# Methods and results

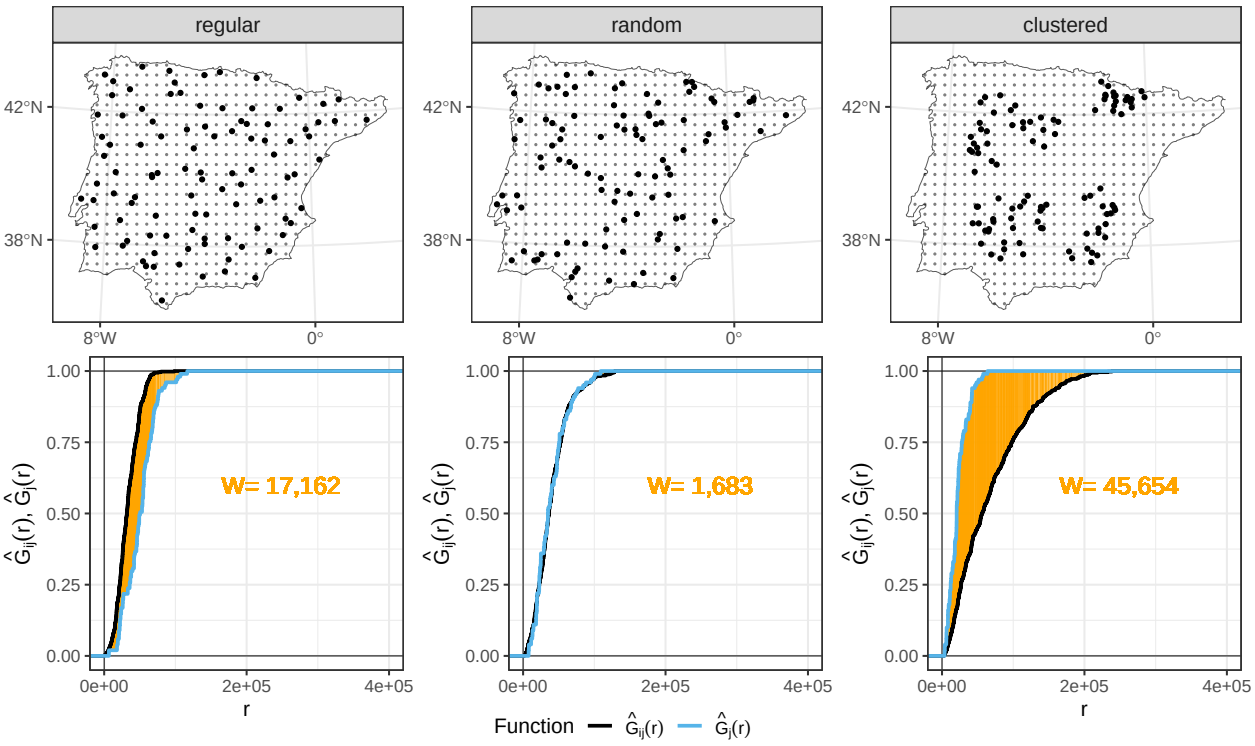
Jan Linnenbrink & Carles Milà

2024-01-25

This file contains the code to reproduce figures 1, 3, 5, 6 & 7 of the paper "kNNDM: k-fold Nearest Neighbour Distance Matching Cross-Validation for map accuracy estimation" by J Linnenbrink, C Milà, M Ludwig & H Meyer.

