

Predicting likely establishment locations of non-native pests from past reports of first establishments

The potential of generative models

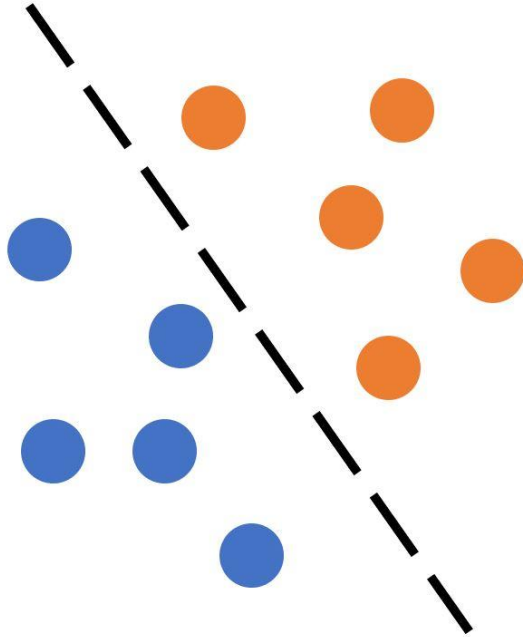
Robbert van den Dool, Alejandro Morales, Wopke van der Werf, Bob Douma



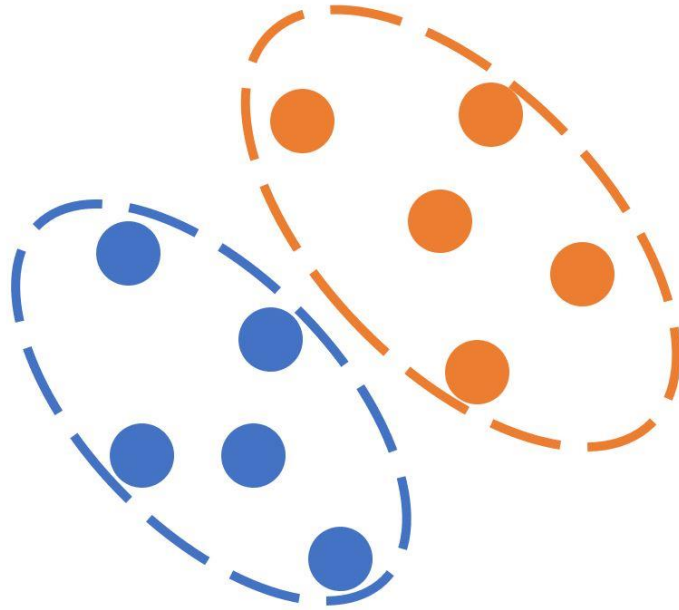
Generative models vs discriminative models

- Discriminative models: model for relative probability of presence based on site characteristics (one model)
- Generative models: two models, one for site characteristics of presence locations and one for site characteristics of all reference locations. Relative probability of presence calculated from this by Bayes' rule (two models)
- Discriminative models: well-developed, e.g. maxent
- Generative models: little explored; potential unclear; better or worse than discriminative?

Discriminative



Generative



Possible advantages and disadvantages of generative models; why we did this

- Advantage: greater transparency – site characteristics of reported presence locations, sample locations (e.g. from GBIF), and all locations (grid) are modelled separately
- Disadvantage: not a single step model; harder to optimize predictive performance
- We were curious about the potential of generative models for modelling first establishment locations of non-native forest pests
- Co-funded by EU project HOMED

Possible advantages and disadvantages of generative models; why we did this

- Advantage: greater transparency – site characteristics of reported presence locations, sample locations (e.g. from GBIF), and all locations (grid) are modelled separately
- Disadvantage: not a single step model; harder to optimize predictive performance
- We were curious about the potential of generative models for



**HOLISTIC MANAGEMENT OF EMERGING FOREST
PESTS AND DISEASES**

The problem

- Invasions of non-native pests are on the rise and they are very damaging
- EU member states are required to conduct survey, but untargeted survey is inefficient and sampling resources are limited
- There is a need for identifying locations more likely to have newly established non-native pests, so these can be surveyed with priority
- Need for identifying site conditions that increase the probability of first establishment

The data

- Data is available on first establishment locations (e.g. Branco et al., 2019; Neobiota 52: 25-46)
- Many of those data are voluntary reports by citizens, e.g. citizen data
- Reported establishment locations not only indicate where establishment has happened, but also where that was reported
- The probability of reporting varies as there is a sampling bias (some locations have a higher chance of visitation than others)
- What drives sampling bias?
- What is the effect of sampling bias?

