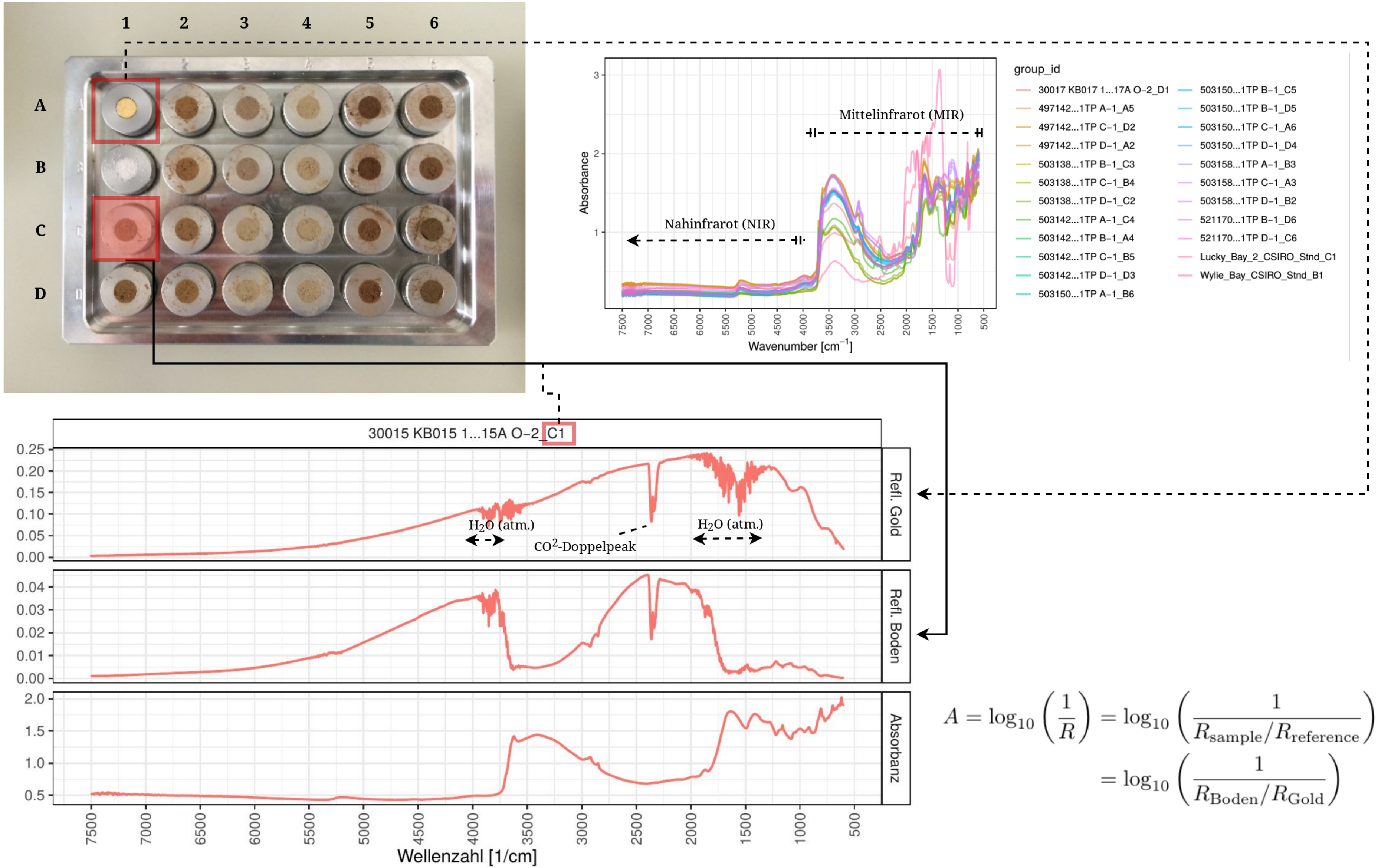
Development of a Swiss National Soil Spectral Model Library using data-driven modeling

Improving the accuracy and cost-efficiency of soil monitoring in Switzerland with mid-IR spectroscopy and rs-local

