Multi-model analysis for projecting the global distribution of *Halyomorpha halys*

Senait D. Senay

International Science & Technology Practice & Policy (InSTePP)
University of Minnesota

Senait D. Senay¹, ⁴&⁶, Craig B. Phillips³, ⁴, Ursula Torres², John Kean³, ⁴, Robert C. Venette⁵ and Susan P. Worner², ⁴

¹ International Science & Technology Practice & Policy (InSTePP), Dept. of Applied Economics, University of Minnesota, 1994 Buford Ave., 248E Ruttan Hall, St. Paul, MN 55108, USA
²BioProtection Research Centre, PO Box 85084, Lincoln University, Lincoln 7674, New Zealand
³AgResearch Limited, Private Bag 4749, Christchurch 8140, New Zealand
⁴Better Border Biosecurity
⁵Northern Research Station, Forest Service, United States Department of Agriculture, 1992 Folwell Ave., St. Paul, MN 55108
⁶ssenay@umn.edu
Standardized multi-factor modelling: Why do we need it?

- There are numerous sources of uncertainty whenever prediction is involved.

- In a correlative SDM setting, quality of presence data, model type, model parametrization and model evaluation are known to be major sources of uncertainty.

- Similar to climate studies, using various models rather than a single model allows us to better estimate best and worst case scenarios, confidence interval and consistency between models.

- An ensemble modelling framework can be useful in weeding out weak models. [as there are no universally best models]
Background – *Halyomorpha halys*

*Halyomorpha halys*

- Native to China, N. & S. Korea and Japan
- Established in the US (1990’s)
- Established in Europe (2000’s)
- Established in Canada (2014)
- Multi-host pest { more than 100 host species recorded}
- Currently expanding its invaded range

General setting

- Geographic extent: Global and NZ
- Occurrence dataset: 3,419
- Environmental covariates: 35 bioclim variables from CLIMOND + Elevation layer
  - Spatial resolution = 10” Global resolution & 30’ NZ extent (interpolated from CLIMOND data)

Research design & Methods

- Two types of presence datasets = Initial and Refined
- Two types of modelling approaches = Presence-only (MaxEnt) & Presence-Absence (7 model ensemble MMF)
- Standard modelling procedures
  - Niche analysis results used to inform modelling decisions
  - Background and pseudo-absence datasets were methodologically generated
  - Centralized variable selection carried out using a random forest classifier
  - Multiple-model evaluation methods used.
Methods - Modelling

**Presence-only (MaxEnt)**
- Software = MaxEnt v.3.3.3k
- Features (functions) = LQPH
- Background points = All locations (37k+ and 40k+ for initial and refined scenarios respectively)
- A moderate regularization multiplier = $\beta=2$ (Radosavljevic and Anderson, 2014)
- 10-fold cross-validation

**Presence-absence (MMF)**
- A 7 model Framework
  - QDA, LOG, NB, CART, RF, ANN, SVM
- Built-in customized model specific parametrization
- Top 3 models selected based on Kappa score and are combined to give weighted average, best and worst scenarios
- 10-fold cross-validation

\[
P_{kA} = \frac{\sum (\kappa_{mi} \times P_{mi})}{\sum \kappa_{mi}} \tag{1}
\]

Where: $\kappa$ is the Kappa score of the $i$th high performing model ($m$), and $P$ is the predicted probability value for a given cell by the $i$th model ($m$).

\[
P_{EB} = \min ([P_{m_i}]) \tag{2}
\]

\[
P_{EW} = \max ([P_{m_i}]) \tag{3}
\]

Where: $[P_{m_i}]$ is a vector of predicted probabilities for a given cell (P$_{m_1}$, ..., P$_{m_n}$) by high performing models $m_1$ ... $m_n$. 
**Result - Presence datasets**

**Initial dataset (hh.set1)**
- Total = 3,189
- After duplicates were removed = 492
- Unique to this dataset = 22
- Removed locations include: Alaska, Florida and Arizona.

