Trustworthy species distribution modelling

Reliability audits for inference from imperfect observations

Kristian Miok

Faculty of Computer and Information Science, University of Ljubljana

Joint work with Lucian Pârvulescu, Blaž Škrlj and Marko Robnik-Šikonja

2nd SMASHing Workshop · Lanthieri Mansion, Vipava · 20 October 2026

I FEEL SLOVENIA

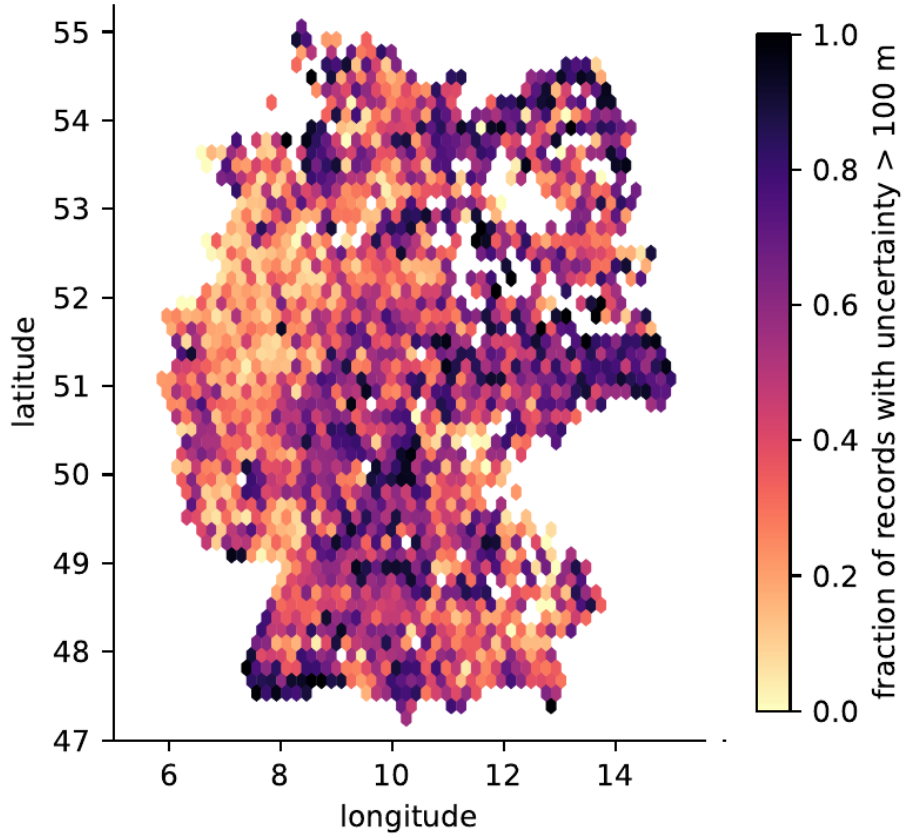
Co-funded by the European Union

SMASH machine learning for science and humanities postdoctoral program

Models are built on data the modeller did not collect

- Aggregated archives: thousands of contributors, instruments and decades in one table
- Each record: a coordinate → the environment at that point → the model learns the niche
- Positional quality ranges from a GPS point to a coordinate inferred from a place name

How far can the predictions be trusted — and where exactly do they fail?

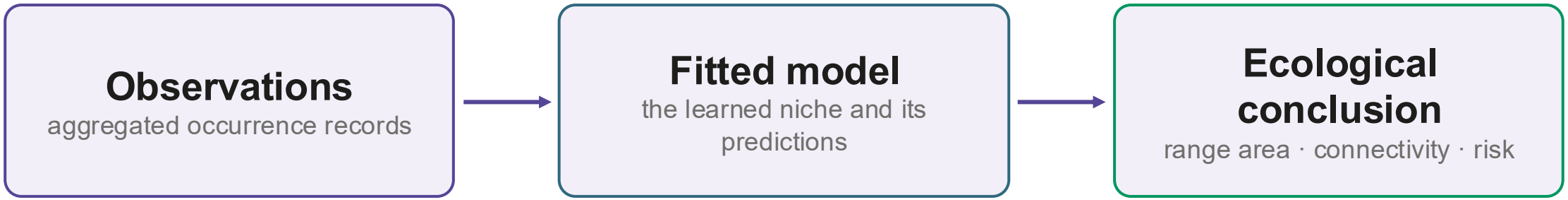


GBIF, Germany — 240,862 records in seven taxon groups: share of records with coordinate uncertainty above 100 m

