Results of simulation 2: above-ground biomass (AGB)

Jan Linnenbrink

today

This file contains the code to reproduce the figures of the appendix of the paper "kNNDM: k-fold Nearest Neighbour Distance Matching Cross-Validation for map accuracy assessment" by Jan Linnenbrink, Carles Milà, Marvin Ludwig and Hanna Meyer. The appendix consists of a second simulation, that is based on the above-ground biomass example presented in de Bruin et al. (2022, https://doi.org/10.1016/j.ecoinf.2022.101665).

Data used in simulation 2

This code reproduces Figure S1: Data used in the AGB simulation: A) shows a subset of the 22 predictors, as well as the response raster. The rasters were linearly stretched to [0,1] for visualization purposes. B) shows one iteration of the different sampling designs. Data from https://doi.org/10.5281/zenodo.6513429.

```
## Reading layer `strata' from data source
##   `/home/janl/git/kNNDM_paper/simulation2_AGB/data/strata.shp'
##   using driver `ESRI Shapefile'
## Simple feature collection with 100 features and 1 field
## Geometry type: MULTIPOLYGON
## Dimension:     XY
## Bounding box:  xmin: 2633250 ymin: 1386000 xmax: 5603250 ymax: 3401000
## CRS:           NA
```
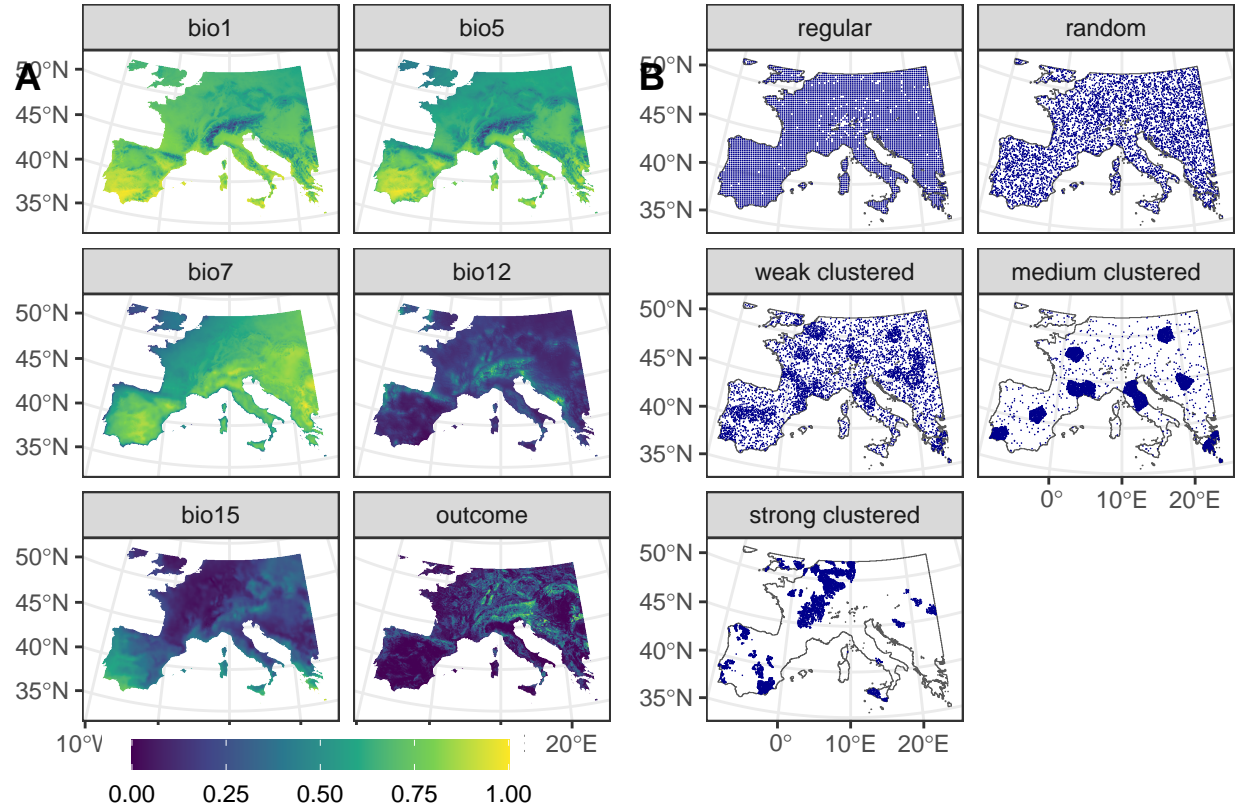
## Comparison of the performance of different CV methods