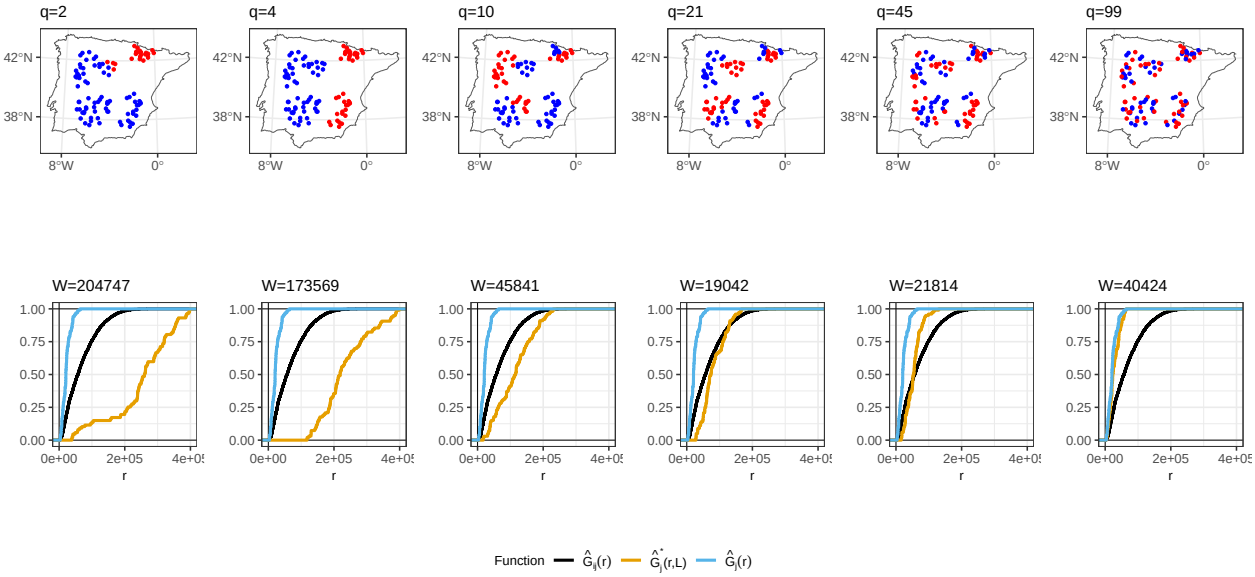
## Example: different clustering, NND ECDFs and W statistics

This code reproduces figure 1, where different configurations of training samples and their corresponding nearest neighbour distance (NND) empirical cumulative distribution functions (ECDF) are shown. Also, the Wasserstein statistic is shown in orange.

