Spatial Upscaling

Amanda Cole

2023-12-17

2.1 Literature

Random cross-validation is a type of cross-validation (a technqiue where a dataset is split into equally sized subsets or folds) where data is randomly divided into folds[1]. In random cross-validation, data is assumed to be independent and identically distributed and no structure or order of the datset is taken into account. Random division of data into folds causes an issue when structural dependencies, such as spatial autocorrelation (when values of variables that are geographically close to one another are similar) are present in the dataˆ2. In cases where data are geographically clustered and geographically close variables are correlated, spatial cross-validation should be used. In spatial cross-validation, data is divided into folds based on geography. Because spatial cross-validation takes into account the spatial relationships of the data, this can improve model accuracy and applicability[2].

An alternative to measuring distance as a geographical distance in a Euclidean space that considers the task of spatial upscaling on environmental covariates more directly could be to measure distance by classes within which the data are generated or temporal proximity. In this dataset for example, leaf nitrogen content will differ based on plant species meaning that although two areas of forest may be farther apart in terms of geographical distance, they will present leaf nitrogen content values that are more similar than than an area of forest and a cropland that are close together in terms of geographical distance. In this dataset, covariates of mean temperature and mean daily irradiance will differ temporally.

2.2 Random Cross-Validation

I prepared the data as outlined in the exercise[4] and saved the prepared data to the data folder in the repository. I then split the dataset into training and test sets (70/30% split respectively). I used the {caret} package to perform a 5-fold cross-validation with the target variable of LeafN and predictor variables of elevation, mean annual temperature, mean annual precipitation, atmospheric nitrogen deposition, mean annual daily irradiance, and species. Hyperparameters were set as mtry=3 and min.node.size=12.

```
# Load data
dfs <- readRDS(here::here("data-raw/dfs.rds"))

# Split dataset into training and testing sets
set.seed(456)  # for reproducibility
split <- rsample::initial_split(dfs, prop = 0.7)
df_train <- rsample::training(split)
df_test <- rsample::testing(split)

# Filter out any NA to avoid error when running a Random Forest
df_train <- df_train |> tidyr::drop_na()
df_test  <- df_test  |> tidyr::drop_na()
```

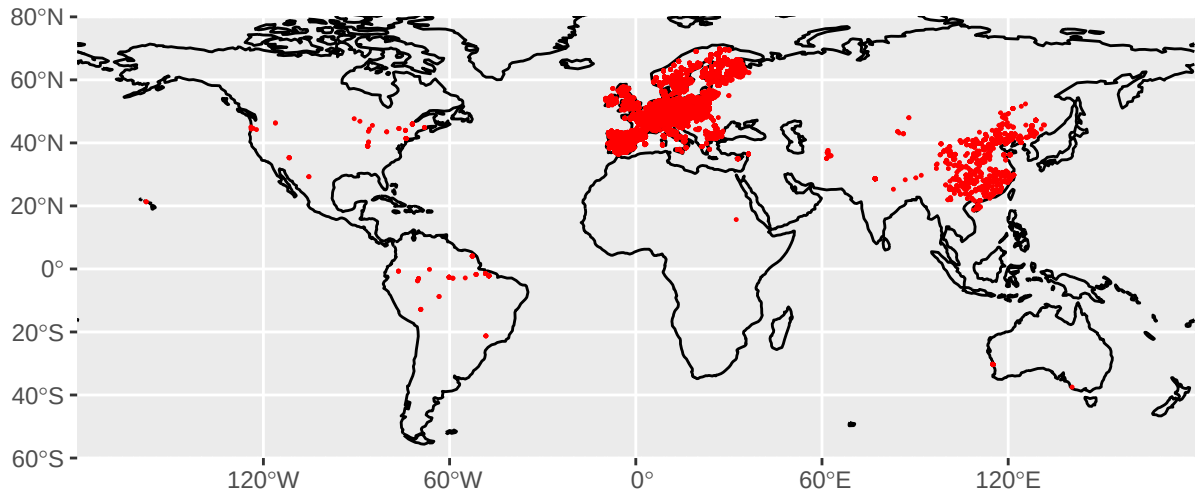
```r
pp <- recipes::recipe(leafN ~ elv+map+mat+ndep+mai+Species, data = df_train) |>
  recipes::step_center(recipes::all_numeric(), -recipes::all_outcomes()) |>
  recipes::step_scale(recipes::all_numeric(), -recipes::all_outcomes())

modcv <- caret::train(
  pp,
  data = df_train |>
    drop_na(),
  method = "ranger",
  trControl = caret::trainControl(method = "cv", number = 5, savePredictions = "final"),
  tuneGrid = expand.grid( .mtry = 3,
                          .min.node.size = 12,
                          .splitrule = "variance"),
  metric = "RMSE",
  replace = FALSE,
  sample.fraction = 0.5,
  num.trees = 100,
  seed = 456                # for reproducibility
)
```

