Thesis

Lukas Prader

13th September 2023

Contents

1 Introduction

(citations missing)
The topic of invasive species has become more and more important, even more so with recent changes in climate and habitats due to human influence. In order to deal with invasive species and the impact they can have on existing ecosystems, species distribution models (SDMs) have been used to predict the potential habitat and with that the threat of an emerging invasive species. Especially time-partitioned models can provide more insight into the process of invasion. Ensemble models are often used in applications where models are projected, since aspects of different modelling approaches can be combined to hopefully gain a more complete model of the species distribution.
There are certain challenges when using SDMs in invasive research though. For example, the invading species is usually not currently in equilibrium with its environment, leading to underestimations of the final niche occupied. A similar difficulty is modelling the invaded range with data from a species native range since the native niche can be quite different from the realized invaded niche. Nonetheless, SDMs are frequently used in trying to predict the spread of invasive species, which is why it is important to gain a better understanding of invasion processes and an SDM's ability to accurately represent it.
Harmonia axyridis, also known as the harlequin or Asian lady beetle, is an invasive species already established in non-native habitats all around the world. The invasion process has already been studied extensively and there have been SDMs modelling the invaded range in Europe as well, though mostly on a national scale.
One goal of this thesis is to look into the limitations of models built early in the invasion process of a species. By iterating over the years of the invasion, model performance can be evaluated with consideration to the current state of invasion. Computing the occupied niche separately for each year also provides more insight into the invasion process and might help to understand modelling limitations for certain years.

2 Methods

(citations missing)

2.1 Datasets

In order to also have some kind of influence of the human nature of first introduction, land cover data was used from the Copernicus Land cover Classification dataset with yearly resolution starting from 2002 up to 2020.
All traditional 19 bioclim variables were obtained from the CHELSA V2.1 climatologies datasets too, using the 1981-2010 time frame for all years from 2002 to 2010 as well as the MPI-ESM 1.2 ssp370 scenario 2011-2040 for all years from 2011 to 2022. For occurrence data, all global occurrences of *Harmonia axyridis* were downloaded from the GBIF database.

2.2 Data preparation

All bioclim and land cover layers were resampled to a matching resolution of 30 arc seconds and cropped to two spatial extents, Europe and the presumed native range referencing (Orlova-Bienkowskaja, Ukrainsky & Brown, 2015). The presence-only points from GBIF were subset to the afore mentioned spatial extents and then checked for missing values for latitude, longitude, year or coordinate uncertainty. No occurrences after 2022 were used, also no points with a coordinate uncertainty larger than 1 km. In Europe, the initial cut off year for presences was 1991, since this is the year of in vasion according to the EASIN website. Afterwards, using the library 'CoordinateCleaner', all remaining data points were again checked for common errors or biases in the respective subset. To prepare the data for modelling, pseudoabsences were generated for each year. For this, the subset of a specific year was taken and used to generate pseudo-absences limited to a radius around the occurrence points. For each presence point, a buffer circle was generated with a maximum distance of 18 km. This buffer circle also excluded any area closer than 1 km to any presence point of any year to limit contradiction in environmental space.

2.3 Model building

For each year, the following Models were computed: General Linear (GLM), General Additive (GAM), Boosted Regression Trees (BRT) and Maximum Entropy (MAXENT). A model for a specific year always included all points from past years as well. The iterative models that were built use only data points from Europe, though there was one model created only with native occurrences and predicted for each year in Europe. For all used occurrence points prior to 2002 and after 2020, the land cover data of 2002 or 2020 was used as a substitute.
Variance inflation factors were used to select the variables used for model building. For this, a GLM was computed only using Europe data from 2002.

This model included all 19 bioclim variables and their squared values as separate variables. For land cover variables, a PCA was computed from the 2002 data. PCA axes were included in the model until a cut off of 90% of explained variance was reached. Variance inflation factors were computed for this GLM and the variable with the highest VIF was dropped until none of the remaining variables had a VIF greater than 10. Quadratic versions of bioclim variables were always dropped before their linear counterparts, even if their VIF was lower.

