Machine-Learning
Model Assessment & Interpretation

Alexander Brenning

Department of Geography, Friedrich Schiller University Jena

OpenGeoHub Summer School 2023, Poznan

Machine-Learning Model Assessment

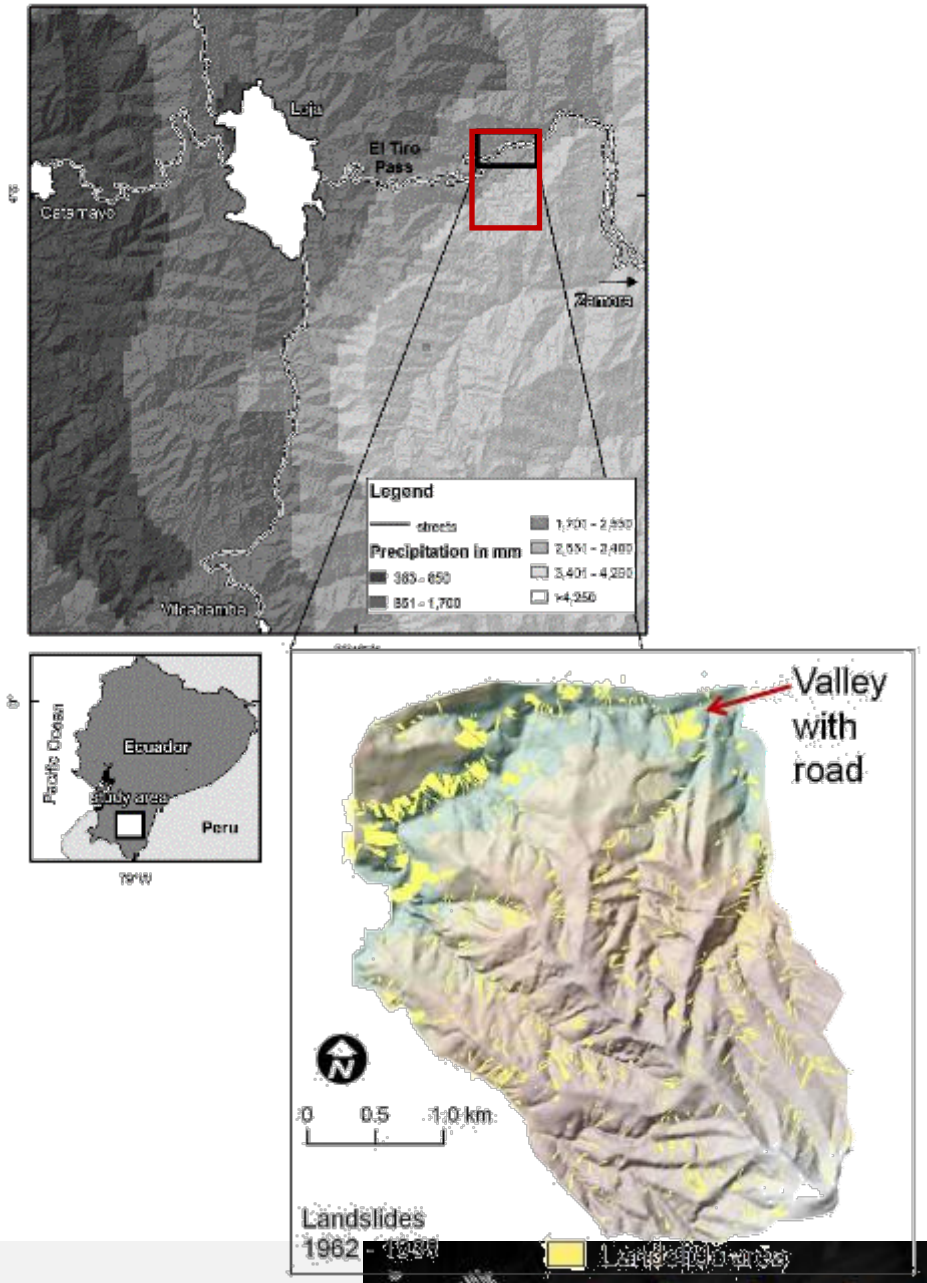
Alexander Brenning

Department of Geography, Friedrich Schiller University Jena

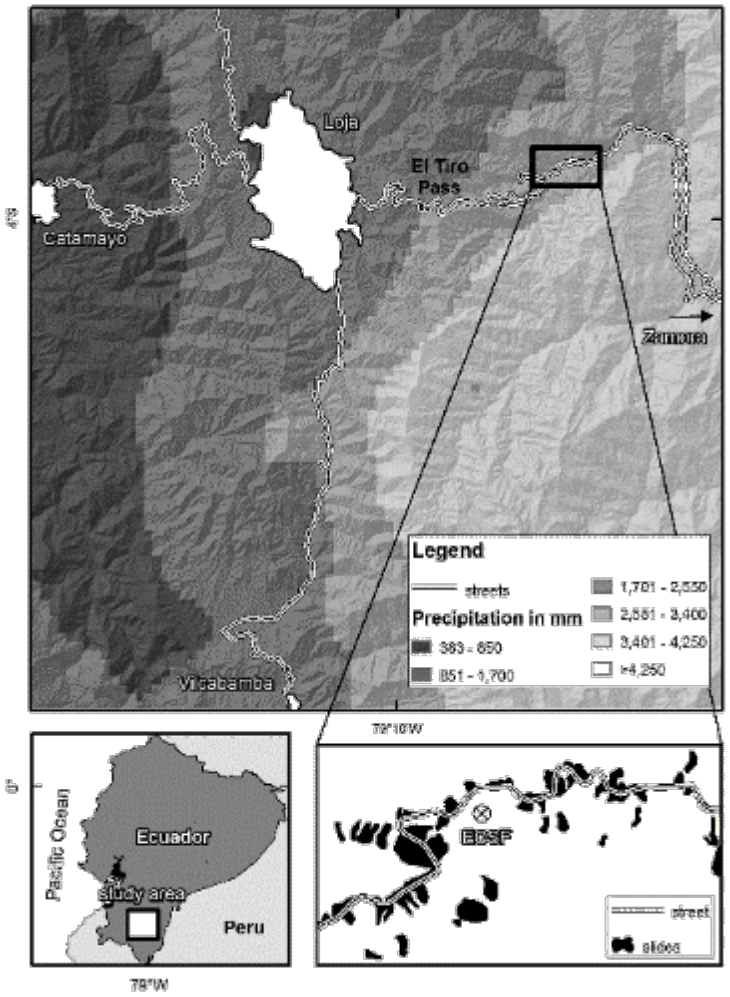
OpenGeoHub Summer School 2023, Poznan

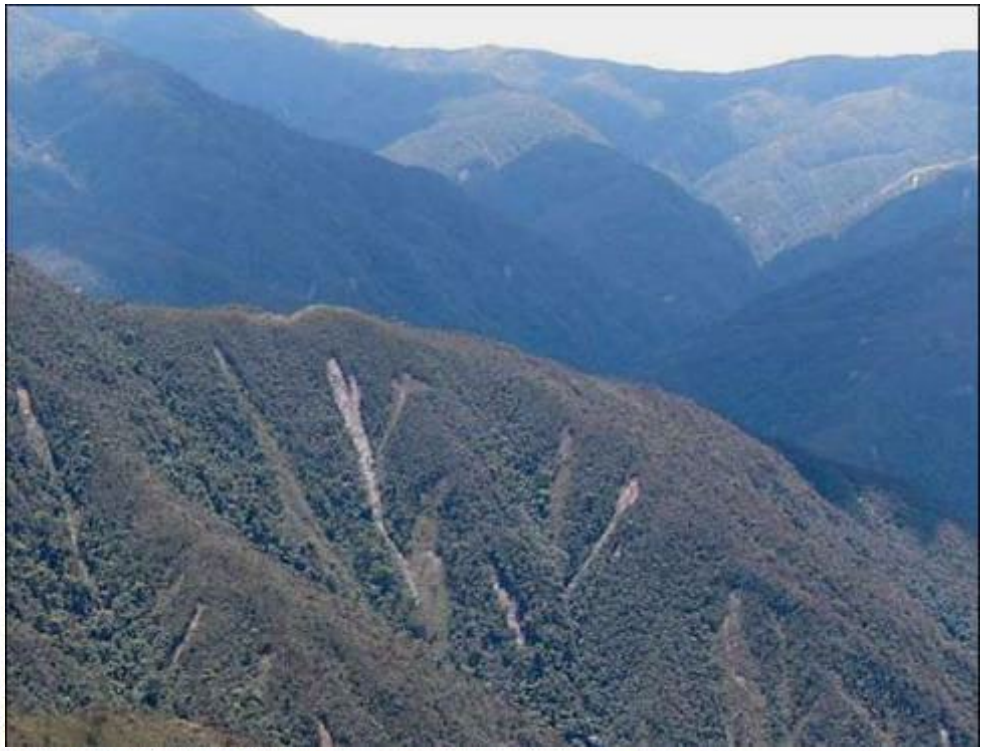
Case Study 1: Landslide Susceptibility

- Goal:
  - **Prediction**: Identify landslide-prone areas, and/or
  - **Analysis**: Identify preparatory factors
- Response:
  - Landslide inventory (presence / absence)
- Predictors:
  - Terrain attributes (e.g., slope angle), land use, distance to road, rock type
- Two-class problem, “soft” classification
- Step-by-step tutorials: **RSAGA** package vignette and **GeocompR** book Chapter 12



Landslides in Southern Ecuador

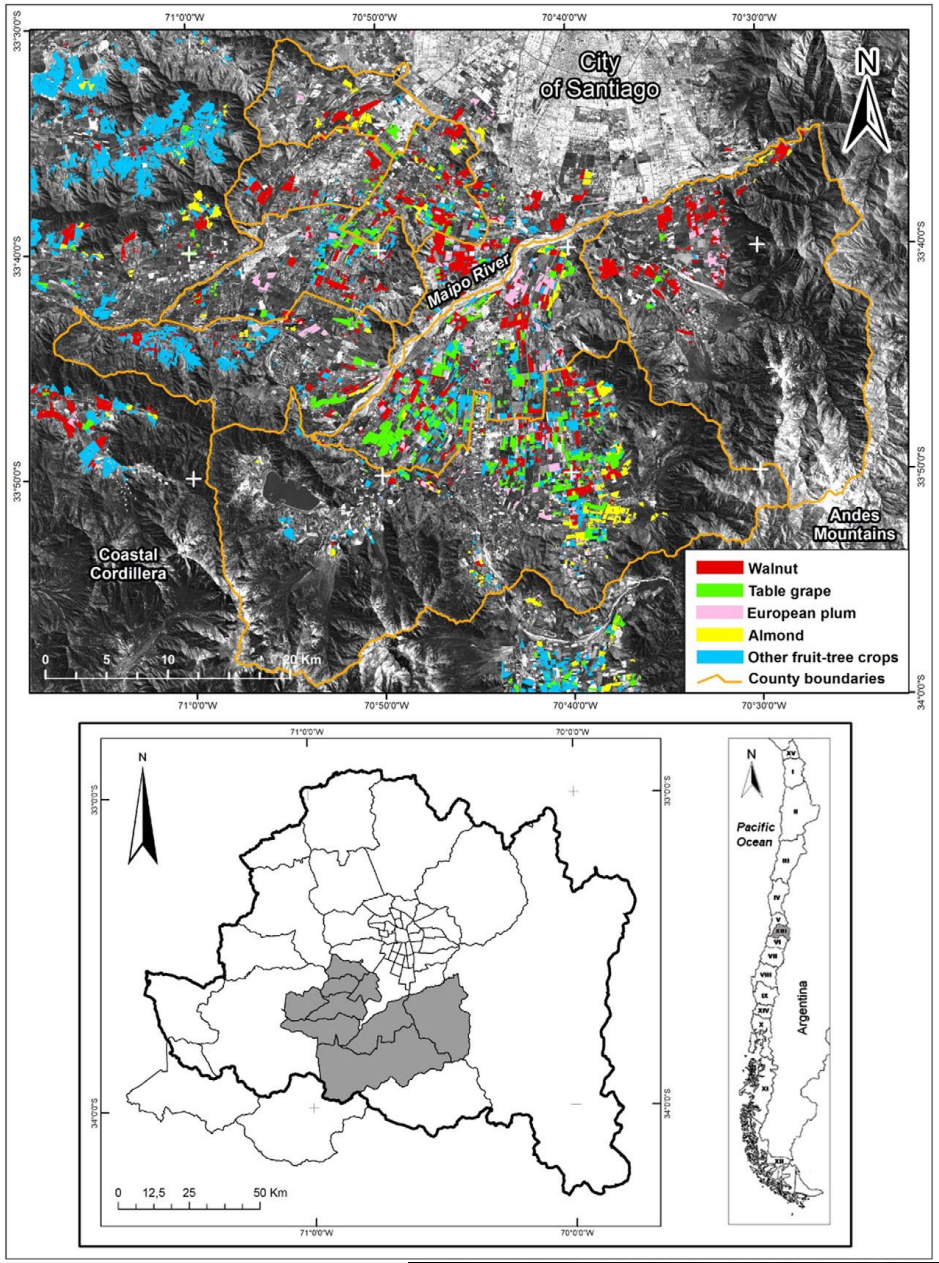




FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA

Case Study 2: Crop Classification

- Goal: Predict crop type based on multitemporal Landsat satellite data
- Data for 7713 raster cells within 400 fields
- Response: 4 fruit-tree crop types
- Predictors: satellite image time series:
  - 6 spectral bands x 8 satellite image dates
  - NDVI and NDWI temporal profiles, i.e. 8 “vegetation index” and 8 “water index” variables
- Subsample of data used by Peña & Brenning (2015) in *Remote Sensing of Environment*
- Characteristics:
  - Multiclass prediction
  - Observations (raster cells) grouped by field



Machine-Learning Modeling

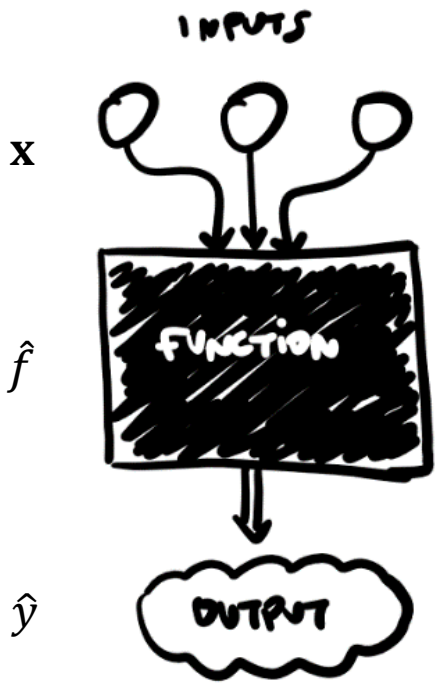
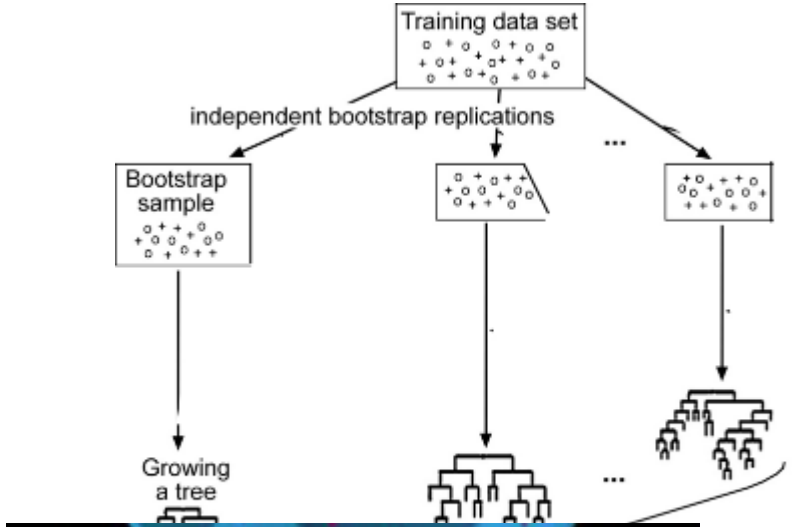
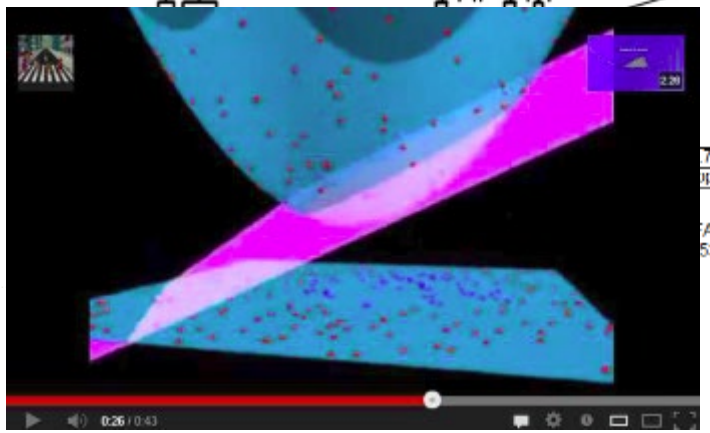
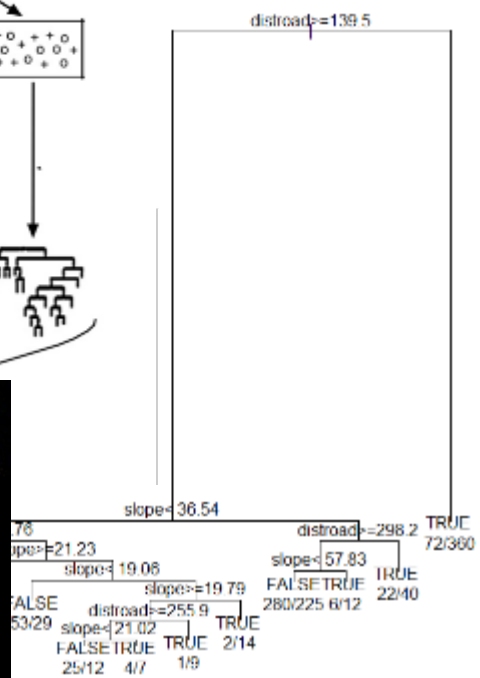
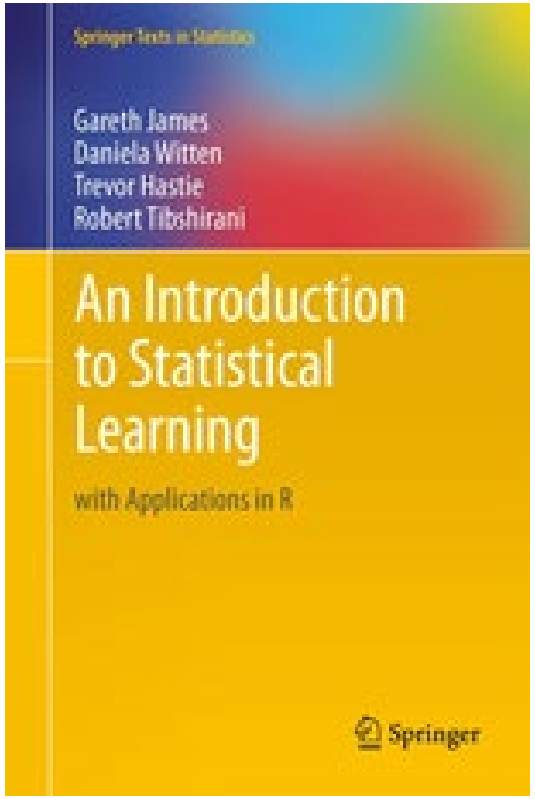


