Classifiers

Assessing Model Accuracy

Alexander Brenning

Department of Geography, Friedrich Schiller University Jena

Geo 408B

What do we need to assess a model‘s accuracy?

A performance measure

An estimation procedure

What do we need to assess a model‘s accuracy?

A performance measure

- An overall numerical measure of the goodness of our predictions of class membership
- E.g., in classification: overall accuracy, kappa coefficient, AUC, Brier score, sensitivity, specificity, …
- In regression: bias, RSE, RMSE, …

…and of course **suitably sampled data**….

- Ideally: random sampling

An estimation procedure

- We don‘t just “calculate” our performance measure – we estimate it (in the statistical sense of “estimation”)
- We need to start thinking about bias and precision of our estimates.

Training Set Approach

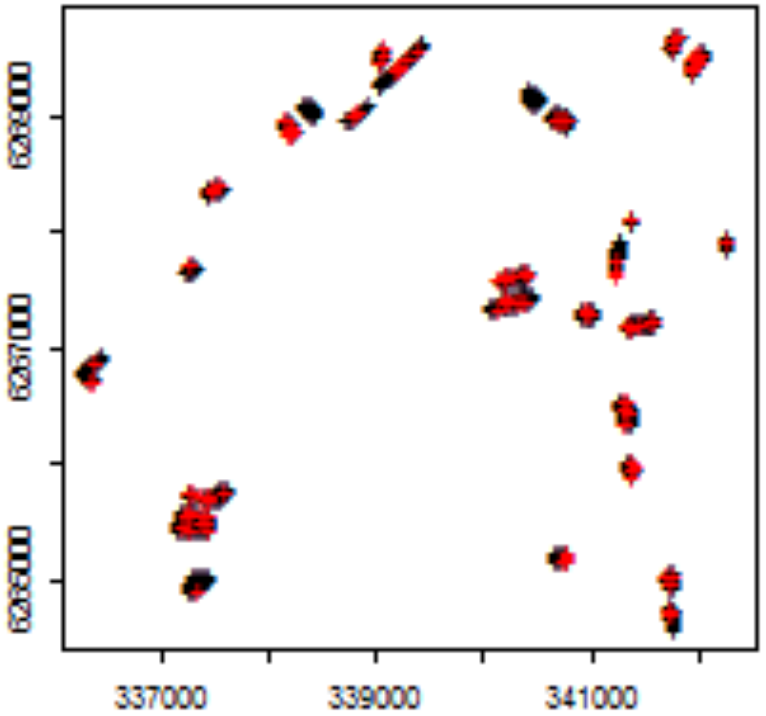
- The classifier‘s performance is assessed on the same data set on which it was trained.
- The resulting error rate is referred to as the **apparent error rate** or **resubstitution error rate**.

- Resubstitution error rates will be **overoptimistic**, especially for very flexible prediction models.

ibm.com

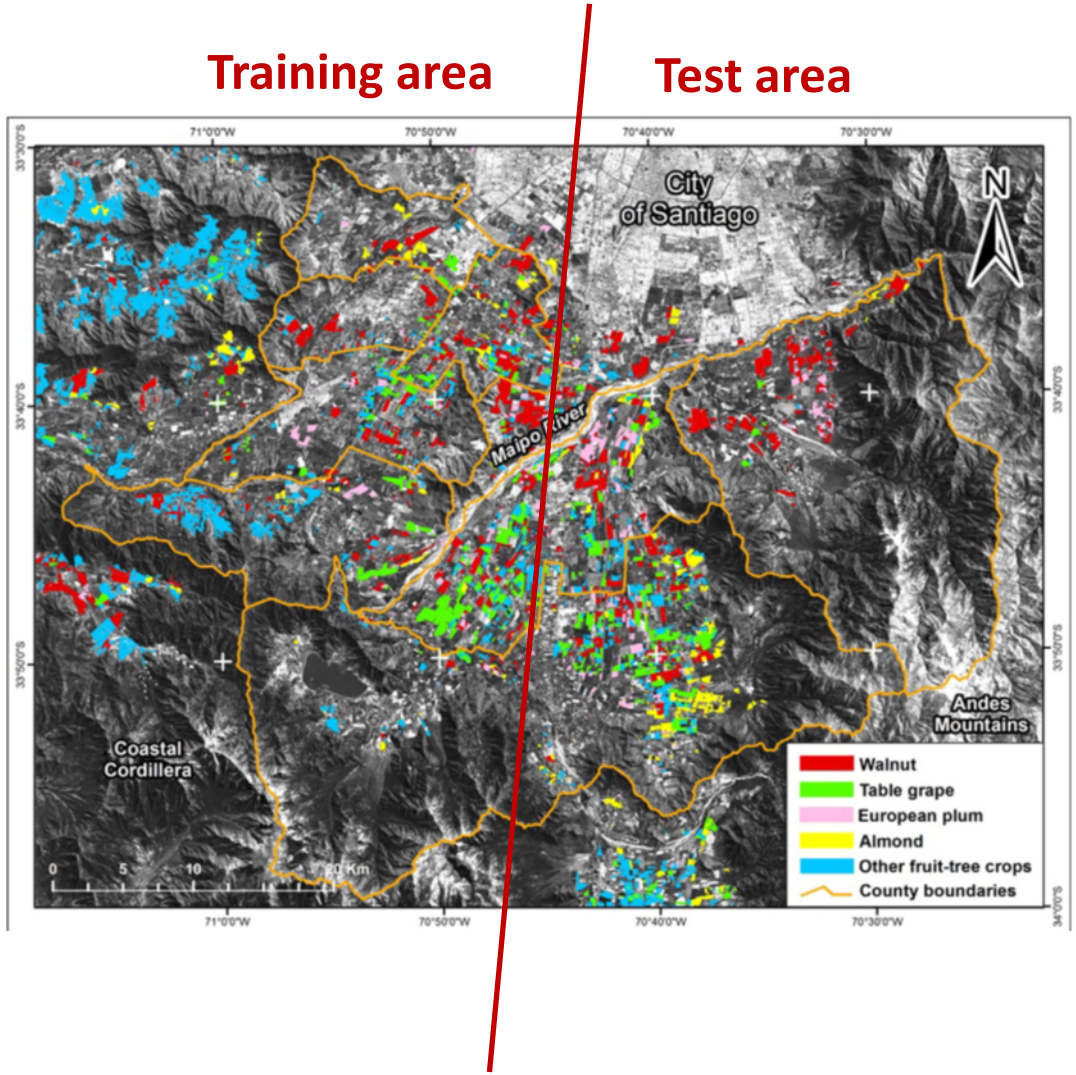
Test Set Approach

- Randomly split the data set into two disjoint sets: a training set and a test set, or hold-out set.
- yields unbiased error estimates
  - (assuming that your data was a random sample)
- But here‘s the catch:
  - Retain a large data set for testing? → Precise error estimator, but unstable classifier.
  - Use a large portion of the data for training? → Poor error estimator.



