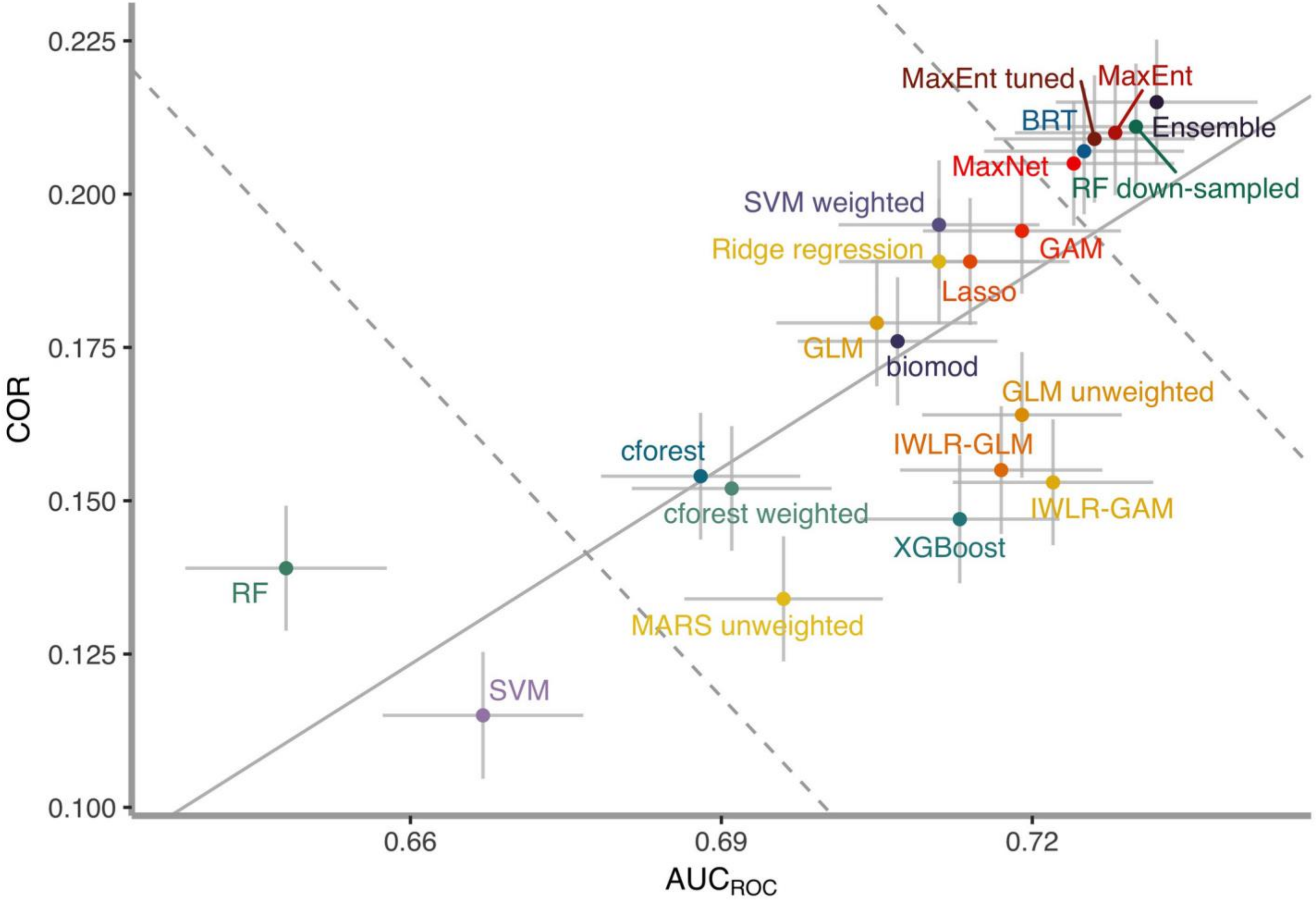
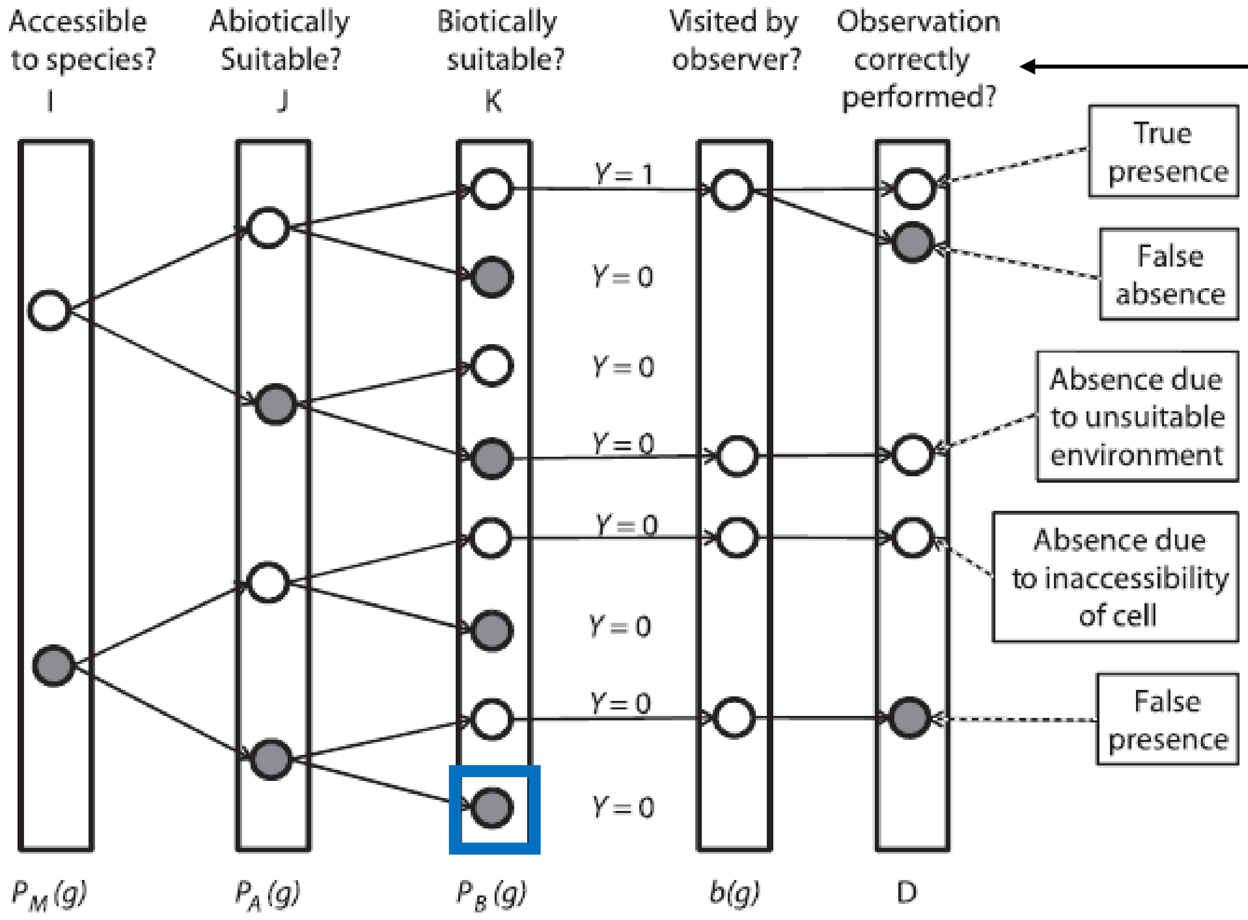
(Pseudo-)Absences for
Species distribution modelling

