# PREFERENTIAL SAMPLING AND INFERENCE FROM PRESENCE-ONLY DATA

Ameur M. Manceur[1,2*] and Ingolf Kühn[2,3]

[1] Department Computational Landscape Ecology, Helmholtz Centre for Environmental Research – UFZ, Permoserstr. 15, 04318 Leipzig, Germany
* Correspondence author. Phone: +49 341 235 1464. Fax: +49 341 235 1939. E-mail: ameur.manceur@ufz.de
[2] Department Community Ecology, Helmholtz Centre for Environmental Research – UFZ, Theodor-Lieser-Str. 4 06120 Halle, Germany
[3] German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Deutscher Platz 5e, 04103 Leipzig, Germany

## Abstract

Databases on species occurrence in space are often collated from various sources and result in presence-only datasets. The rate of false-negatives can be influenced by the environment: the sampling effort can co-vary with the predictors of species occurrence (preferential sampling). This work aims at inferring the probability of occurrence and the prevalence of a species, based on presence-only data biased by preferential sampling. The Bayesian Image Restoration (BIR) model allows inference based on an environmental mean model, a spatial neighborhood component, and a model of the sampling effort. The latter relies on the number of 'control species' (selected by expert advice) observed. An observed absence is more likely to be a false-negative when the environment is conducive, the neighboring grid cells are occupied, and the number of control species is small (low sampling effort). With an artificial species whose true prevalence is 0.31 and a realistic sampling scenario with an observed prevalence of 0.18, the modeled prevalence was 0.32 (0.30-0.34). The estimation was robust to a failure of the assumptions, even with a misspecified environmental mean model. The restored map was visually close to the true map. The method is fit for the purpose of restoring maps and drawing inference when data is biased by preferential sampling. The link between the sampling effort and the number of control species is crucial, and a sensitivity analysis of the sampling effort model is recommended. BIR applied to *Asarum europaeum* L. resulted in different results compared to a spatial model with no sampling effort, and a significant interaction between rainfall and temperature appeared to be an artifact of observer bias. Given the pervasiveness of imperfect sampling in ecology, the benefits of using a priori knowledge from experts on sampling bias and Bayesian estimation are expected to outweigh the risks of model-based inference.