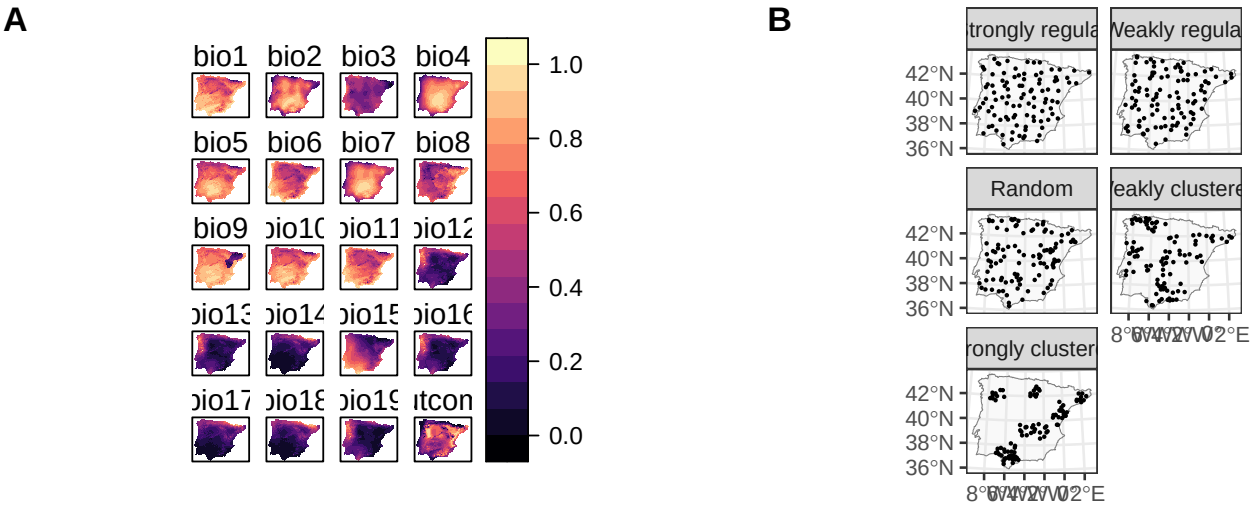



## kNNDM workflow

This code reproduces the kNNDM workflow shown in figure 3. Several numbers of clusters $q$ are compared regarding their W statistic (bottom row).