Image source: thatsoftwaredude.com

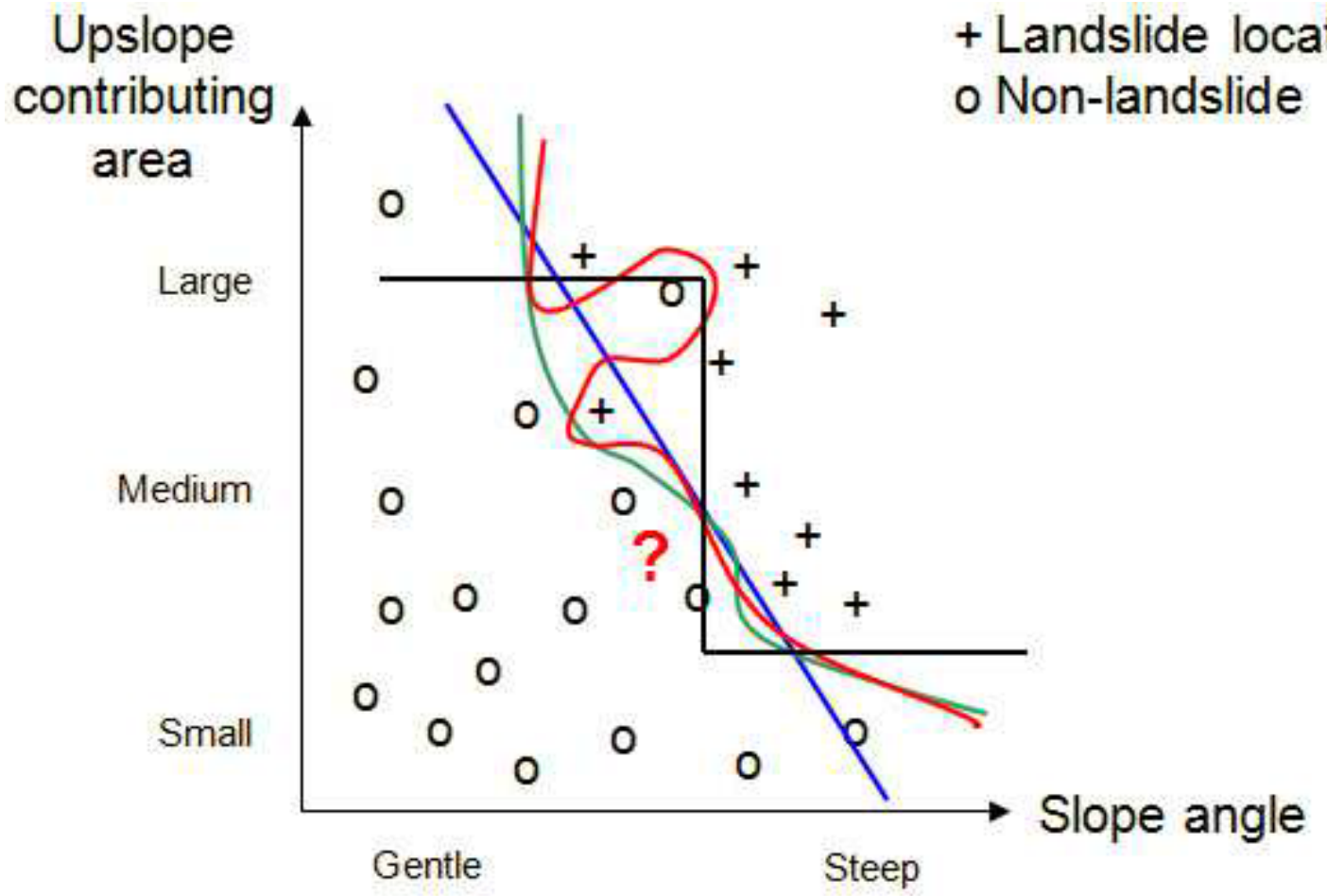








How Classifiers Work (Basically…)



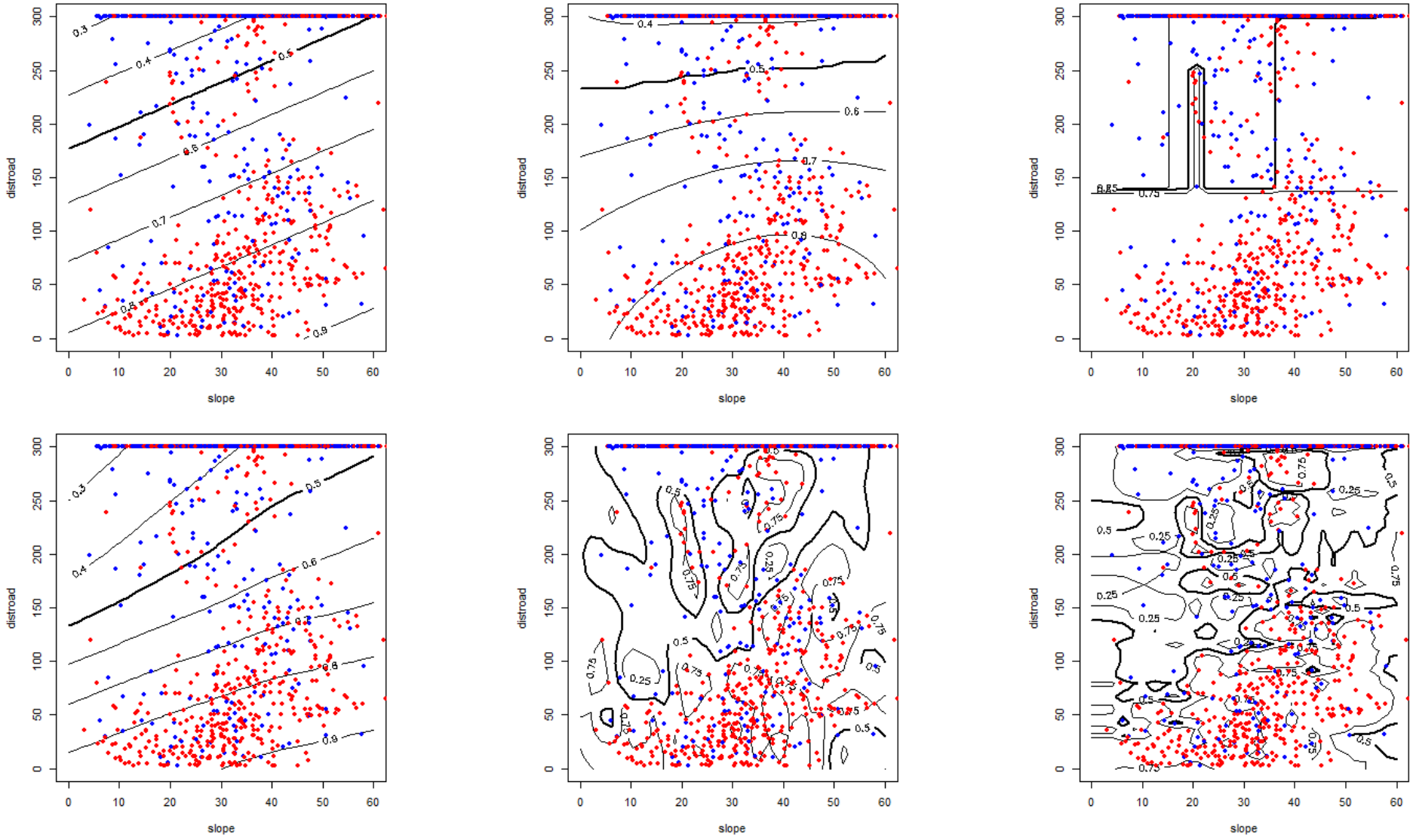
Model Predictions in Feature Space: Comparison

For illustration only:

Using only slope and distroad as predictors

Contours are lines of equal predicted "probability"

Points are landslide (red) and non-landslide (blue) observations



Words of Wisdom

All models are wrong, but some are useful

George E. P. Box (1919-2013)

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

A performance measure

An estimation procedure

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

- Ideally: random sampling

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

A performance measure

- An overall numerical measure of the goodness of our predictions
- E.g., in classification: overall accuracy, kappa coefficient, AUC, sensitivity, specificity, …
- In probability estimation: Brier score
- In regression: bias, precision, RMSE, …
- Performance measure must match the objectives of our analysis

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*.

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

- Ideally: random sampling

Training Set Approach

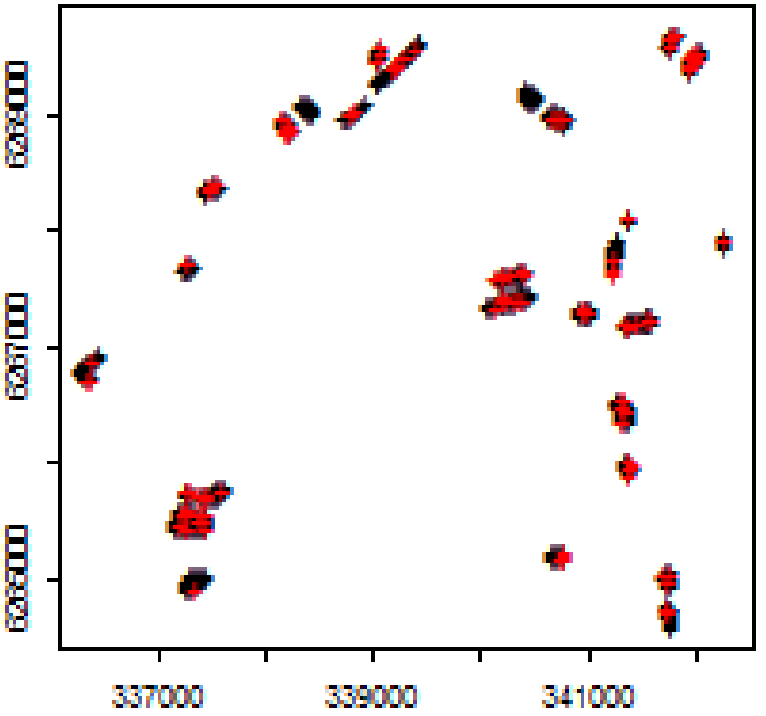
- The classifier‘s performance is assessed on the same data set on which it was trained.
- The resulting error rate is referred to as the **apparent error rate** or **resubstitution error rate**.

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

ibm.com

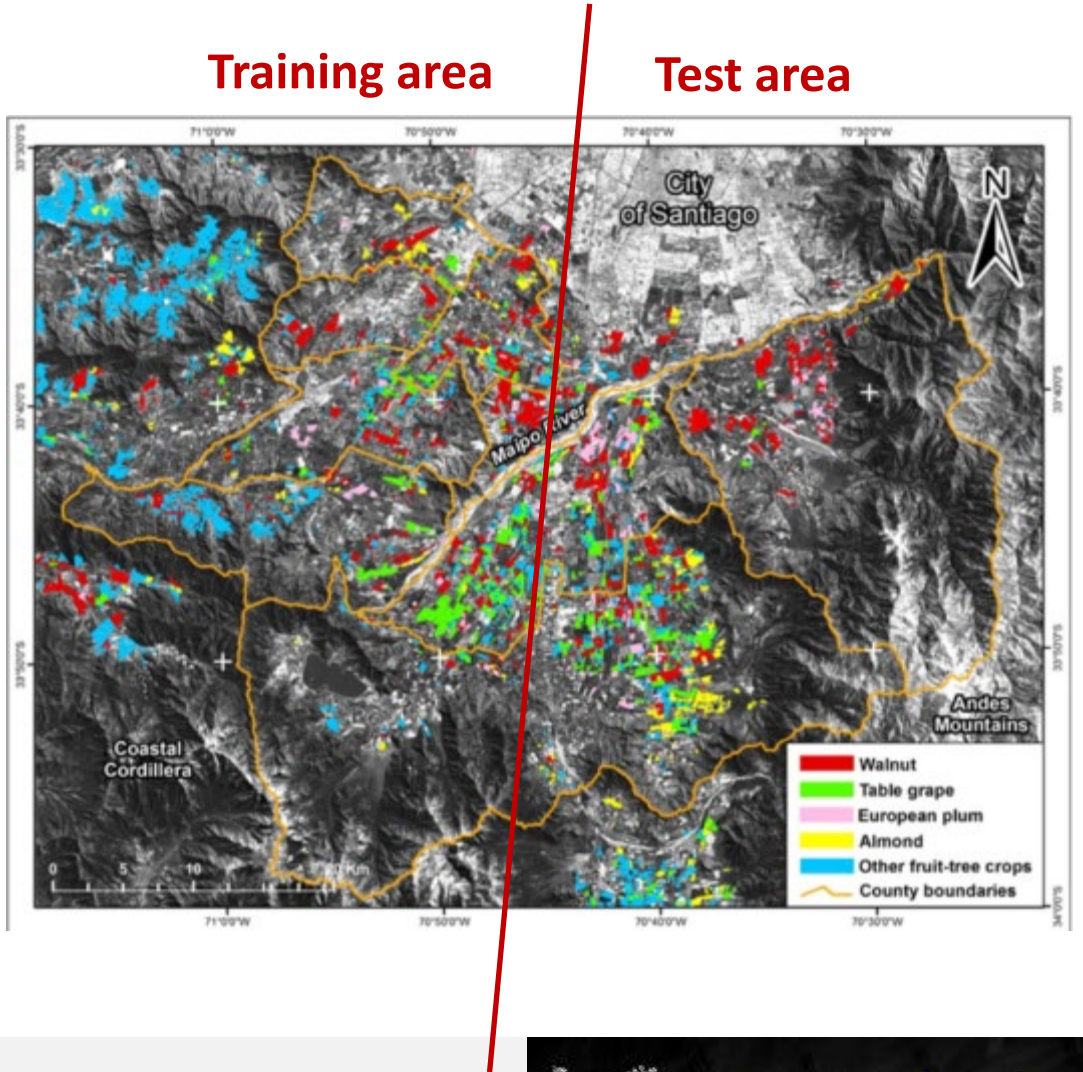
Test Set Approach and Design-Based Estimators

- Randomly split the data set into two disjoint sets: a training set and a test set, or hold-out set.
  - Or: Obtain a *new* random sample of data that is only used for performance estimation, not for model training → **design-based approach** using **statistical sampling theory**
- yields **unbiased** error estimates
  - (assuming that your data was a random sample)
- But there are no free lunches:
  - 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.
  - Does not measure ability to transfer model to new realizations of the random field