Definitions & notation

Site conditions, e.g. distance to nearest city, # of people in NUTS region, forest cover in 1x1 km², etc.: $f(\theta)$, where f is probability density and θ is a vector of site characteristics

Sampling: $S = 1$ (sampling) or $S = 0$ (no sampling)

Presence: $P = 1$ (presence) or $P = 0$ (absence)

Reporting: $R = 1$ (report available) or $R = 0$ (no report)

For voluntary reporters, only reports of presence are entered, no reports of absence

Definitions & notation (2)

$f(\theta|S = 1)$: site characteristics at sampling locations: analyse, e.g., GBIF data

$f(\theta|P = 1)$: site characteristics at presence locations: this is what we want to know (but we don't know)

$f(\theta|R=1)$: site characteristics at reporting locations: analyse, e.g., Branco's dataset

Definitions & notation (3)

$\Pr(S = 1|\theta)$: probability of sampling at sites with characteristics θ

$\Pr(P = 1|\theta)$: probability of presence at sites with characteristics θ

$\Pr(R = 1|\theta)$: probability of reporting of pest presence at sites with characteristics θ

$P_f(S|\theta)$ What is the probability of supply at places with characteristics θ ?

$P(P|\theta)$ What is the probability of presence at places with characteristics θ ?

$P(R|\theta)$ What is the probability of reports at places with characteristics θ ?

$$P(R=1|\theta) = P(P=1|\theta) * P(S=1|\theta)$$

$$P(P=1|\theta) = \frac{P(R=1|\theta)}{P(S=1|\theta)}$$

$$P(R=1|\theta) f(\theta) = f(\theta|R=1) P(R=1)$$

$$P(R=1|\theta) = \frac{P(R) f(\theta|R=1)}{f(\theta)}$$

unknown

$$P(R=1|\theta) = \frac{f(\theta|R=1)}{f(\theta)}$$

$$P(S=1|\theta) f(\theta) = f(\theta|S=1) P(S=1)$$

$$P(S=1|\theta) = \frac{f(\theta|S=1) P(S=1)}{f(\theta)}$$

$$P(P=1|\theta) = \frac{P(R) \frac{f(\theta|R=1)}{f(\theta)}}{P(S) \frac{f(\theta|S=1)}{f(\theta)}}$$

$$= \frac{P(R) f(\theta|R=1)}{P(S) f(\theta|S=1)}$$

$$P(P=1|\theta) \sim \frac{f(\theta|R=1)}{f(\theta|S=1)}$$

To cut a long story short:

We use Bayes' rule:

$$\Pr(Y = 1|\theta) = \frac{f(\theta|Y = 1) \Pr(Y = 1)}{f(\theta)}$$

And obtain the result:

$$\Pr(P = 1|\theta) \sim \frac{f(\theta|R = 1)}{f(\theta|S = 1)}$$

i.e. the probability of presence, given theta, is proportional to the ratio of the probability density of theta at the reporting locations and the probability of theta at the sampling locations