Differences between Cross-Validated and true RMSE for the first supplementary simulation (Figure S2). kNNDM [tuned] refers to the kNNDM split that yielded the lowest W statistic among different number of folds $k \in [2, 4, 6, ..., 20]$.
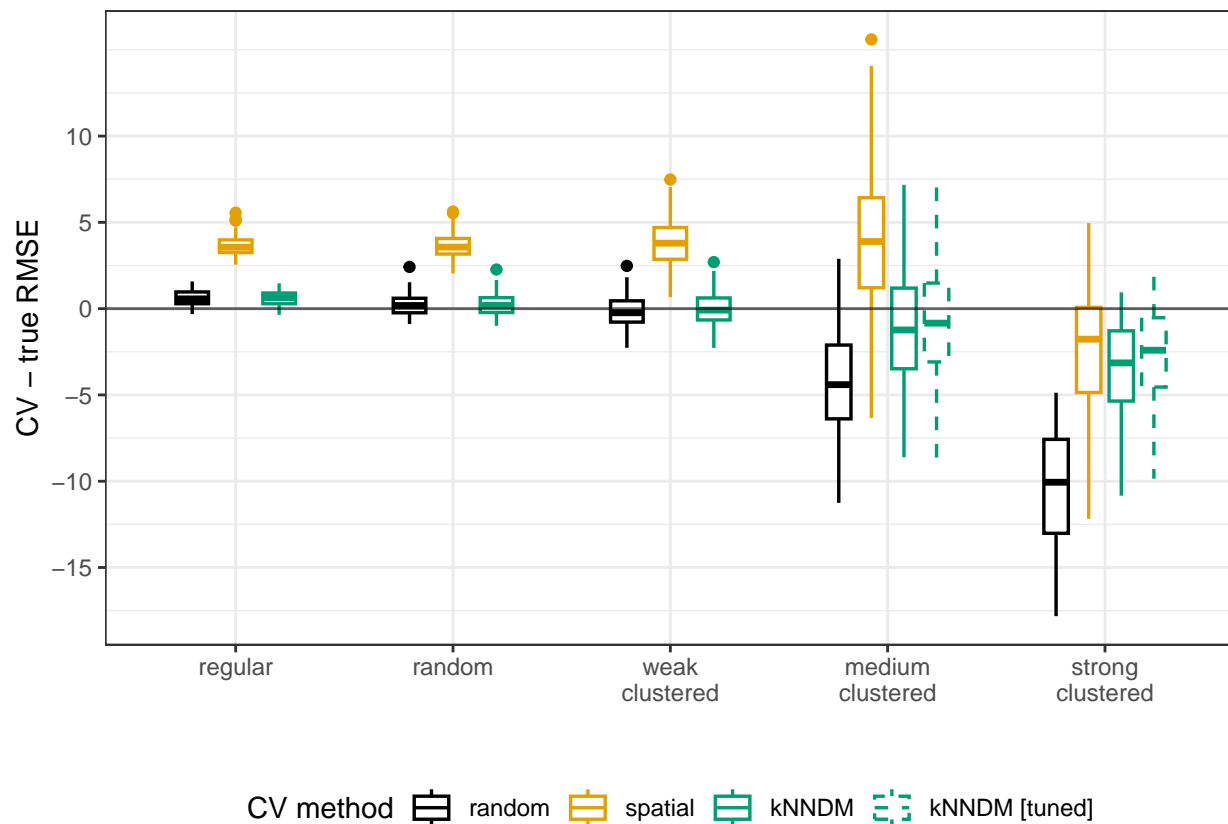
Association CV - True error and W statistic

The following code reproduces figure S3, which shows the relationship between the absolute value difference between the Cross-Validated and true RMSE with the W statistic in the second simulation. Here, the W statistic explained 28% of the variation in the absolute value differences.

```
## [1] "Rsquared for RMSE: 0.28"
```
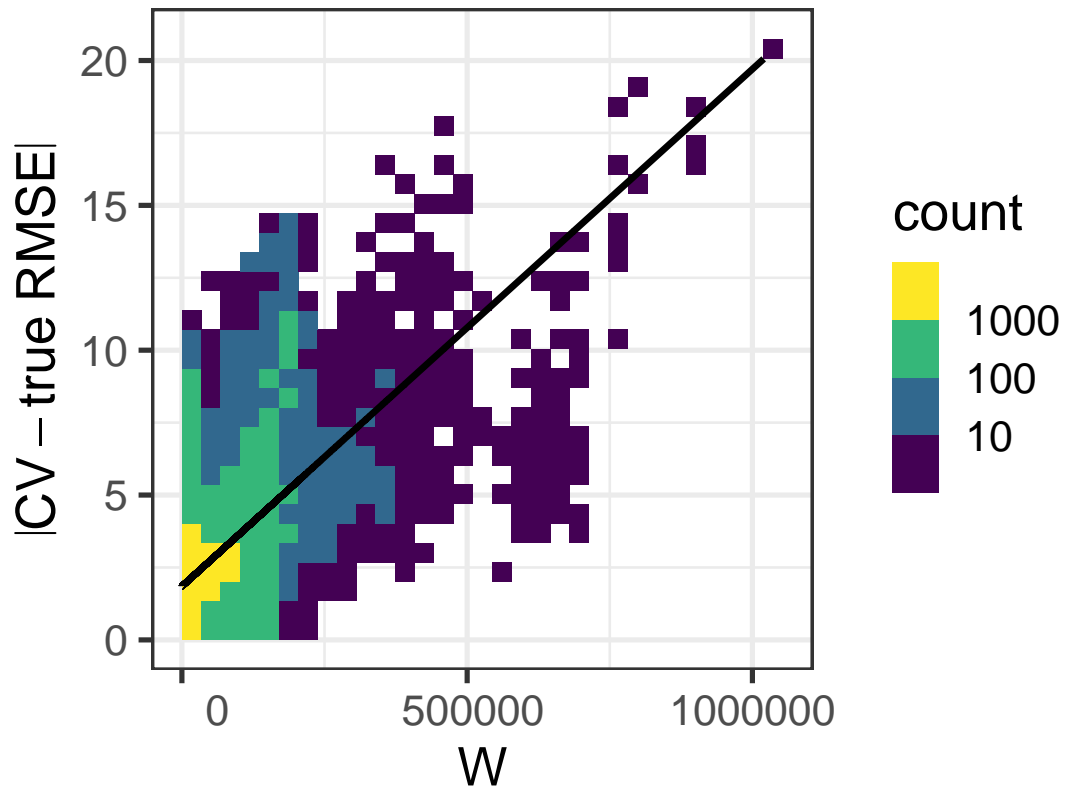
Different numbers of k

Following, figure S4 is reproduced, which shows the influence of different numbers of k on the difference between the Cross-Validated and true RMSE (upper row), and on the W stat (lower row). Note that the values of the W statistic were log-scaled.