BioSpace25 - Biodiversity insight from Space
10 - 14 February 2025 | ESA-ESRIN | Frascati - Italy

The challenges of broad-scale biodiversity intactness estimation

Jakob Nyström[1], Lisa Mandle[2], Jeffrey Smith[3], Tobias Andermann[1]

1. Uppsala University; 2. Natural Capital Project, Stanford University; 3. Princeton University.

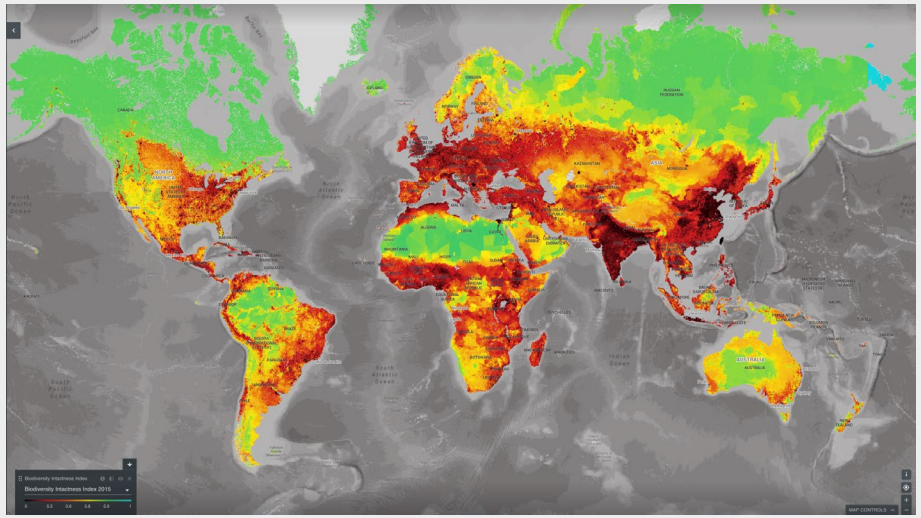
High demand for data products showing state of biodiversity in relation to anthropogenic pressures

BII: Biodiversity intactness index

Predicts | Natural History Museum

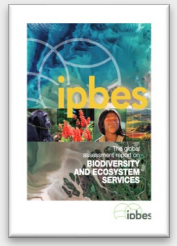
GLOBIO
Global biodiversity model for policy support

SEED



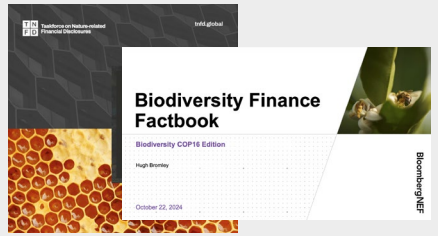
Example: BII intactness heatmap

Measure state of biodiversity

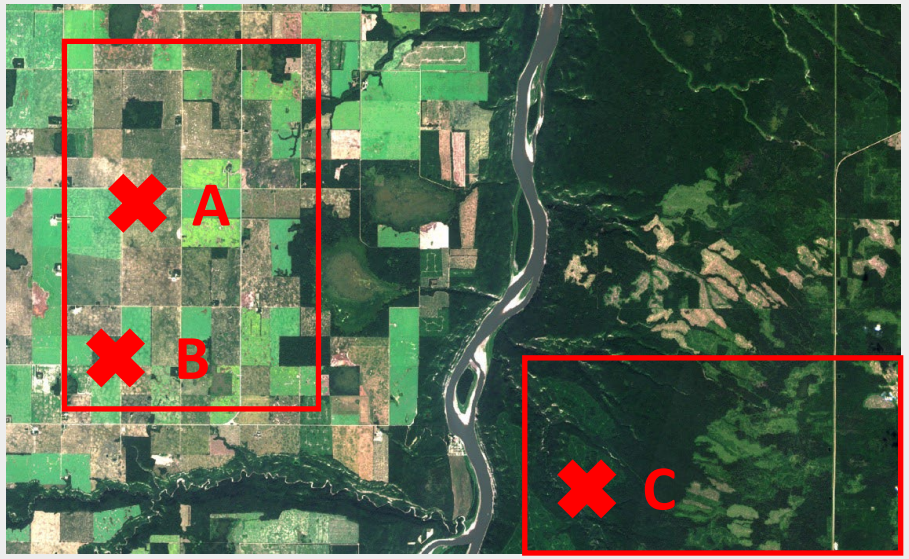


Monitor biodiversity targets

Quantify corporate impacts



Biodiversity intactness: Estimated diversity of a site relative to primary vegetation reference site



*Data from multiple **studies***

*Sites sometimes organized in **blocks***

Site A: Cropland

 0  3  1  2

Site C: Primary vegetation

 3  1  4  3

Relative abundance: 0.55

Compositional similarity: 0.47

How well can such models predict on unseen data?