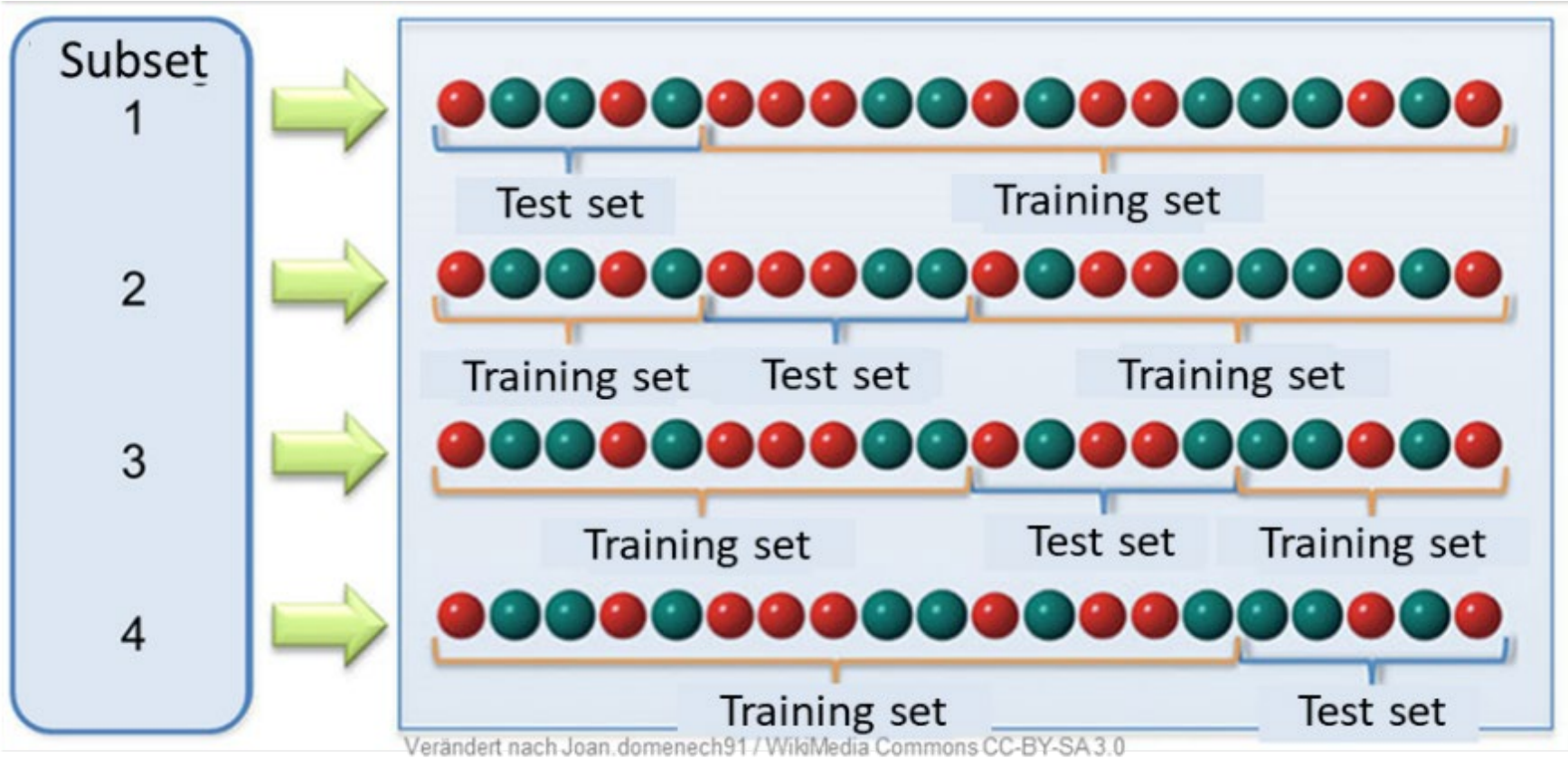


Test *Area* Approach

- 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

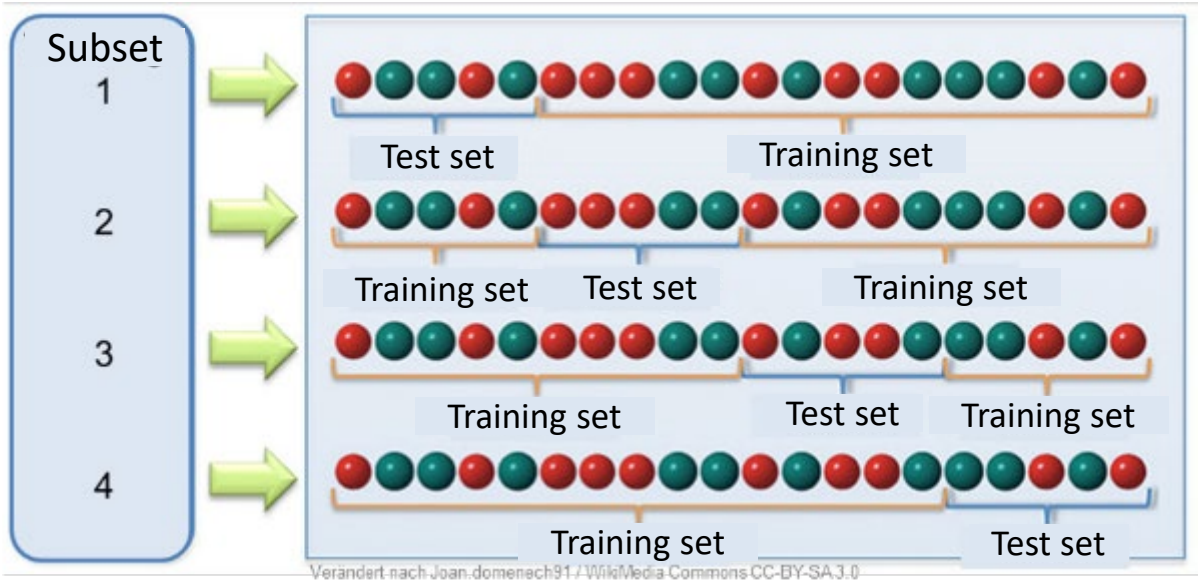


Verändert nach Joan.domenech91 / WikiMedia Commons CC-BY-SA 3.0

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.
- 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 $k = N$: **Leave-one-out cross-validation** (LOO-CV)



Applicable when the goal is

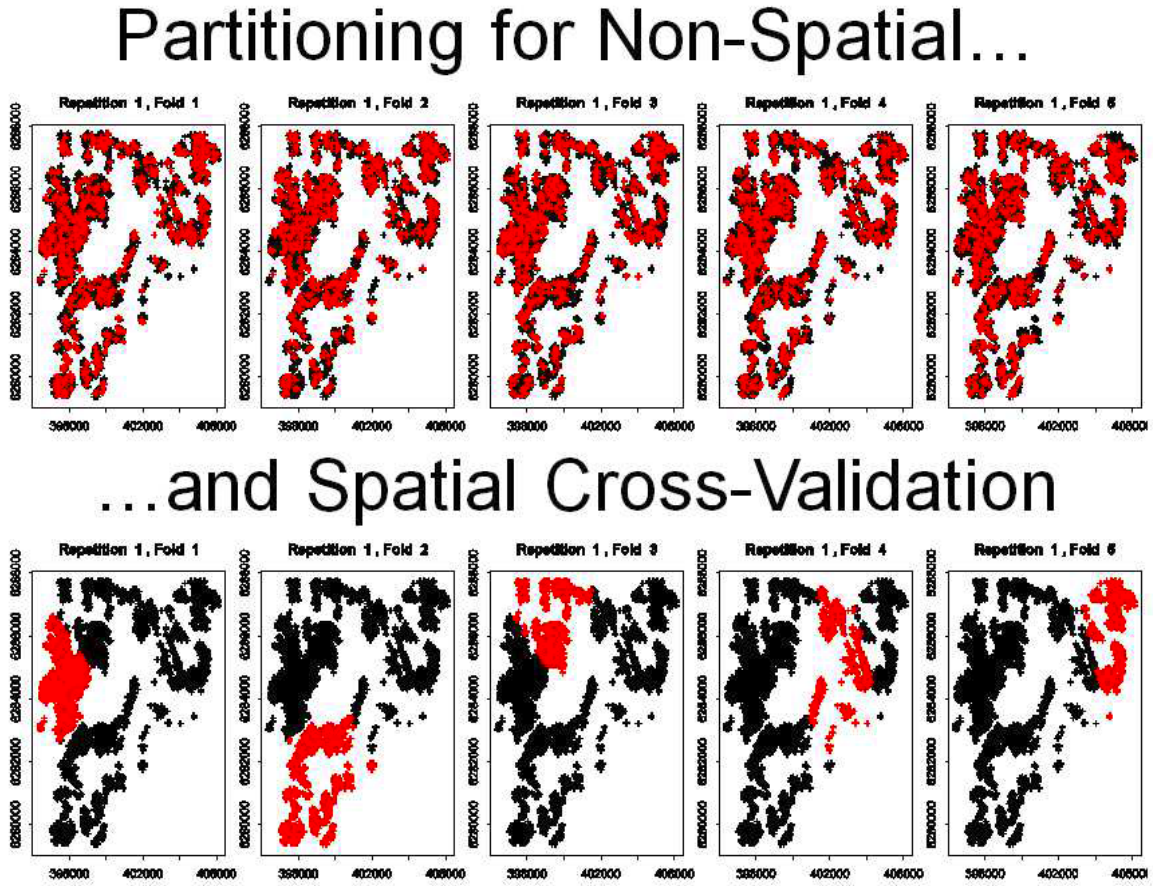
- to predict *within* an area,
- *not* to generalize to new realizations of the random field or to transfer the model to another area.

Spatial Cross-Validation

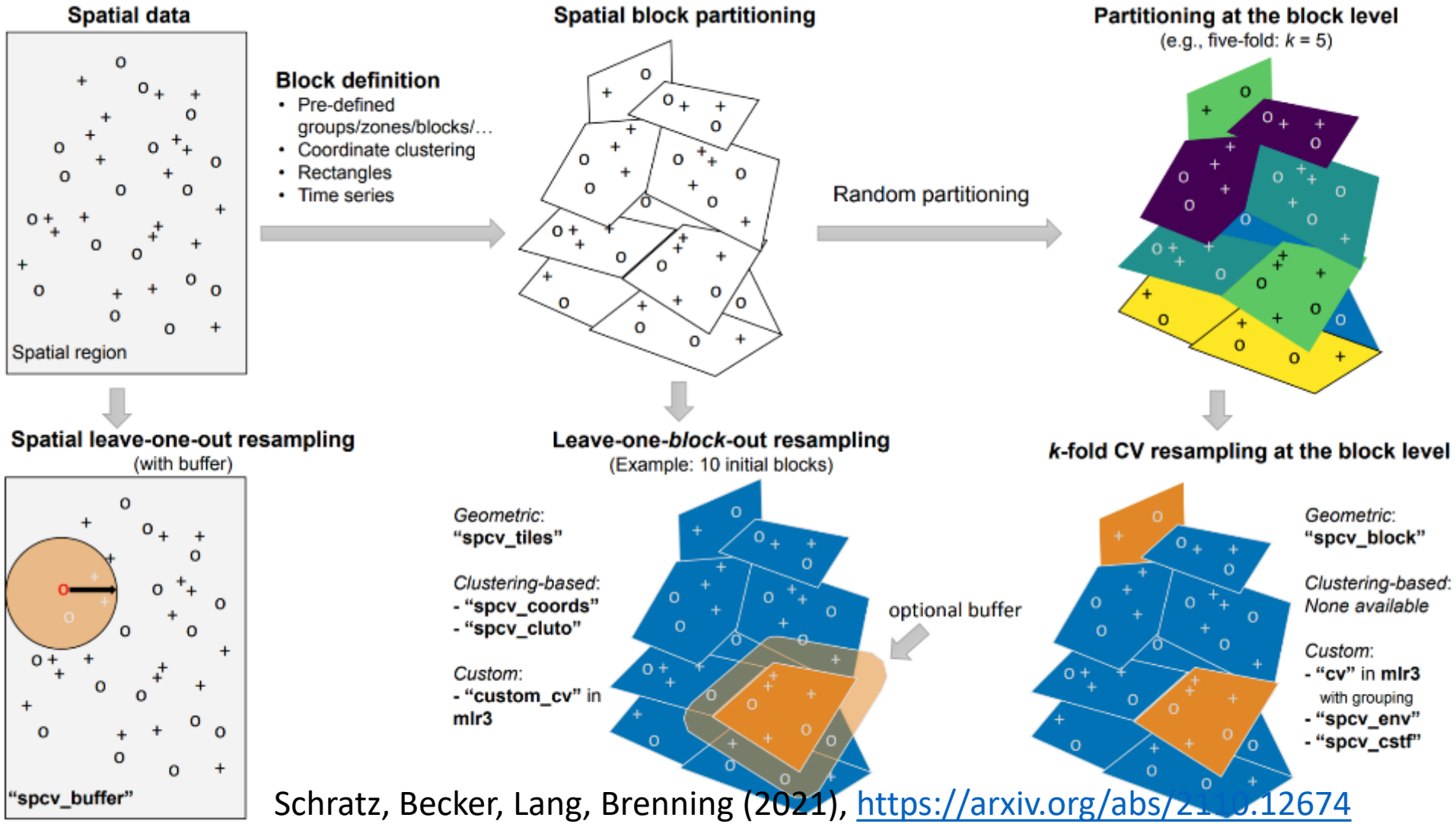
- Divide the study area into disjoint subregions (**blocks**)
  - E.g. using *k*-means clustering of coordinates (Ruß & Brenning, 2010)
- Or: use existing blocks, e.g. agricultural fields

Distinguish between:

- Leave *one* block out at a time (**leave-one-block-out CV**)
- Partition the the blocks, and leave partition out (**CV at the block level**)



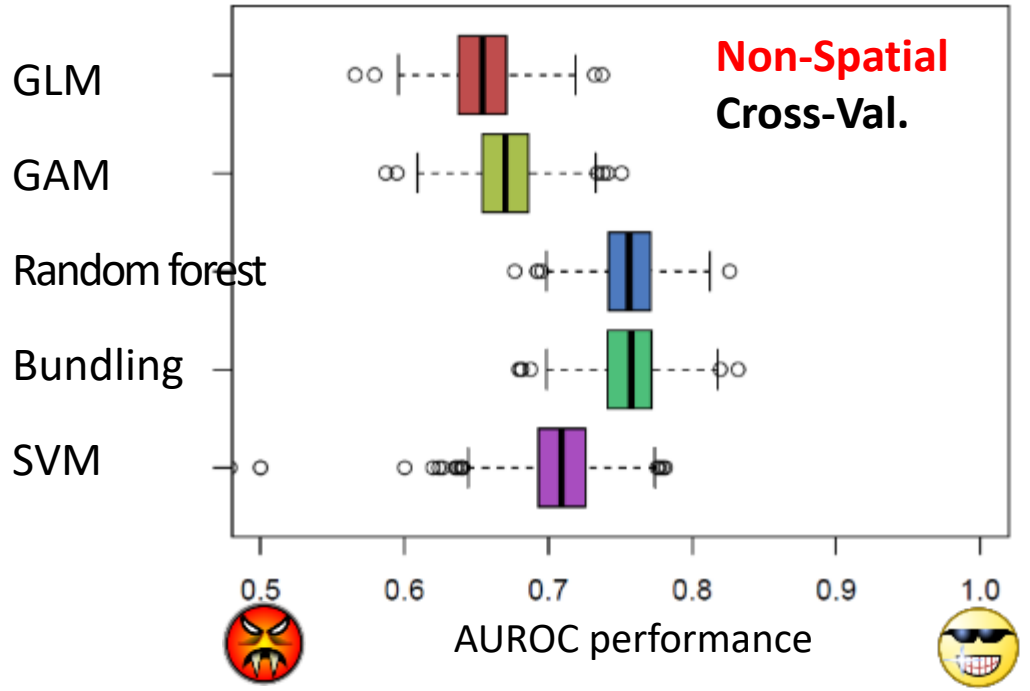
Overview of Spatial CV Strategies

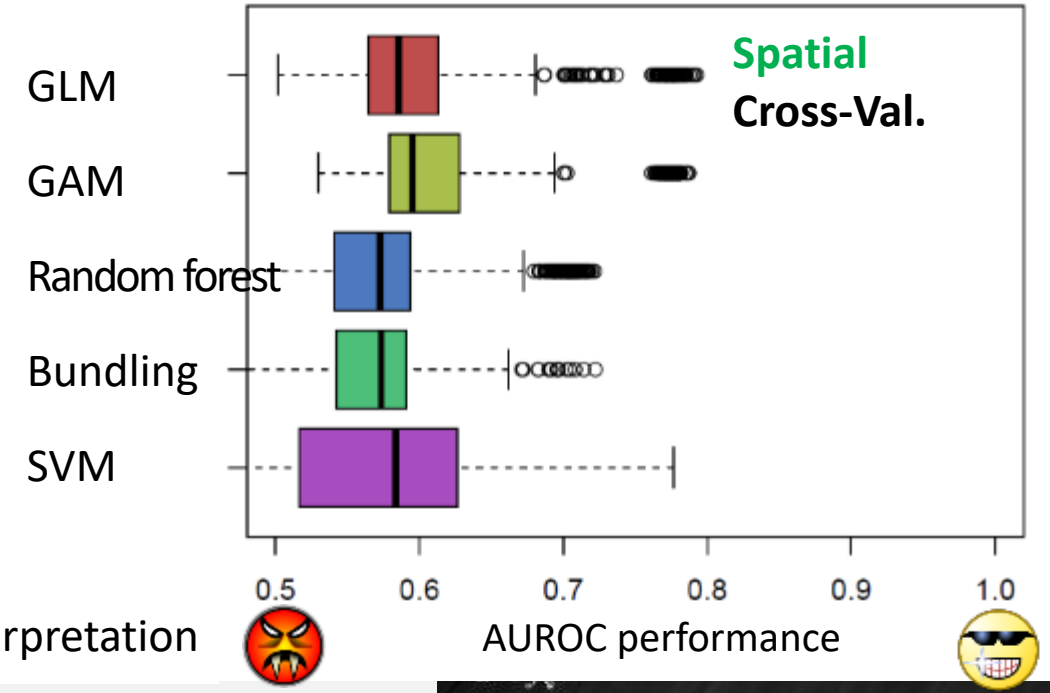


Schratz, Becker, Lang, Brenning (2021), https://arxiv.org/abs/2110.12674

Model Performance: Landslide Susceptibility

- Case study from Ecuadorian Andes
- Non-spatial CV results are over-optimistic
- Spatial CV reveals overfitting to training data
  - Here: leave-one-block out, using k-means clustering to define blocks
- Simpler ML models are more transferable, better able to generalize from training sample