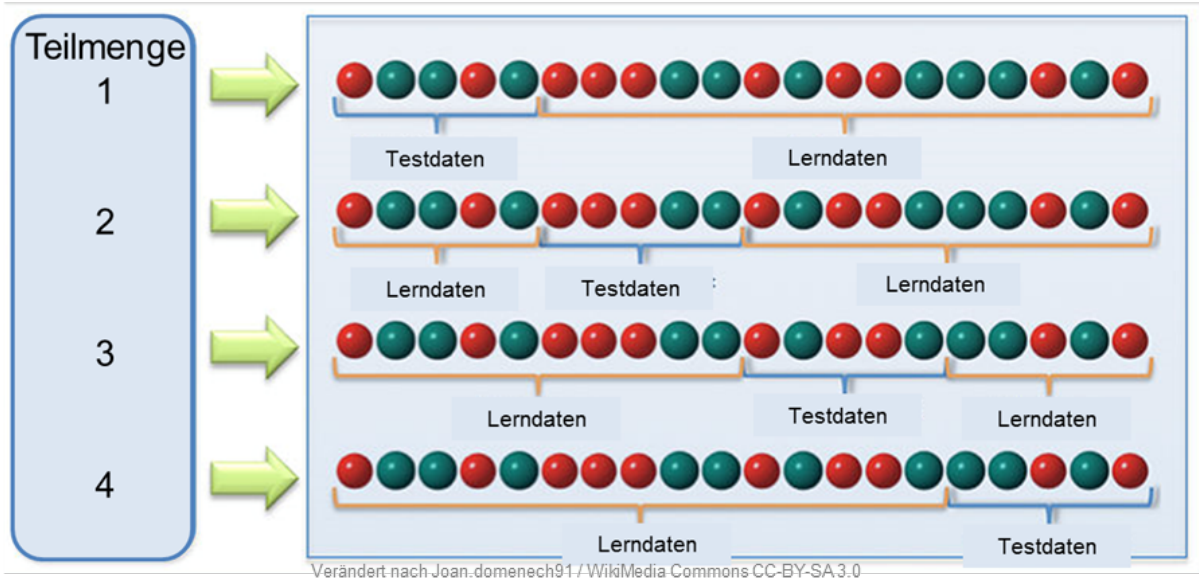
Test *Area* Approach

- Spatial version of the test-set approach: Split the study area into spatially disjoint training and test areas.
- Problem:
  - Training and test areas may have different distributions of, e.g., geological background, topographic characteristics, etc.
  - Test-area error estimates may therefore be biased and not representative for the entire study region.



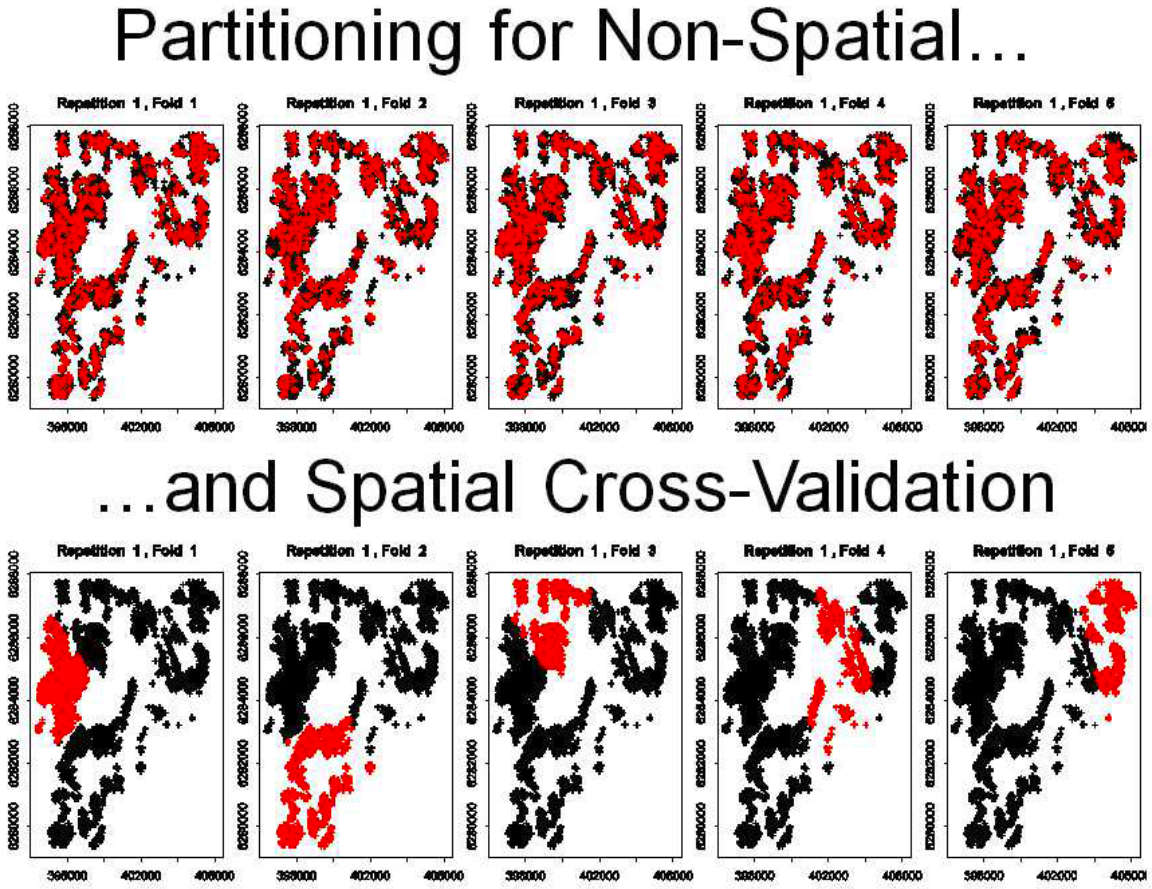
k-fold Cross-Validation

- Randomly partition the sample into *k* equally-sized disjoint subsets.
  - Usually *k* = 10 or *k* = 5.
- Train the classifier on the data from all but one of these subsets,
- and test it on the held out set (or fold).
- Repeat this for all *k* partitions in order to use the entire data set for testing.
  - Also repeat this procedure *r* times using different random partitionings.
- Special case of *k* = *N*: **Leave-one-out cross-validation** (LOO-CV)



Spatial Cross-Validation

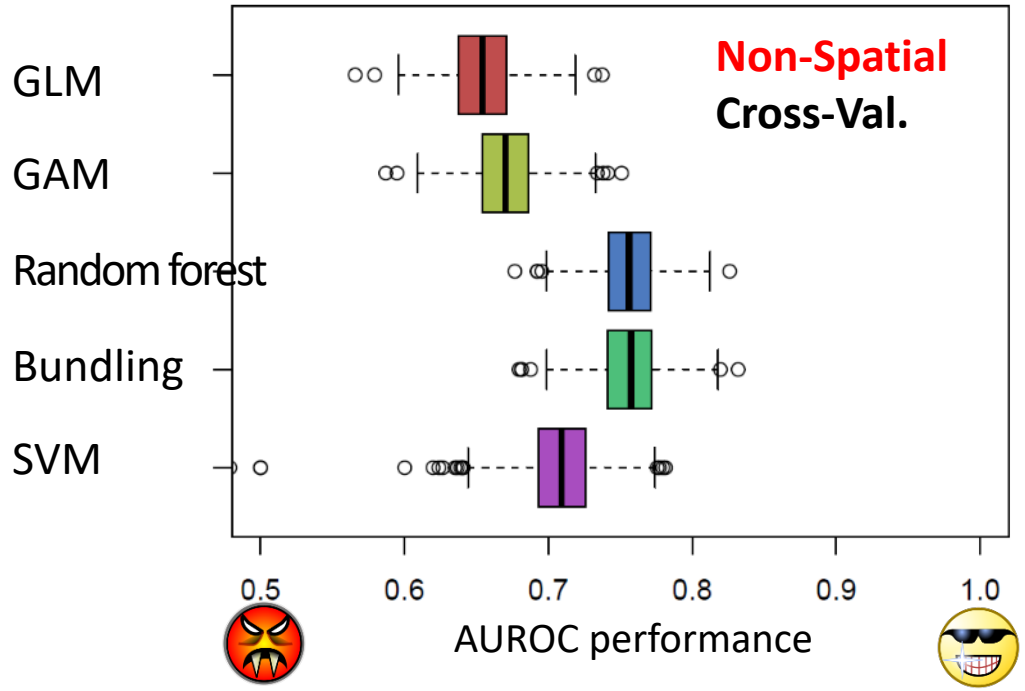
- Divide the study area into disjoint subregions
  - E.g. using *k*-means clustering of coordinates (Ruß & Brenning, 2010)
- Perform cross-validation at the level of the subregions
  - I.e. leave out one subregion at a time
- If data is grouped, perform cross-validation at the group level.
  - E.g. at the field level when there are multiple data points within a field