The RMSE and R^2 across validation folds are as follows:

```
##       RMSE  Rsquared      MAE Resample
## 1 2.457810 0.7811883 1.601820    Fold1
## 2 2.542865 0.7660813 1.634175    Fold2
## 3 2.458005 0.7701473 1.631721    Fold3
## 4 2.322459 0.7907585 1.574781    Fold4
## 5 2.339475 0.7794291 1.532173    Fold5
```

2.3 Spatial Cross Validation

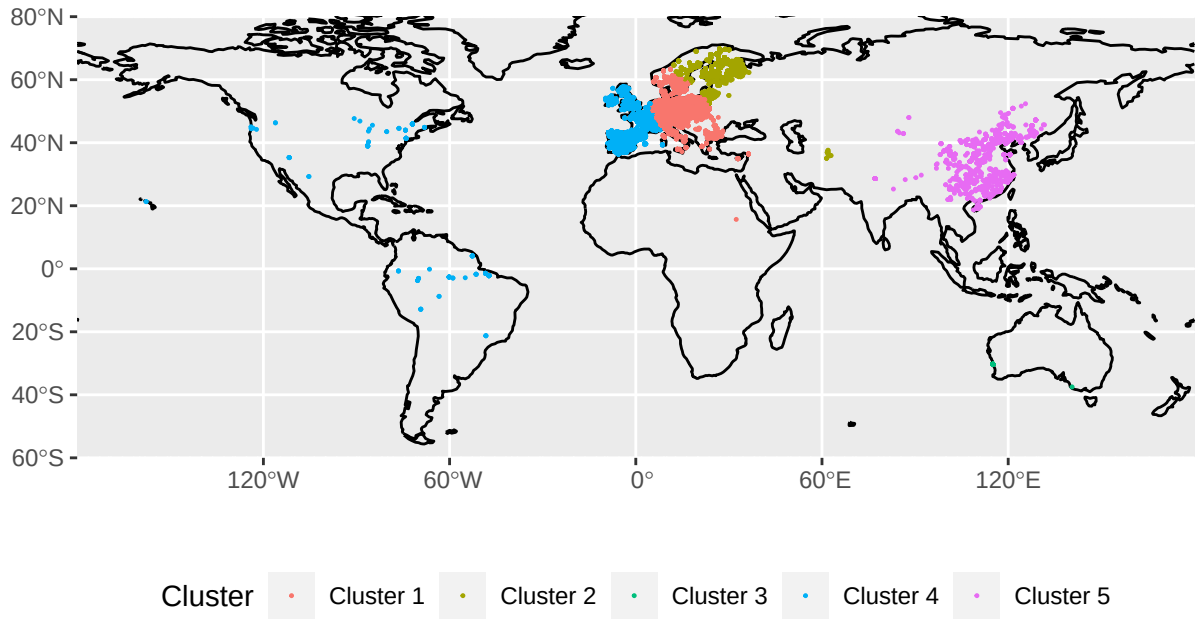


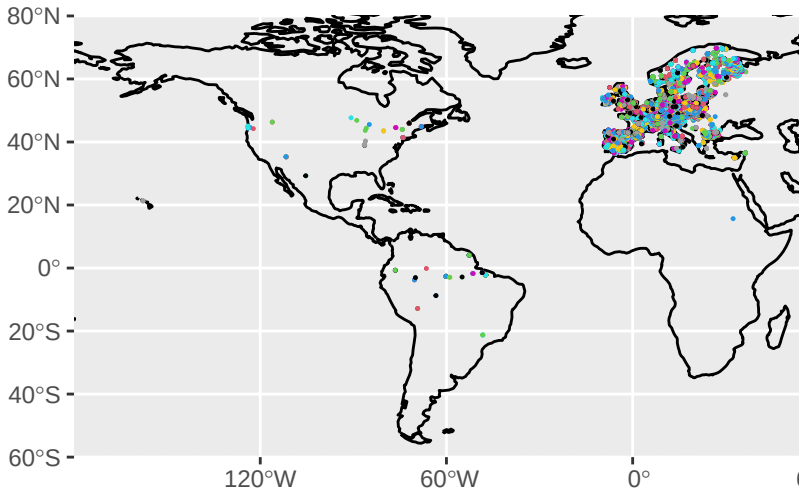
As shown in the above, data are heavily geographically clustered. Using random cross-validation on a dataset that is heavily geographically clustered may produce a model that has poor generalisability, i.e. is not able to make predictions for new locations. Using spatial cross-validation will allow us to produce a model that has an improved ability to make predictions for new locations.

To perform a spatial cross-validation, I first used the k-means algorithm, setting k=5, to identify geographical clusters of data considering latitude and latitude of the data points as follows:

```
dfs <- as.data.frame(dfs, cell = TRUE)
clusters <- kmeans(
  dfs[, 2:3],
  centers = 5
)
dfs$cluster <- clusters$cluster
```

The points are plotted on a global map below with the 5 clusters shown in pink, orange, green, blue, and dark yellow.


80°N
60°N
40°N
20°N
0°
20°S
40°S
60°S
120°W
60°W
0°
60°E
120°E
Cluster
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5



The distribution of LeafN by cluster is shown below:

I used the {purrr} package to split the data into five folds corresponding to geographical clustered identified above.

```
# create folds based on clusters
# assuming 'dfs' contains the data and a column called 'cluster' containing the
# result of the k-means clustering
group_folds_train <- purrr::map(
  seq(length(unique(dfs$cluster))),
  ~ {
    dfs |>
      select(cluster) |>
