Results of simulation 2: above-ground biomass and soil organic carbon stock

Jan Linnenbrink

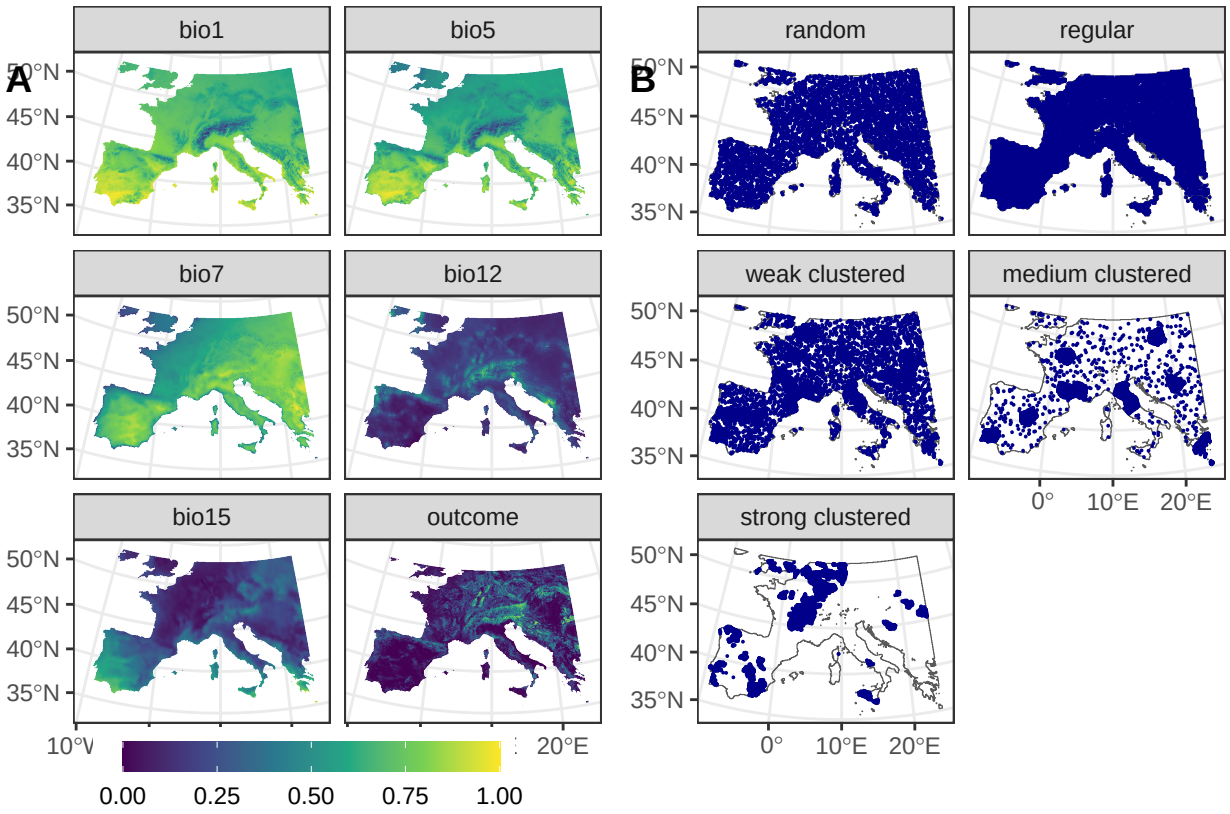
today

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).