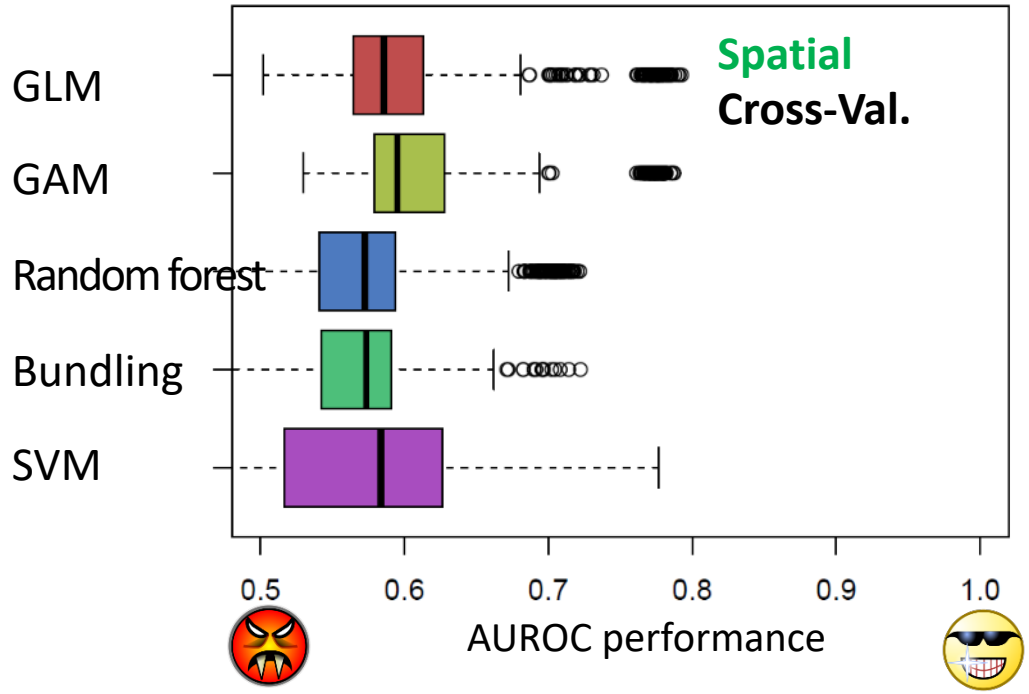


Model Performance: Landslide Susceptibility

- Non-spatial C.V. results are over-optimistic
- Spatial C.V. reveals overfitting to training data
- Simpler methods more transferable

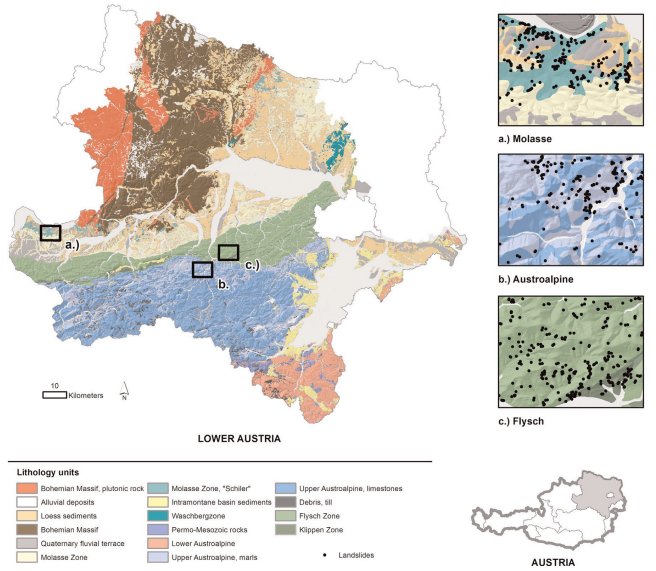
Compare Brenning (2005) in NHESS

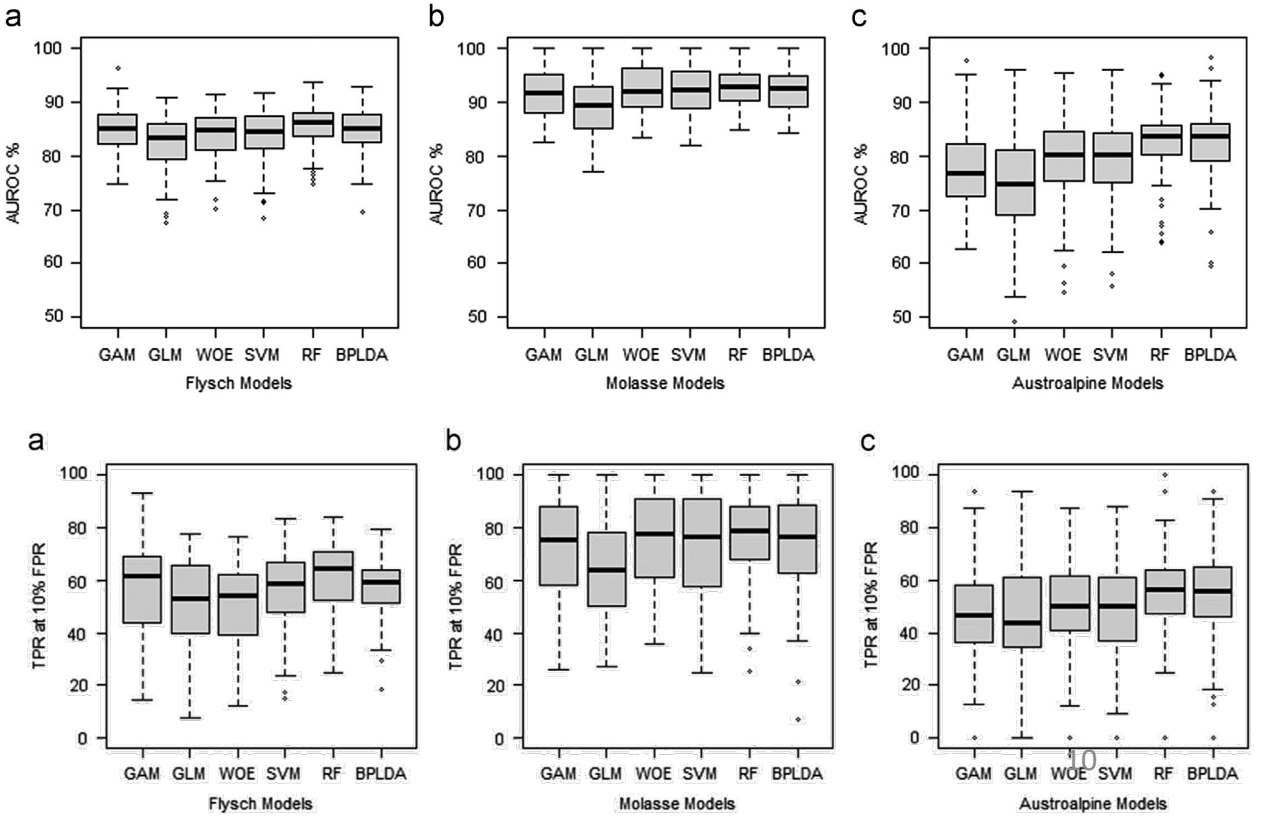




Example:
Landslides in Lower Austria

- Comparative study using data from different geological units in Lower Austria (193-285 landslides each)
- No significant differences, although tree ensembles showed best average performances