⚠ **Key gap:** Model-based intactness indicators not tested for predictive performance

? How well do current models generalize to unseen data?

? Are there approaches that are spatially more granular, and generalize better?

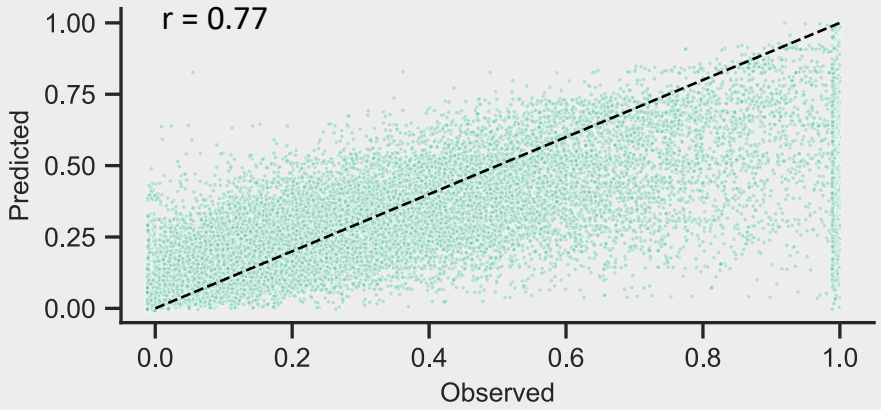
Testing shows limited generalization capabilities

Model example: BII

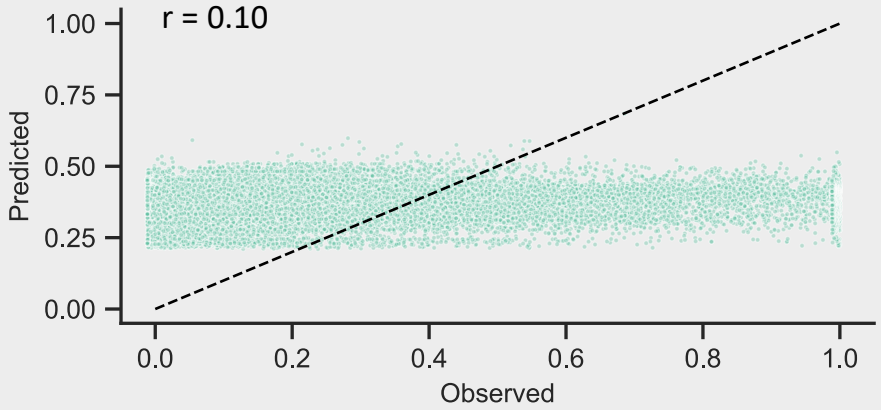
- BII alpha diversity model[1]
- PREDICTS data: 680 studies and 26,753 sites (many species groups)

Note: r = Pearson correlation

In-sample *predictions (training data)*



Out-of-sample *predictions (test data)*

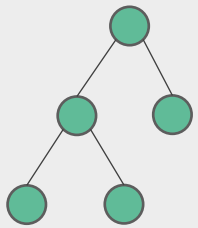


1. Following De Palma et al 2021.

Why lack of generalization? How can we improve it?

Limitations of the BII model

- Linear mixed model with study-block random effects → ***explanatory*** model
- Does not work well as ***predictive*** model
  - Random effects can't be used on test data
  - Fixed effects averaged across all data



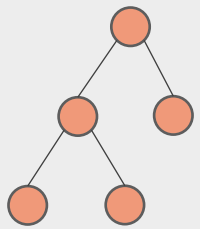
Population level *fixed effects*

Study *random slopes / intercepts*

Spatial block *random intercepts*

Potential improvements explored

- **New model family:** Bayesian hierarchical models
- **New hierarchy:** Biogeographic (biomes, realms)
- **New predictors:** Bioclimatic, topographic