I FEEL SLOVENIA
Co-funded by the European Union

SMASH
machine learning for science and humanities postdoctoral program

Reliability propagation: error travels — and often grows



AUDIT

inspect the observations before they go in

CALIBRATE

correct the predictions that come out

PROPAGATE

carry the residual uncertainty into the decision

Error is not absorbed on the way. It travels — and often grows.

Today: the first link, in depth.

I FEEL SLOVENIA
Co-funded by the European Union

SMASH
machine learning for science and humanities postdoctoral program

When cleaning becomes biasing

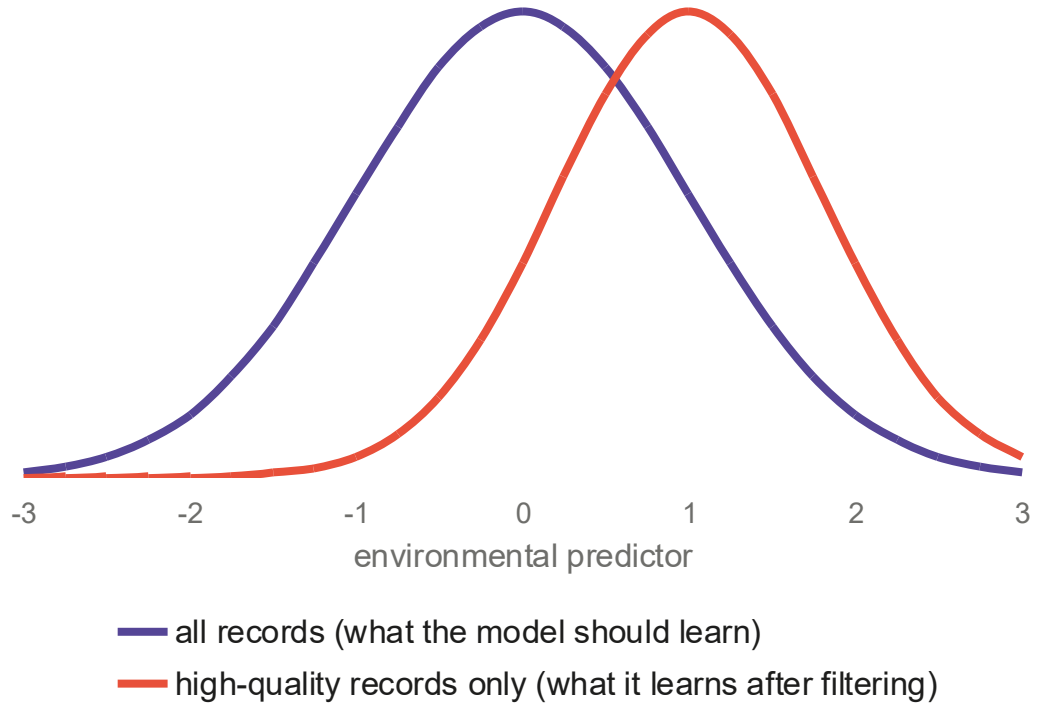
1. The standard step: keep the records flagged high-quality, drop the rest

2. Neutral only if quality is unrelated to the predictors

3. Otherwise the removal is non-random across predictor space: the learned distribution shifts — covariate shift, induced by the cleaning itself

Low-SNR detections · low-quality samples · noisy transcripts — the same step in every field

Illustration: the distribution the model learns from



Schematic, not data: low-quality records cluster in part of predictor space, so dropping them moves the whole distribution.