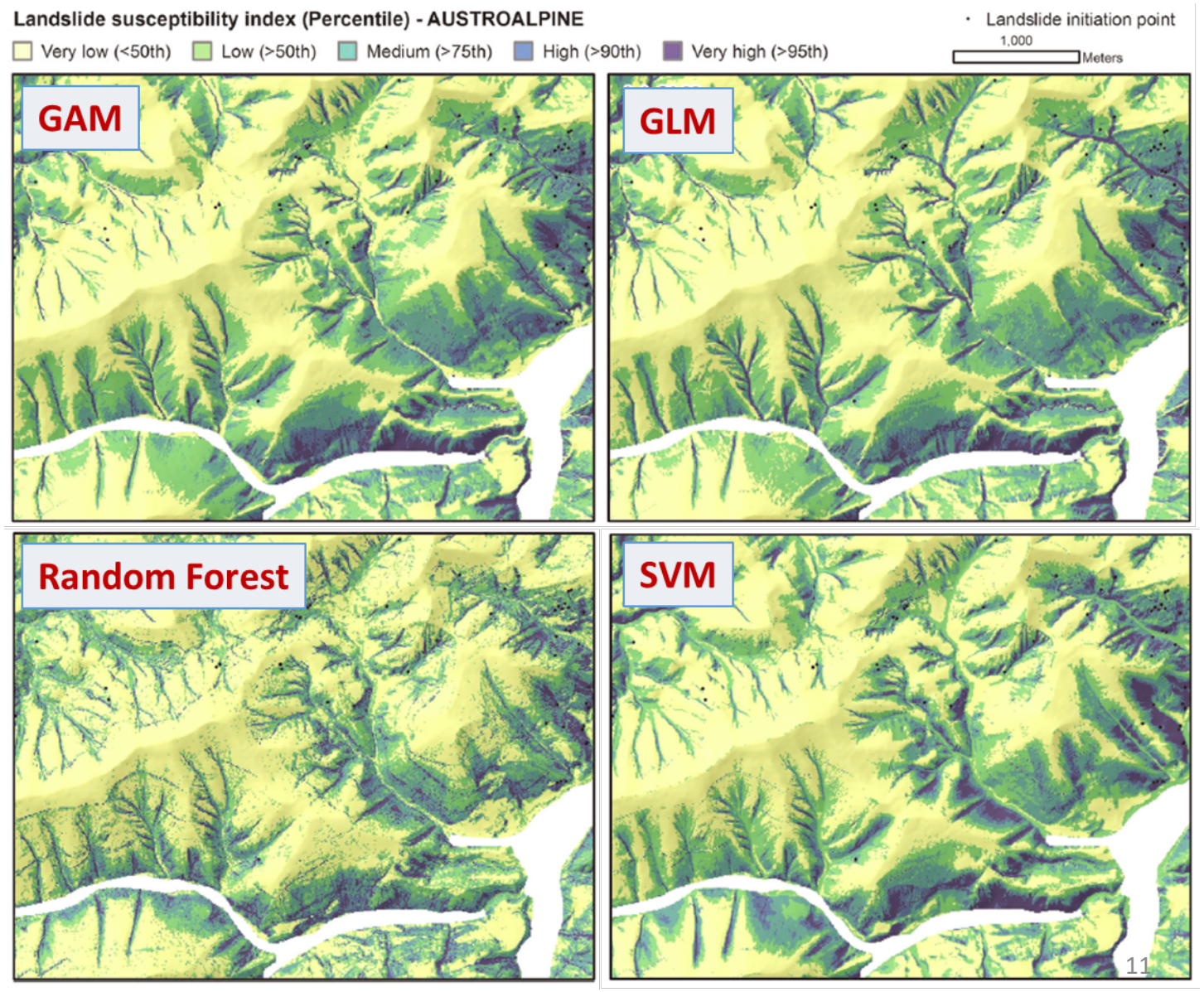




Goetz et al. (2015)

Visual Comparison: Lower Austria

- Also consider qualitative criteria (Steger *et al.*, 2015)

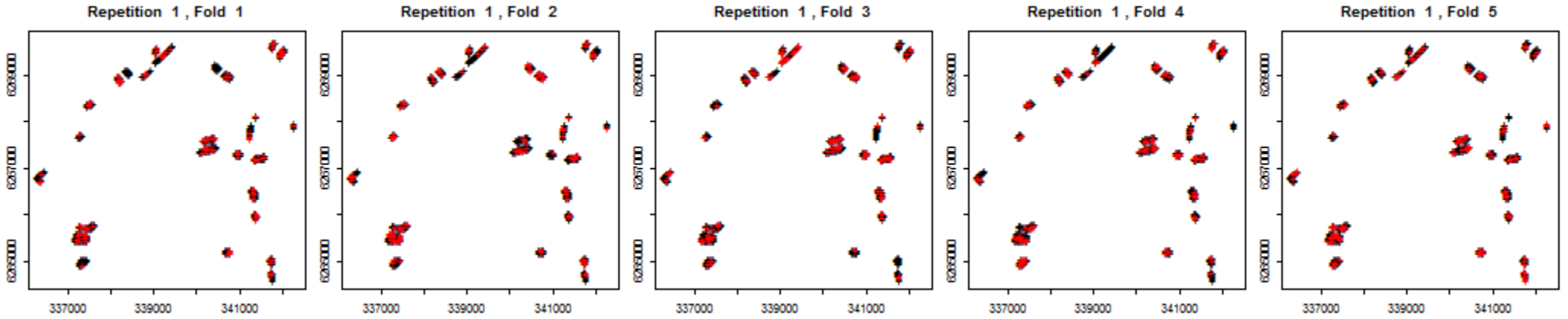


Goetz et al. (2015)

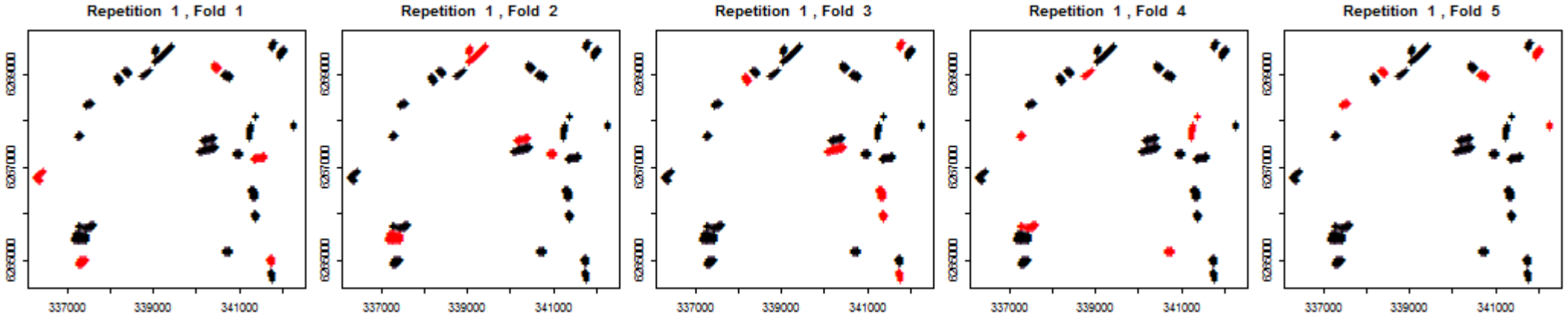
Spatial Cross-Validation: Resampling at the Field Level

