

Model validation strategies

Single data set

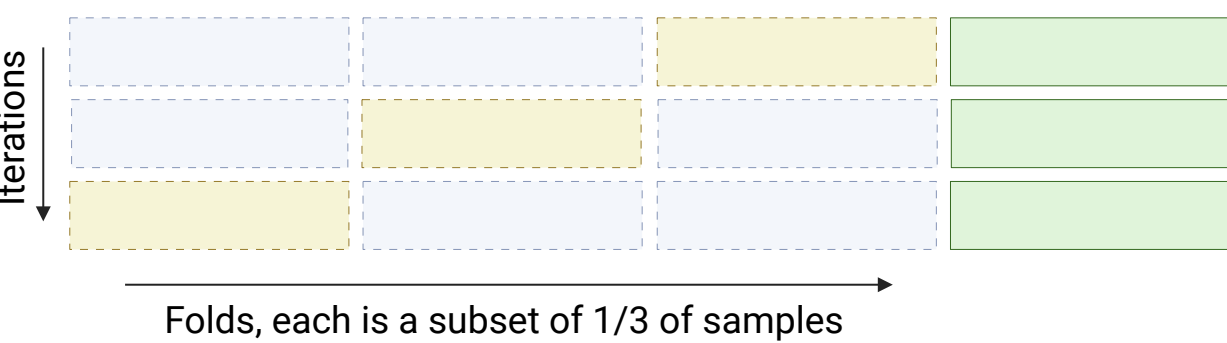
- A. Shuffled train-test split**
Randomly split all samples into training and testing sets.



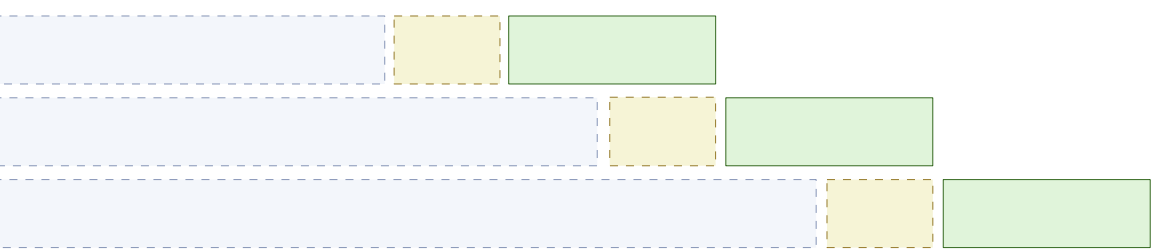
- B. Time-aware train-test split**
Chronologically all samples into training and testing sets.



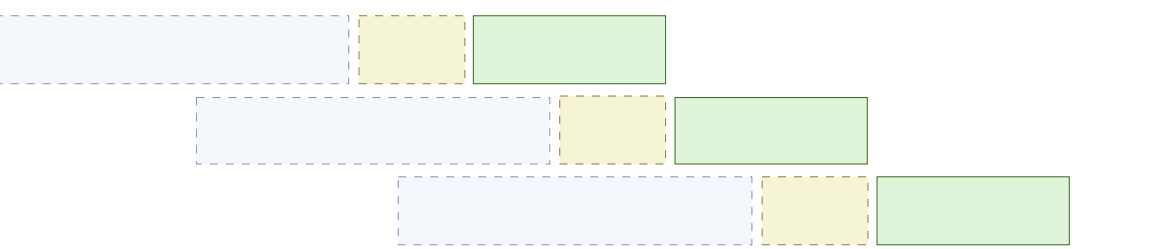
- C. K-fold cross-validation (k=3)**
Randomly split all samples into k folds, at each iteration train on all k-1 folds and validate on the left out hold. Finally, test on the test fold. The chronological sequence may not be maintained within cross-validation folds.