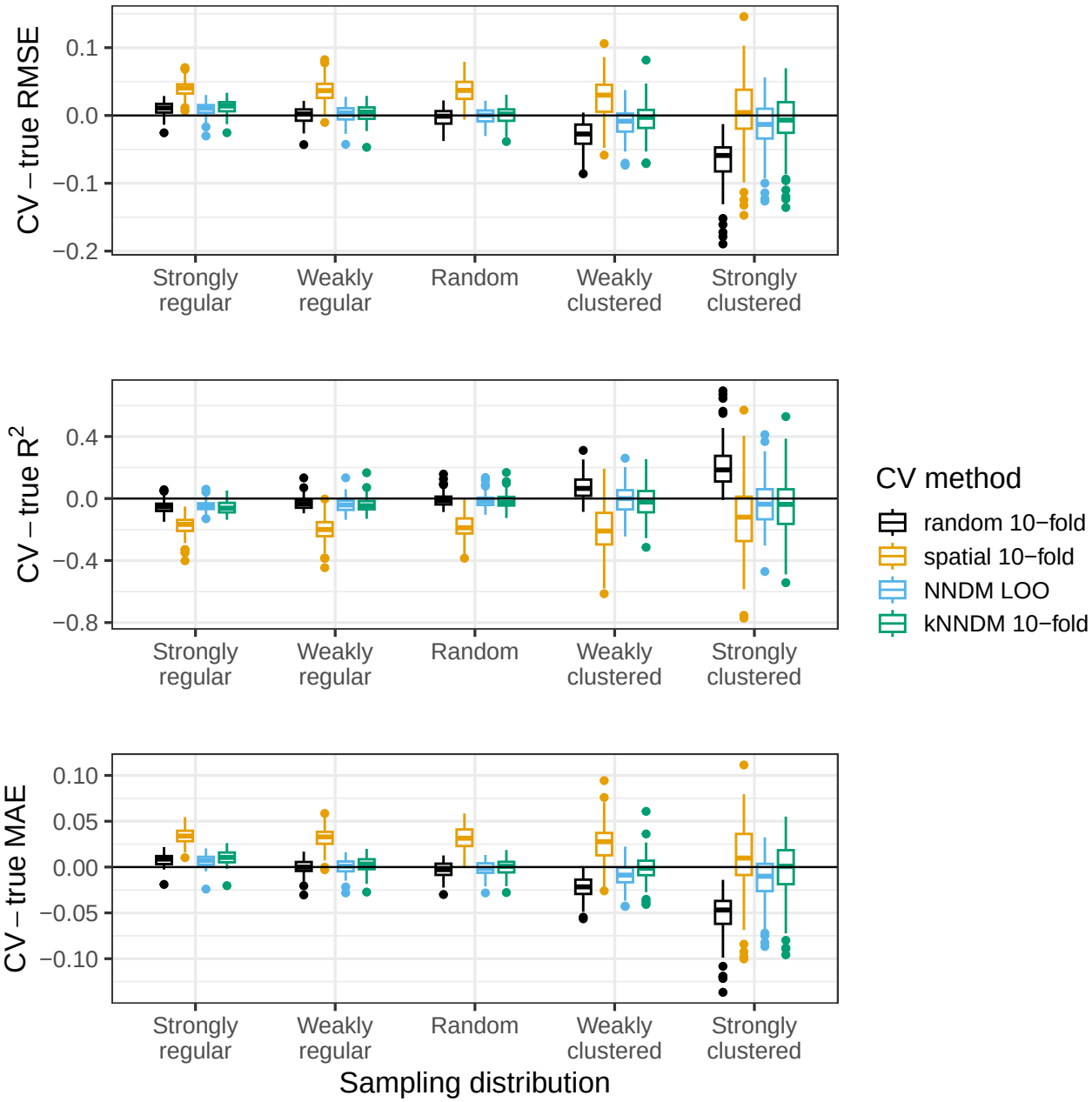



## Simulation data




## Simulation results

The following code reproduces figure 5, and shows the differences between cross-validated and true RMSE, R^2 and MAE for different sampling distributions.

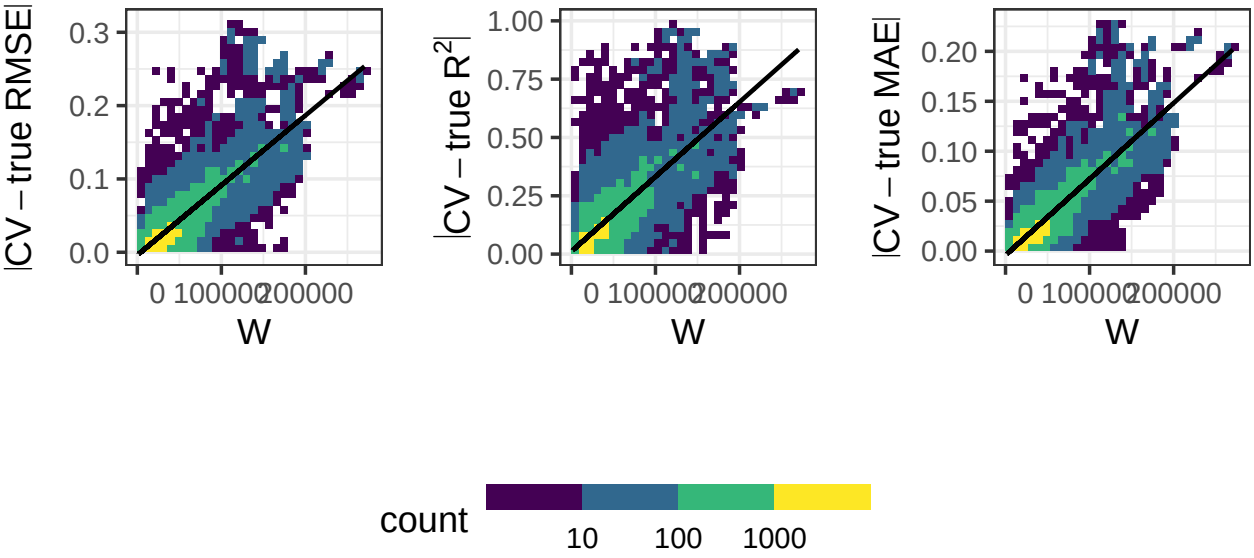



```
## # A tibble: 4 x 4
## # Groups:   name [4]
##   name       dsample     mean     sd
##   <chr>      <chr>       <dbl>  <dbl>
## 1 mae_knndm  sclust   -0.00310 0.0307
## 2 mae_nndm   sclust   -0.0126  0.0257
## 3 rmse_knndm sclust   -0.00837 0.0425
## 4 rmse_nndm  sclust   -0.0146  0.0377
```

