

Zuo W, Lao N, Geng Y, Ma K 2008. GeoSVM: an efficient and effective tool to predict species' potential distributions. Journal of Plant Ecology 1(2): 143-145.

Appendix Table A1. Longitude and latitude coordinates for collection localities of *Delphastus catalinae* (n=14)

Longitude	Latitude
-82.4572	27.95058
-61.7305	12.05337
-61.4381	10.66076
-80.1495	26.0112
-80.4776	25.46872
-82.5723	27.52143
-82.3248	29.65163
-75.5197	10.38733
-75.5557	10.40119
-115.625	33.12716
-120.046	34.6853
-76.26	-10.7098
-118.451	33.38698
-119.695	34.01592

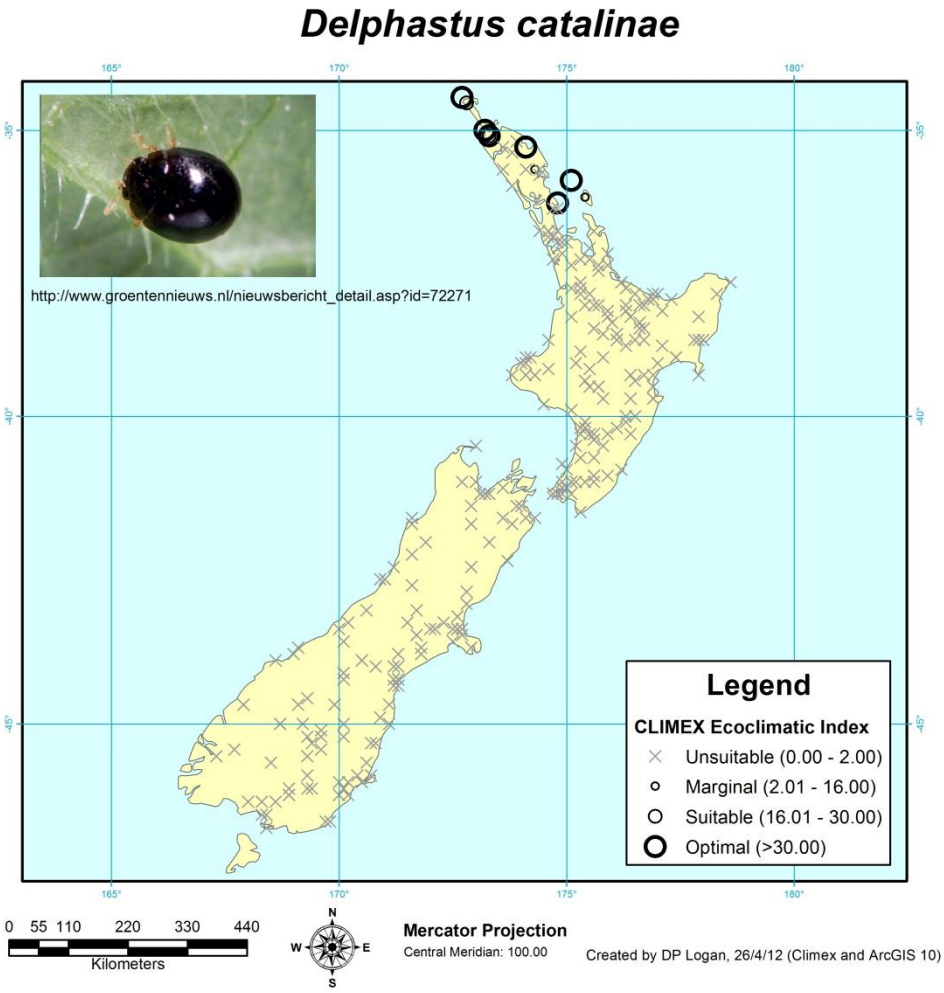
Appendix Table A2. Longitude and latitude coordinates for collection localities of *Macrolophus melanotoma*/*M. pygmaeus* (n=23)

Longitude	Latitude
21.92715	39.36564
21.35617	37.79329
23.31663	38.3405
23.09904	38.375
2.448347	41.53506
2.205816	41.45544
2.375238	41.51667
-1.41278	38.22861
-1.7	38.22472
-1.985	38.275
-16.61	28.39278
-7.455	39.41972
2.408611	41.55417
-1.49692	37.49517
-1.93381	38.05769
-1.7835	38.10878
-1.14042	37.84397
-1.47722	37.60061
-1.06019	37.95044
-1.31872	38.14708
-1.72747	38.17933
-16.8078	28.36564
-16.6102	28.39281
-8.89806	37.19425