Available data

Reports: 273 from literature

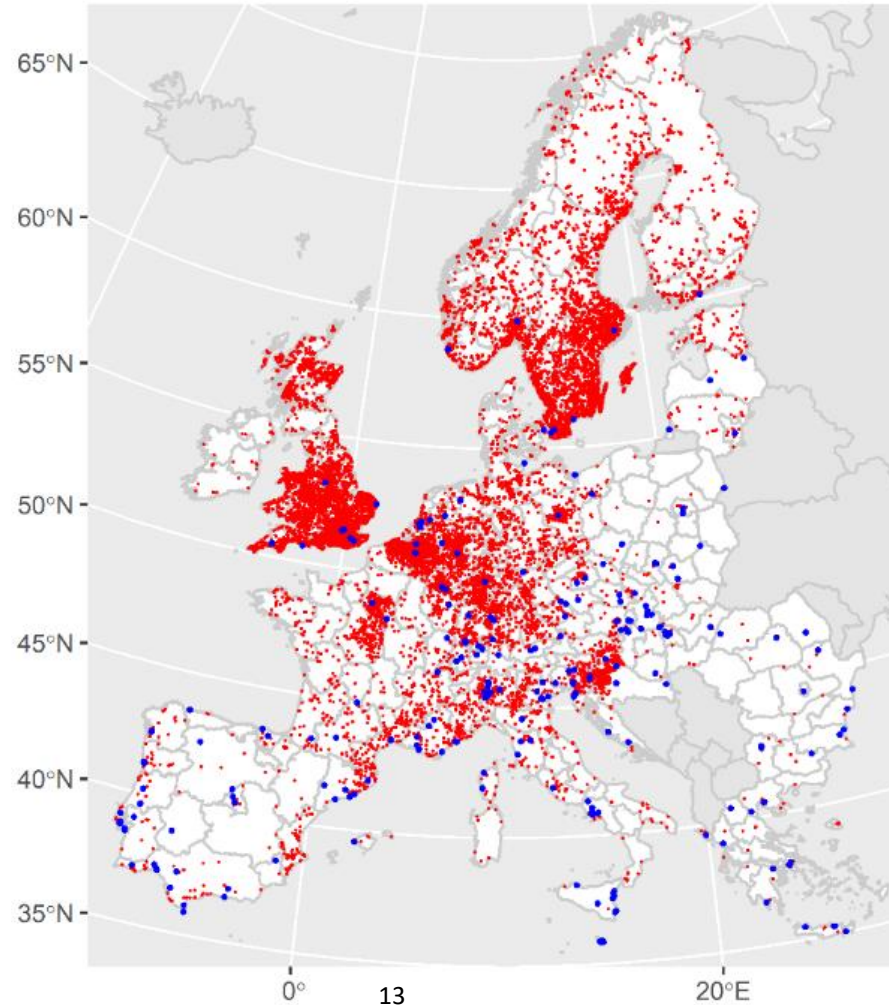
From Branco et al. 2019