I FEEL SLOVENIA
Co-funded by the European Union

SMASH
machine learning for science and humanities postdoctoral program

A diagnostic that needs no ground truth

Table: the predictors at each record, plus its quality flag

↓

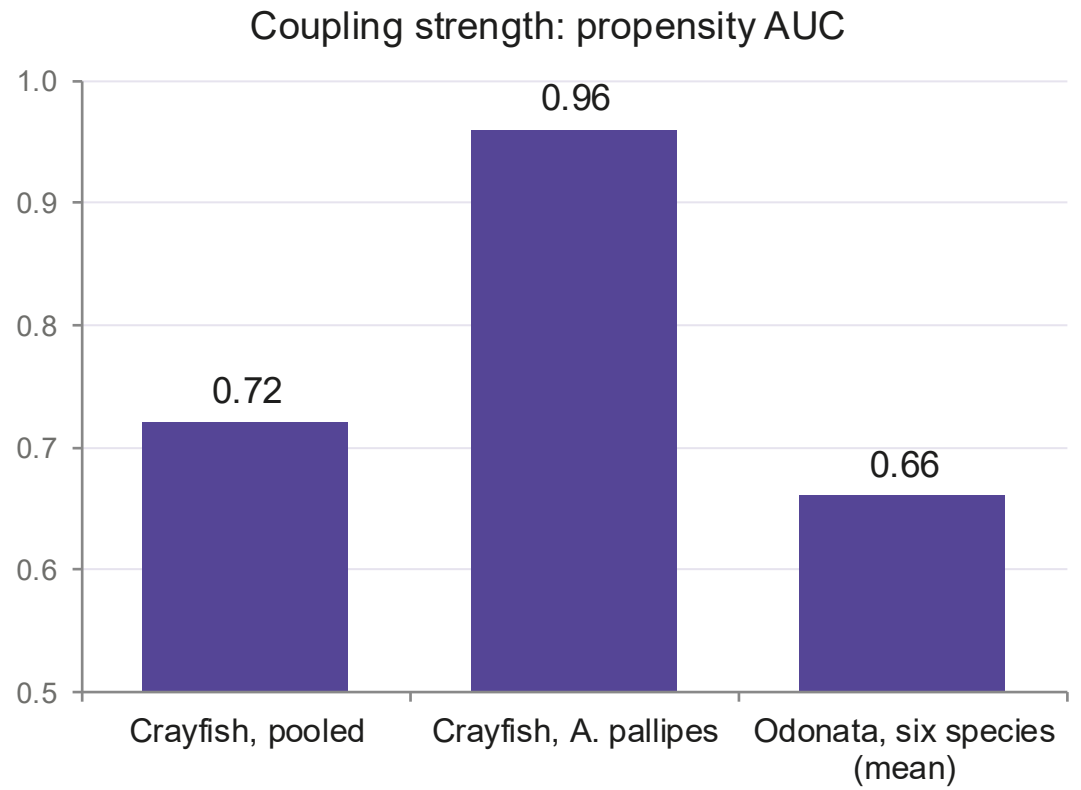
Fit a classifier for P(high quality | predictors) — the propensity