Population level *hyperparameters*

Biome *varying slopes / intercepts*

Realm *varying slopes / intercepts*

Cross-validation approaches: No size that fits all

Goal of cross-validation

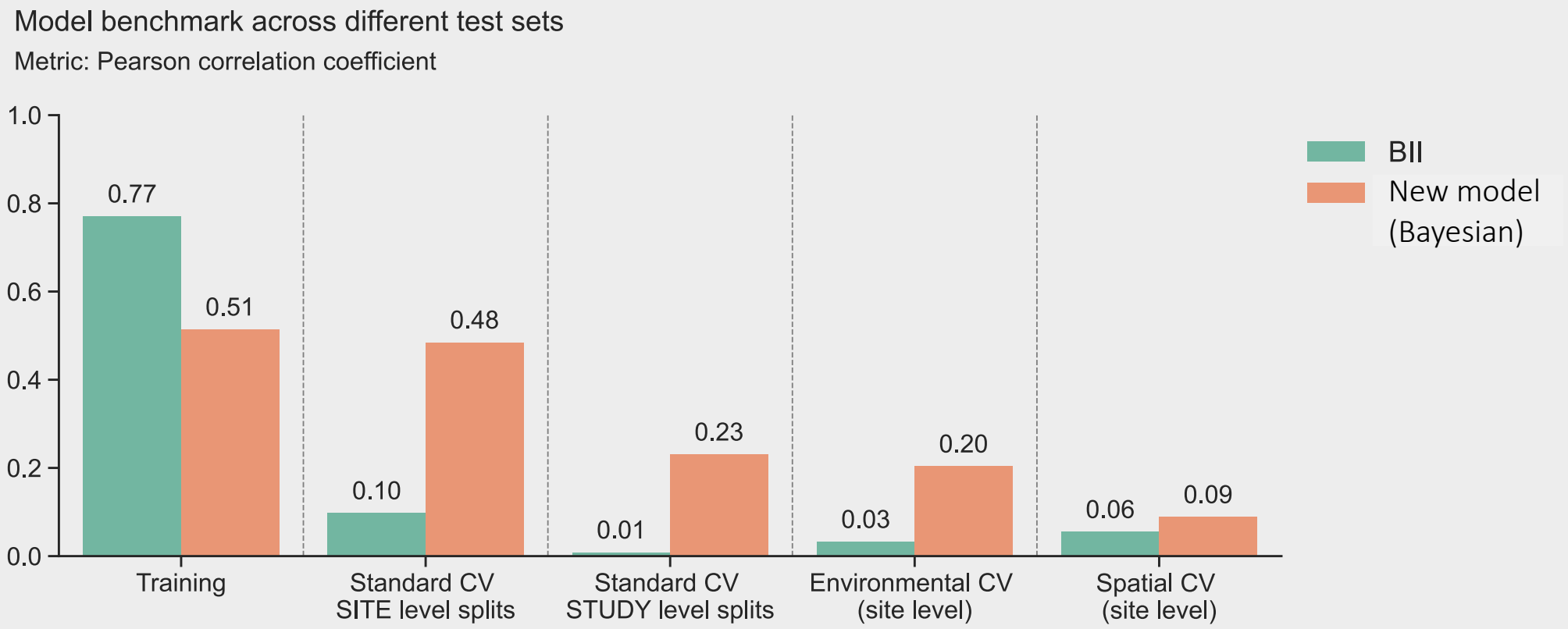
- Simulate model accuracy on new, unseen sites
- Approximate ground-truthing via new data collection

Challenges

- Data from multiple studies
- Studies range from 2-754 sites; vary greatly in scope

- **Standard CV:** Interpolation in well-sampled areas
- **Cross-study CV:** Generalization of model learnings across data sources
- **Environmental CV:** Extrapolation to new biotic and abiotic conditions
- **Spatial CV:** Extrapolation to new areas

Standard CV would suggest decent accuracy, but other approaches pinpoint fundamental limitations



Note: Stratified sampling is used, with a minimum of 10 studies in each biogeographical stratum. Standard CV (site and study splits) uses stratification at the biome-realm level. Spatial and environmental CV adds ecoregion, since clustering can otherwise produce too distant clusters.

Future work and recommendations

Check out our group webpage

Future work

- ***Question:*** *Given data limitations, can we build "good enough" predictive models?*
  - **Testing:** CV methods for heterogeneous, imbalanced, multi-study data
  - **Data:** More biodiversity data (GBIF, …), better (continuous) land cover data
  - **Model:** Combining top-down and bottom-up approaches?

Recommendations

- Support massive, standardized biodiversity data collection using scalable methods, e.g. eDNA
- Develop time series of higher-resolution land cover and habitat condition maps, going back in time