### Association CV - True error and W statistic

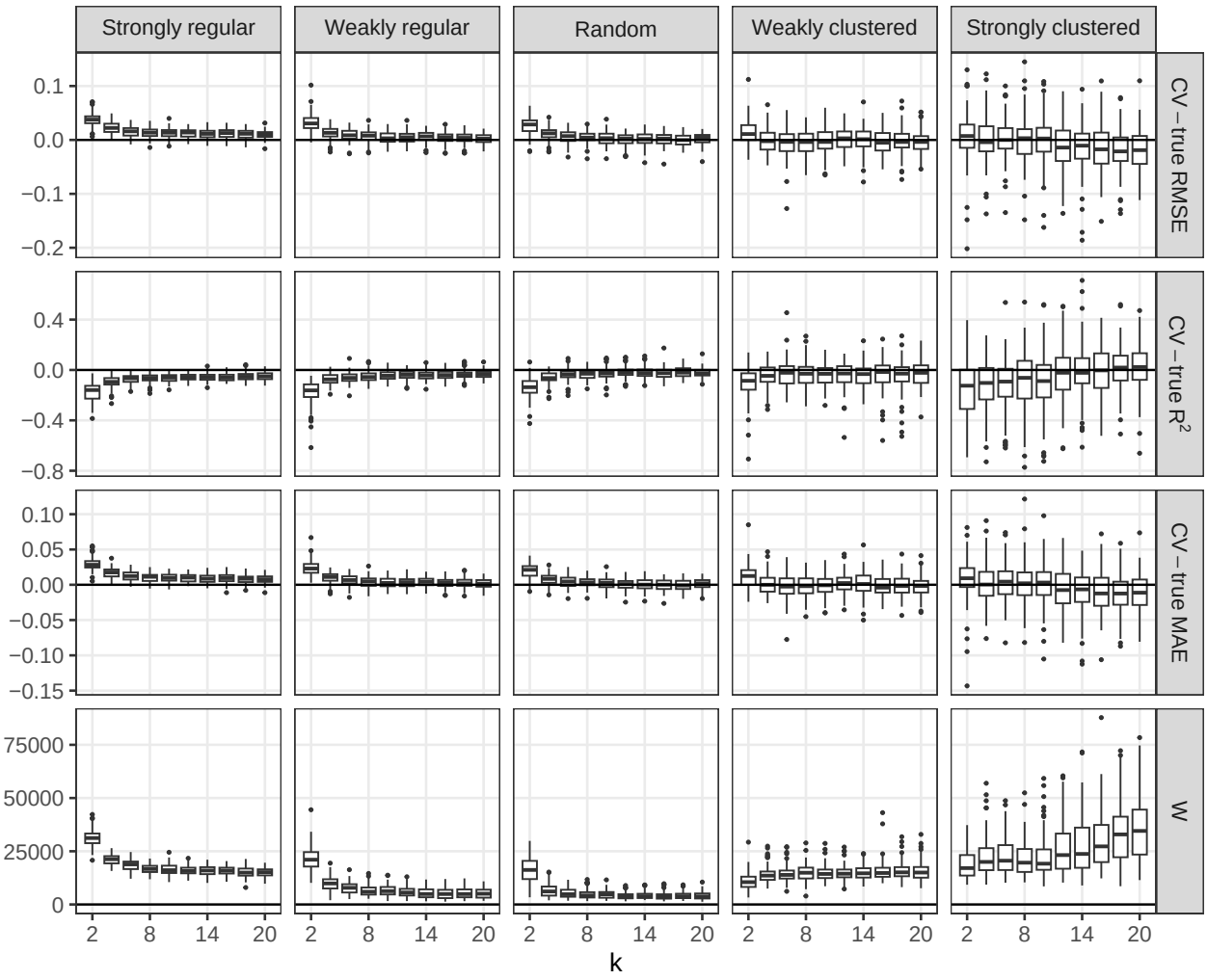
The following code reproduces figure 6, which shows the association between the absolute value difference between the CV and true map accuracy statistics and W statistic.

```
## [1] "Rsquared for Rsquared: 0.6 ; Rsquared for MAE: 0.73 ; Rsquared for RMSE: 0.66"
```