Workflow

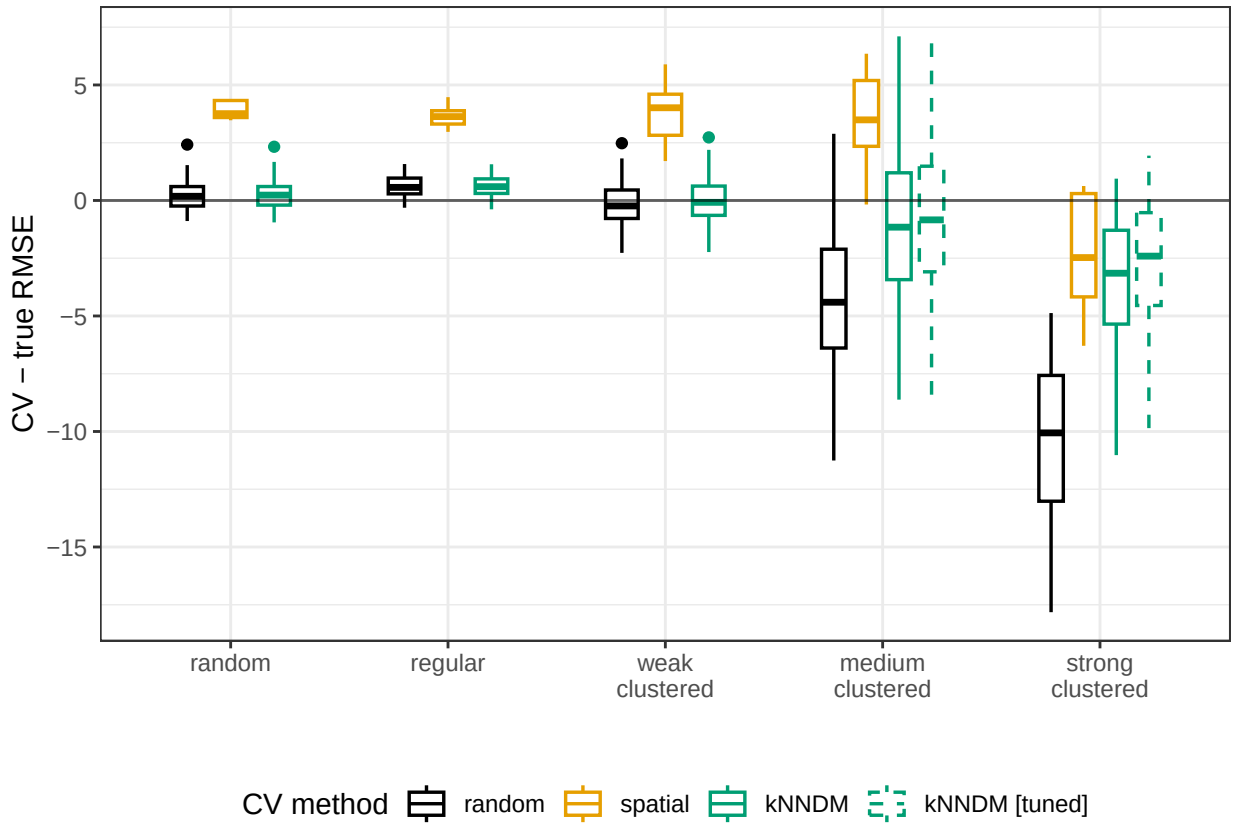
This code reproduces the workflow of the second simulation shown in Figure S1. A) shows 5 of the 22 predictors, as well as the outcome (the above-ground biomass map). B) shows one example of the 5 different realizations of the 5,000 sample points. The raster datasets and the AOI shapefile were not uploaded on Github due to their large size, but can be downloaded from https://zenodo.org/record/6513429.

```
## Reading layer `strata' from data source
##   `/home/janl/CrossValidation/data_sim_AGB/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
```



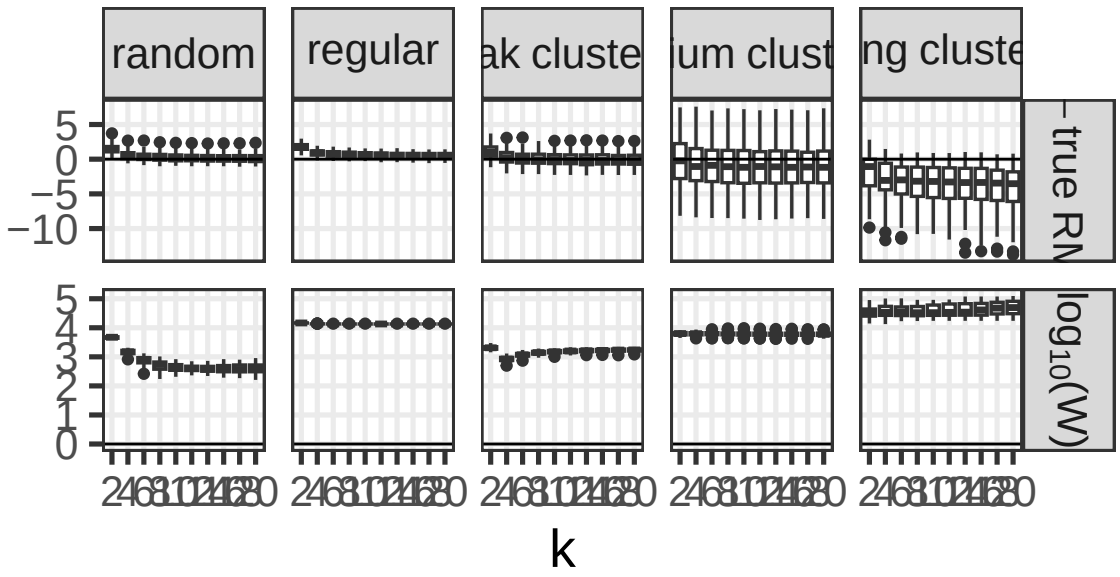
Comparison of the performance of different CV methods