↓

Read off the AUC (coupling strength) and the axes that carry it

AUC ≈ 0.5 → filtering is random: safe

AUC > 0.5 → filtering displaces the data — along exactly those axes



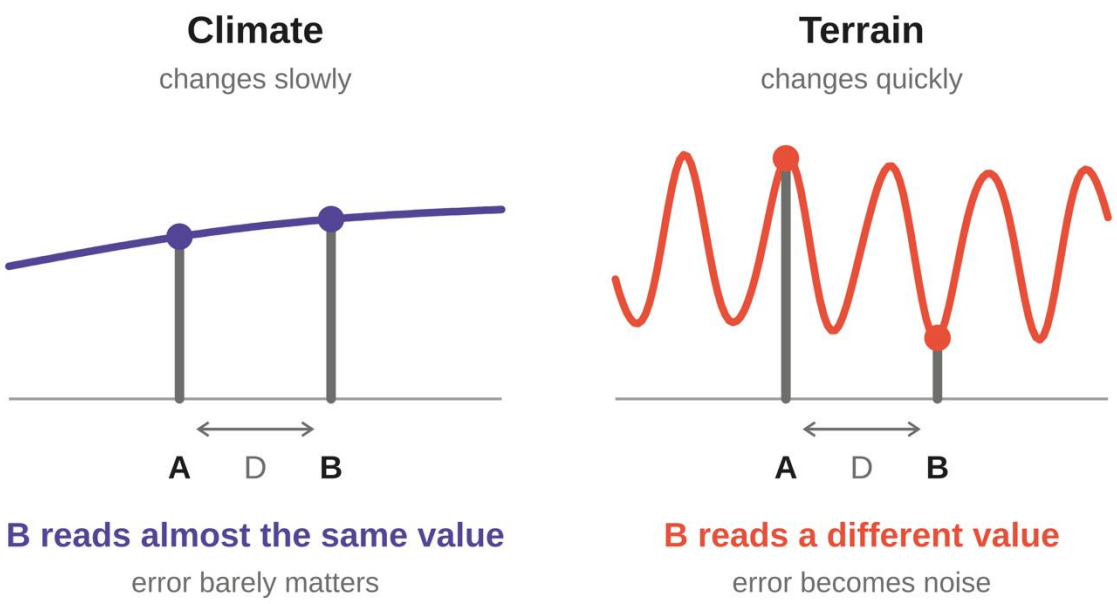
Bars start at 0.5 = chance (filtering random). Crayfish: curated global database, 7 species; Odonata: open aggregator, Germany.

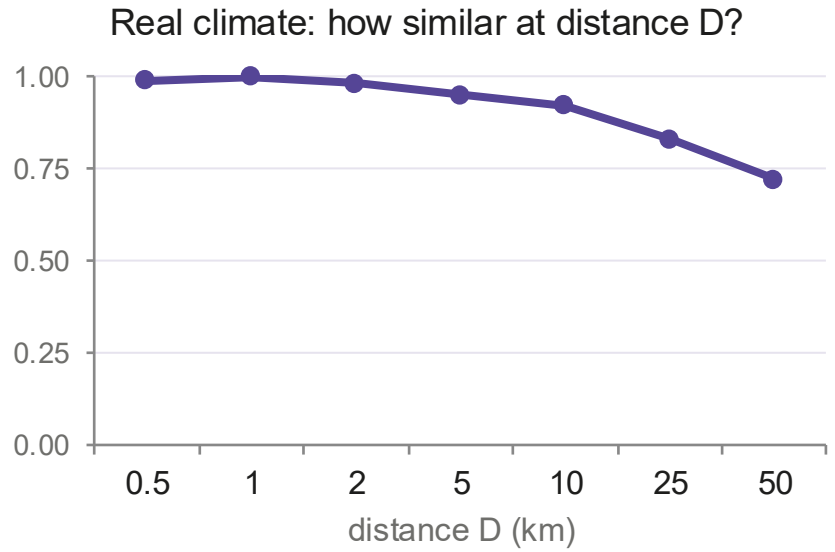
I FEEL SLOVENIA
Co-funded by the European Union

SMASH
machine learning for science and humanities postdoctoral program

Same error, different damage

A record is really at A, but the database says B. The model reads the environment at B. Whether that hurts depends on one thing: how much the environment changes over the error distance D.





- Dragonflies: errors of a few hundred metres
- Crayfish: tested up to 25 km — climate still similar

Small error or slowly changing environment → the moved record is still useful.