2.4 Analysis

All SDM models of each year were tested for their accuracy on predicting the occurrences of the following year and the final year of 2022. For each year, the occupied niche was also computed by running a PCA analysis on the bioclim variables. The niche can then be visualized by plotting a dynamic occurrence density grid for the first two PCA axes(Broennimann et al. 2011).

3 Background

This section aims to give sufficient background on *Harmonia axyridis* and the theory necessary to understand this thesis.

3.1 Invasion theory

The Invasive Alien Species Regulation of the European Union from 2014 uses the following definitions for alien species and invasive alien species [1]:

> (1) 'alien species' means any live specimen of a species, subspecies or lower taxon of animals, plants, fungi or micro-organisms introduced outside its natural range; it includes any part, gametes, seeds, eggs or propagules of such species, as well as any hybrids, varieties or breeds that might survive and subsequently reproduce;
>
> (2) 'invasive alien species' means an alien species whose introduction or spread has been found to threaten or adversely impact upon biodiversity and related ecosystem services;

This already shows the main concerns associated with invasive species, non-native species that have an often negative impact on native ecosystems. Main goals in research are to find out which species have potential to become invasive, what habitat will be susceptible to invasion by those species, how fast the species will invade the new area and what impact its invasion will have on the native ecosystem [2]. To this end, many theories have been created to describe invasion processes. The invasion of a species can generally be described with four stages [3]:

1. Transport: Leaving the native range, arriving at a new location
2. Introduction: Existing in specific locations (captivity / cultivation)
3. Establishment: Existing outside of areas of introduction in the wild
4. Spread: Sustaining establishment and dispersing to new environments

The impacts of invasive alien species can be numerous, ranging from food web changes to reductions in habitat and species richness, hydrology and nutrient cycle changes, enhanced invasion of other species and mass extinctions [4]. For example intraguild predation, the predation of species using similar resources, can create completely new stable states of an ecological system [5].

3.2 Species Distribution Modelling

Species Distribution Models (SDMs) attempt to connect species distribution data, i.e. occurrence at a location, to the environmental and spatial properties of said location [6]. They have been shown to generate substantial insight into the ecological requirements of species and, as niche models, can be used to predict the potential habitat of a species [7]. There has been considerable

debate on the capabilities and limitations of SDMs, especially when used for prediction. In general, SDMs are made with the (ideal) assumption that the species is in environmental equilibrium [6], implying that its ecological niche is not currently changing. If these models are now used to predict new, unsampled areas, there actually is no measure to assess their accuracy, since no data is presently available for that area [7]. There is also no guarantee that the biotic interactions sampled in the study area will reflect the final interactions in the new area [6]. All of these issues apply especially to the prediction of invasive species, since there might be limited data in the invaded range, the species is often not currently at equilibrium and interactions with native species are completely new [8]. Despite all these challenges, SDMs have been used numerous times to provide insight into the invasive potential and the invasion dynamics of alien species [9]. One way of gaining more insight into the invasion process is to create models with data from different time periods during the invasion [10]. For example, data from a time period early in the invasion process can be used to build models which then are evaluated against data from a later time period [11].

3.3 Modelling methods

A diverse array of mathematical models have been developed to describe ecological processes. In species distribution modelling, several modelling methods have emerged with differing strengths and weaknesses, including the following:

Generalized Linear Models (GLM) [12]:

GLMs are the generalized extension of a classical linear model, meaning a linear combination of all used variables in the form of

$$Y = \beta_1 x_1 + \beta_2 x_2 + \ldots = \sum_i \beta_i x_i, \tag{1}$$

where Y is the final response value, x_i are the independent variables used in the model and β_i are the parameters that need to be estimated for each variable. This can now be generalized for non-linear relationships between the variables and the response value, for example the expectation values of a statistical distribution. A link function g (non-linear), is applied to the response in order to match the results of the (still linear) combination of model variables to a statistical distribution (i.e. normal, binomial, Poisson, gamma), resulting in

$$E = g(Y) = \sum_i \beta_i x_i. \tag{2}$$

With this transformation, $g(Y)$ now represents the expectation value E for the model in the chosen distribution.