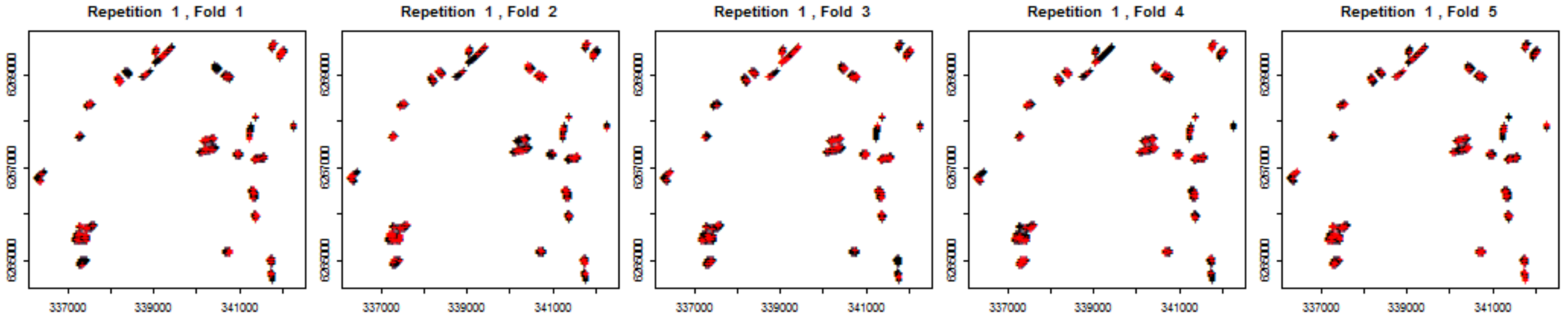


Compare Brenning (2005) in NHESS

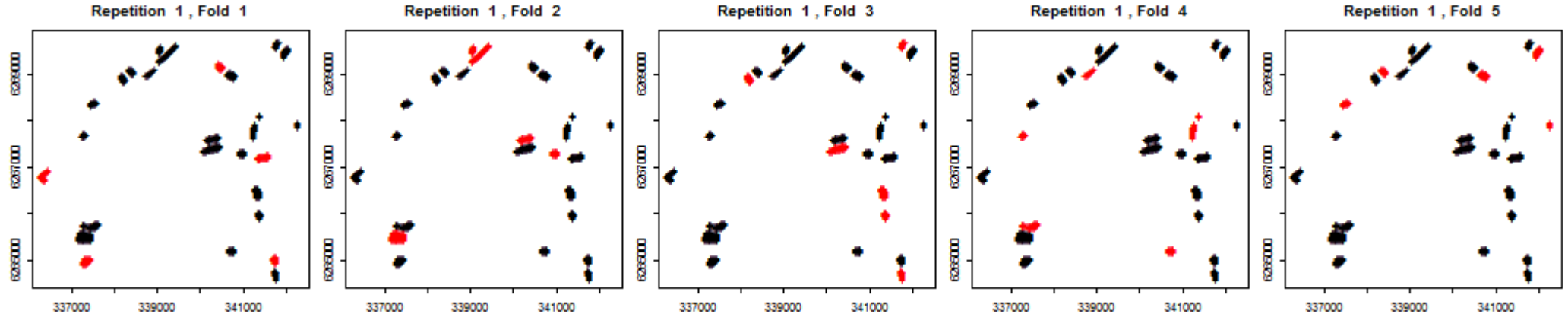
Cross-Validation at the Field Level:
Resampling Fields, not Grid Cells

Figures show a small subsample of the crop classification data

Ordinary (grid cells)

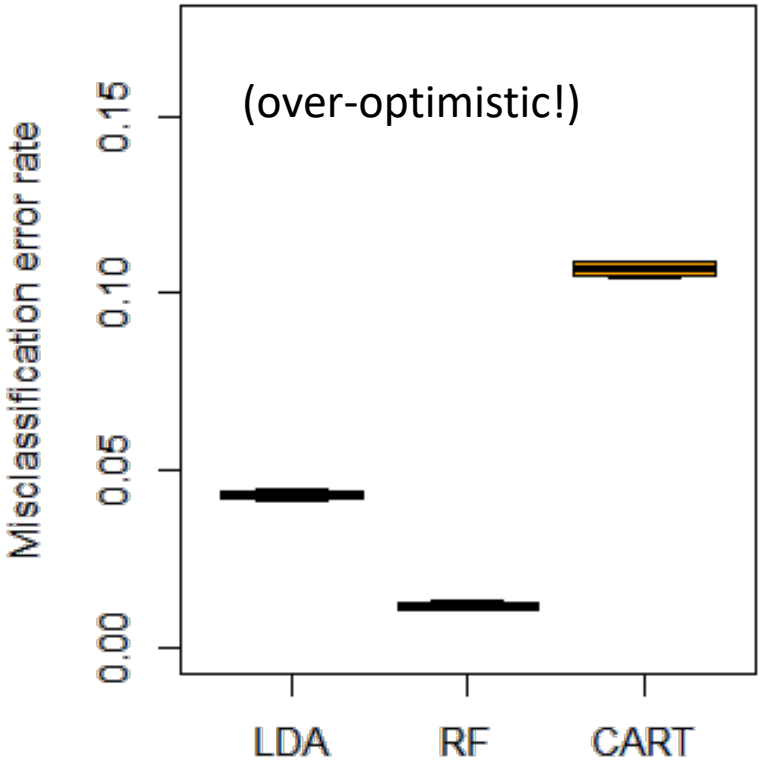


Spatial (fields)

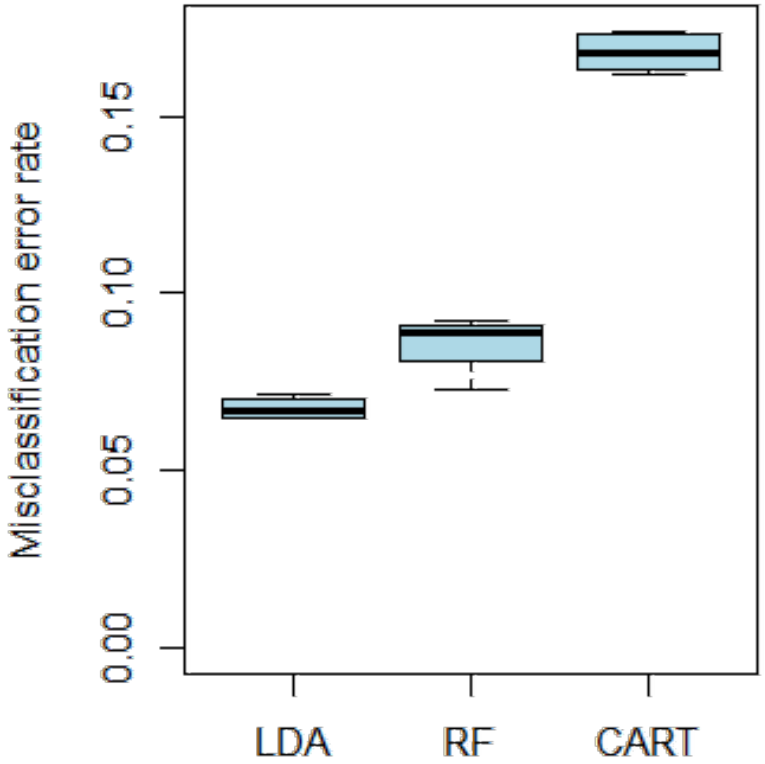


Example: Crop Classification
Misclassification Error Rate

Random CV
(resampling grid cells)



Field-level CV
(resampling fields)

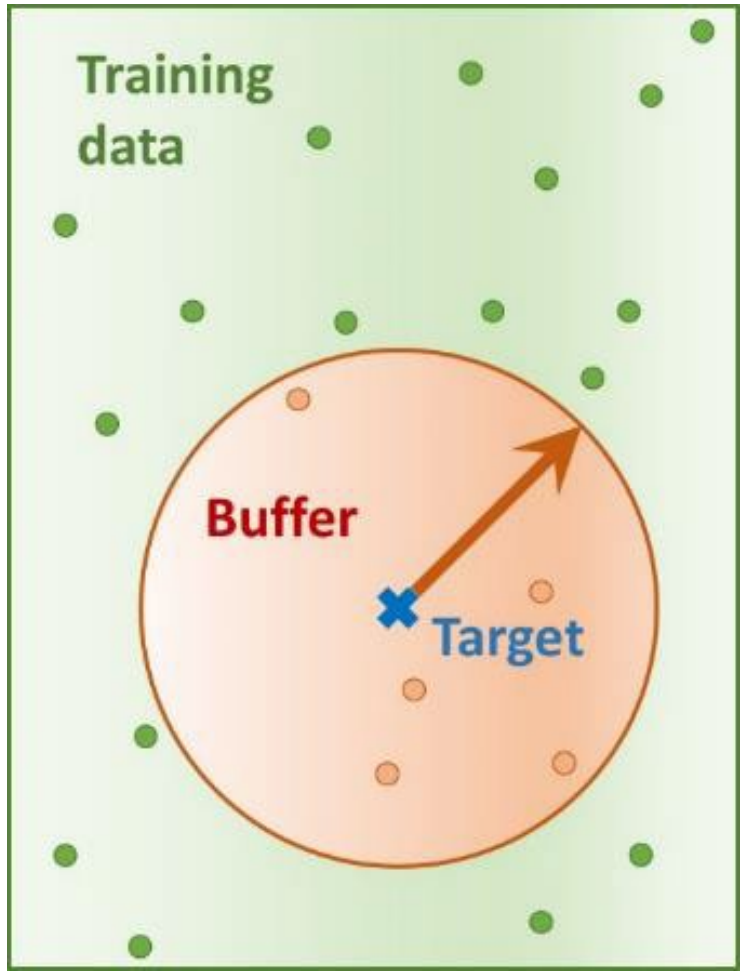


Example: Crop Classification
Misclassification Error Rate (MER) Estimates

Classifier	Apparent MER	Random CV MER	Field-level CV MER
LDA	0.044	0.043	**0.068**
CART	0.111	0.107	0.168
Random Forest	0.000	0.012	0.086

5-fold cross-validation
at the field level: pixels from same field will jointly be either in the training set or in the test set

A Distance-Based Approach

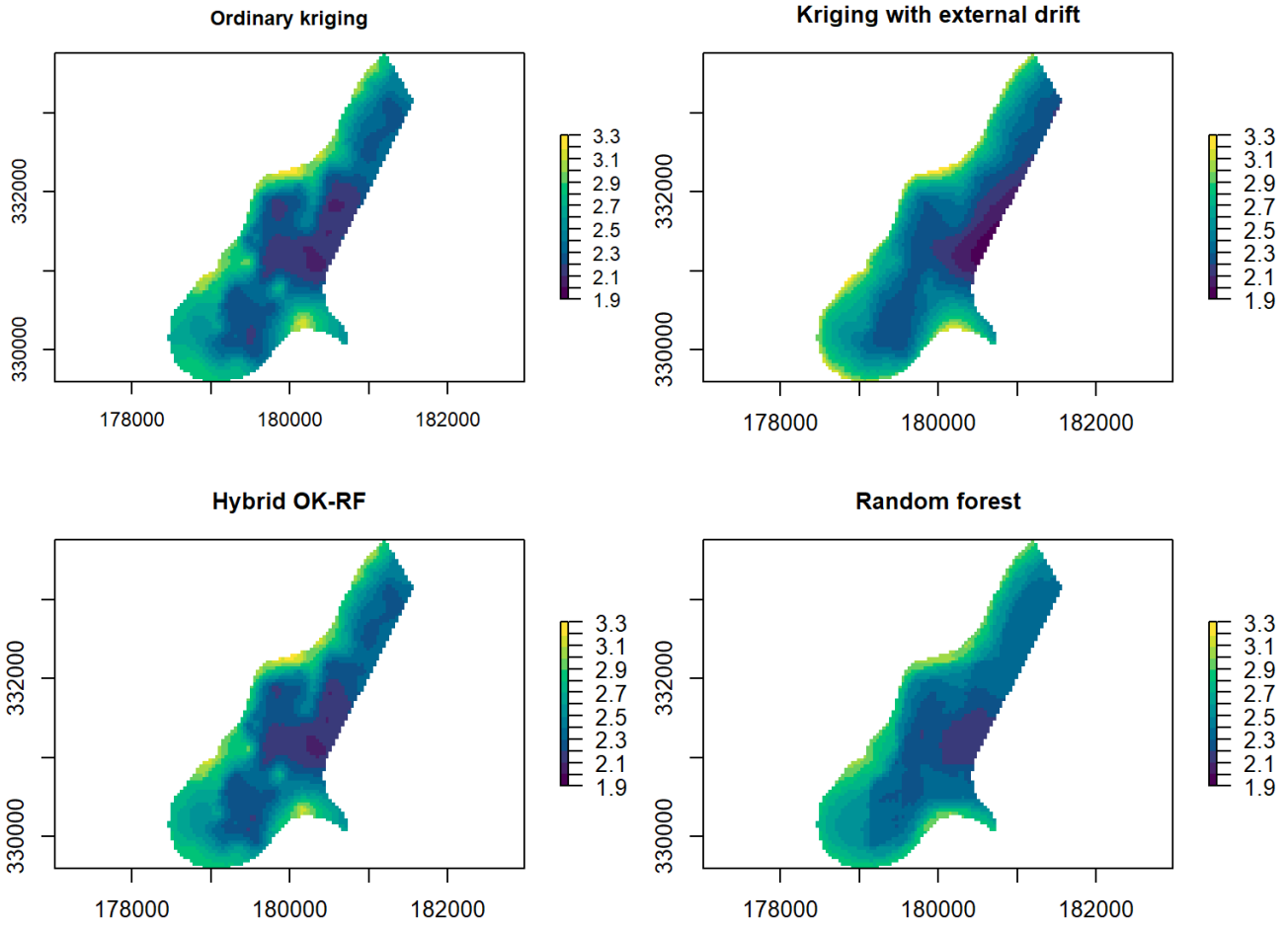


- Artifically create prediction situations with a desired prediction distance *r* by using LOO with a buffer.
- Error measures become a function of prediction distance!

→ **Spatial prediction error profile (SPEP)**

Brenning (2023) in *IJGIS*

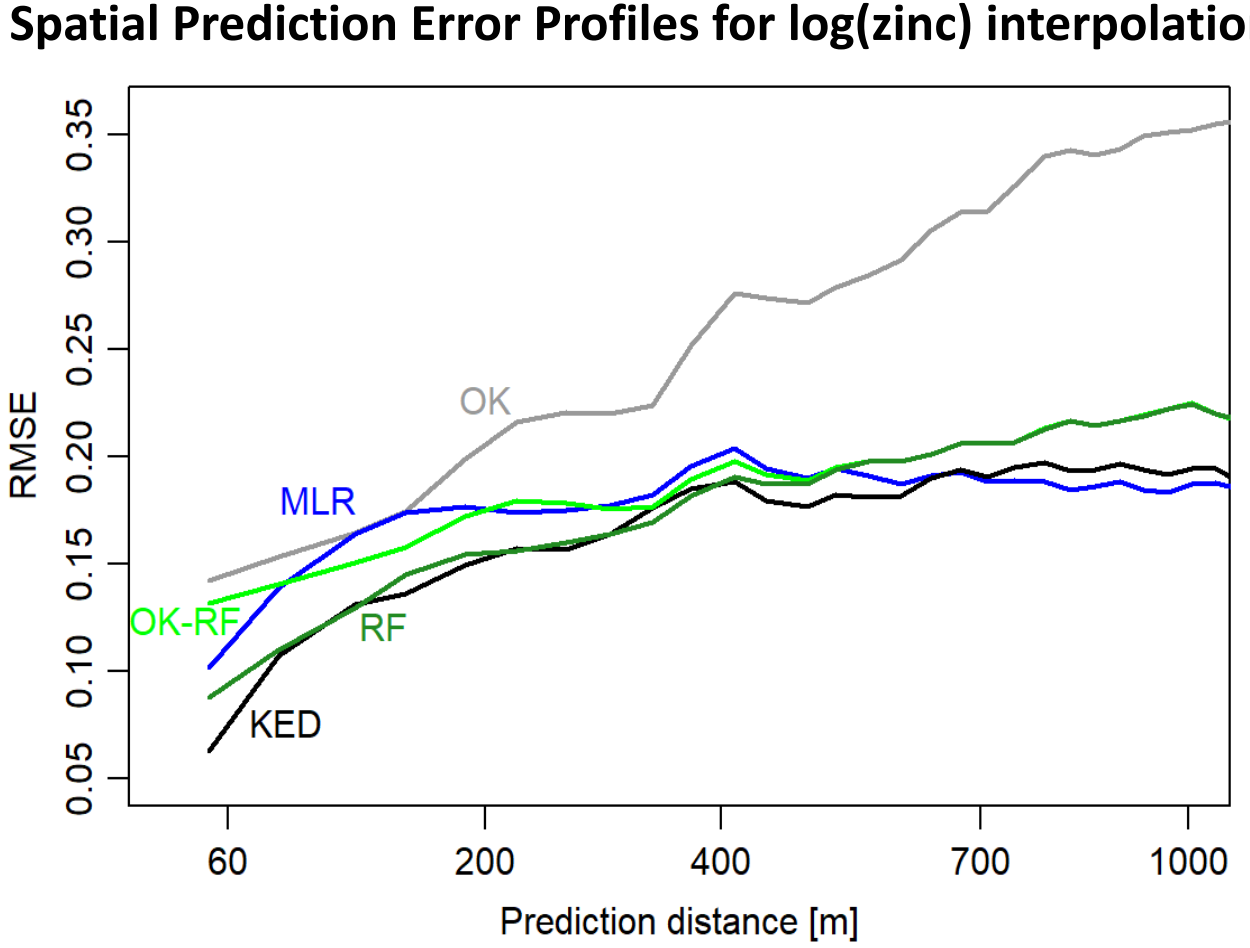
Example: Meuse dataset (log-zinc interpolation)



- 155 point measurements on a floodplain in NL

Brenning (2023) in *IJGIS*

Example: Meuse dataset (log-zinc interpolation)