Figures show a small subsample of the crop classification data

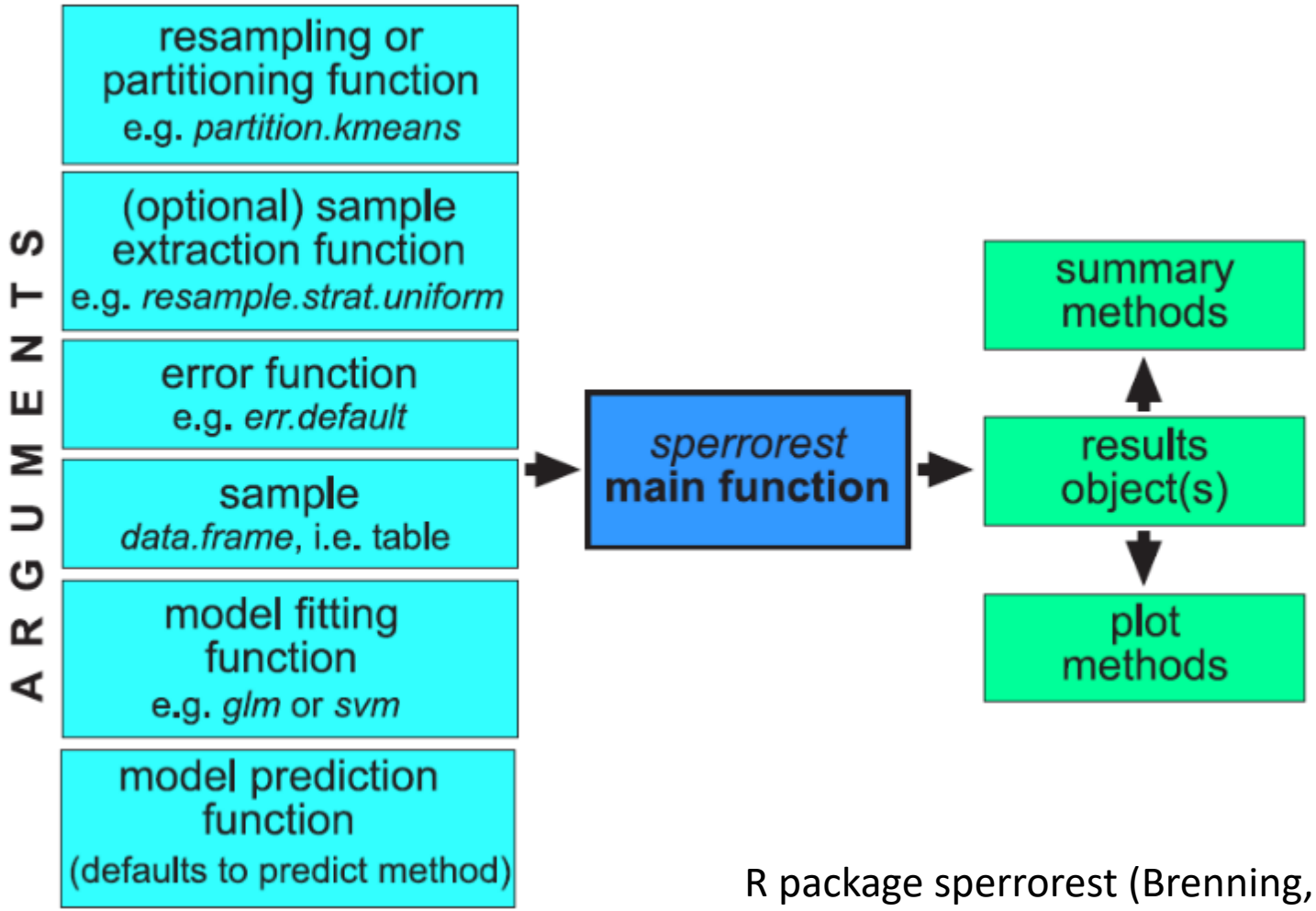
Ordinary



Spatial

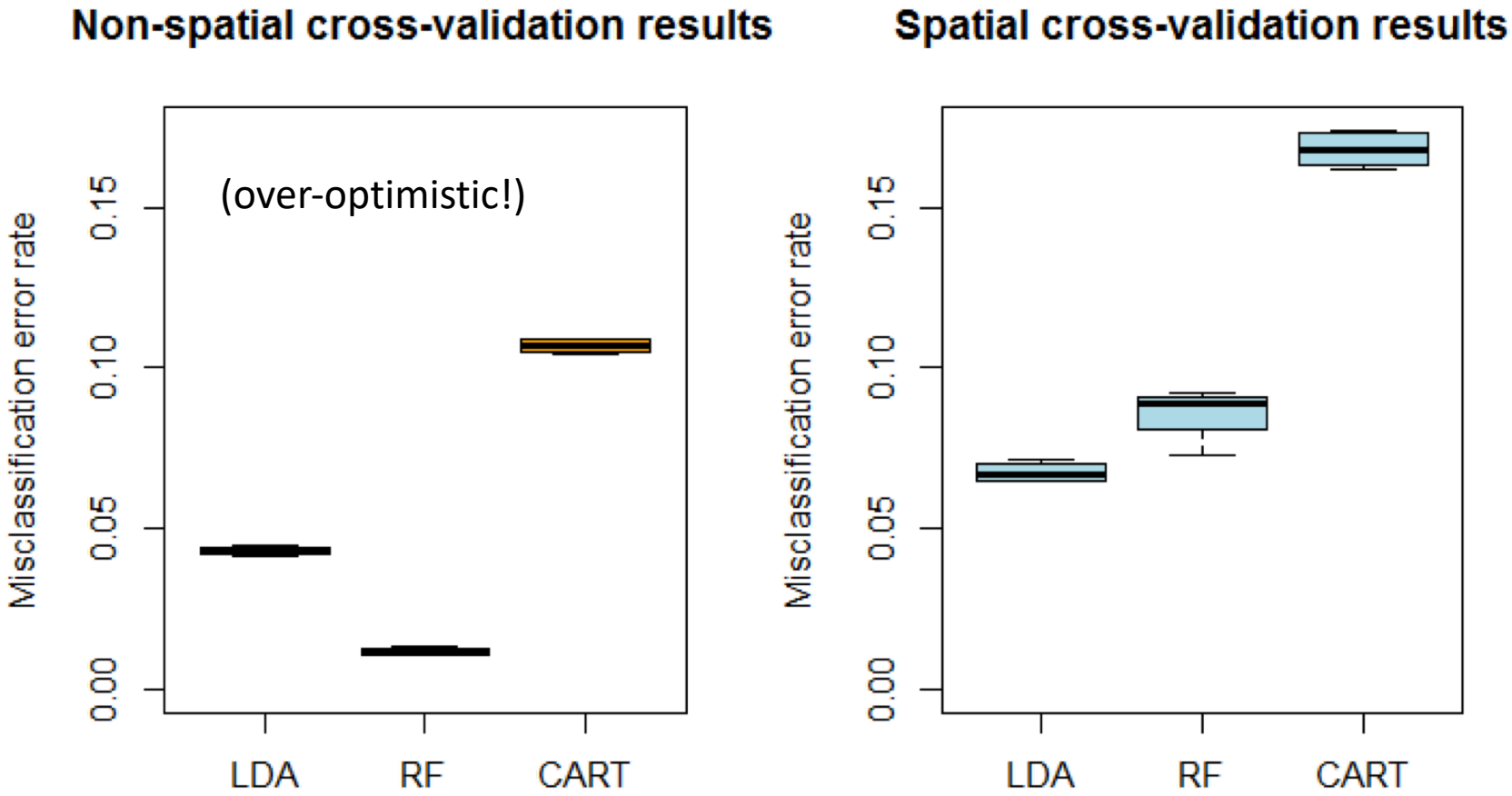


Spatial Cross-Validation in R



R package sperrorest (Brenning, 2012)