Differences between Cross-Validated and true RMSE for the second simulation (Figure S2). kNNDM [tuned] refers to the kNNDM split that yielded the lowest W statistic among different numbers of k [2,20].



Different numbers of k

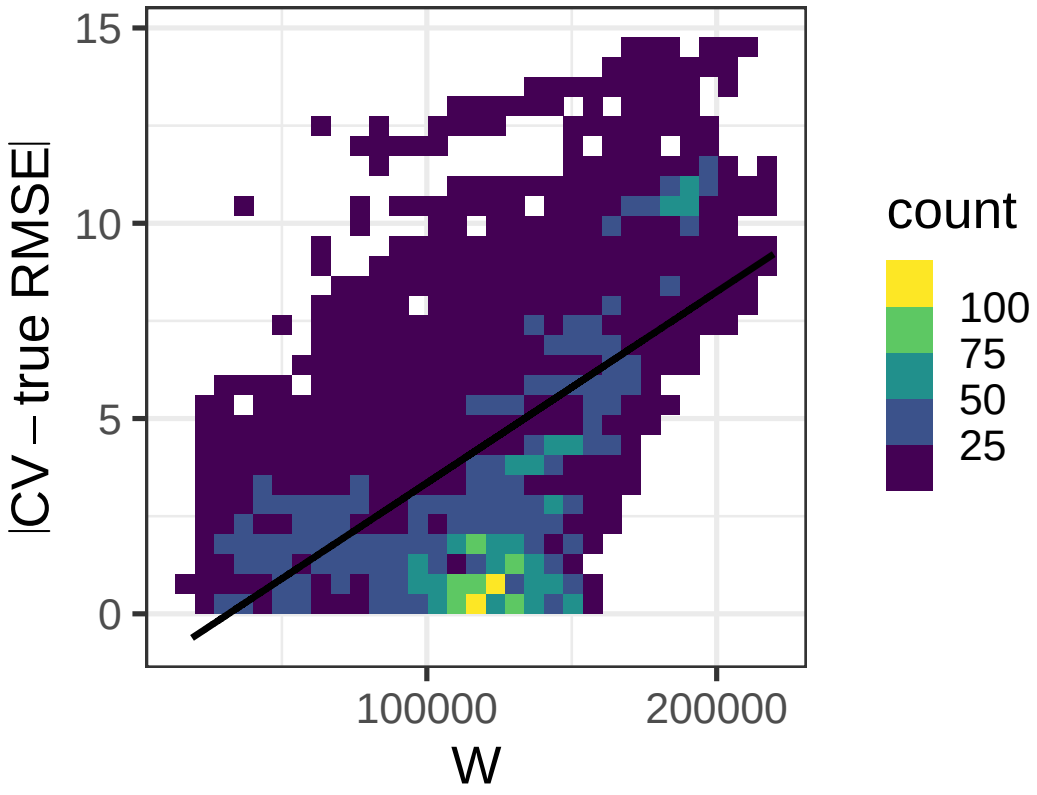
Following, figure S3 is reproduced, which shows the influence of different numbers of k on the difference between the Cross-Validated and true RMSE (upper row), and on the W stat (lower row). Note that the values of the W statistic were log-scaled.



Association CV - True error and W statistic

The following code reproduces figure S4, which shows the relationship between the absolute value difference between the Cross-Validated and true RMSE with the W statistic in the strongly clustered design of the second simulation. Here, the W statistic explained 37% of the variation in the absolute value differences.

```
## [1] "Rsquared for RMSE: 0.37"
```


15
10
5
0
|CV – true RMSE|
100000
200000
W
count
100
75
50
25