      mutate(idx = 1:n()) |>
      filter(cluster != .) |>
      pull(idx)
  }
)

group_folds_test <- purrr::map(
  seq(length(unique(dfs$cluster))),
  ~ {
    dfs |>
      select(cluster) |>
      mutate(idx = 1:n()) |>
      filter(cluster == .) |>
      pull(idx)
  }
```

```
)
```

I then used the {ranger} package to fit a random forest model with hyperparameters as mtry=3 and min.node.size=12 and performed a 5-fold cross-validation with the clusters as folds to predict for target variable LeafN.

```
target <- dfs$leafN

train_test_by_fold <- function(dfs, idx_train, idx_val){

  mod <- ranger::ranger(
    x =  dfs[idx_train, 2:9],
    y =  target[idx_train],
  )

  pred <- predict(mod,
                  data = dfs[idx_val, 2:9]
  )

  tmp <- dfs[idx_val,]
  tmp$preds <- predictions(pred)

  rsq <- yardstick::rsq(tmp, "leafN", "preds") # the R-squared determined on the validation set
  rmse <- yardstick::rmse(tmp, "leafN", "preds") # the root mean square error on the validation set

  return(tibble(rsq = rsq, rmse = rmse))
}

out <- purrr::map2_dfr(
  group_folds_train,
  group_folds_test,
  ~train_test_by_fold(dfs, .x, .y)
) |>
  mutate(test_fold = 1:5)
```

The RMSE and R^2 across the five folds are as follows:

```
## # A tibble: 5 x 3
##   rsq$.metric $.estimator $.estimate rmse$.metric $.estimator test_fold
##   <chr>       <chr>            <dbl> <chr>        <chr>           <int>
## 1 rsq         standard       0.542   rmse         standard            1
## 2 rsq         standard       0.357   rmse         standard            2
## 3 rsq         standard       0.00120 rmse         standard            3
## 4 rsq         standard       0.543   rmse         standard            4
## 5 rsq         standard       0.00978 rmse         standard            5
## # i 1 more variable: rmse$.estimate <dbl>
```