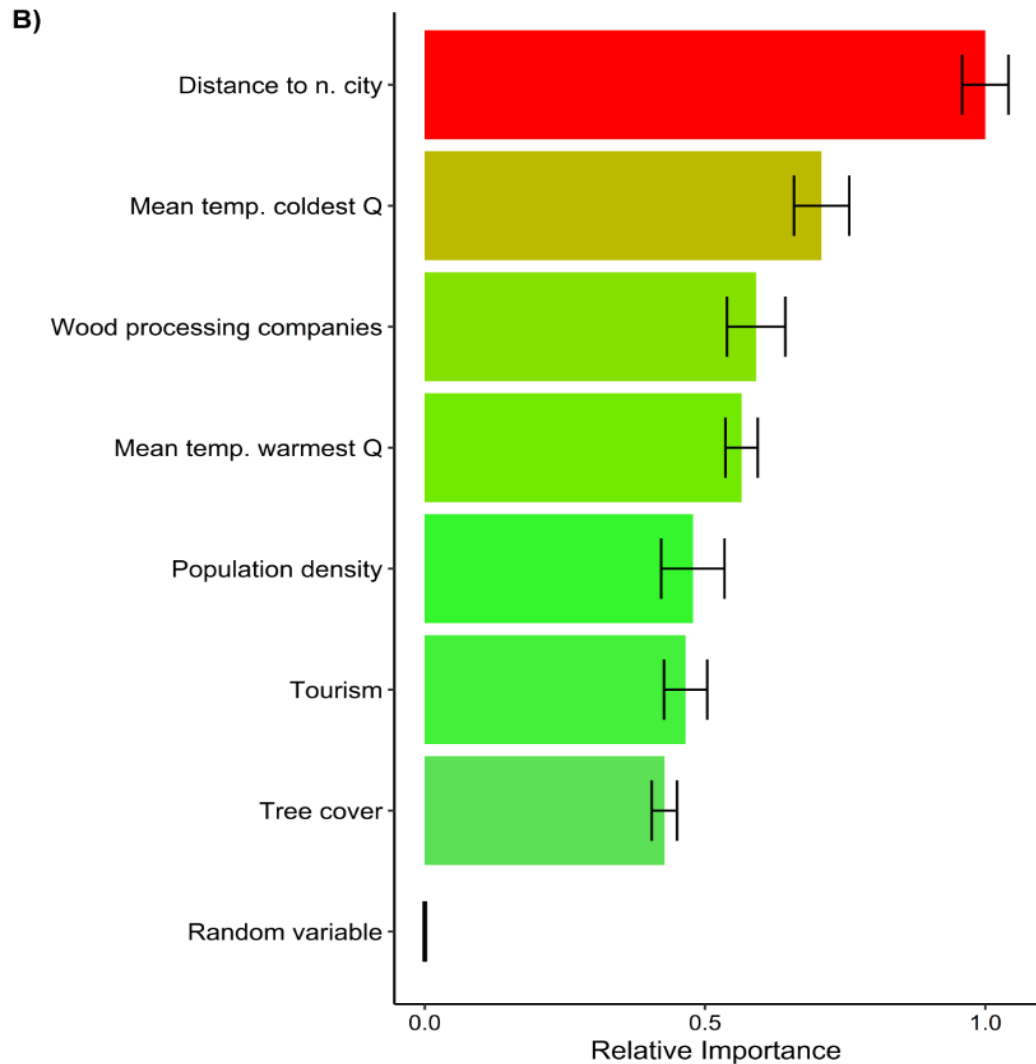
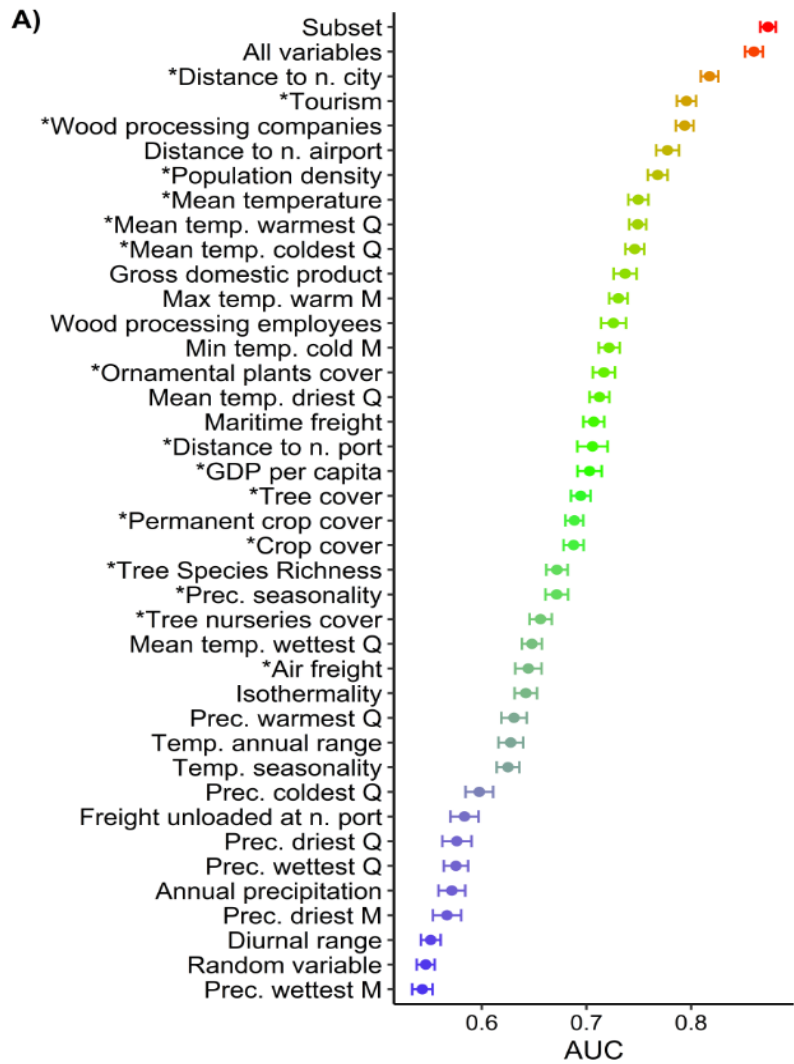
- > 1950
- Feeding on woody plants
- First records in countries