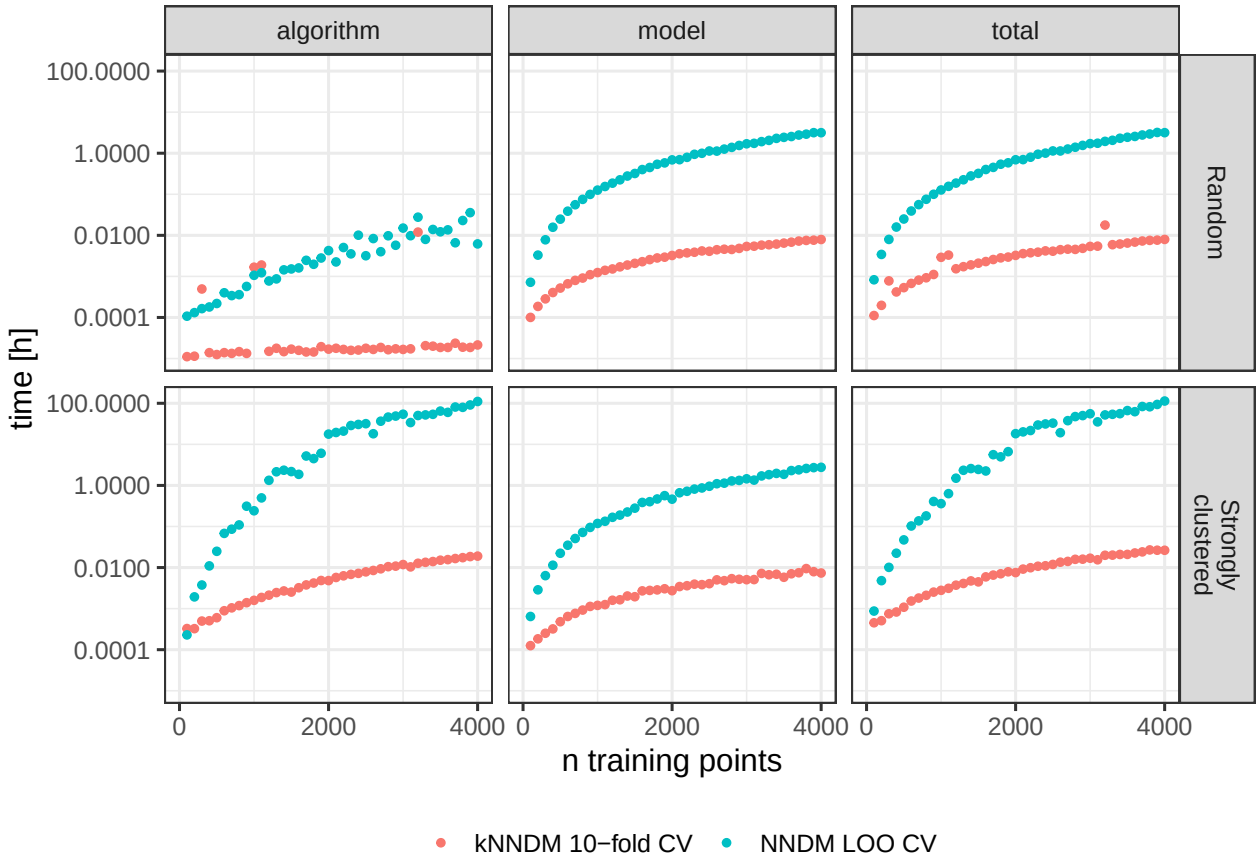

### Different numbers of k

CV error estimates for kNNDM CV with different numbers of $k$ (first three rows). The respective W stat is shown in the fourth row.




## Computational time

The following code reproduces figure 8, which compares NNDM LOOCV and kNNDM CV regarding their computational time requirements.


algorithm
model
total
Random
Strongly clustered
time [h]
100.0000
1.0000
0.0100
0.0001
0
2000
4000
n training points
kNNDM 10−fold CV
NNDM LOO CV