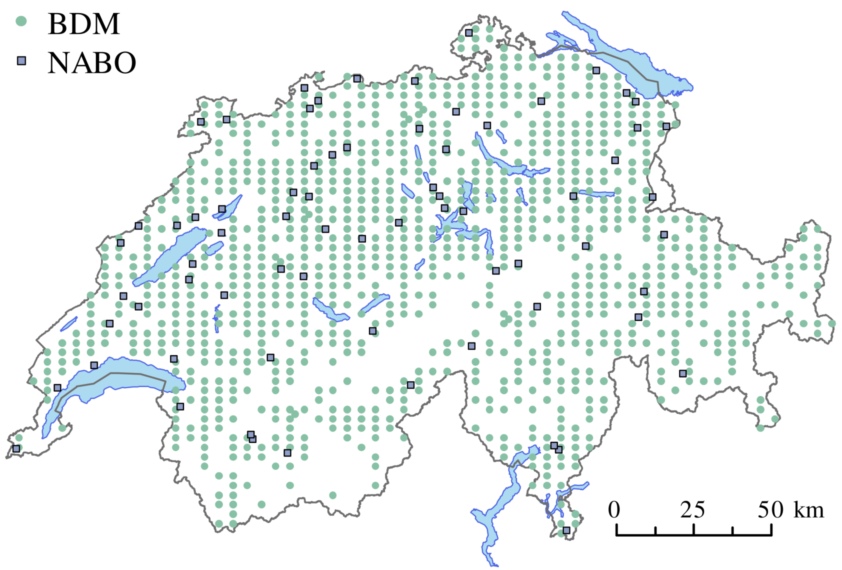
Philipp Baumann[1], Anatol Helfenstein[2], Andreas Gubler[3], Reto Meuli[3], Armin Keller[4], Juhwan Lee[5], Raphael A. Viscarra Rossel[5], and Johan Six[1]
philipp.baumann@usys.ethz.ch | github.com/philipp-baumann
Update: 2020-05-06)

Mid-infrared spectroscopy: cost-effective



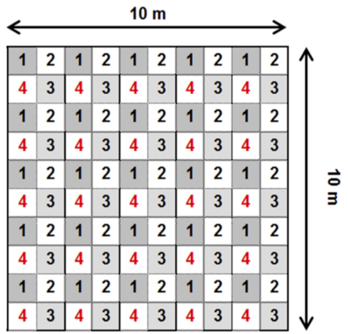
$$A = \log_{10}\left(\frac{1}{R}\right) = \log_{10}\left(\frac{1}{R_{\text{sample}}/R_{\text{reference}}}\right)$$
$$= \log_{10}\left(\frac{1}{R_{\text{Boden}}/R_{\text{Gold}}}\right)$$

Scale up soil monitoring with soil spectral libraries?