**Refined dataset (hh.set2)**
- Total = 3,384
- After duplicates were removed = 547
- Unique to this dataset = 77
- Added locations include: More locations in Japan
Stability of the inferred climatic niche of *Halyomorpha halys*

- PCA performed on the 36 predictors with respect to the *H. halys* distributional data.
- Eco-regions were used to delimit the invaded and native ranges.
- A kernel density function using a 100*100 cell was applied to obtain the smoothed density of occurrences in each cell in the PCA environmental space (based on 1st by 2nd component matrix)

**Stability** = 0.63
**Unfilling** = 0.89
**Expansion** = 0.37
Background data

- **Variable importance analysis used to delimit optimum geographic distance to buffer background data**
- **Same distances identified both for the initial and refined scenario**
**Pseudo-absence points**

- One-Class Support Vector Machines (OCSVM) used to environmentally profile the background data into different classes based on how dissimilar they are to the presence locations.
- K-means clustering used to select centroids of clusters generated from the most dissimilar OCSVM class.
### Variables selected using a Random forest classifier

Table 2. Variables chosen for modelling in declining order of importance. (A) Based on the initial presence dataset, (B) Based on the refined presence dataset.

<table>
<thead>
<tr>
<th>(A) Selected variables for the Liberal scenario</th>
<th>Df</th>
<th>Deviance</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td></td>
<td>717.67</td>
<td>735.67</td>
</tr>
<tr>
<td>Annual mean radiation (W m-2) [var 20]</td>
<td>1</td>
<td>728.64</td>
<td>744.64</td>
</tr>
<tr>
<td>Mean temperature of warmest quarter (°C) [var 10]</td>
<td>1</td>
<td>729.53</td>
<td>745.53</td>
</tr>
<tr>
<td>Annual mean temperature (°C) [var 01]</td>
<td>1</td>
<td>739.55</td>
<td>755.55</td>
</tr>
<tr>
<td>Max temperature of warmest week (°C) [var 05]</td>
<td>1</td>
<td>743.12</td>
<td>759.12</td>
</tr>
<tr>
<td>Mean moisture index of warmest quarter [var 34]</td>
<td>1</td>
<td>746.93</td>
<td>762.93</td>
</tr>
<tr>
<td>Precipitation of warmest quarter (mm) [var 18]</td>
<td>1</td>
<td>755.74</td>
<td>771.74</td>
</tr>
<tr>
<td>Highest weekly radiation (W m-2) [var 21]</td>
<td>1</td>
<td>761.79</td>
<td>777.79</td>
</tr>
<tr>
<td>Isothermality (Bio02 + Bio07) [var 03]</td>
<td>1</td>
<td>785.46</td>
<td>801.46</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(B) Selected variables for the Conservative scenario</th>
<th>Df</th>
<th>Deviance</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td></td>
<td>611.64</td>
<td>637.64</td>
</tr>
<tr>
<td>Annual mean temperature (°C) [var 01]</td>
<td>1</td>
<td>611.65</td>
<td>635.65</td>
</tr>
<tr>
<td>Radiation of warmest quarter (W m-2) [var 26]</td>
<td>1</td>
<td>611.7</td>
<td>635.7</td>
</tr>
<tr>
<td>Max temperature of warmest week (°C) [var 05]</td>
<td>1</td>
<td>612.08</td>
<td>636.08</td>
</tr>
<tr>
<td>Annual mean radiation (W m-2) [var 20]</td>
<td>1</td>
<td>613.42</td>
<td>637.42</td>
</tr>
<tr>
<td>Precipitation of warmest quarter (mm) [var 18]</td>
<td>1</td>
<td>613.66</td>
<td>637.66</td>
</tr>
<tr>
<td>Isothermality (Bio02 + Bio07) [var 03]</td>
<td>1</td>
<td>615.05</td>
<td>639.05</td>
</tr>
<tr>
<td>Mean moisture index of warmest quarter [var 34]</td>
<td>1</td>
<td>615.11</td>
<td>639.11</td>
</tr>
<tr>
<td>Highest weekly radiation (W m-2) [var 21]</td>
<td>1</td>
<td>619.34</td>
<td>643.34</td>
</tr>
<tr>
<td>Mean temperature of wettest quarter (°C) [var 08]</td>
<td>1</td>
<td>625.51</td>
<td>649.51</td>
</tr>
<tr>
<td>Mean temperature of coldest quarter (°C) [var 11]</td>
<td>1</td>
<td>720.22</td>
<td>744.22</td>
</tr>
<tr>
<td>Temperature seasonality (C of V) [var 04]</td>
<td>1</td>
<td>803.22</td>
<td>827.22</td>
</tr>
</tbody>
</table>
Evaluation – confusion matrix based