Sampling

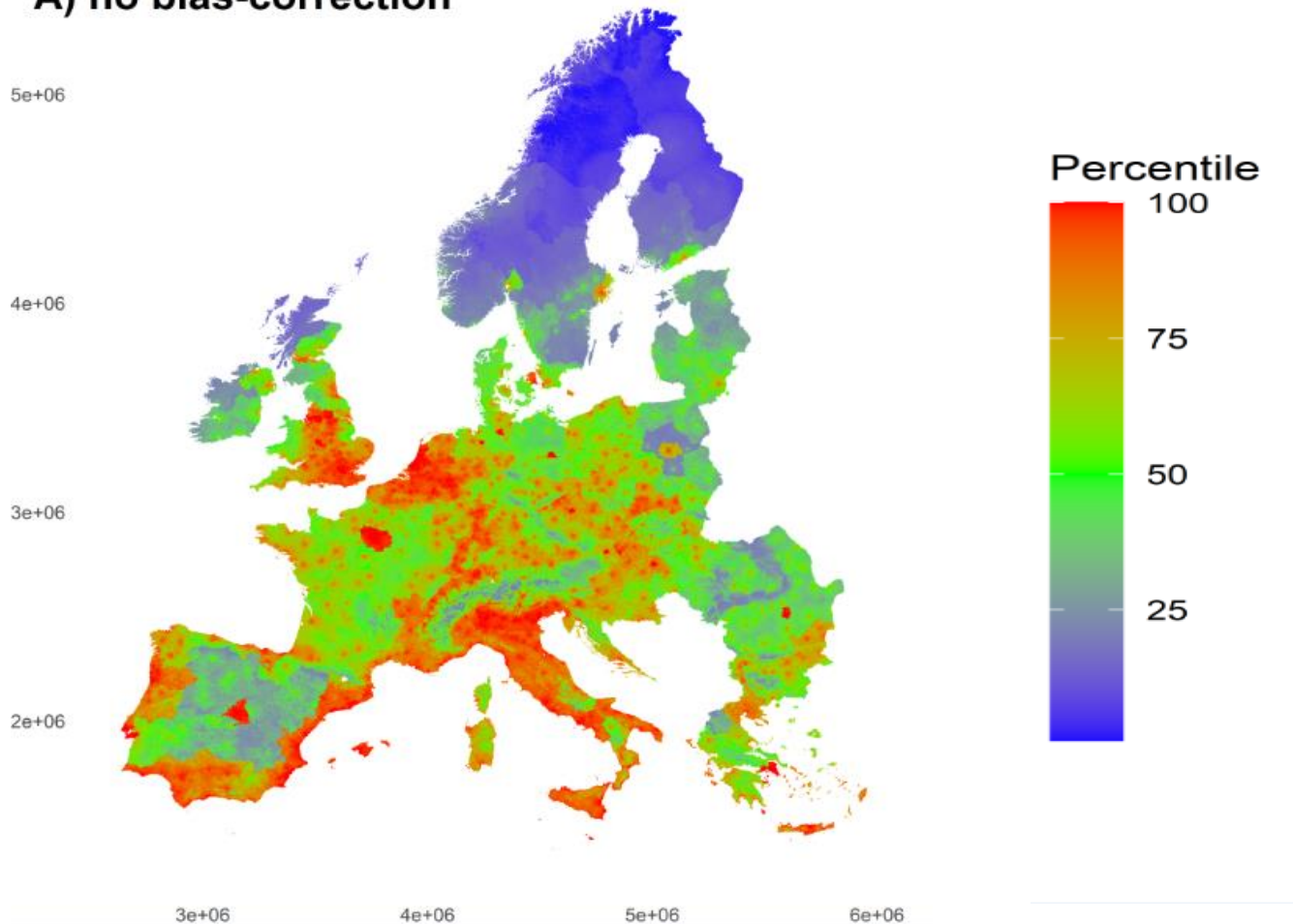
All (human) observations in gbif.org
on same genera (55,000),