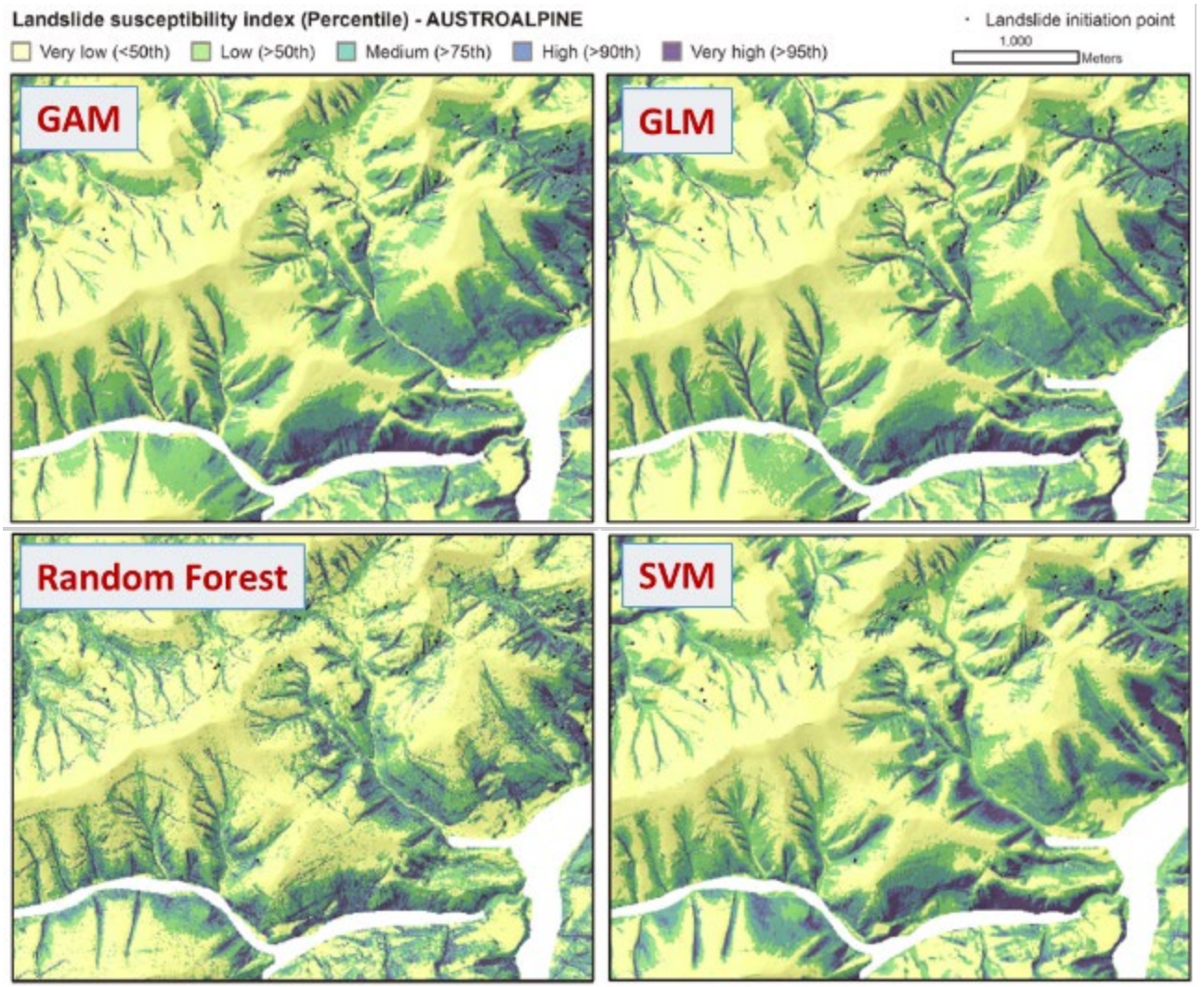


Brenning (2023) in *IJGIS*

Visual Comparison: Lower Austria

- Observe patterns created by different landslide susceptibility models!

→ Also consider qualitative aspects! (Steger *et al.*, 2015)



Goetz et al. (2015)

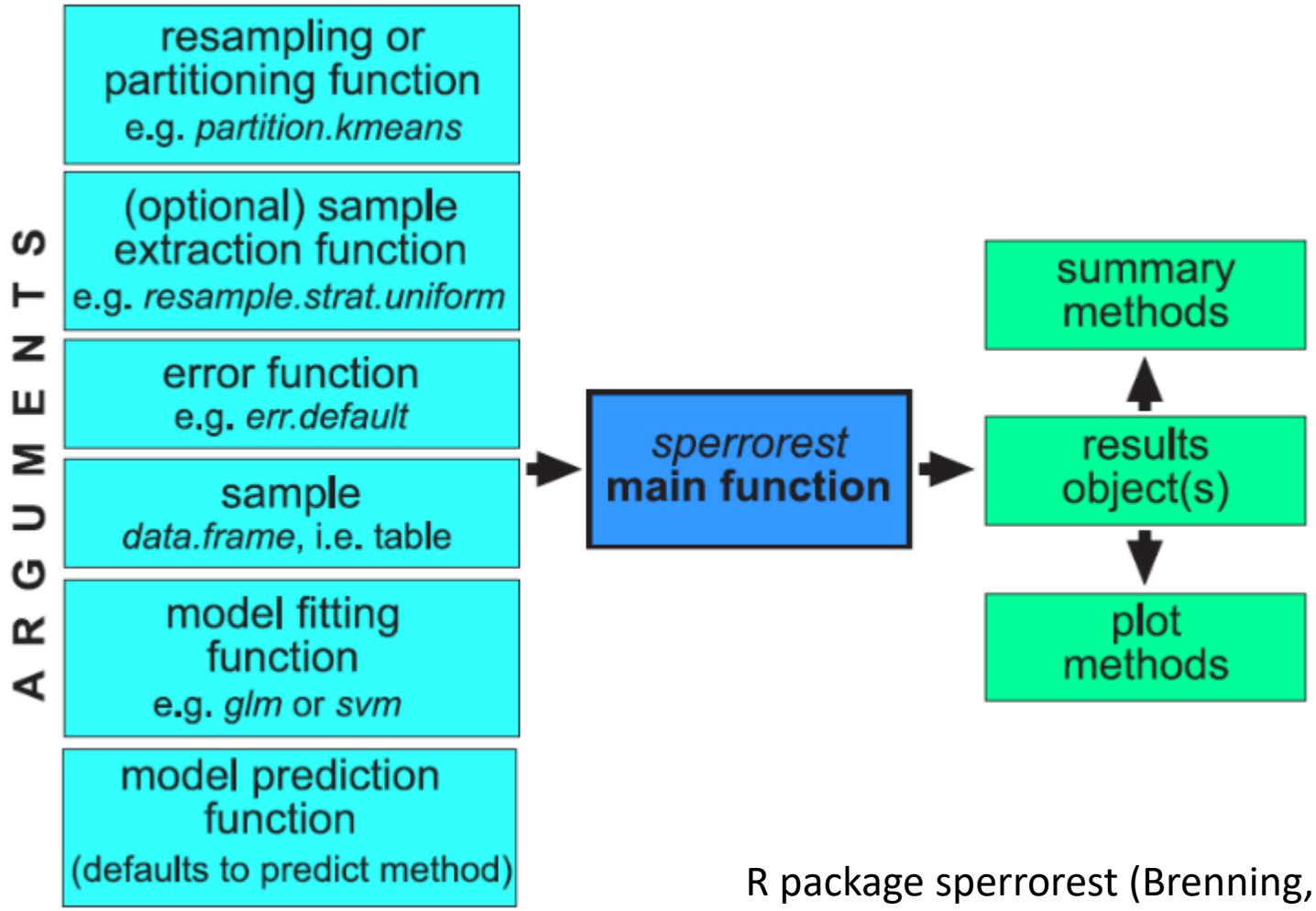
Lessons Learned

- In predictive modelling, we can be pragmatic about the type of model used – as long as it provides good predictions.
- CV helps us to reduce bias in model assessments.
- Spatial CV accounts for spatial dependence.
- Spatial prediction error profiles provide detailed insights into predictive behaviour.

What Can Go Wrong?

- The type of (re-) sampling used for model assessment must be consistent with the prediction task at hand.
  - E.g., range of prediction distances in spatial model application
  - E.g., forecasting vs. hindcasting for a given prediction horizon
- Never use the same test set for hyperparameter tuning and model assessment.
  - This will lead to an over-optimistic model assessment.
  - Use nested CV.
- Use appropriate error measures that match the prediction task.
  - Different types of misclassification can have different consequences (‘costs’).
- It’s not all about the numbers.
  - Consider qualitative aspects. (Which also brings us to model interpretation…)

(Spatial) Cross-Validation in R in *sperrorest*



For an overview of R packages and an intro to spatial CV using mlr3 and mlr3spatiotempcv: Schratz et al. (2022), https://arxiv.org/abs/2110.12674

R package sperrorest (Brenning, 2012)

Machine-Learning Model Interpretation

Alexander Brenning

Department of Geography, Friedrich Schiller University Jena

OpenGeoHub Summer School 2023, Poznan

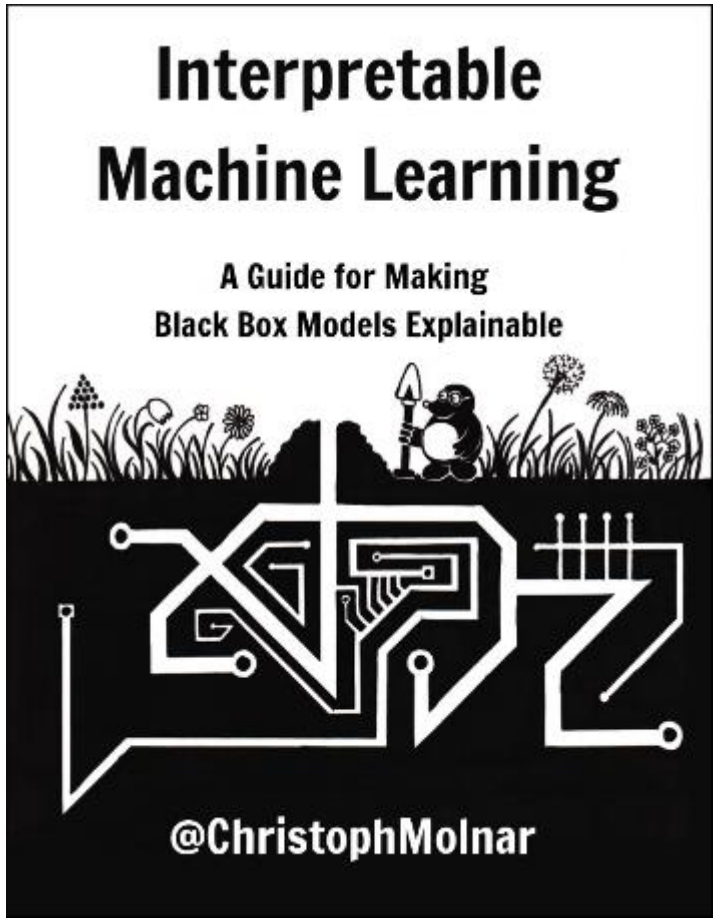
Interpretable Machine Learning

- What relationships are represented in my model?
- Which variables really contributed to the model, to its predictions?

- Intrinsically interpretable models versus black-box models
- Model-specific versus **model-agnostic** interpretation tools
- Feature summary statistics versus visualization
- Dataset-level versus **instance-level** (or local) interpretation

Biecek, Przemyslaw. 2018. *DALEX: Descriptive mAchine Learning Explanations*. https://pbiecek.github.io/DALEX/

For a comprehensive overview, please consult this book:



https://christophm.github.io/interpretable-ml-book/

Permutation-based Variable Importance (PVI)

- The PVI measures a predictor's overall contribution to the model's predictive skill.
- It is based on **how the performance measure changes when a predictor is permuted**.

Algorithm:

Input: Trained model.

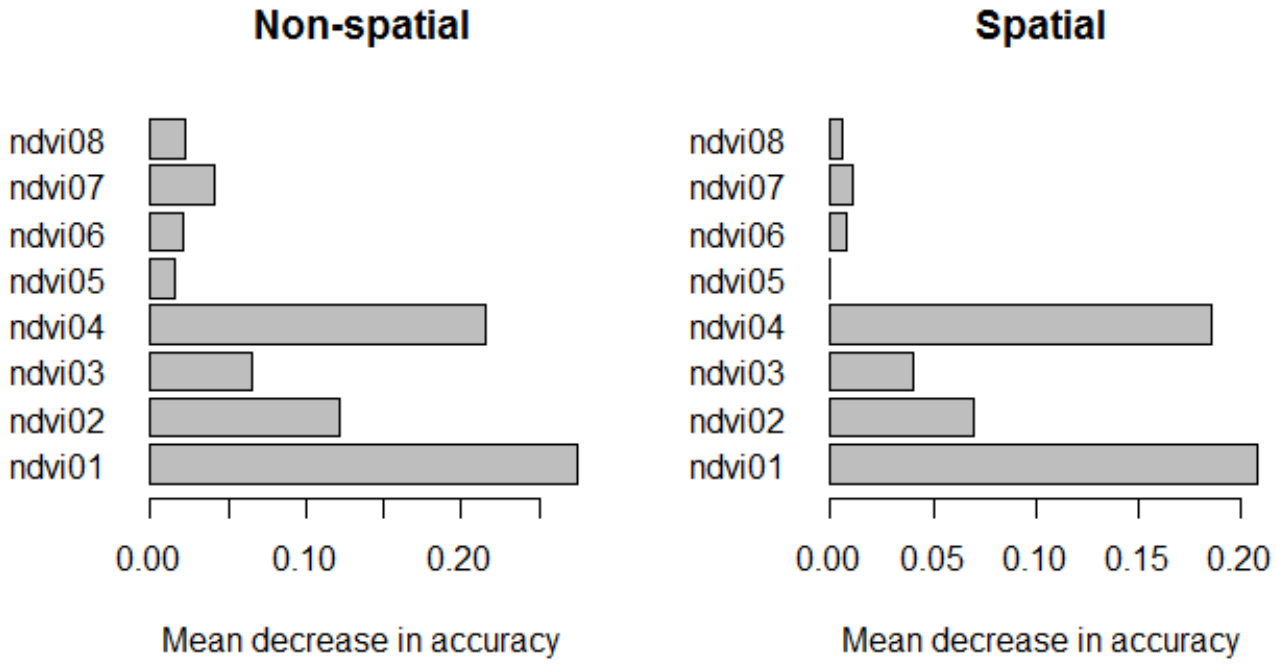
1. Assess its accuracy on the test set.
2. Permute a predictor on the test set, and use this partly "messed-up" data for prediction and accuracy assessment. Repeat this many times, using different random permutations of the predictor.
3. Calculate the mean difference between "regular" and "messed-up" accuracy.

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

Spatial PVI

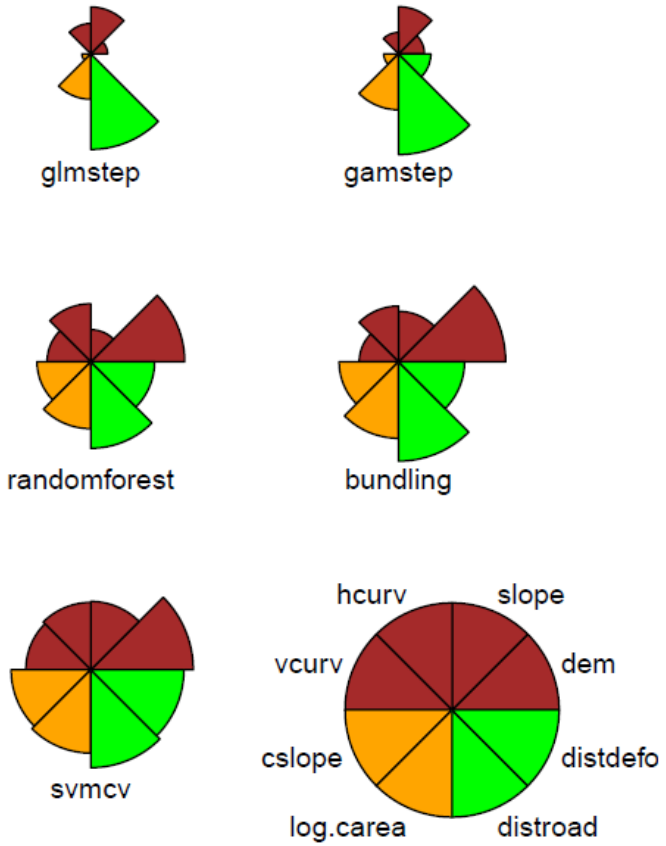
- "Standard" PVI ignores spatial dependence or grouping as well as the prediction horizon in the model's application
- Embed PVI assessment within a spatial CV to assess a variable's ability to contribute to *generalizable* or *transferable* predictive capabilities.
  - `sperrorest` package
- **Spatial variable importance profiles** extend this concept by creating distance-related estimates using spatial LOO-CV