Generalized Additive Models (GAM) [13]:

A GAM can be understood as an even further extension of a GLM. Here, variables are not estimated with a linear parameter any more, but are each contained in a "smooth function" $s()$:

$$E = \sum_i s_i(x_i) \tag{3}$$

The smooth functions are estimated non-parametrically, meaning that the form of $s()$ is not strictly predefined in comparison to GLMs, where β_i is clearly a simple parameter to be estimated. The smooth function could for example be estimated using splines, piecewise defined polynomial functions. Both GLMs and GAMS are fundamental advancements in statistical regression and have been used heavily in ecological research [14].

Boosted Regression Trees (BRT) [15]:

BRTs are a form of machine learning model using multiple decision trees and boosting to create a well performing model. Tree based models can be used for classification or regression tasks. Decision trees fit a constant value to intervals of all variables, either the most probable class (classification) or the mean response (regression) of all observed values in that interval.

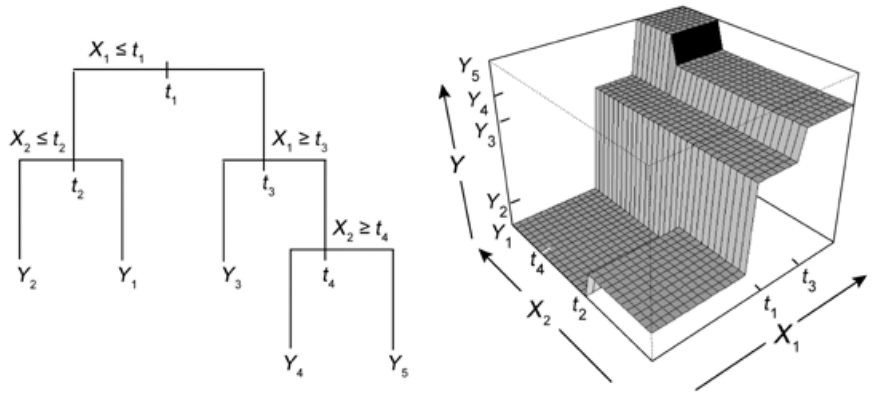

Figure 1: Visualization of a decision tree (left) and the resulting response surface for the two variables X_1 and X_2. (modified from [15])

Multiple decision trees can be combined using boosting to generate a tree ensemble with much better performance. Here, a first tree is built trying to maximize the prediction accuracy for the fitted values. All following trees are generated with emphasis on values that are not well described by the current amount of trees. This can be done by applying weights to those values for the next tree. When using BRTs for regression, boosting can be understood as optimizing the residuals of the fit function, the difference between the observed value and the estimated value. Each tree usually only contributes a very small amount to the total result, scaled by a so-called learning rate. BRTs have the potential to strongly overfit the data, meaning that the resulting function is very accurate for the used training data, but much less so for new data not used in the training process.

Maximum Entropy Method (MaxEnt) [16, 17]:

Maximum Entropy at its core is an optimization method commonly used in machine learning. It follows the concept that as few constrains as possible should be used when modelling a distribution that is currently unknown. This is achieved by maximizing the ”entropy” of the estimation, here meaning a measurement for the amount of constraints influencing the result. In MaxEnt SDMs, variables are added in form of ”features” representing different constraints put on the model for a specific variable:

- Linear feature: The mean of the response should be close to the observed value.
- Quadratic feature: The variance of the response should be close to the observed value.
- Product feature: The covariance of two variable responses should be close to the observed value.
- Threshold feature: The proportion of response values above a given threshold should be close to the observed value.
- Binary feature: The proportion of response values belonging to a certain binary value should be close to the observed value (for categorical data).
- Hinge feature: A piecewise linear function imposing constraints similar to the threshold feature.

MaxEnt models also use regularization, a process that puts penalties on complex models and produces a simpler model (fewer and less complex coefficients). It has recently been shown that MaxEnt models can be fit using infinitely weighted logistic regression, making it very similar to GLMs and GAMs in its concept, apart from the use of "features".