I FEEL SLOVENIA
Co-funded by the European Union

SMASH
machine learning for science and humanities postdoctoral program

Coupling is the rule, danger is the exception

115,191

crayfish · curated global database · 7 species

42,271

dragonflies · open aggregator · Germany · 6 species

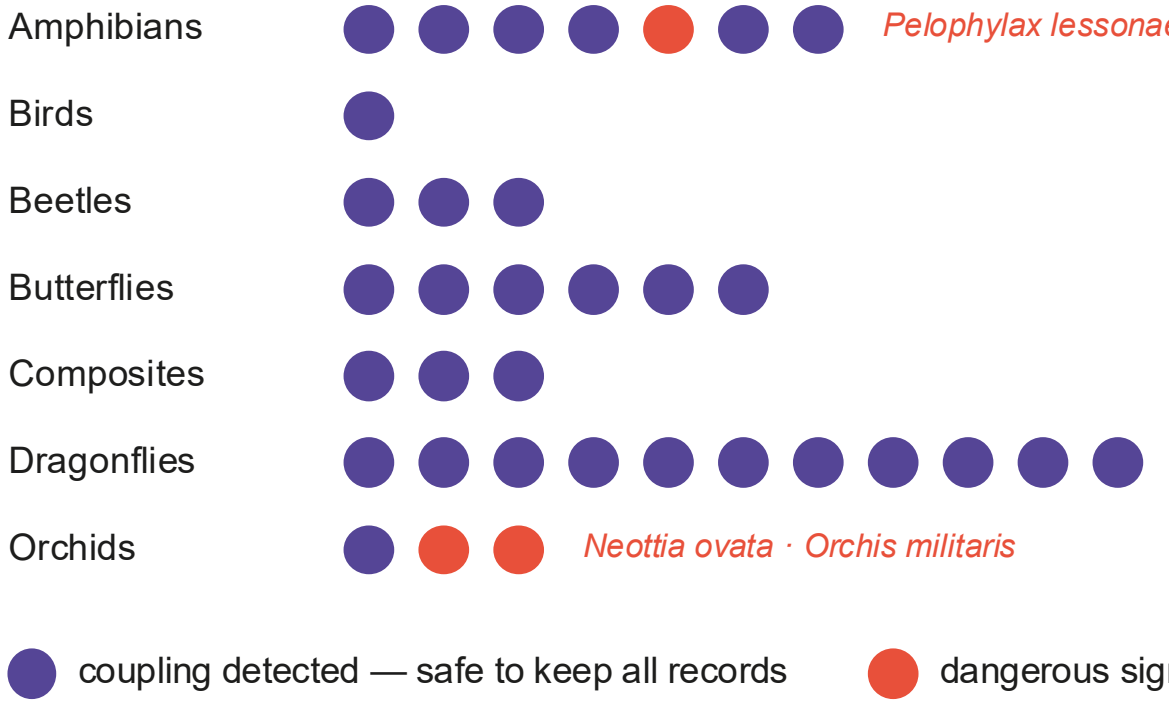
240,859

survey · 7 taxon groups · Germany · 34 units

We ran the diagnostic on all of it — no ground truth needed.

Almost everything is coupled. About 3 of 34 are dangerous. The diagnostic tells them apart.

Survey, Germany: 34 analysis units



Dangerous = kilometre-scale uncertainty on an environment that changes quickly. Crayfish and dragonflies: coupled, but protected (previous slide).