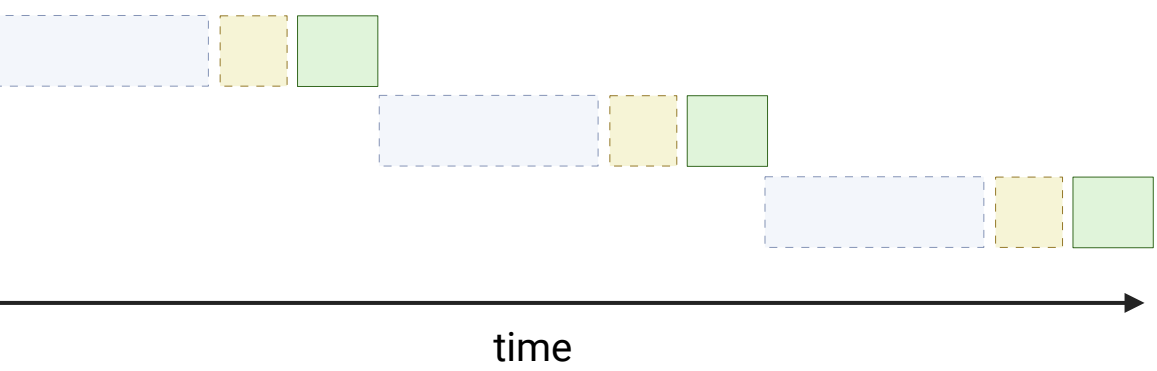
- D. Nested cumulative time-aware cross-validation**
Chronologically split all samples into training, validation and testing sets. The outer loop represents going forward in time, at each iteration the best model is selected using the validation set and its true performance estimated on the test set. The size of the sets grows bigger at each step.



- E. Nested rolling time-aware cross-validation**
Chronologically split data set into training, validation and testing sets. The size of the sets remains constant through the use of a rolling or sliding window.

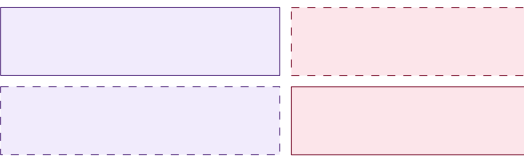


- F. Time-aware k-fold cross-validation (k=3)**
Chronologically split data set into non overlapping training, validation and testing sets. All the sets are therefore fully independent without no information leaking during model selection. Requires large volumes of data.

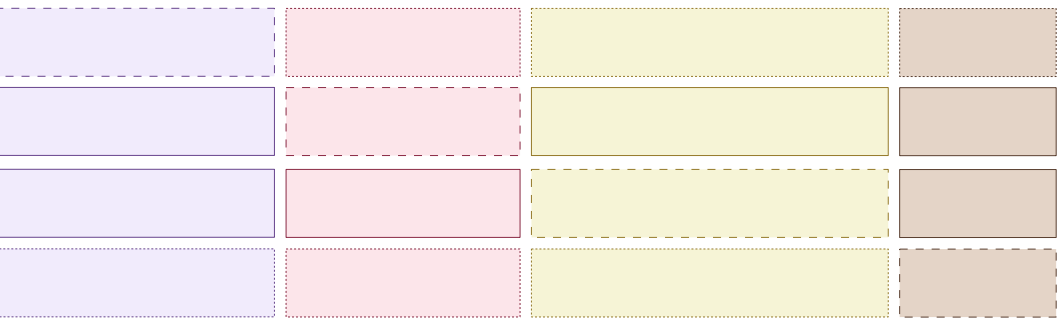


Multiple data sets

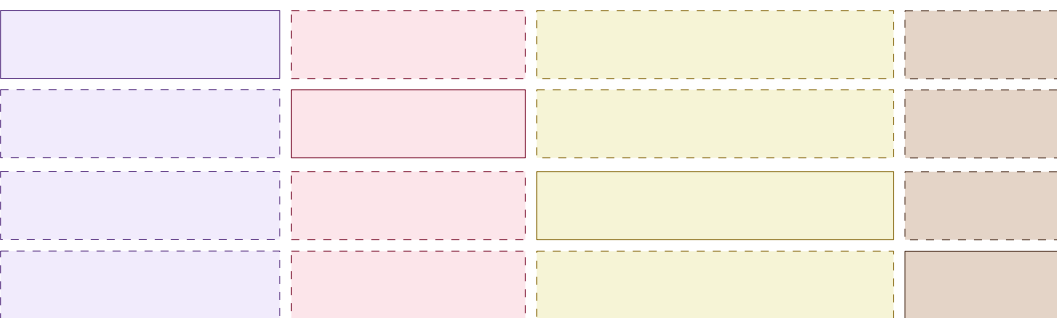
- G. Cross data sets**
Train using one dataset and test on the other.



- H. Train on one data set test on others**
Train on one data set and test its performance on the others.



- I. Leave one data set out**
Train on all but one data set and test on the left-out data set.



Legend