Why do we need them?



The problem with absence data



Especially the case with CS list data

In most cases absence data is not available
-> generation of Pseudo-Absences is necessary

Pseudo-Absence vs. Background data

- **background** data: characterise the whole landscape
  - sample points to shorten the computational time
  - risk of introducing false absences (Anderson, 2010)
  - overfitting to conditions found near the presences (Anderson, 2010)
  - inflate test statistics (Iturbide et al., 2015)

- **pseudo-absence** data: characterise the landscape where the species does not occur
  - a priori decisions about the suitability of regions needed (may influence the modelling results) (Stokland et al., 2011)

- **study region:** restricted calibration area + prediction into the larger study area leads to more realistic predictions (Anderson, 2010)
  - extrapolation occurs mainly in regions where presence is unlikely
  - delineation based on topography or coarse distributional patterns

- Background/PA data can be sampled with the same geographic bias as the training data (Phillips et al., 2009) -> but: many assumptions

Strategies for sampling of pseudo-absences

- **sampling within areas that are**
  - at a geographical distance from the training points or
  - environmental dissimilar to the training points (Senay et al., 2013)

- **distance in geographical space:**
  - choosing the ideal distance based on the model performance (trial and error)
  - based on expert knowledge
  - based on telemetry data derived movement characteristics

- **distance in environmental/predictor space:**
  - identification of least suitable areas using presence-only methods
  - based on expert knowledge

Strategies for sampling of pseudo-absences

- **Three step method** (Senay et al., 2011):
    1. Specify geographical extent:
        - increasing the distance until the contribution of the most important variable decreases

    2. Classify these areas based on their environmental similarity to presence sites
        - no similarity -> potential pseudo-absence points

    3. Cluster these areas in environmental space
        - number of clusters = number of presences available
        - centroids of the clusters are projected into geographical space

          -> pseudo-absence points

- **R-package mopa** (Iturbide et al., 2015): provides 3-step PA-sampling method
    - Pseudo-Absences provide generally higher AUC than background points; but: poor CV-strategy + weak test metric

Strategies for sampling of PA / BG data

- **package *flexsdm*:**

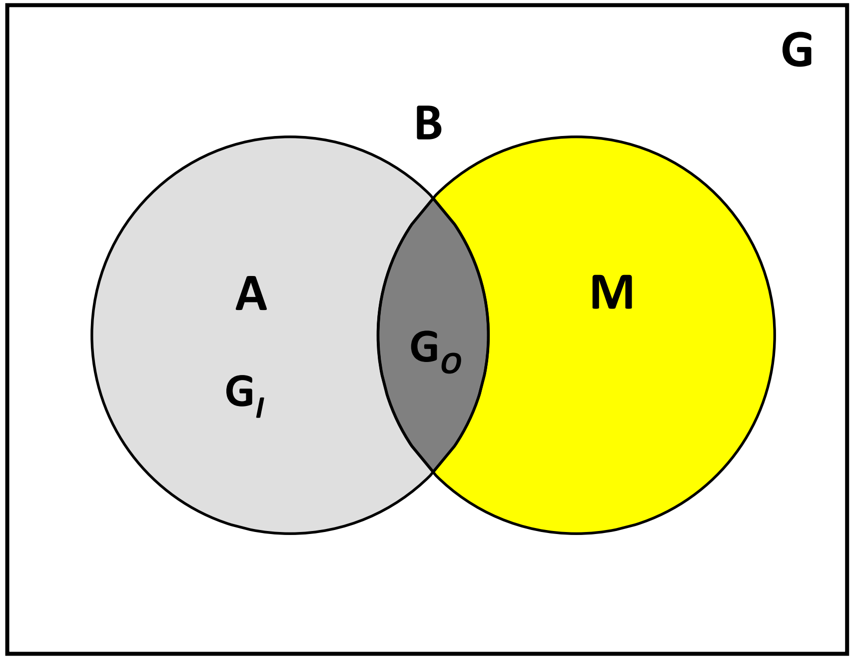
  - background points:
    - random
    - thickening: more sampling towards presence sites
    - bias: using raster layer representing sampling effort