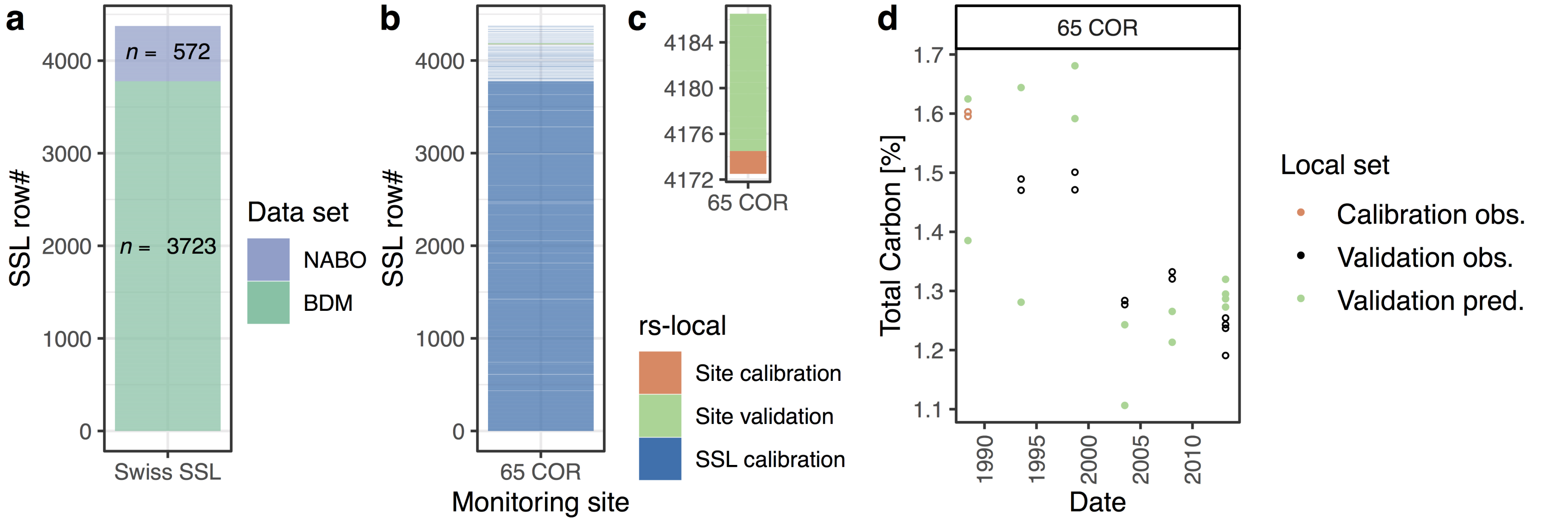
BDM: Biodiversity Monitoring Switzerland (Federall Office for the Environment; soil analysis: NABO, Agroscope)
– 1094 locations (one sampling event)

NABO: Swiss Soil Monitoring Network (NABO)
– 71 sites (subset of agricultural locations)



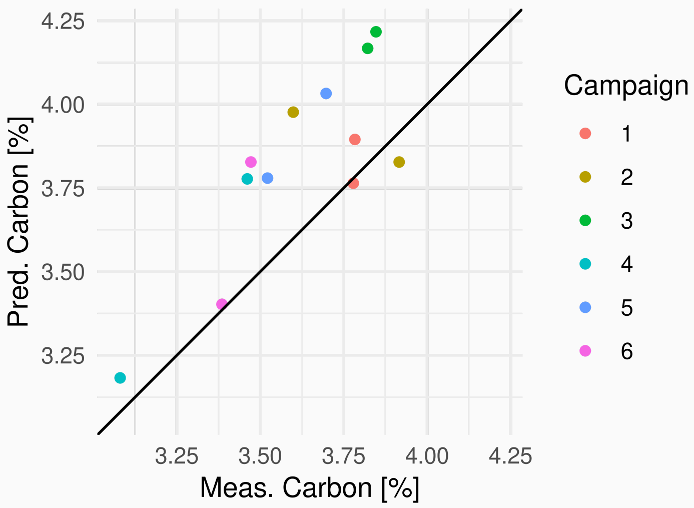
Sampling scheme for a NABO site

- Sampling in 5 year intervals since 1985
- Area: 10m x 10m
- One to four spectral measurements per year (pooling replicate samples)