Example: Crop Classification
Misclassification Error Rate



Example: Crop Classification
Misclassification Error Rate (MER) Estimates

Classifier	Apparent MER	Non-spatial CV	Spatial CV
LDA	0.044	0.043	**0.068**
CART	0.111	0.107	0.168
Random Forest	0.000	0.012	0.086

Spatial CV: 5-fold cross-validation at the field level

i.e. pixels that belong to the same field will jointly be either in the training set or in the test set

Variable Importance

Variable Importance

- Which variables really contributed to the model?
  - In generalized linear models, look at:
    - Model coefficients and derived quantities such as odds ratios
    - Statistical inference, i.e. hypothesis tests on model coefficients

- **Permutation-based variable importance**: general-purpose computational measure of variable importance

Variable Importance

- Which variables really contributed to the model?
  - In generalized linear models, look at:
    - Model coefficients and derived quantities such as odds ratios
    - Statistical inference, i.e. hypothesis tests on model coefficients
- **Permutation-based variable importance**: general-purpose computational measure of variable importance

Algorithm:

1. Train the model, and assess ist accuracy on test set.
2. Permute a predictor in the test set, and use this partly messed-up data for prediction and accuracy assessment.
3. Calculate the difference between “regular” and “messed-up” test accuracy.

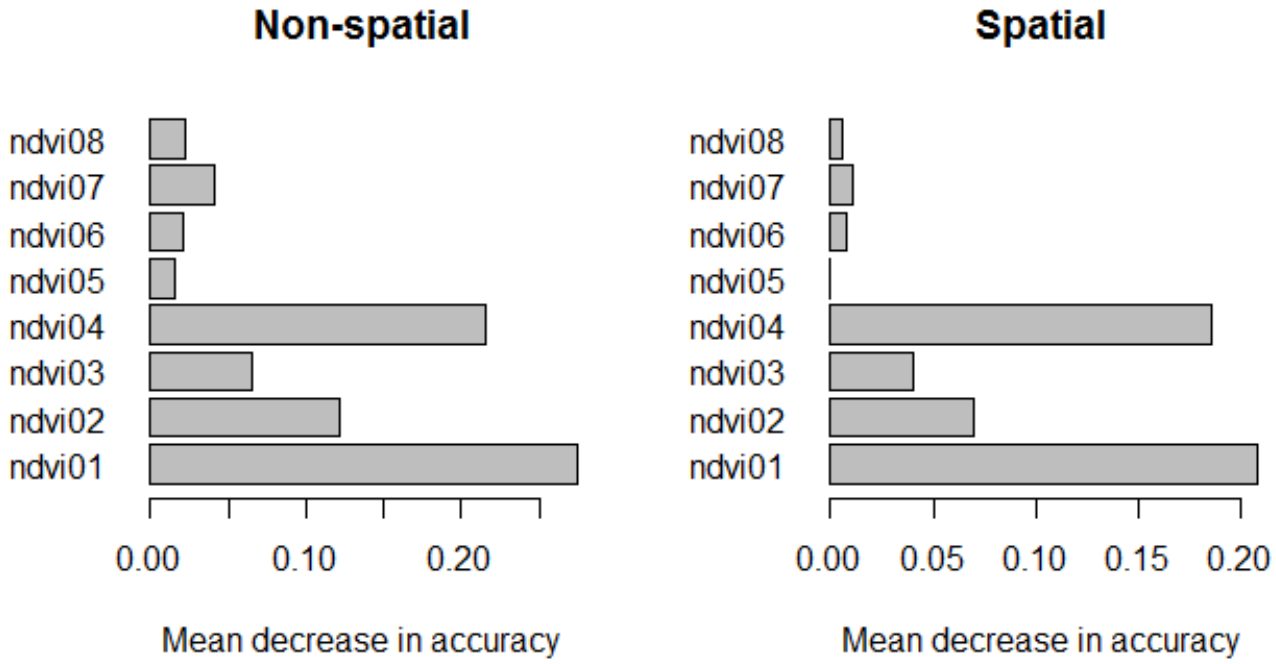
Repeat this for each variable, and for each cross-validation training / test set combination.

Use many random permutations.

Spatial Variable Importance

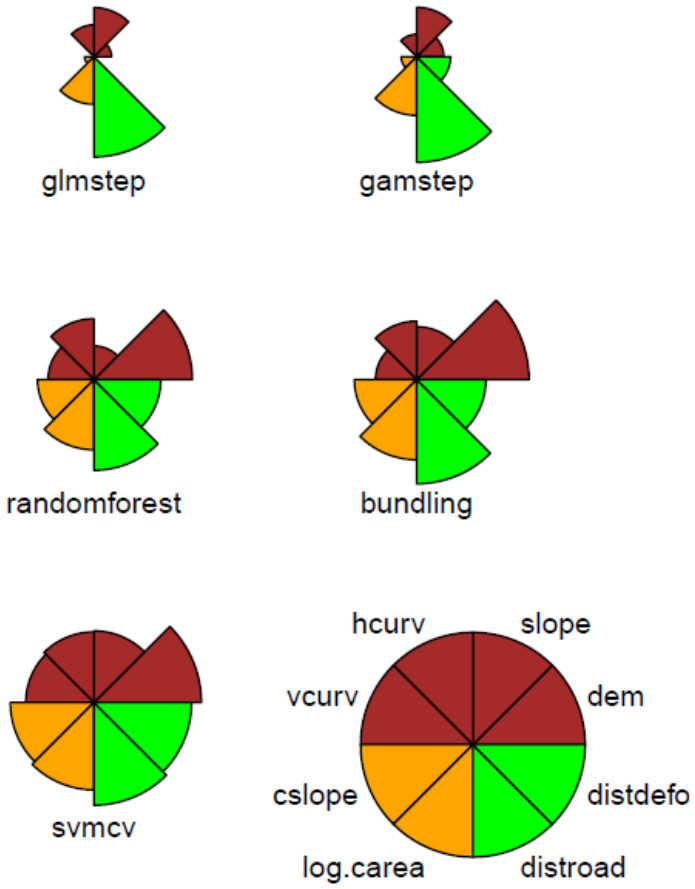
- "Standard" permutation importance ignores spatial dependence or grouping
- Embed variable importance assessment within a spatial cross-validation to assess a variable's ability to contribute to *generalizable* predictive capabilities.
  - `sperrorest` package