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

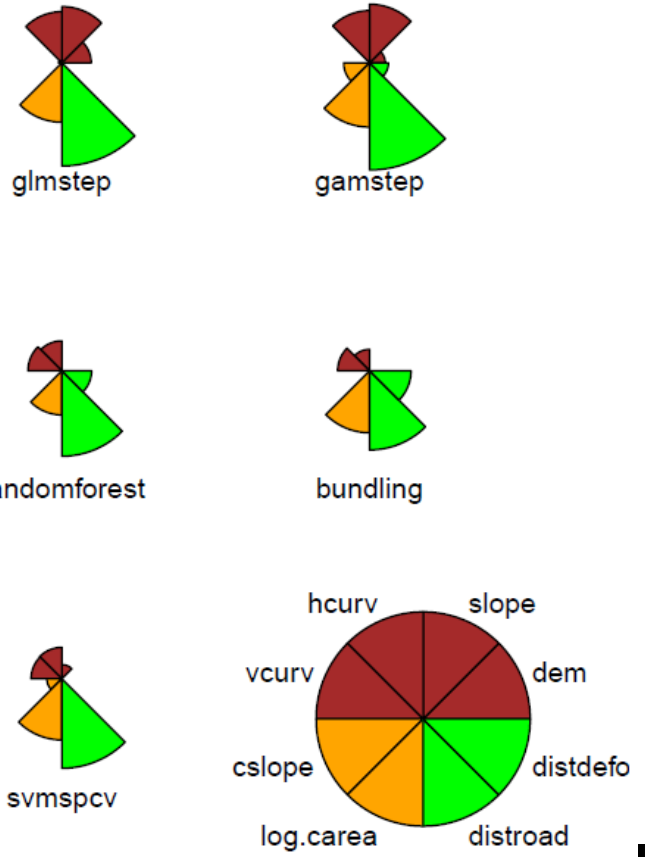


Example: Landslides in Ecuador

Non-spatial AUROC importance



Spatial AUROC importance

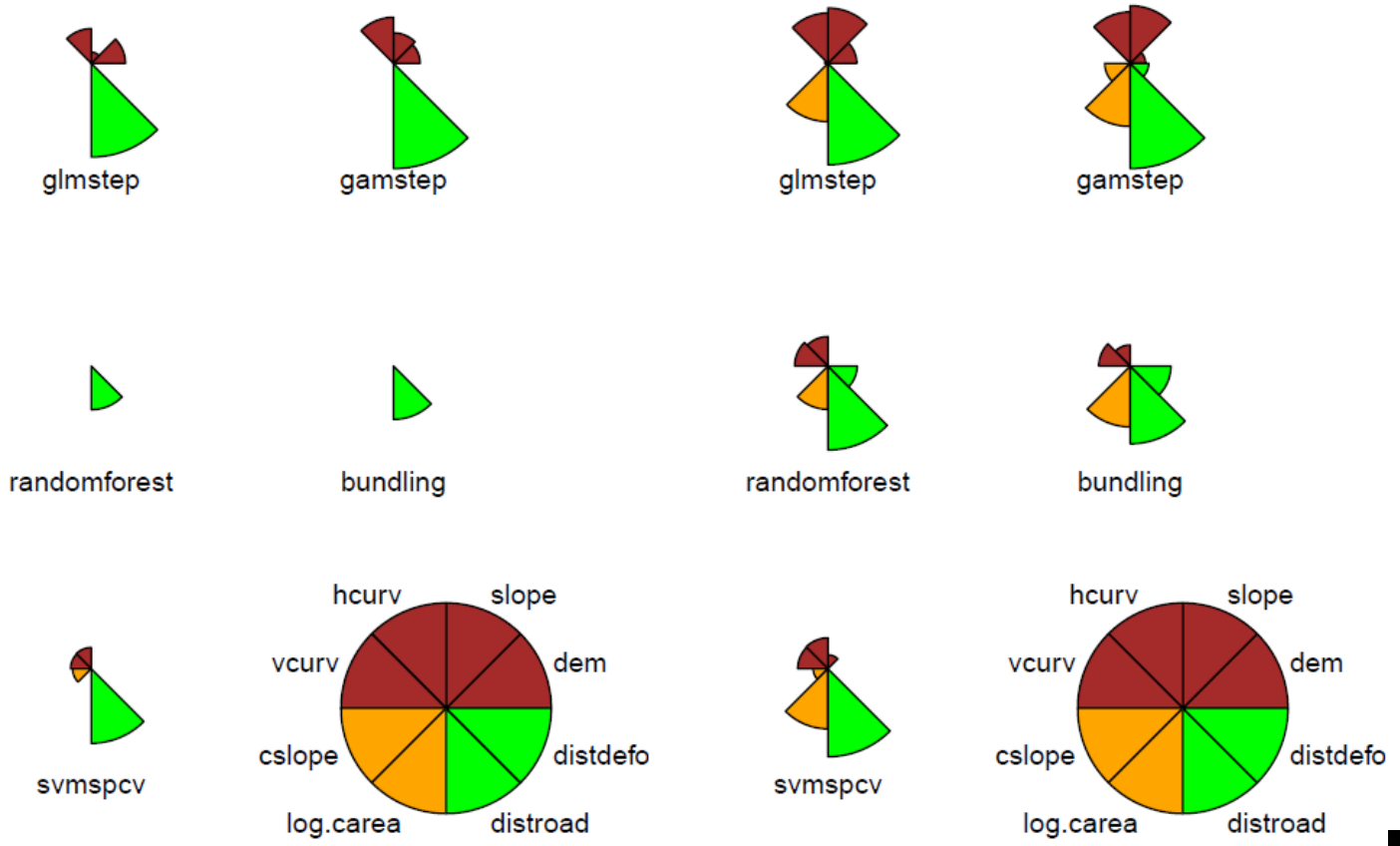


Example: Landslides in Ecuador

Variable importance also varies with the performance criterion used!

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

Non-spatial TPR90 importance



Spatial TPR90 importance

PVI Criticism

Mock Example: Crop Classification

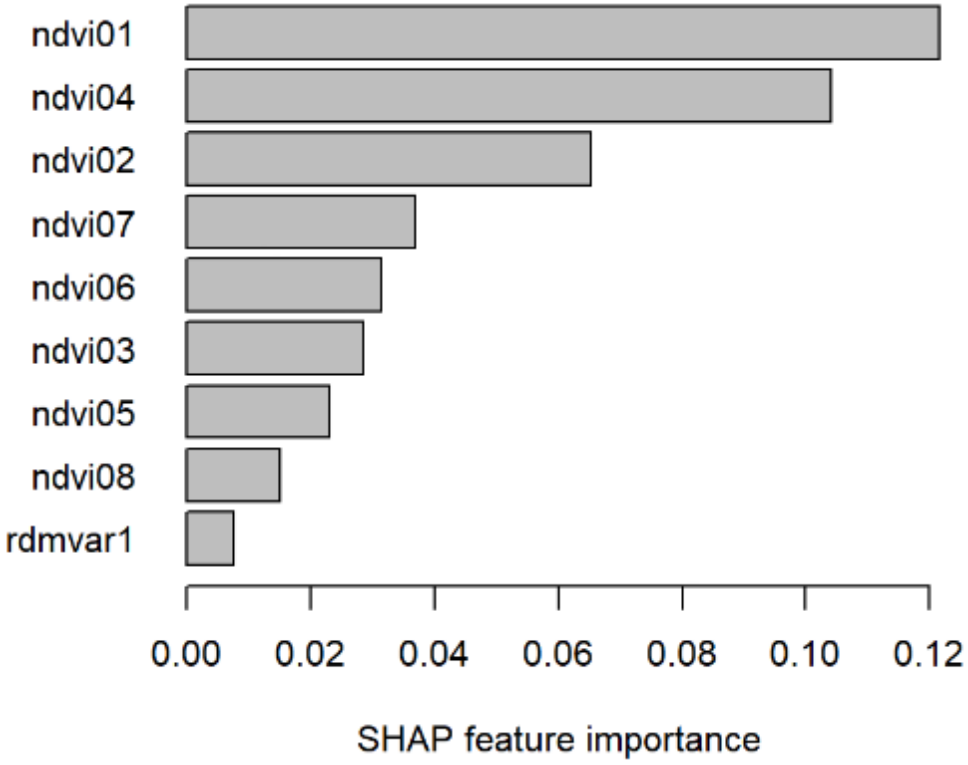
- A model uses, among many other variables, **NDVI on June 1st** and **NDVI on June 8th** as predictors of crop type.
- These two variables are strongly correlated.
- The model is therefore trained on a sample that does *not* include pixels with "brown" June 1st and "green" June 8th. Such pixels don't exist.
- However, the permutation *generates* instances with **(permuted) "brown" June 1st and (correct) "green" June 8th**.
- Predictions are made for this data in order to estimate the decrease in accuracy. This is our basic ingredient in estimating the PVI of NDVI on June 8th!

- In a PVI assessment, we use our model to make predictions for combinations of predictor values that do not exist in our training sample.
  - This is extrapolation. Extrapolation is usually bad.
- Even worse, permuted predictor values may result in implausible combinations.
  - E.g., pixels that suddenly turn green
  - Or pregnant men…
- Also, when features are correlated, their PVI is spread across these features.

- *Aren't there better ways to assess variable importance?*

SHAP Feature Importance

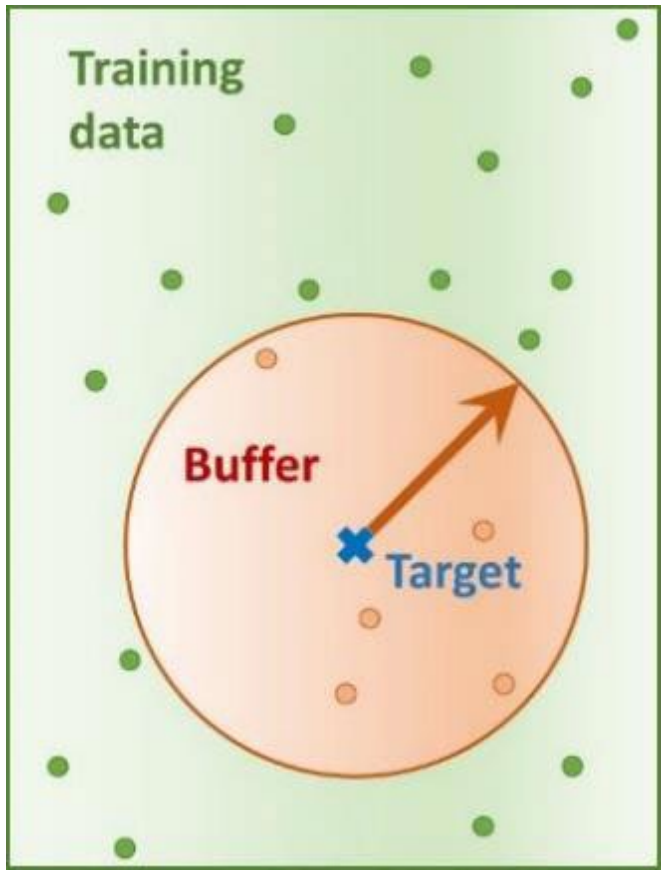
- Shapley values represent the contribution of each feature to the prediction based on game theory.
- How should the „payout“ fairly be distributed among features, or groups of features?
- SHAP feature importance is mean of all absolute Shapley values.
- Criticism: Samples from marginal distribution; ignores dependence among features.
- Model-based versions: KernelSHAP, TreeSHAP

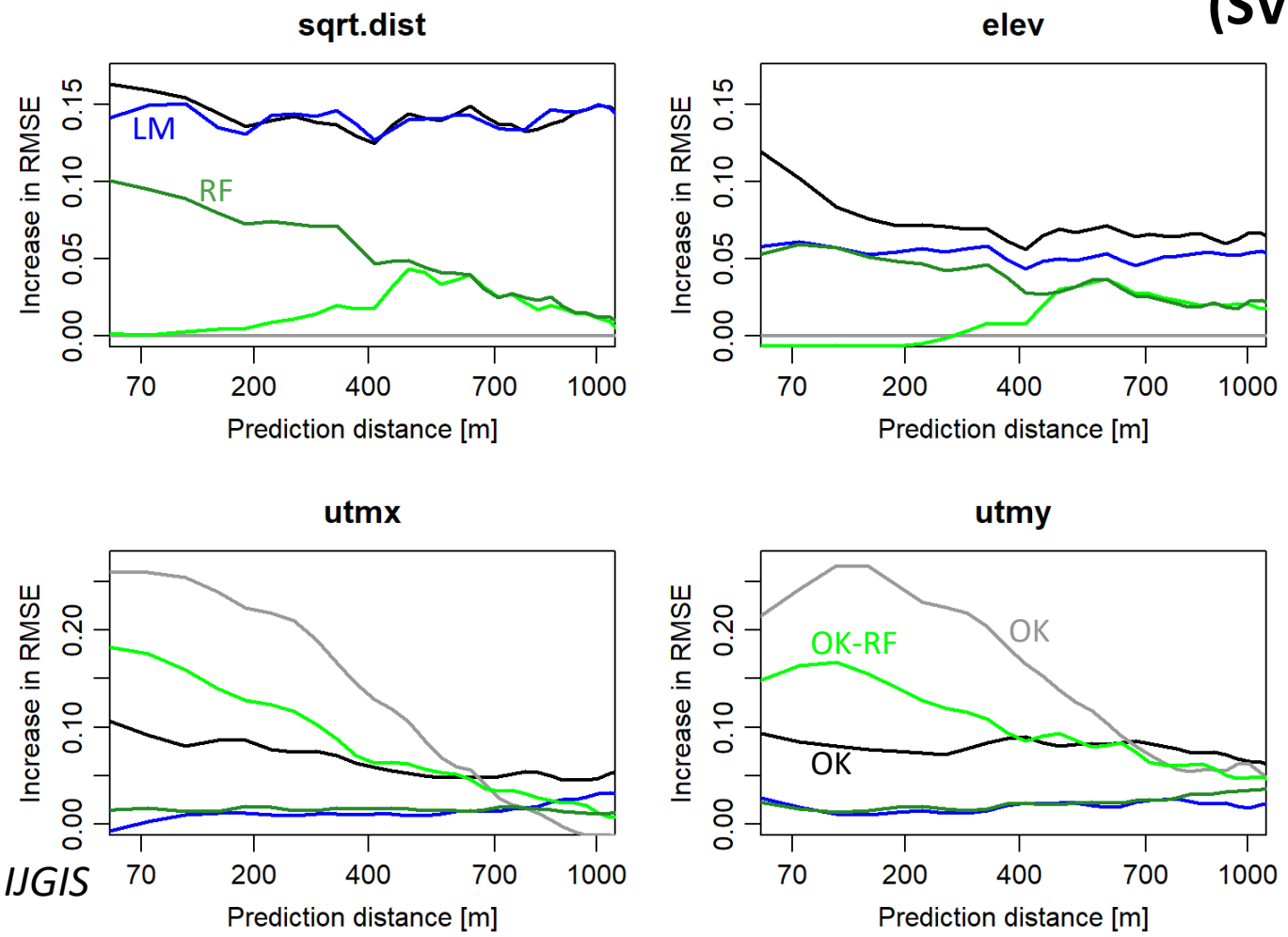


A Distance-Based Approach

Meuse data: log-zinc interpolation

Spatial variable importance profiles (SVIPs)





Brenning (2023) in *IJGIS*

Partial Dependence Plot (PDP)

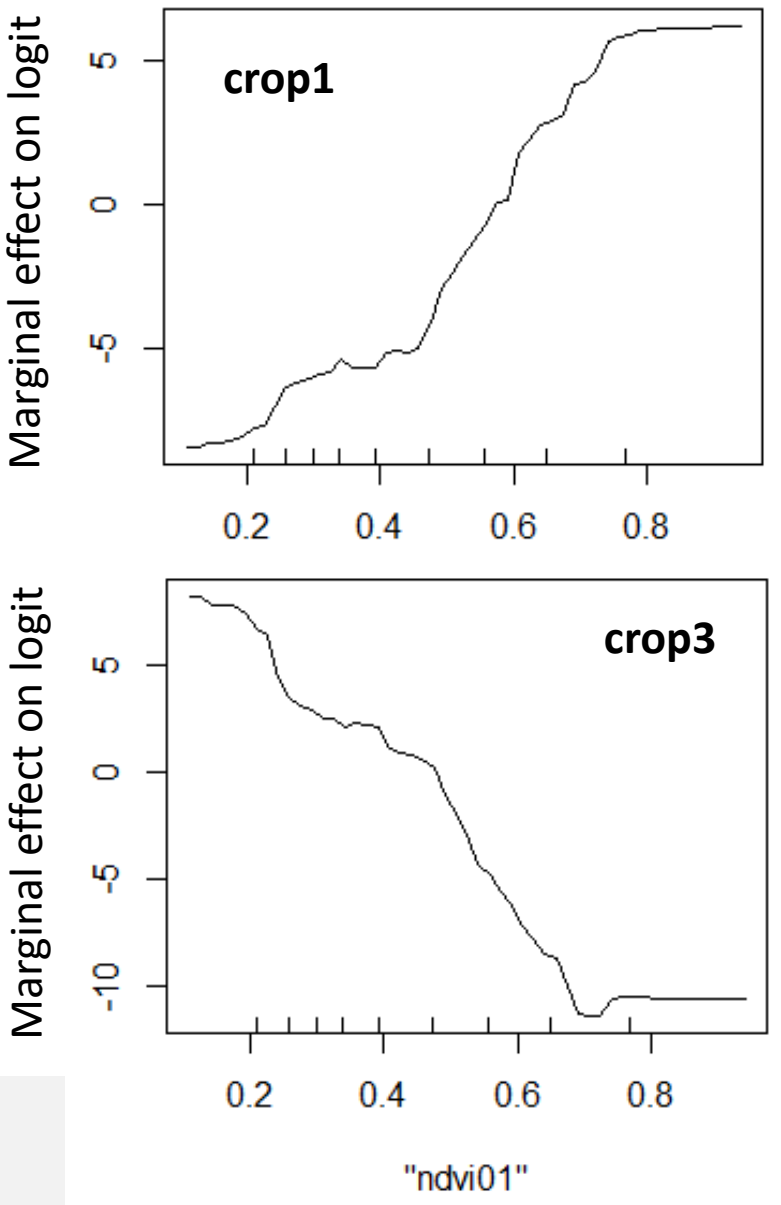
- In GAM modelling, we were able to plot the transformation functions. Can we create similar plots for more complex models?
- **Partial dependence function**:

$$\hat{f}_{x_S}(x_S) = \frac{1}{n}\sum_{i=1}^{n}\hat{f}(x_S, x_C^{(i)})$$

where $\hat{f}$ is the fitted model, x_s the selected variable, and x_c all other predictors. The sum is over all observations.

- In R: function `partialPlot` in package `randomForest`, or package `pdp` more generally

Crop Classification: Random Forest PDPs



PDP Criticism

- PDPs visualize the marginal effect of no more than two predictors simultaneously.
  - All (other) interactions are hidden.
- PDPs, like PVI, assume that the predictors are independent.
- PDPs, unlike PVI, are calculated on the training sample.