3.4 Harmonia axyridis

Harmonia axyridis, also known as the Harlequin ladybird or multicoloured Asian lady beetle, is of the family of the Coccinellidae and has its native origin in Asia [18]. At first widely introduced as a control species against pest aphids, *H. axyridis* has turned out to be a highly invasive species reaching an almost global distribution [19]. In America, the species was introduced as early as 1916 (California) and in 1988, first populations outside intended release were found [20]. Usage of *H. axyridis* for biological control in Europe dates back as far as 1990 (France) [21]. First invasive occurrences were confirmed in multiple countries during the early 2000s, including Germany (2000), Belgium (2001), the Netherlands (2002) and the United Kingdom (2003) [18]. The first confirmation in Austria, where it was never used for biological control, was 2006 [22]. It has been shown that all established invasive populations outside of North America have their origin in the first established population in eastern North America, with the European populations being significantly influenced by the used biocontrol strain [23].
The impact of *H. axyridis* on invaded areas is diverse. In some contexts, the ladybird has been shown to have a negative impact on the diversity and abundance of native ladybird species [18]. Many studies show intra guild predation and direct interspecific competition in favour of *H. axyridis* [24]. It has also been shown that the species feeds on a variety of damaged fruit crops, for example grapes, apples, stone fruit and berry crops [25]. The aggregating behaviour of *H. axyridis*, mostly as a strategy for overwintering, is also a cause of disturbance, since private homes and facilities are invaded by large amounts of beetles at a given time [26].

References

[1] C. o. t. E. U. European Parliament. "REGULATION (EU) No 1143/2014 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 22 October 2014 on the prevention and management of the introduction and spread of invasive alien species". In: *Official Journal of the European Union* L 317 (2014), pp. 35–55. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32014R1143.

[2] N. Shigesada and K. Kawasaki. *Biological invasions: theory and practice*. Oxford University Press, UK, 1997. URL: https://books.google.at/books/about/Biological_Invasions_Theory_and_Practice.html?id=Ri-hle_zdpsC&redir_esc=y.

[3] T. M. Blackburn et al. "A proposed unified framework for biological invasions". In: *Trends in ecology & evolution* 26.7 (2011), pp. 333–339. DOI: 10.1016/j.tree.2011.03.023.

[4] D. Simberloff et al. "Impacts of biological invasions: what's what and the way forward". In: *Trends in ecology & evolution* 28.1 (2013), pp. 58–66. DOI: 10.1016/j.tree.2012.07.013.

[5] G. A. Polis, C. A. Myers and R. D. Holt. "The ecology and evolution of intraguild predation: potential competitors that eat each other". In: *Annual review of ecology and systematics* 20.1 (1989), pp. 297–330. DOI: 10.1146/annurev.es.20.110189.001501.

[6] J. Elith and J. R. Leathwick. "Species distribution models: ecological explanation and prediction across space and time". In: *Annual review of ecology, evolution, and systematics* 40 (2009), pp. 677–697. DOI: 10.1146/annurev.ecolsys.110308.120159.

[7] M. B. Araújo and A. Guisan. "Five (or so) challenges for species distribution modelling". In: *Journal of biogeography* 33.10 (2006), pp. 1677–1688. DOI: 10.1111/j.1365-2699.2006.01584.x.

[8] K. P. Mainali et al. "Projecting future expansion of invasive species: comparing and improving methodologies for species distribution modeling". In: *Global change biology* 21.12 (2015), pp. 4464–4480. DOI: 10.1111/gcb.13038.

[9] N. E. Zimmermann et al. "New trends in species distribution modelling". In: *Ecography* 33.6 (2010), pp. 985–989. DOI: 10.1111/j.1600-0587.2010.06953.x.

[10] R. D. Briscoe Runquist et al. "Species distribution models throughout the invasion history of Palmer amaranth predict regions at risk of future invasion and reveal challenges with modeling rapidly shifting geographic ranges". In: *Scientific Reports* 9.1 (2019), p. 2426. DOI: 10.1038/s41598-018-38054-9.

[11] M. Barbet-Massin et al. "Can species distribution models really predict the expansion of invasive species?" In: *PloS one* 13.3 (2018), e0193085. DOI: 10.1371/journal.pone.0193085.