I FEEL SLOVENIA
Co-funded by the European Union

SMASH
machine learning for science and humanities postdoctoral program

Against known truth, filtering costs more than it saves

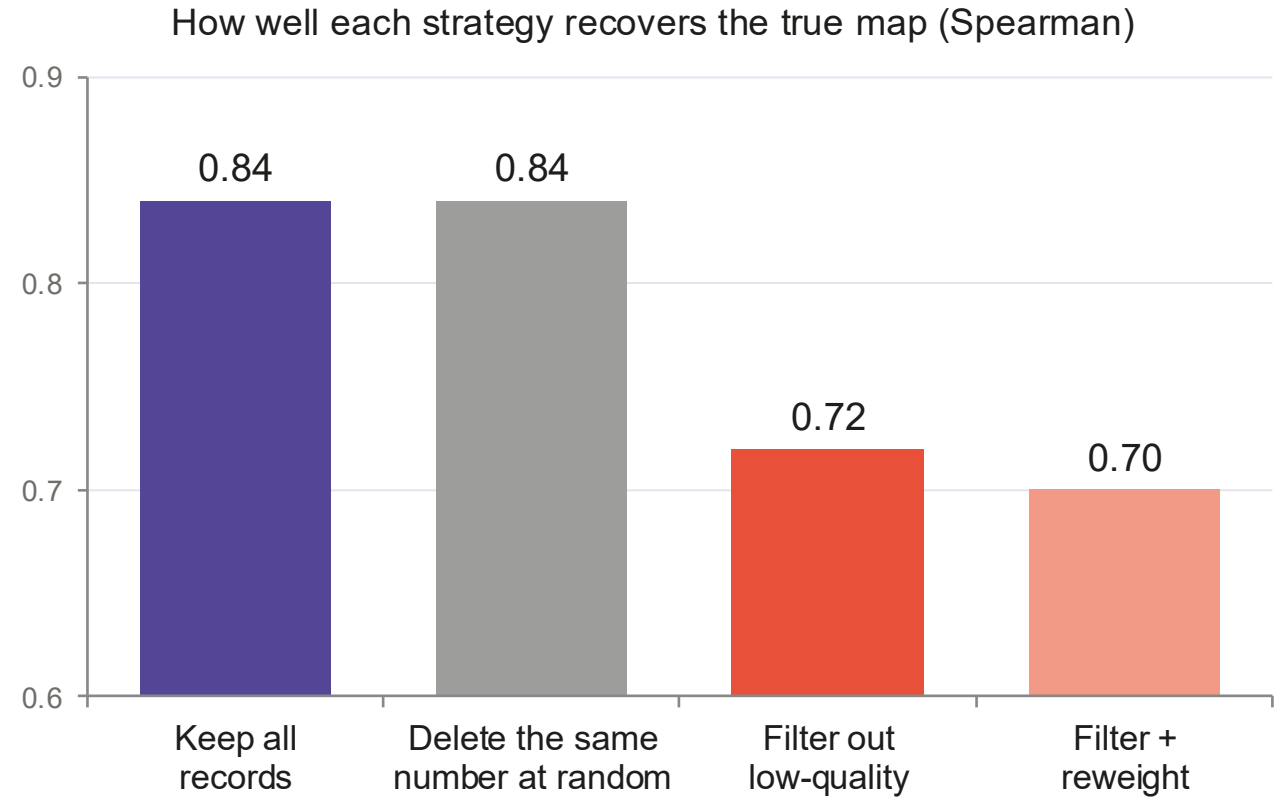
The test

Take precise crayfish records — the truth. Move some of them on purpose, the way imprecise records are off in real data. Now score every strategy: how well does its map agree with the true one?

0.12

worse than deleting the same number of records at random. Same amount of data lost — different records lost. That gap is the bias.

A moved record is still useful. A hole in the data is a hole.



Axis starts at 0.6 to show the gap. Keeping all records with corrected coordinates scores 0.82–0.83; same picture for every kind of error we tried.