Inaccuracy < 1km

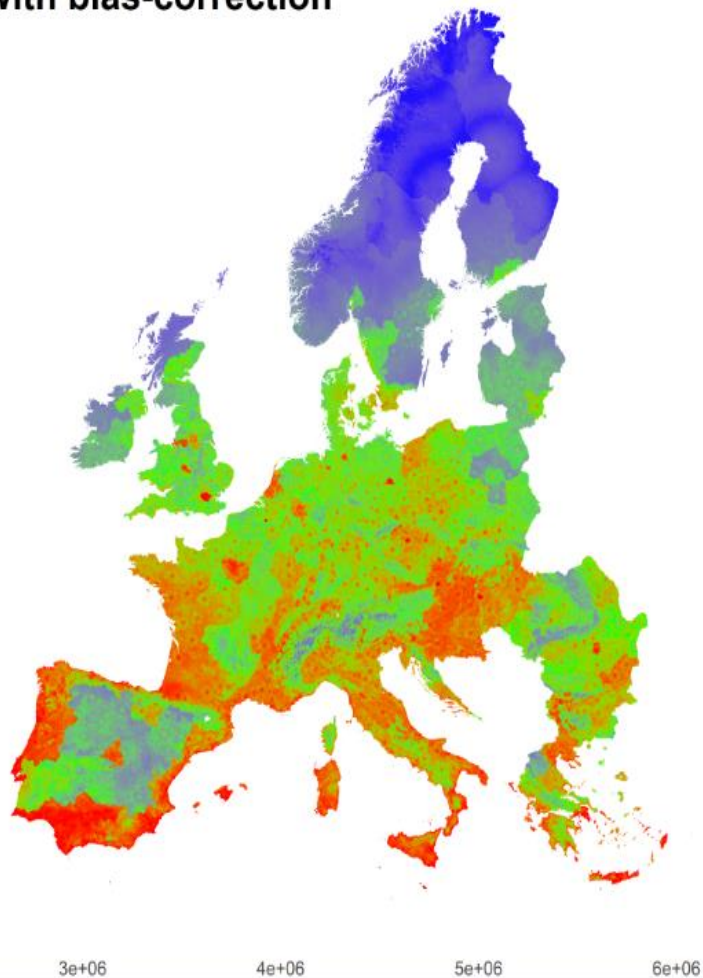




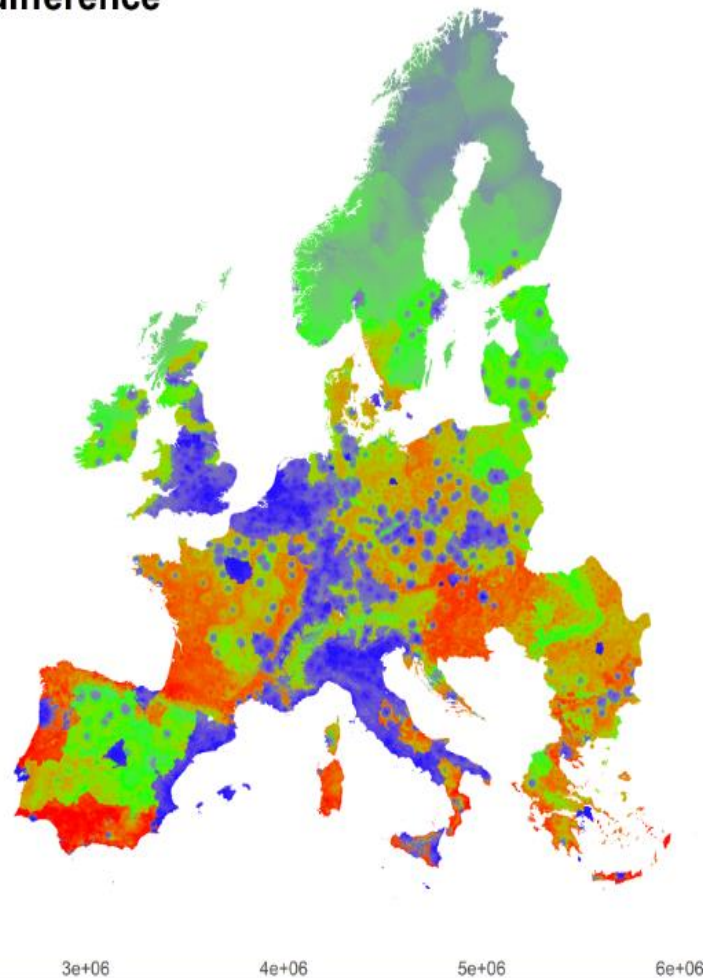
A) no bias-correction



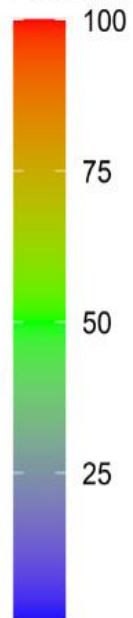
B) with bias-correction



C) difference



Percentile



100

75

50

25

3e+06

4e+06

5e+06

6e+06

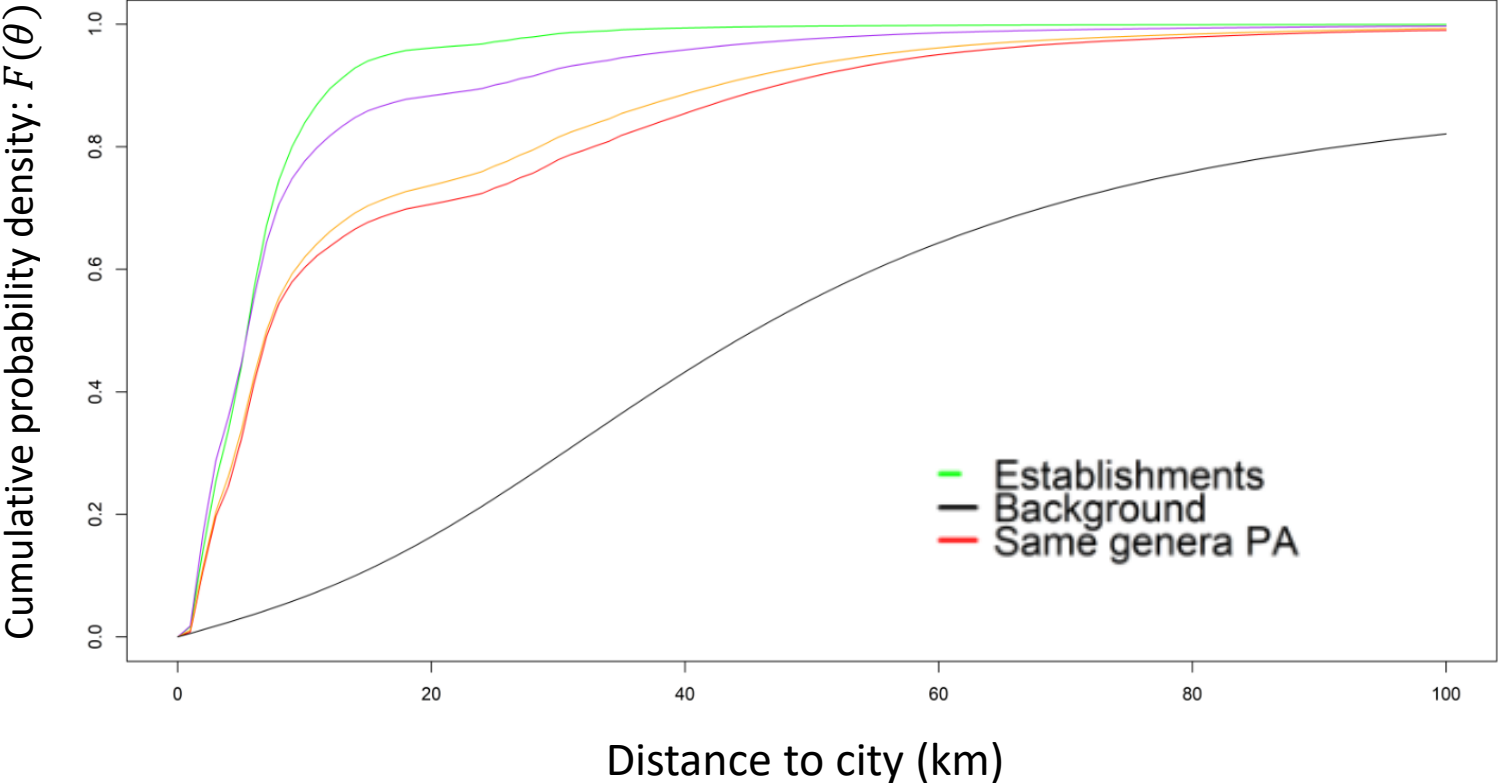
3e+06

4e+06

5e+06

6e+06

Cumulative probability density functions for reported presence (green), all locations in Europe (black), GBIF reporting for same genera (red), and GBIF reporting for same genera (purple)



Main findings of Robbert's thesis

- Generative modelling is as good in prediction as discriminative modelling with Maxent
- Shown by benchmarking study with 226 species (Chapter 2)
- The city effect in first establishment locations is to a large extent due to sampling bias (Chapter 3)
- Accounting for this bias can nullify the city effect
- Warmer regions in Europe are more at risk of non-native pest establishment
- When grouping species according to their traits, the area of origin was the most important. Species from the southern hemisphere (Neotropic, Afrotropic, Indomalaya, Australasia) were found at warmer temperatures and mainly invaded southern Europe (Chapter 4)

Mixed cropping as a means to sustainable agriculture

[Home](#) / [Mixed cropping as a means to sustainable agriculture](#)



Postgraduate Course

Mixed cropping as a means to sustainable agriculture: Pro's, Con's and solutions

Monday 23 - Friday 27 September 2024

Park Hotel De Bosrand, Ede, the Netherlands

Thank you!