Simple Example: Crop classification using Random Forest, only NDVI variables

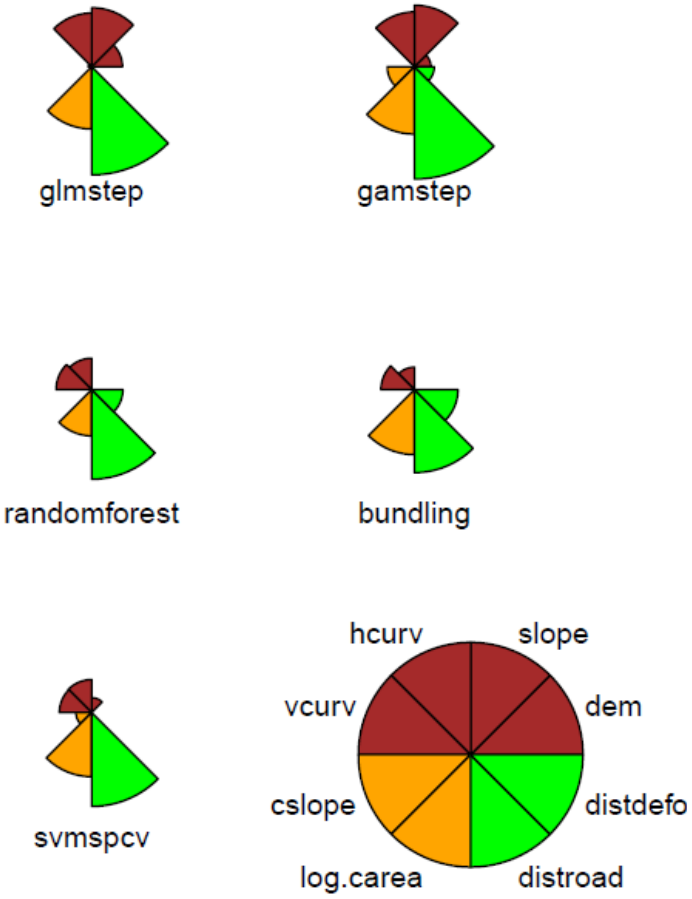


Example: Landslides in Ecuador

Non-spatial AUROC importance



Spatial AUROC importance



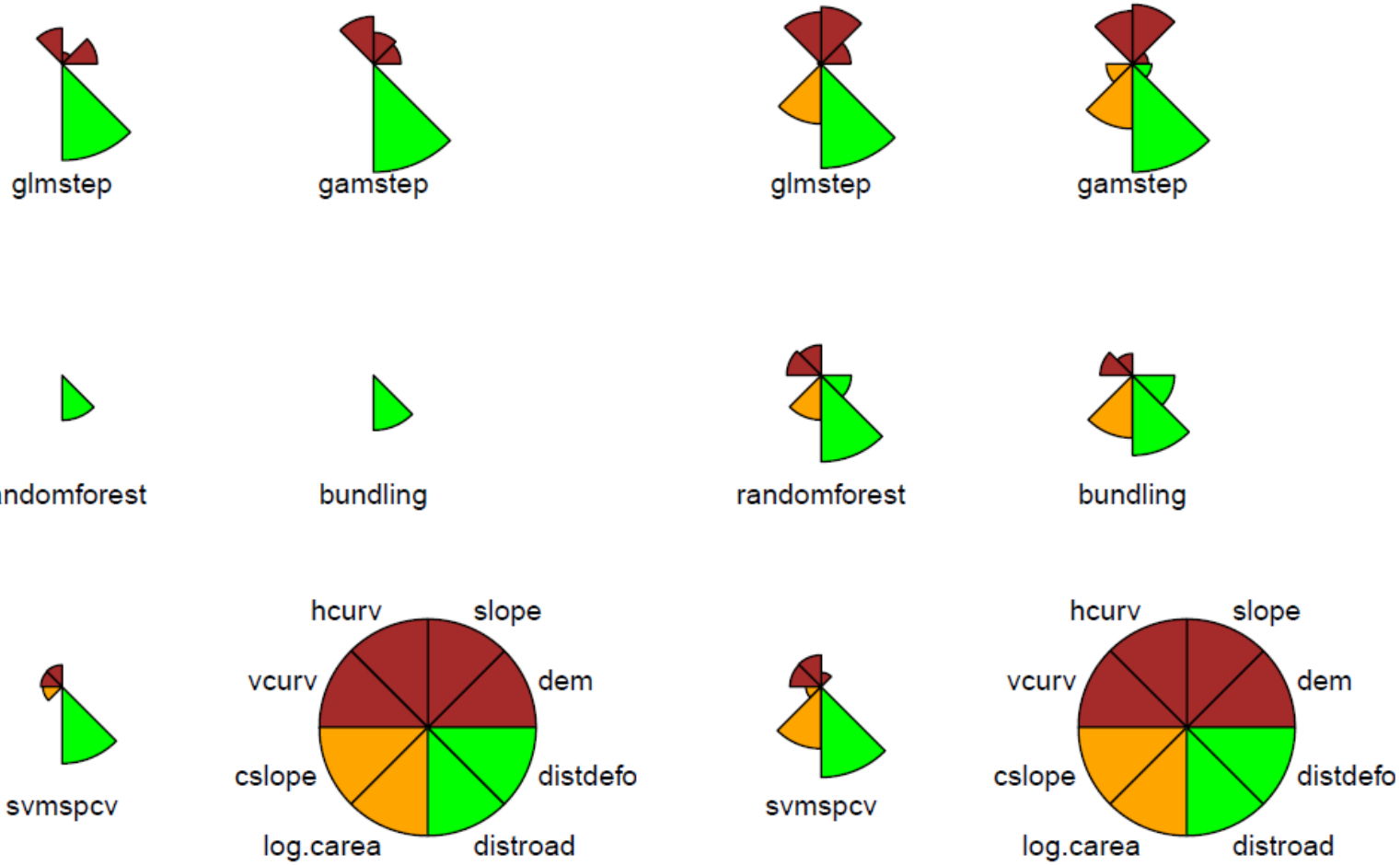
Example: Landslides in Ecuador

Non-spatial TPR90 importance

Spatial TPR90 importance

Variable importance also varies with the performance criterion used!

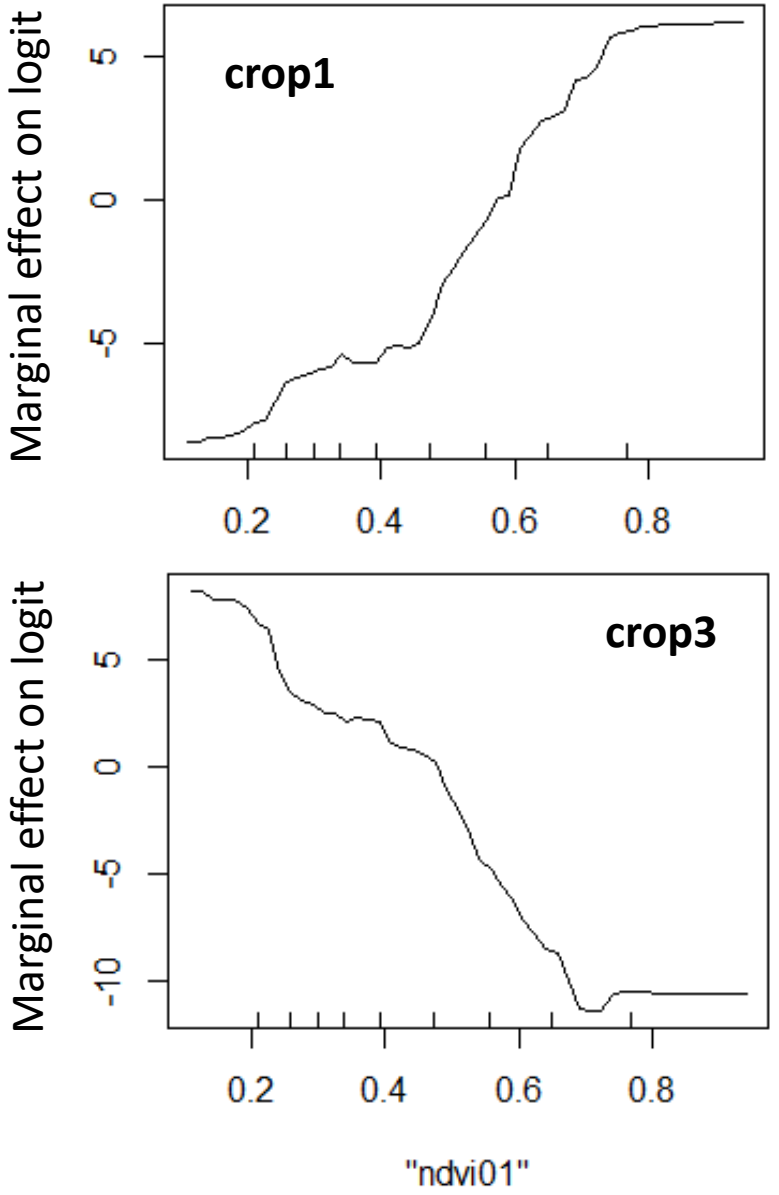
TPR90: true positive rate (sensitivity) at a 90% specificity



Analyzing Models

- In general, nonlinear models with higher-order interactions among variables are hard to understand…
- Additive models: plot the nonlinear transformations – their effects add up
- More complex models,
  - Plot 1D transects of the prediction function
  - Plot marginal effect of a predictor on predicted class probability – e.g. `partialPlot` for `randomForest`
- *(But why use such complex models to interpret relationships?)*

Crop Classification:
Random Forest Marginal Plots



Lessons Learned

- In predictive modelling, we can be pragmatic about the type of model used – as long as it provides good predictions.
- More flexible models tend to overfit to the training data.
- This may remain undetected if spatial autocorrelation isn‘t taken into account in accuracy assessments.
- Always use nested cross-validation to tune hyperparameters.
- Use appropriate error measures that match the objectives of your study.