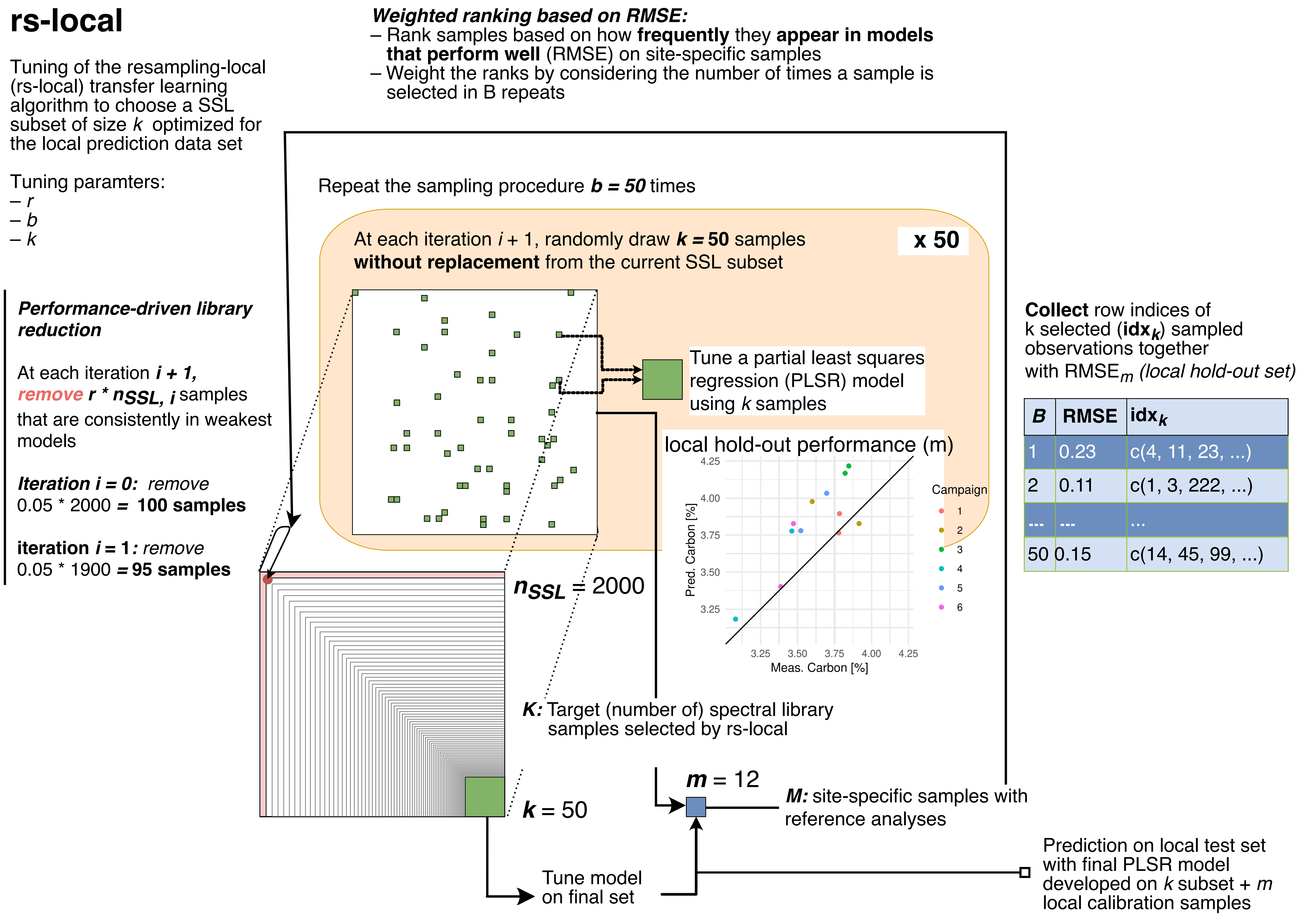
Instance-based transfer learning vs. general ML

- Cubist hold-out predictions for one site
  - For example: location "70 DIS"
  - Field campaigns every 5 years since 1985
  - Hold-out samples are grouped by location
  - *Overall bias (entire SSL) close to zero, but there is "site" bias*



- *Transfer learning*:
  - Transfer from knowledge in an source problem or domain to a target domain
- "Resampling(rs)-local" as a form of *Instance-based transfer learning*
  - Lobsey, C. R., Viscarra Rossel, R. A., Roudier, P., & Hedley, C. B. (2017). rs-local data-mines information from spectral libraries to improve local calibrations: rs-local improves local spectroscopic calibrations. European Journal of Soil Science. https://doi.org/10.1111/ejss.12490
  - "Brute-force peeling"