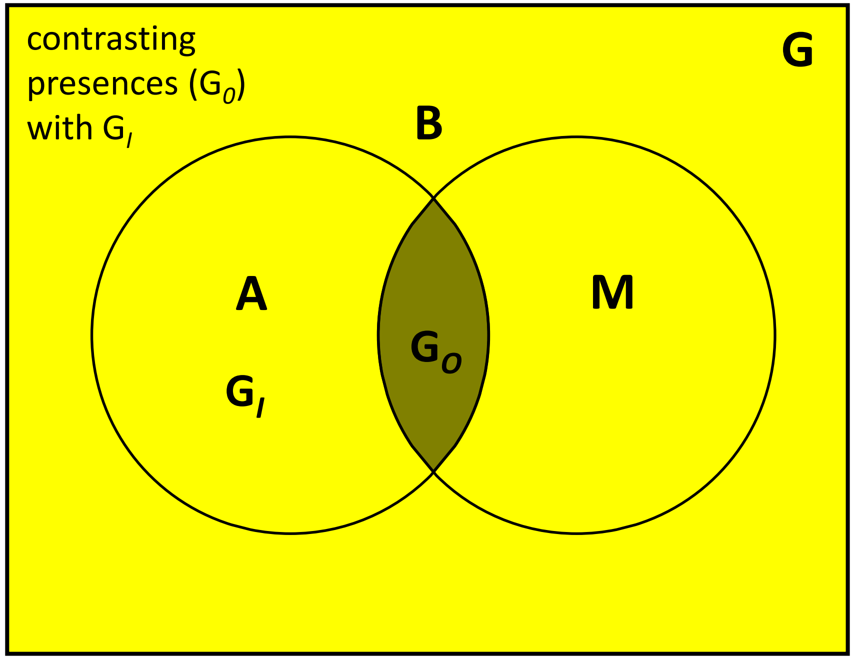
  - pseudo-absence points:
    - random
    - env_const: distance in environmental space
    - geo_const: distance in geographical space
    - geo_env_const: combination of geo & env constrained (2-step)
    - geo_env_km_const: + distributing PAs in env-space using k-means (3-step)

Pseudo-Absence vs. Background data

pseudo-absence data: estimation of the potential distribution (ENM) (Iturbide et al., 2015)

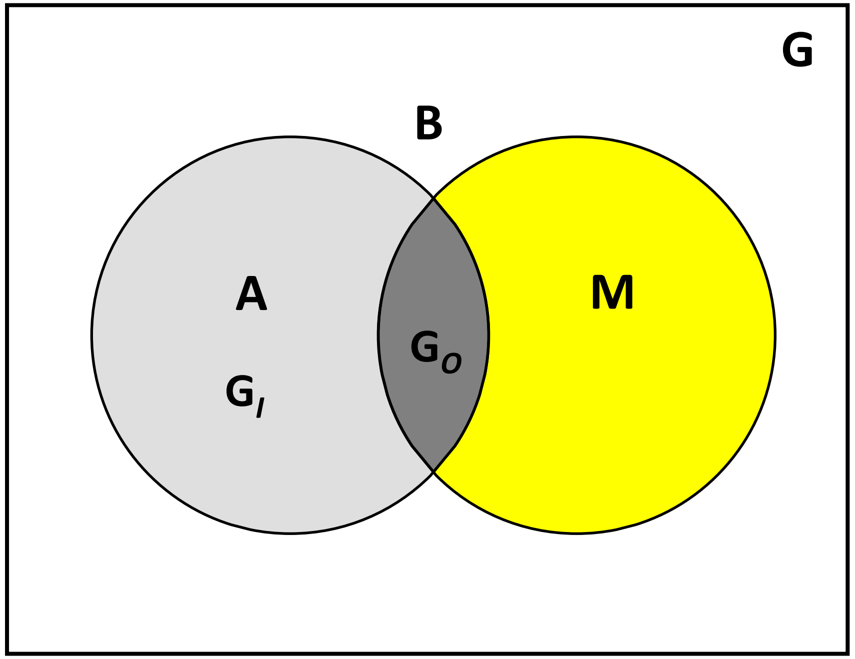
background data: estimation of the realized distribution (SDM) (Iturbide et al., 2015)





Pseudo-Absence vs. Background data

- **Calibration area:** should be restricted to the area that is accessible to the species (M)
  - Distinction between PA & BG-data difficult



Pseudo-Absence vs. Background data

Algorithms	Type	Weighting
Maxent, BRT, ENFA	BG [1]; BRT: see RF [2]	unequal
Random Forest	BG [1] or PAg (when small sample size) [2] or PAe (when sample size is large) [2] class overlap -> better PA?	equal weight by down-sampling of bg-points
GLM, GAM	PAg [2;3] (when climatically biased* presences) or BG [3]	equal weight
ANN, SVM, NB	pseudo-absence	

*can result from spatially biased sampling / when not the whole range of suitable environmental conditions was sampled [2]

[1] Valavi et al. (2021)
[2] Barbet-Massin et al. (2011)
[3] Senay et al. (2011)

BG: Background-Points
PAg: Pseudo-Absence points based on distance in geographical space
PAe: Pseudo-Absence points based on distance in environmental space