The random cross-validation had a higher RSQ, which describes how the variability in the target variable is captured by the model, than the spatial cross-validation and a lower RMSE, which describes the magnitude of errors[1]. While we would expect that spatial cross-validation would result in a model with better predictive ability due to the heavy geographic clustering of the data, our results indicate that the random cross-validation results in a model with a higher ability to make accurate predictions and explain the variability

of the target variable. Ludwig et al., 2023 notes that there are ongoing discussions in the field regarding cross-validation strategies[2]. Wadoux et al., 2021 found in their study that standard cross-validation led to smaller bias than spatial cross-validation[3].

#2.4 Environmental Cross-Validation

```
dfs <- as.data.frame(dfs, cell = TRUE)
clusters <- kmeans(
  dfs[, 5:6],
  centers = 5
)
dfs$cluster <- clusters$cluster

group_folds_train <- purrr::map(
  seq(length(unique(dfs$cluster))),
  ~ {
    dfs |>
      select(cluster) |>
      mutate(idx = 1:n()) |>
      filter(cluster != .) |>
      pull(idx)
  }
)

group_folds_test_env <- purrr::map(
  seq(length(unique(dfs$cluster))),
  ~ {
    dfs |>
      select(cluster) |>
      mutate(idx = 1:n()) |>
      filter(cluster == .) |>
      pull(idx)
  }
)
```

```
target <- dfs$leafN

train_test_by_fold <- function(dfs, idx_train, idx_val){

  mod <- ranger::ranger(
    x =  dfs[idx_train, 2:9],
    y =  target[idx_train],
  )

  pred <- predict(mod,
                  data = dfs[idx_val, 2:9]
  )

  tmp <- dfs[idx_val,]
  tmp$preds <- predictions(pred)

  rsq <- yardstick::rsq(tmp, "leafN", "preds") # the R-squared determined on the validation set
  rmse <- yardstick::rmse(tmp, "leafN", "preds") # the root mean square error on the validation set
```

```
  return(tibble(rsq = rsq, rmse = rmse))
}

out <- purrr::map2_dfr(
  group_folds_train,
  group_folds_test,
  ~train_test_by_fold(dfs, .x, .y)
) |>
  mutate(test_fold = 1:5)
```

```
## # A tibble: 5 x 3
##   rsq$.metric $.estimator $.estimate rmse$.metric $.estimator test_fold
##   <chr>       <chr>            <dbl> <chr>        <chr>           <int>
## 1 rsq         standard         0.754 rmse         standard            1
## 2 rsq         standard         0.823 rmse         standard            2
## 3 rsq         standard         0.509 rmse         standard            3
## 4 rsq         standard         0.920 rmse         standard            4
## 5 rsq         standard         0.831 rmse         standard            5
## # i 1 more variable: rmse$.estimate <dbl>
```

Overall, the environmental cross-validation had roughly equivalent RMSE to the environmental cross-validation and had a higher RSQ than the spatial cross-validation. This indicates that the environmental cross-validation explains more variability in the target variable than the spatial cross-validation and that the model performs better when evaluated on subsets of environmental covariates than spatial. The better performance by the environmental cross-validation indicates and the environmental characteristics have a larger explanatory role than spatial characteristics and that the model is more sensitive to environmental features than spatial.

References [1] Benjamin Stocker, et al. 2019. "Applied Geodata Science (v1.0)." Zenodo. https://doi.org/10.5281/zenodo.7740560. [2] Ludwig, Marvin, Alvaro Moreno-Martinez, Norbert Hölzel, Edzer Pebesma, and Hanna Meyer. 2023. "Assessing and Improving the Transferability of Current Global Spatial Prediction Models." Global Ecology and Biogeography 32 (3): 356–68. https://doi.org/10.1111/geb.13635. [3] Wadoux, A. M.-C., et al. 2021. "Spatial cross-validation is not the right way to evaluate map accuracy." Ecological Modelling, 457, 109692.