rs-local in a nutshell


rs-local
Tuning of the resampling-local (rs-local) transfer learning algorithm to choose a SSL subset of size k optimized for the local prediction data set
Tuning paramters:
– r
– b
– k
Performance-driven library reduction
At each iteration i + 1, remove r * nSSL, i samples that are consistently in weakest models
Iteration i = 0: remove 0.05 * 2000 = 100 samples
iteration i = 1: remove 0.05 * 1900 = 95 samples
Weighted ranking based on RMSE:
– Rank samples based on how frequently they appear in models that perform well (RMSE) on site-specific samples
– Weight the ranks by considering the number of times a sample is selected in B repeats
Repeat the sampling procedure b = 50 times
At each iteration i + 1, randomly draw k = 50 samples without replacement from the current SSL subset
x 50
Tune a partial least squares regression (PLSR) model using k samples
local hold-out performance (m)
Pred. Carbon [%]
Meas. Carbon [%]
3.25
3.50
3.75
4.00
4.25
Campaign
1
2
3
4
5
6
nSSL = 2000
K: Target (number of) spectral library samples selected by rs-local
k = 50
m = 12
M: site-specific samples with reference analyses
Tune model on final set
Prediction on local test set with final PLSR model developed on k subset + m local calibration samples
Collect row indices of k selected (idxk) sampled observations together with RMSEm (local hold-out set)
B | RMSE | idxk
1 | 0.23 | c(4, 11, 23, ...)
2 | 0.11 | c(1, 3, 222, ...)
--- | --- | ...
50 | 0.15 | c(14, 45, 99, ...)

RS-local tuning

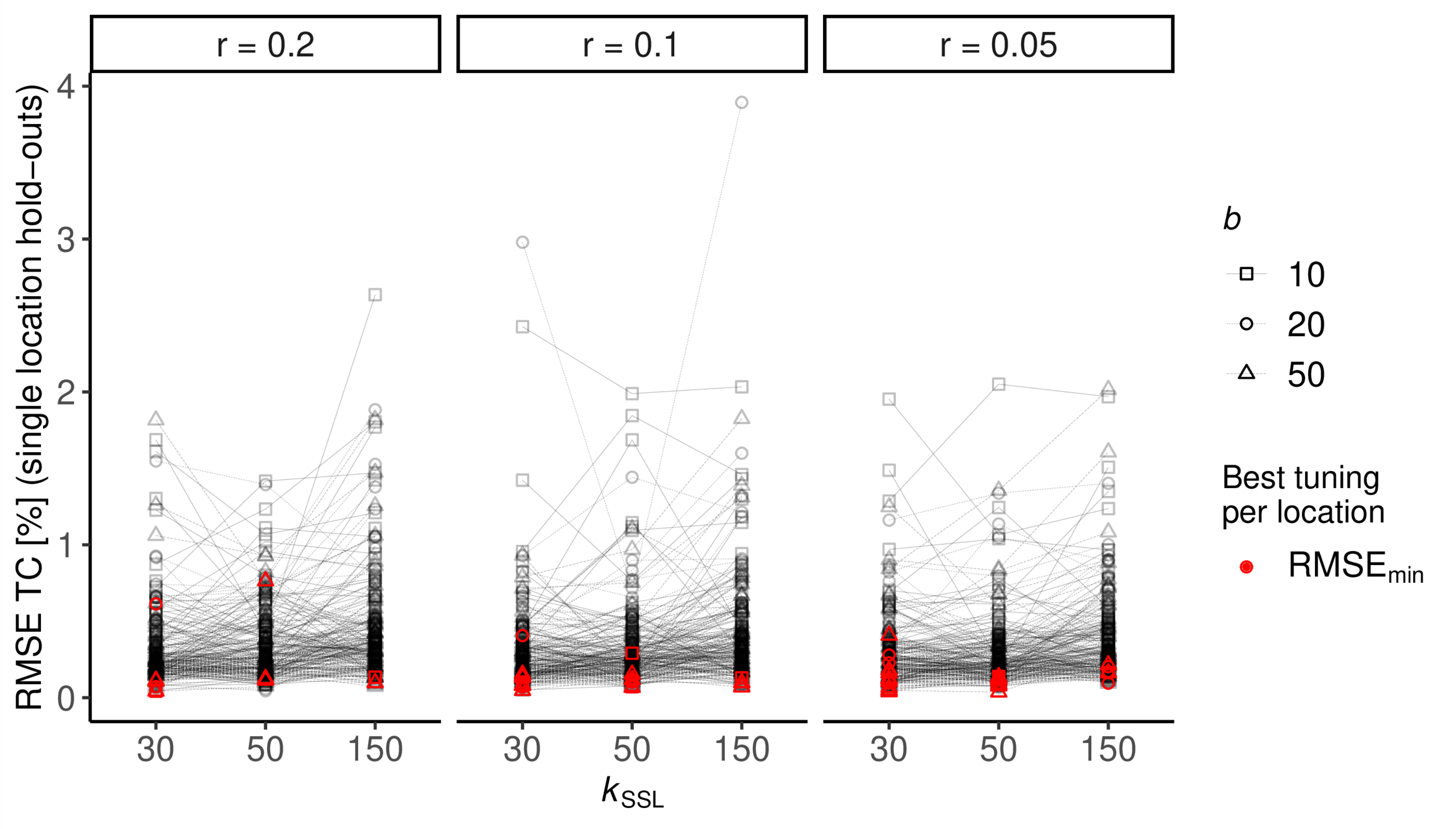
- Let's wrap and tune for each NABO site hold-out:

```
(grid_rslocal <- dials::grid_regular(
  rslocal_k %>% value_set(c(30L, 50L, 150L)),
  rslocal_b %>% value_set(c(10L, 20L, 50L)),
  rslocal_r %>% value_set(c(0.2, 0.1, 0.05))
))
```

```
## # A tibble: 27 x 3
##    rslocal_k rslocal_b rslocal_r
##        <int>     <int>     <dbl>
##  1        30        10       0.2
##  2        50        10       0.2
##  3       150        10       0.2
##  4        30        20       0.2
##  5        50        20       0.2
##  6       150        20       0.2
##  7        30        50       0.2
##  8        50        50       0.2
##  9       150        50       0.2
## 10        30        10       0.1
## # … with 17 more rows
```