(A) MMF: MMFMaxEnt

AUC QDA = 0.951
AUC LOG = 0.916
AUC NB = 0.98
AUC CART = 0.941
AUC RF = 0.984
AUC SVM = 0.975
AUC NNET = 0.981

(B) MMF: MMFMaxEnt

AUC QDA = 0.958
AUC LOG = 0.949
AUC NB = 0.956
AUC CART = 0.958
AUC RF = 0.986
AUC SVM = 0.986
AUC NNET = 0.983

MaxEnt

Average Sensitivity vs. 1 - Specificity for Halymorpha halys_initial

Mean (AUC = 0.961) ■
Mean +/− one std dev ■
Random Prediction ■

Average Sensitivity vs. 1 - Specificity for Halymorpha halys_reined

Mean (AUC = 0.981) ■
Mean +/− one std dev ■
Random Prediction ■
Evaluation – Response curves (initial)

RF

- Annual mean temperature (°C) [var 10]
- Mean temperature of warmest quarter (°C) [var 10]
- Annual mean temperature (°C) [var 01]
- Max temperature of warmest week (°C) [var 05]

SVM

- Mean moisture index of warmest quarter [var 34]
- Precipitation of warmest quarter (mm) [var 18]
- Highest weekly radiation (W m⁻²) [var 21]
- Isothermality (Bio02 + Bio07) [var 03]

NNET

- Annual mean temperature (°C) [var 10]
- Isothermality (Bio02 + Bio07)
- Max temperature of warmest week (°C) [var 05]
- Mean temperature of warmest quarter (°C)

MaxEnt

- Annual mean temperature (°C) [var 10]
- Isothermality (Bio02 + Bio07)
- Max temperature of warmest week (°C) [var 05]
- Mean temperature of warmest quarter (°C)
Evaluation – Response curves (refined)

RF

SVM

Annual mean temperature (°C) [var 01] 1
Radiation of warmest quarter (W m-2) [var 26] 2
Max temperature of warmest week (°C) [var 05] 3
Annual mean radiation (W m-2) [var 20] 4
Precipitation of warmest quarter (mm) [var 18] 5
Isothermality (Bio02 + Bio07) [var 03] 6
Mean moisture index of warmest quarter [var 34] 7
Highest weekly radiation (W m-2) [var 21] 8
Mean temperature of wettest quarter (°C) [var 08] 9
Mean temperature of coldest quarter (°C) [var 11] 10
Temperature seasonality (C of V) [var 04] 11
Mean temperature of warmest quarter (°C) [var 10] 12

NNET

MaxEnt
Presence-only *H. halys* potential distribution projections

Best scenario

Average

Worst scenario
Presence-absence *H. halys* potential distribution projections

**Best scenario**

**Weighted average**

**Worst scenario**
H. halys distribution projections for New Zealand

Background data

NZ data

Type 1 novelty and Similarity
Ensemble map – *H. halys* establishment risk (I/O)

\[ P/A \text{ prev. } = 0.5 \]
\[ P-0 \text{ prev. } = 0.37 \]
Conclusion

- Standardized modelling frameworks allow grounded comparisons between different model results.

- Using multiple models in potential distribution predictions could reveal probable but different spatial distribution trajectories.

- While using ensemble models it is important to weed out weaker and low-performing models so as not to include unnecessary variation in predictions.

- *H. halys* is expanding its invasion, however from the fragmented climatically suitable sites predicted in Europe it appears that its expansion in North America might be faster than in Europe. The fact that there are suitable environments that are still not occupied by *H.*halys all over its invaded range increases the likelihood of its expansion.

- Follow up studies on *H. halys* dispersal and effect of control mechanisms on range expansion are highly recommended in order to estimate the potential for eradication or containment.

- *H. halys* proved to be a difficult pest to predict due to:
  1. Its association with urban/human environments
  2. Also possibly because it has variant acclimation potential in different environments.
Thank you!

www.instepp.umn.edu
Evaluation-predicted area

<table>
<thead>
<tr>
<th></th>
<th>&lt;0.2</th>
<th>(0.2-0.4]</th>
<th>(0.4-0.6]</th>
<th>(0.6-0.8]</th>
<th>(0.8-1]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global initial</td>
<td>94.77%</td>
<td>2.89%</td>
<td>1.42%</td>
<td>0.92%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Global refined</td>
<td>93.12%</td>
<td>4.09%</td>
<td>2.15%</td>
<td>0.57%</td>
<td>0.07%</td>
</tr>
<tr>
<td>New Zealand initial</td>
<td>61.85%</td>
<td>35.75%</td>
<td>2.39%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>New Zealand refined</td>
<td>58.30%</td>
<td>41.70%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
</tbody>
</table>