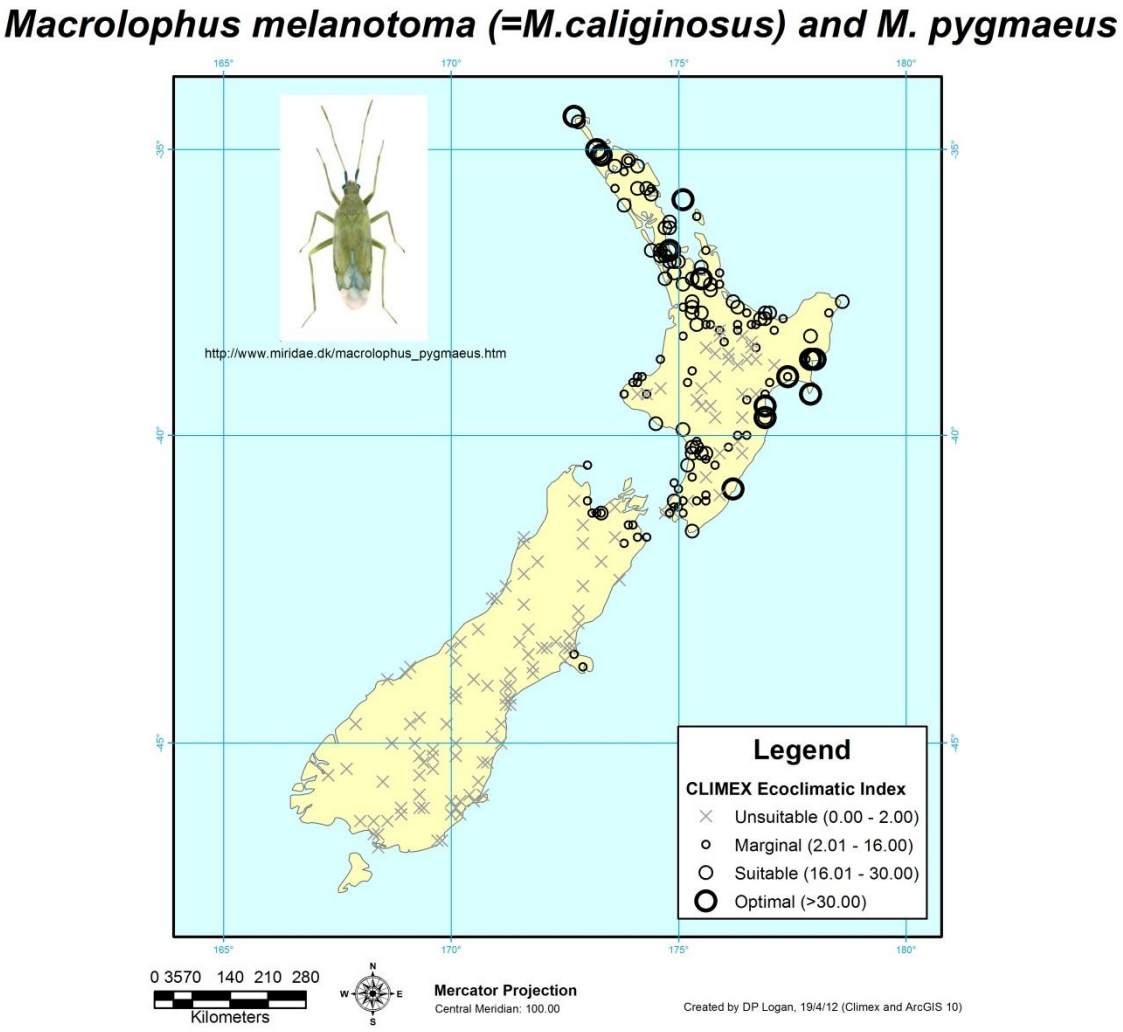
Appendix Table A3. Longitude and latitude coordinates for collection localities of *Nesidiocoris tenuis* (n=30)

Longitude	Latitude
-1.58116	37.4038
15.05702	37.49805
14.70939	36.9203
21.54504	37.75936
22.64264	38.60111
7.557056	44.32646
8.197249	44.06017
34.77053	32.04572
34.76615	32.06342
34.84408	32.16337
35.20685	31.77008
34.81127	31.89277
34.92088	32.44278
35.38357	31.45119
152.3655	-32.5074
120.4358	15.94412
120.7689	15.65806
120.55	16.11667
125.0667	6.216671
120.9876	15.69068
78.83111	24.74083
72.95083	22.55417
85.82452	20.29606
76.9971	20.70388
90.4201	23.9984
31.20214	30.07365
15.96999	-4.38706
-72.5205	9.967492
-76.3866	3.398049
-69.2401	9.648832

Appendix Figure A1. Habitat suitability for *Delphastus catalinae* based on a CLIMEX model (Logan 2012)



Appendix Figure A2. Habitat suitability for *Macrolophus melanotoma /M. pygmaeus* based on a CLIMEX model (Logan 2012)



Appendix Figure A3. Habitat suitability for Nesidiocoris tenuis based on a CLIMEX model (Logan 2012)