Accumulated Local Effects (ALE) Plot

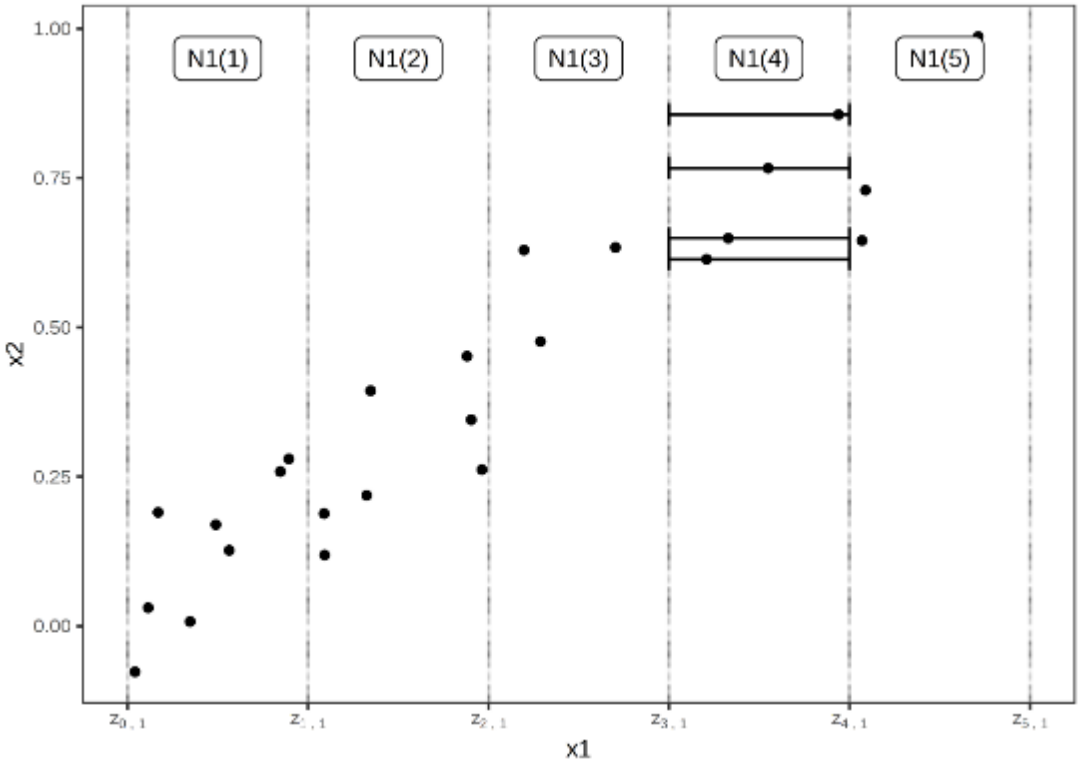
- The ALE plot looks at how the model predictions *change* in a small window of the predictor.
- It averages only over observations *in* that moving window, not over all observations.
- This solves two issues of the PDP and PVI:
  - Nonsensical, or highly unlikely, instances are avoided.
  - The effects of correlated predictors are separated.

$$\hat{f}_{j,ALE}(x) = \sum_{k=1}^{k_j(x)} \frac{1}{n_j(k)} \sum_{i:x_j^{(i)} \in N_j(k)} \left[f(z_{k,j}, x_{\setminus j}^{(i)}) - f(z_{k-1,j}, x_{\setminus j}^{(i)}) \right]$$

accumulate over all intervals, up to the one corresponding to x_j

sum over all features within the **local** neighborhood in the variable of interest, x_j

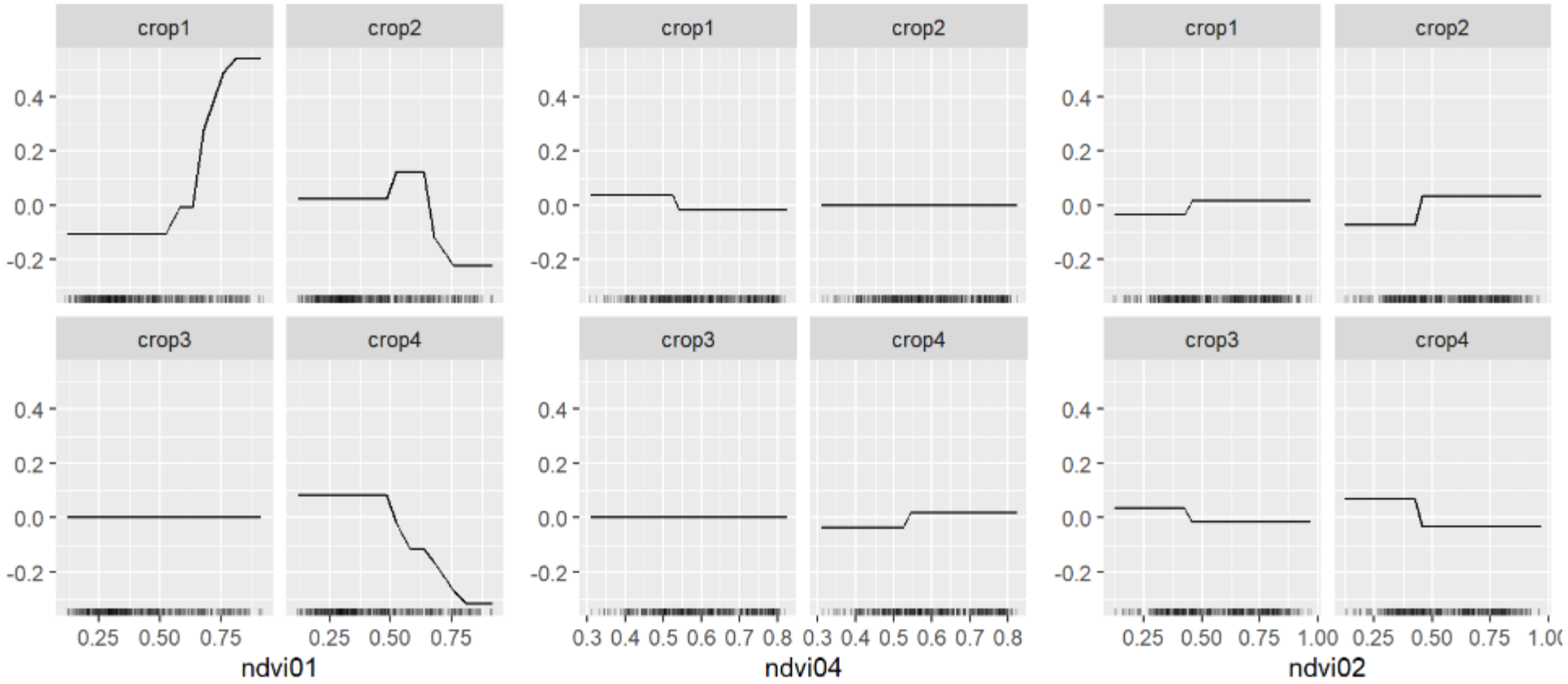
differences in predictions for different predictor values ("**effects**")



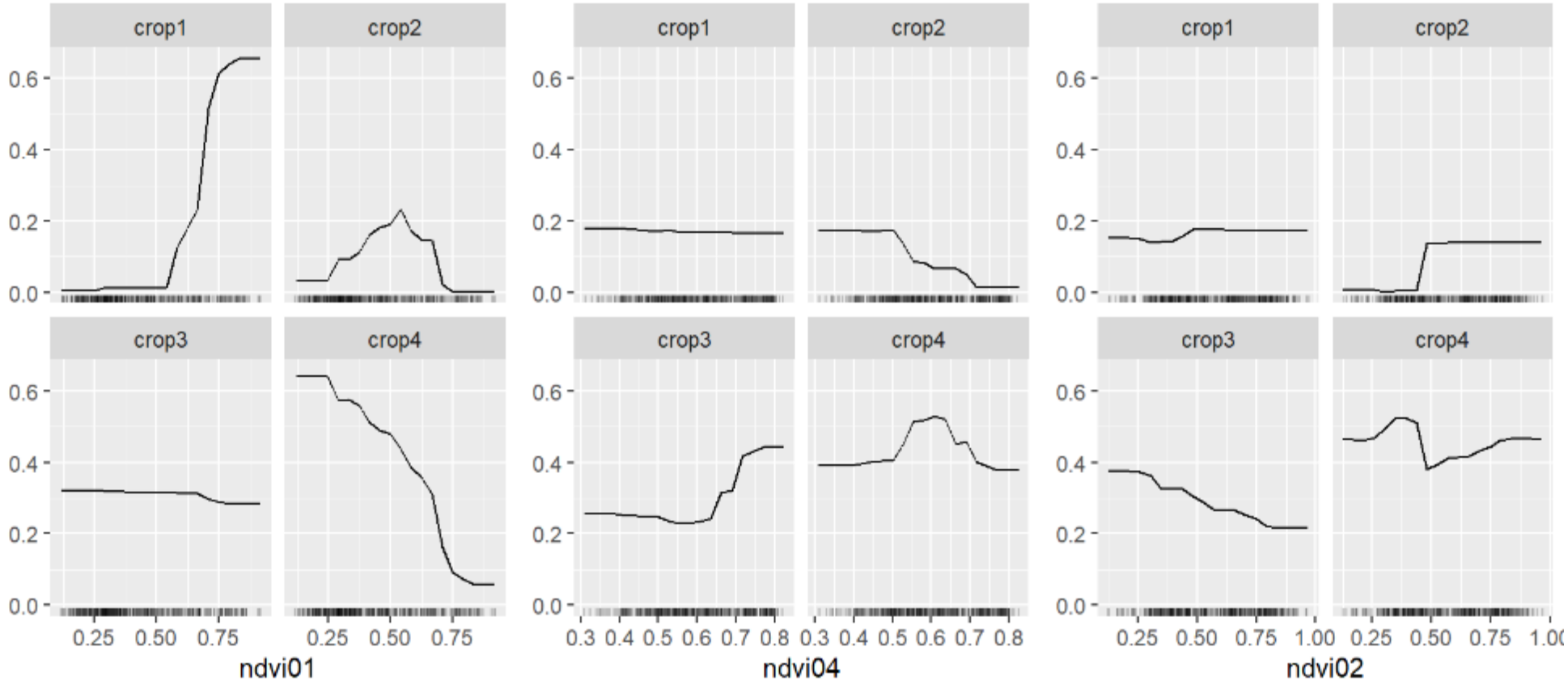
https://christophm.github.io/interpretable-ml-book/ale.html#ale

In R: packages **ALEPlot** , **iml** and **DALEX**

Case study: Crop classification
ALE plots



Case study: Crop classification
PD plots for comparison



ALE Plot Criticism

- ALEPs are calculated on the training sample.
- ALEPs can be highly variable, depending on the size of the local neighbourhood.
- ALEPs are less intuitive than PDPs.

Model Interpretation in High Dimensions

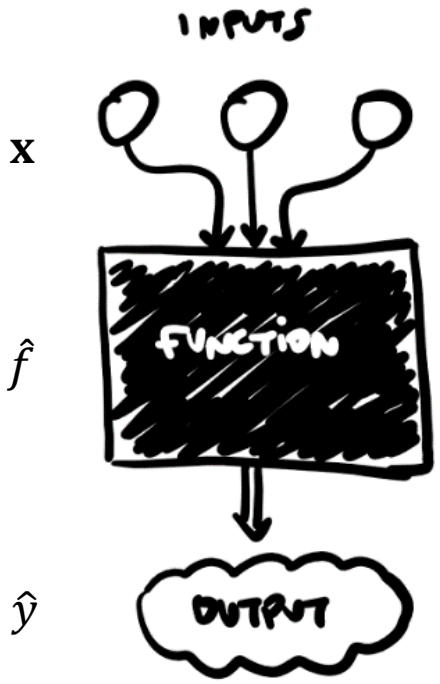
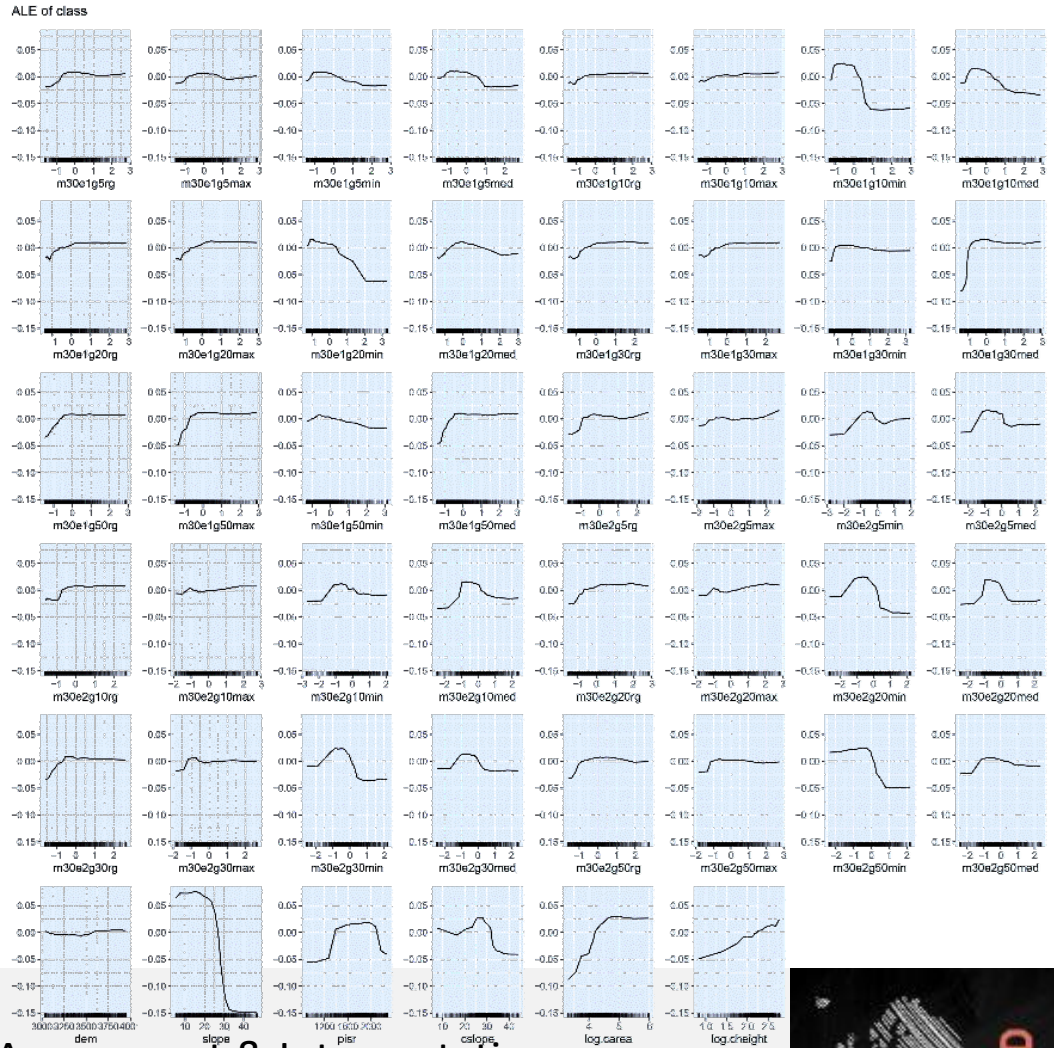


Image source: thatsoftwaredude.com



Model Interpretation in High Dimensions

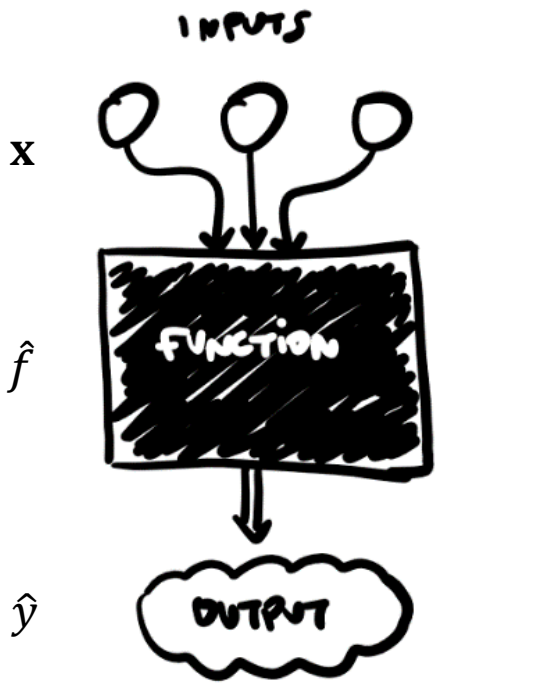
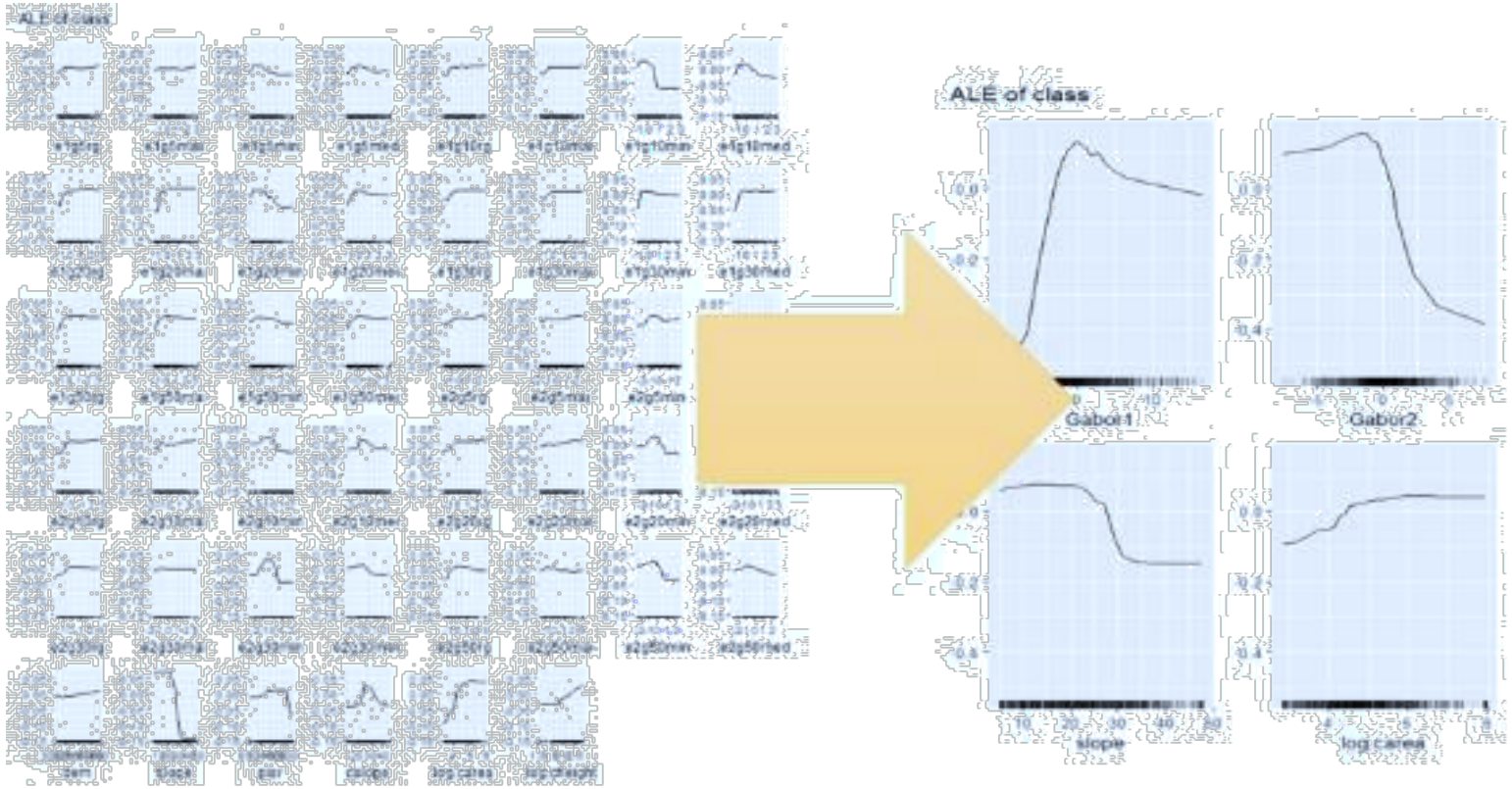