I FEEL SLOVENIA
Co-funded by the European Union

SMASH
machine learning for science and humanities postdoctoral program

Error is amplified downstream, not absorbed

0.03–0.05

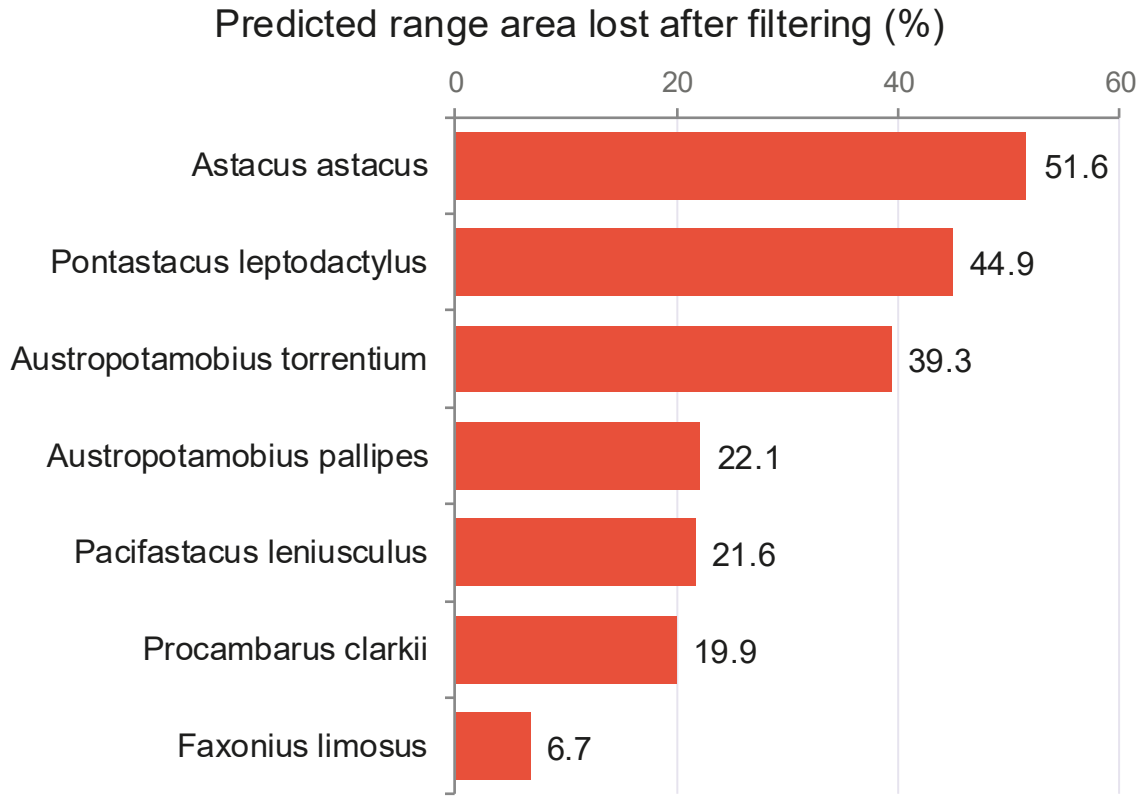
mean change in predicted suitability per cell after filtering

↓

7–52%

change in predicted range area across species (mean 29%)

Range area is the quantity decisions actually use.



Random forest, top-decile threshold · same records before vs after quality filtering · 7 crayfish species

I FEEL SLOVENIA
Co-funded by the European Union

SMASH
machine learning for science and humanities postdoctoral program

The wider reliability layer

AUDIT

- Quality-filter diagnostic: is filtering neutral, and along which axes?
- ingrain R package (v1.0): coordinate uncertainty against the analysis grain

→

CALIBRATE

- Miscalibration is spatially structured: it concentrates at stream-network headwaters
- Adaptive conformal correction of range inflation, with coverage guarantees

→

PROPAGATE

- Residual uncertainty carried into range area and into connectivity in dendritic river networks
- Under construction

Designed to sit alongside biomod2 / flexsdm — not instead of them.

I FEEL SLOVENIA
Co-funded by the European Union

SMASH
machine learning for science and humanities postdoctoral program

Wherever a quality flag correlates with the predictors, cleaning is a modelling decision.

1. Predict your quality flag from your features. If you can, filtering is not neutral.
2. Compare filtered vs keep-all-weighted. Report the difference as a sensitivity analysis.
3. Propagate that difference to the quantity you actually decide on.

Thank you

Kristian Miok · FRI, University of Ljubljana · github.com/KristianMiok

I FEEL SLOVENIA

Co-funded by the European Union

SMASH
machine learning for science and humanities postdoctoral program