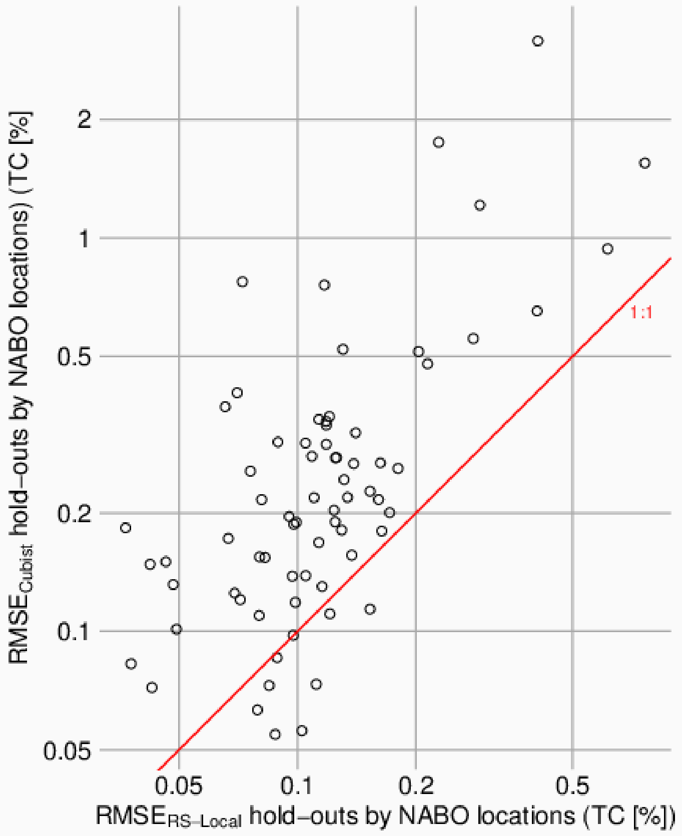
RS-Local: tuning results

- RMSE of Cubist vs. RS-local (modeling per site) across 71 NABO sites



Results RS-Local vs. Cubist (rule-based ML)

- 71 NABO sites (~ 5 days on 48 CPU cores (see previous slide ;-))
- RMSE of Cubist vs. RMSE RS-local (tuned separately for each site)
- RS-local: local m samples from locations only used for selection of k SSL samples, not for prediction; to avoid data leakage into test



Next step:

- Alternative selection strategy for rs-local transfer:
  - Use two local calibration/tuning samples (pooling sample replicate measurement) per NABO monitoring site
  - Further minimize data leakage (selection bias) (see slide 3)