Predictive Performance

What do we need to assess a model‘s accuracy?

A performance measure

- An overall numerical measure of the goodness of our predictions of class membership
- E.g., in classification: overall accuracy, kappa coefficient, AUC, Brier score, sensitivity, specificity, …
- In regression: bias, RSE, RMSE, …

…and of course **suitably sampled data**….

- Ideally: random sampling

An estimation procedure

- We don‘t just “calculate” our performance measure – we estimate it (in the statistical sense of “estimation”)
- We need to start thinking about bias and precision of our estimates.

Confusion Matrix

- The confusion matrix estimates the number of observations (or e.g. the area) corresponding to each combination of predicted and observed class:

$$(\#\{y = i \text{ and } \hat{y} = j\})_{i,j=1,\ldots,m}$$

- The diagonal elements represent correctly classified areas.
- Remember that any proportions calculated from the confusion matrix are proportion estimators!
  - Proportion estimators have standard error $SE(\hat{p})=\sqrt{\frac{\hat{p}\hat{q}}{n}}$ under certain conditions…
- Unbalanced confusion matrix → Predicted areas are biased!
- Example of a simple classification tree model applied to landslides in Ecuador:

Area in km²	Predicted non-lsl.	Predicted landslide	Total
Observed non-lsl.	6.41	3.83	**10.25 (92%)**
Observed landslide	0.38	0.58	**0.95 (8%)**
Total	**6.79 (61%)**	**4.41 (39%)**	**11.20**

Overall accuracy, Misclassification error rate

- The overall accuracy is the probability of correct classification.
- We estimate the **overall accuracy** using

$$\widehat{OA} = \#\{y = \hat{y}\}/n$$

- This is the sum of the diagonal elements of the confusion matrix, divided by the sample size, *n*.

- The **misclassification error rate** is estimated using

$$\widehat{MER} = 1 - \widehat{OA}$$

- Both are proportion estimators and have sampling variability!
- Both give the same weight to all types of misclassification.
- The ***κ* (kappa) index** modifies the OA to account for chance agreement between observation and prediction.

Sensitivity & Specificity (Producer's Accuracy)

Sensitivity

- Proportion of positive observations that are correctly classified (predicted) as positives.

$$\hat{p}(\hat{y} = 1 | y = 1) = \frac{\#\{\hat{y} = 1 \text{ and } y = 1\}}{\#\{y = 1\}}$$

- AKA as true positive rate (TPR) or hit rate.
- *What percentage of diseased persons gets a positive diagnosis?*
- *What percentage of landslide-affected grid cells is predicted as being unstable?*

Specificity

- Proportion of negative observations that are correctly classified as negatives.

$$\hat{p}(\hat{y} = 0 | y = 0) = \frac{\#\{\hat{y} = 0 \text{ and } y = 0\}}{\#\{y = 0\}}$$

- AKA as true negative rate (TNR).
- *What percentage of stable slopes gets classified as stable?*

Positive & Negative Predictive Value (User's Accuracy)

Positive Predictive Value

- Proportion of positive observations that are correctly classified (predicted) as positives.

$$\hat{p}(y = 1|\hat{y} = 1) = \frac{\#\{\hat{y} = 1 \text{ and } y = 1\}}{\#\{\hat{y} = 1\}}$$

- AKA as precision.
- *When a emergency is detected, what is the probability that it's a real emergency?*

Negative Predictive Value

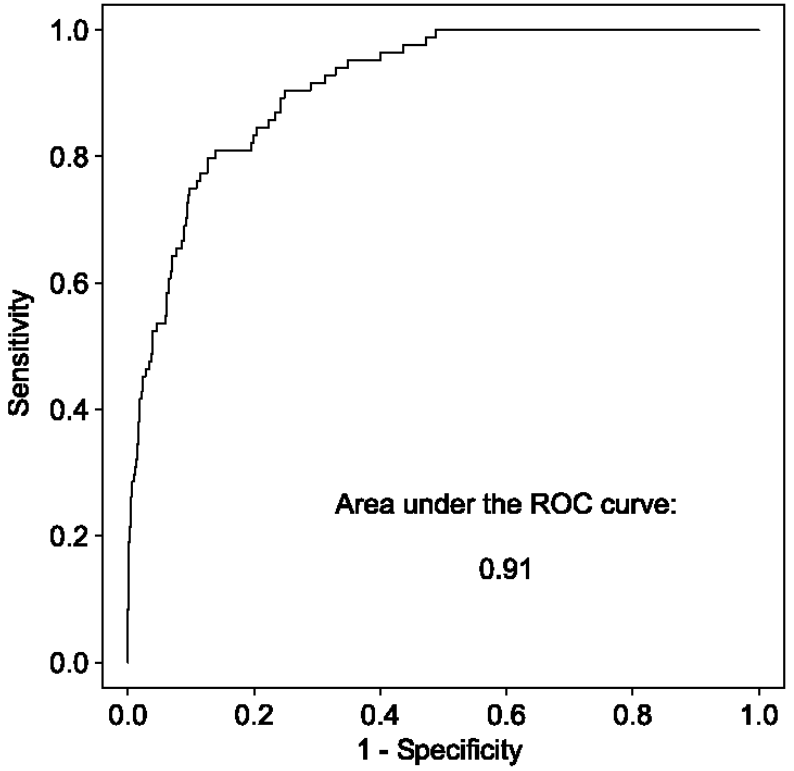
- Proportion of negative observations that are correctly classified as negatives.

$$\hat{p}(y = 0|\hat{y} = 0) = \frac{\#\{\hat{y} = 0 \text{ and } y = 0\}}{\#\{\hat{y} = 0\}}$$

- *If the image classification says that there is no oil spill, what is the probability that there is really no oil spill?*

ROC Curve

- For „soft“ classifiers
  → Sensitivity and specificity depend on a particular decision threshold used
- **Receiver-operating characteristic** (ROC) curve:
  Vary the decision threshold, and plot all possible sensitivities against the respective specificities
- Area under the ROC curve (**AUROC**, or **AUC**)
  Rule of thumb (don‘t take it too seriously!!!):
  0.9-1: excellent
  0.8-0.9: very good
  0.7-0.8: good
  0.6-0.7: average
  0.5-0.6: poor
  <0.5: check your class labels!



example: Brenning et al. (2007), Rock glacier distribution modeling using generalized additive models and terrain attributes

References

- Brenning, A. (2012): Spatial cross-validation and bootstrap for the assessment of prediction rules in remote sensing: the R package ‘sperrorest’. Proceedings, 2012 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 23-27 July 2012, 5372-5375.
- Goetz, J.N., Brenning, A., Petschko, H., Leopold, P. (2015). Evaluating machine learning and statistical prediction techniques for landslide susceptibility modeling. Computers & Geosciences, 81: 1-11.
- Peña, M.A., Brenning, A. (2015). Assessing fruit-tree crop classification from Landsat-8 time series for the Maipo Valley, Chile. Remote Sensing of Environment, 171: 234-244.