[12] J. A. Nelder and R. W. Wedderburn. "Generalized linear models". In: *Journal of the Royal Statistical Society Series A: Statistics in Society* 135.3 (1972), pp. 370–384. DOI: 10.2307/2344614.

[13] T. Hastie and R. Tibshirani. “Generalized Additive Models”. In: *Statistical Science* 1.3 (1986), pp. 297–310. URL: https://hastie.su.domains/Papers/gam.pdf.

[14] A. Guisan, T. C. Edwards Jr and T. Hastie. “Generalized linear and generalized additive models in studies of species distributions: setting the scene”. In: *Ecological modelling* 157.2-3 (2002), pp. 89–100. DOI: 10.1016/S0304-3800(02)00204-1.

[15] J. Elith, J. R. Leathwick and T. Hastie. “A working guide to boosted regression trees”. In: *Journal of animal ecology* 77.4 (2008), pp. 802–813. DOI: 10.1111/j.1365-2656.2008.01390.x.

[16] S. J. Phillips, R. P. Anderson and R. E. Schapire. “Maximum entropy modeling of species geographic distributions”. In: *Ecological modelling* 190.3-4 (2006), pp. 231–259. DOI: 10.1016/j.ecolmodel.2005.03.026.

[17] S. J. Phillips et al. “Opening the black box: An open-source release of Maxent”. In: *Ecography* 40.7 (2017), pp. 887–893. DOI: 10.1111/ecog.03049.

[18] H. E. Roy et al. “The harlequin ladybird, Harmonia axyridis: global perspectives on invasion history and ecology”. In: *Biological invasions* 18 (2016), pp. 997–1044. DOI: 10.1007/s10530-016-1077-6.

[19] P. Brown et al. “Harmonia axyridis in Europe: spread and distribution of a non-native coccinellid”. In: *From biological control to invasion: the ladybird Harmonia axyridis as a model species* (2008), pp. 5–21. DOI: 10.1007/978-1-4020-6939-0_2.

[20] J. B. Chapin, V. Brou et al. “Harmonia axyridis (Pallas), the third species of the genus to be found in the United States (Coleoptera: Coccinellidae)”. In: *Proc. Entomol. Soc. Wash* 93.3 (1991), pp. 630–635. URL: https://www.biodiversitylibrary.org/partpdf/55539.

[21] J. Coutanceau. “Harmonia axyridis (Pallas, 1773): Une coccinelle asiatique introduite, acclimatée et en extension en France”. In: *Bulletin de la Société Entomologique de France* 111 (Jan. 2006), pp. 395–401. DOI: 10.3406/bsef.2006.16343.

[22] W. Rabitsch and R. Schuh. “First record of the multicoloured Asian ladybird Harmonia axyridis (Pallas, 1773) in Austria”. In: *Beiträge zur Entomofaunistik* 7 (2006), pp. 161–164. URL: https://www.zobodat.at/pdf/BEF_7_0161-0164.pdf.

[23] E. Lombaert et al. “Bridgehead effect in the worldwide invasion of the biocontrol harlequin ladybird”. In: *PloS one* 5.3 (2010), e9743. DOI: 10.1371/journal.pone.0009743.

[24] J. K. Pell et al. “Intraguild predation involving Harmonia axyridis: a review of current knowledge and future perspectives”. In: *BioControl* 53 (2008), pp. 147–168. DOI: 10.1007/978-1-4020-6939-0_10.

[25] R. Koch et al. “Phytophagous preferences of the multicolored Asian lady beetle (Coleoptera: Coccinellidae) for autumn-ripening fruit”. In: *Journal of Economic Entomology* 97.2 (2004), pp. 539–544. DOI: 10.1093/jee/97.2.539.

[26] C. Nalepa, G. Kennedy and C. Brownie. “Role of visual contrast in the alighting behavior of Harmonia axyridis (Coleoptera: Coccinellidae) at overwintering sites”. In: *Environmental Entomology* 34.2 (2005), pp. 425–431. DOI: 10.1603/0046-225X-34.2.425.

Appendix