Thesis Outline

Lukas Prader

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.

This jumps quite quickly to a pretty technical topic, it will need more explanation

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. This also comes into play when trying to model 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.
This thesis aims to analyse the spread of *Harmonia axyridis* in detail, as well as creating new insight into the progression of an ensemble SDM iterating over each year of sampling.

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. 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 (Phillips et al. 2009). (correction for overlap with following years?)

2.3 Model building

For each year, the following Models were computed: General Linear (GLM), General Additive (GAM), Multivariate Adaptive Regression Splines (MARS), Boosted Regression Trees (BRT) and Maximum Entropy (MAXENT). (Model choices subject to change) Each model used either only data points from Europe (eu) or from Europe and the native range in Asia (eu+as). A model for a specific year always included all points from past years as well. For all used occurrence points prior to 1992 and after 2020, the land cover data of 1992 or 2020 was used as a substitute. (necessary for after 2020 if only for verification?)
To select the variables used, bioclim variables were chosen using variance inflation values. For the land cover data, a PCA was computed and added to the chosen bioclim variables.

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,

though only using data points from the year in question (Broennimann et al. 2011). These niches were again compared to the following year and 2022.