Image source: thatsoftwaredude.com



High-dimensional feature space: *Interpretation in transformed space*

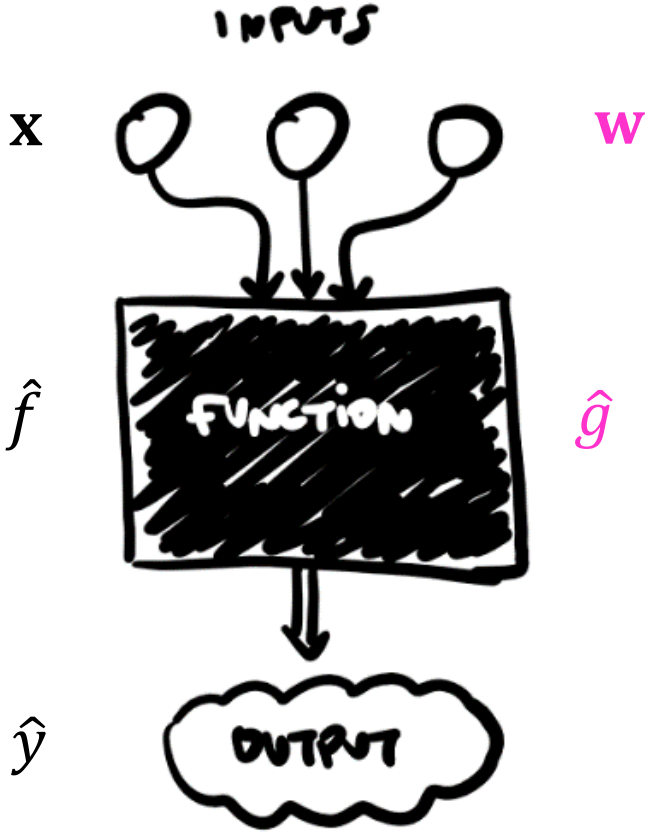


Image source: thatsoftwaredude.com

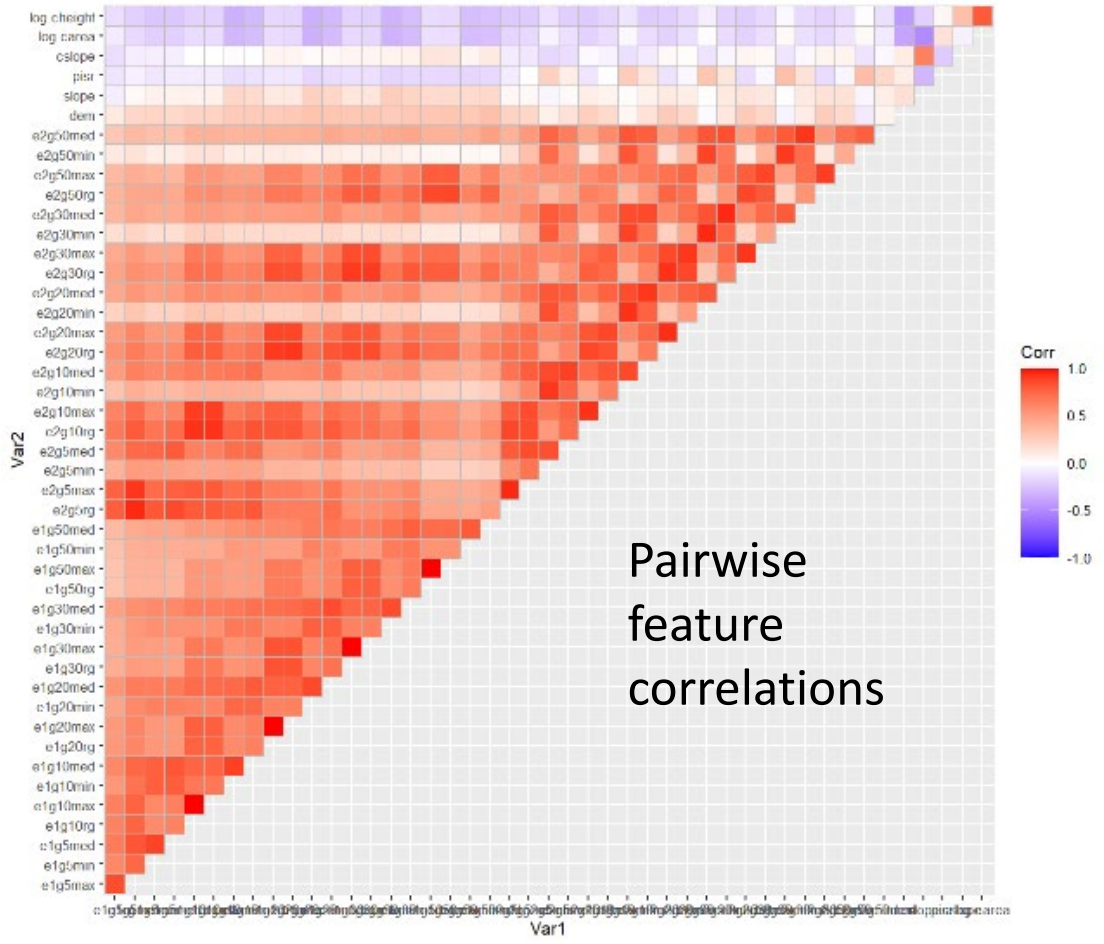
Proposal:

- Use an invertible mapping $T: X \longmapsto W$ onto an interpretation space W
- E.g. PCA or nonlinear embedding
- Now interpret $\hat{g} := \hat{f} \circ T^{-1}$ in interpretation space W
- Does not modify the model $\hat{f}$!
- R package **`wiml`** $\longmapsto$ **`iml`**, **`DALEX`**

Image source: pxfuel.com

Brenning (2023) in *Machine Learning*

Case study: Rock glaciers



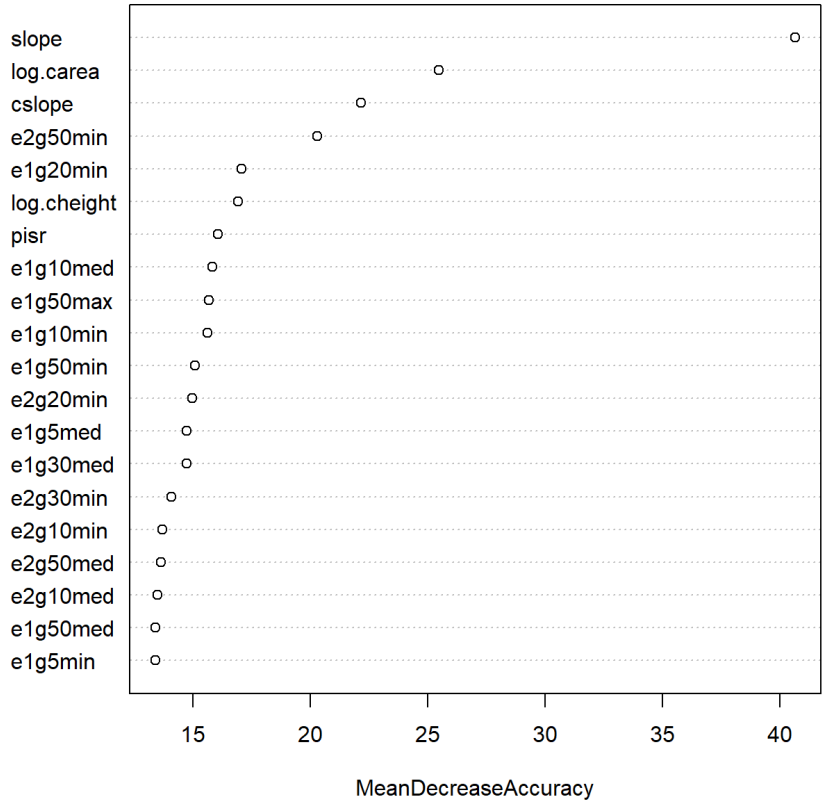
Rock glacier in the Andes….

…and in the Alps

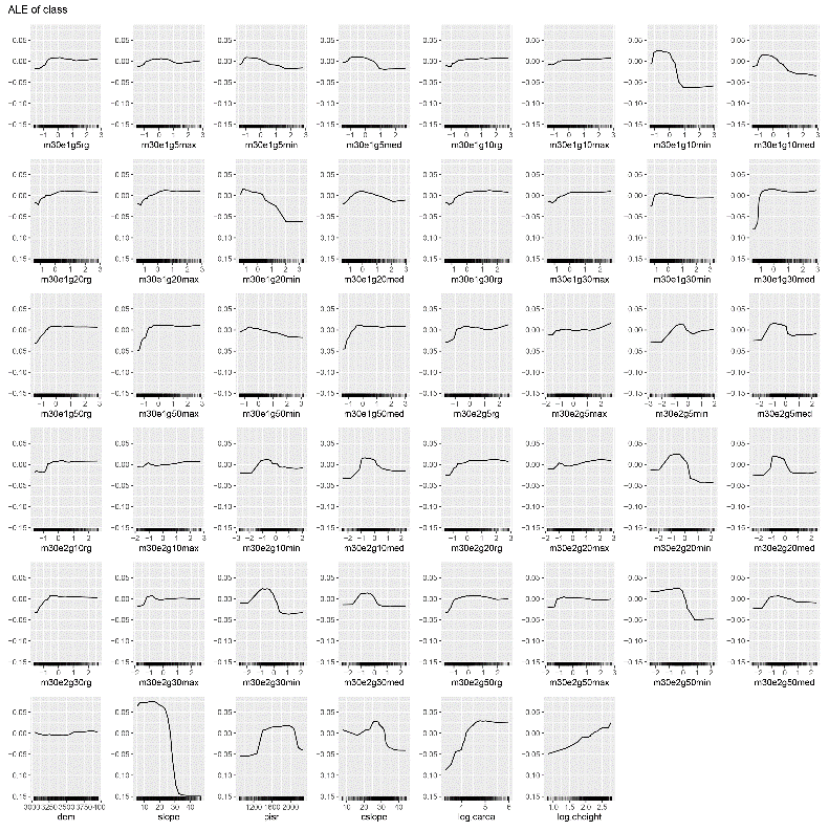
Case study: Rock glaciers

Interpretation, *traditional* style

Permutation importance

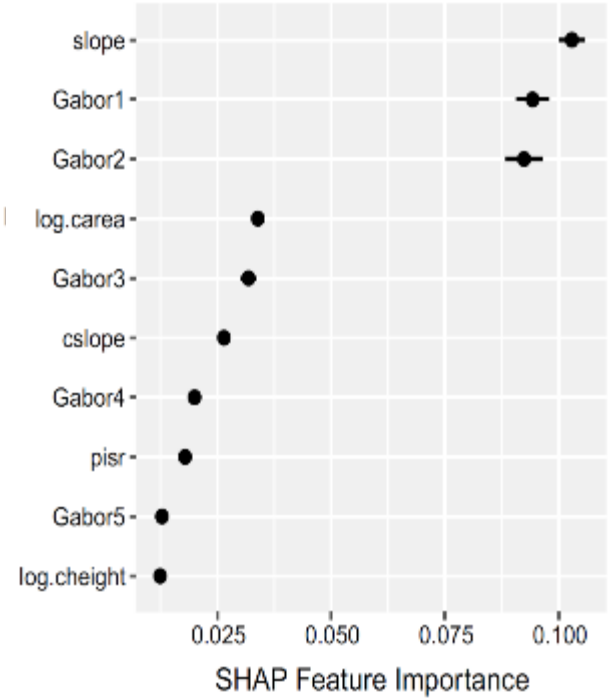


ALE plots

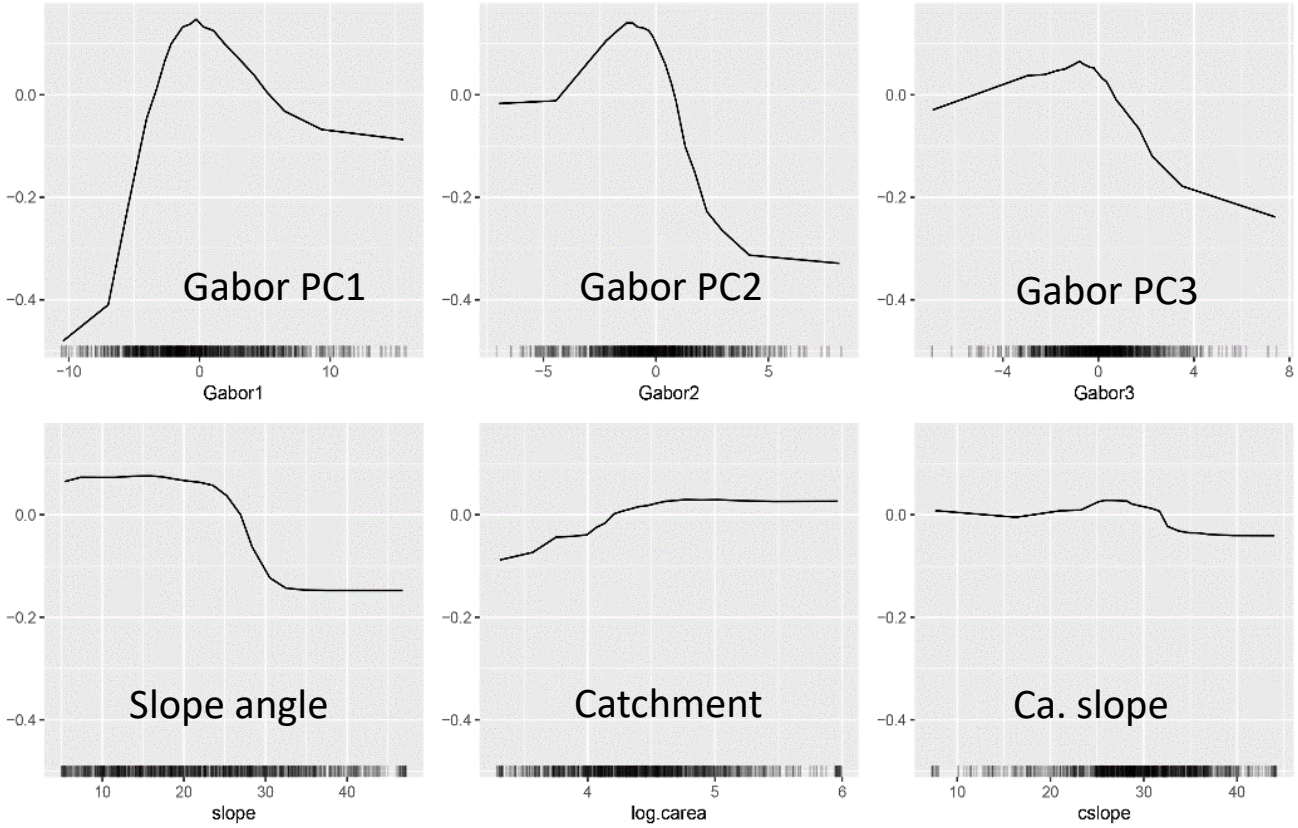


Case study: Rock glaciers
Interpretation in *transformed* space

SHAP importance

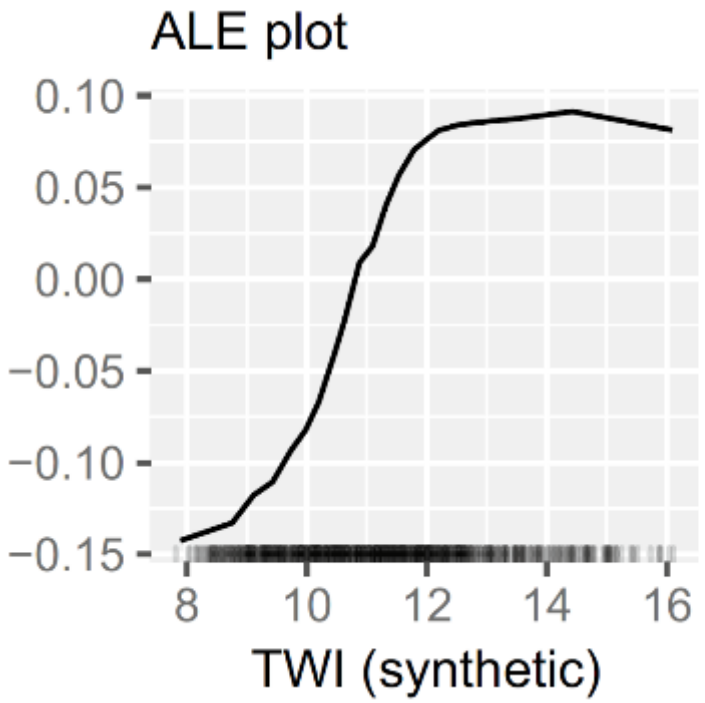


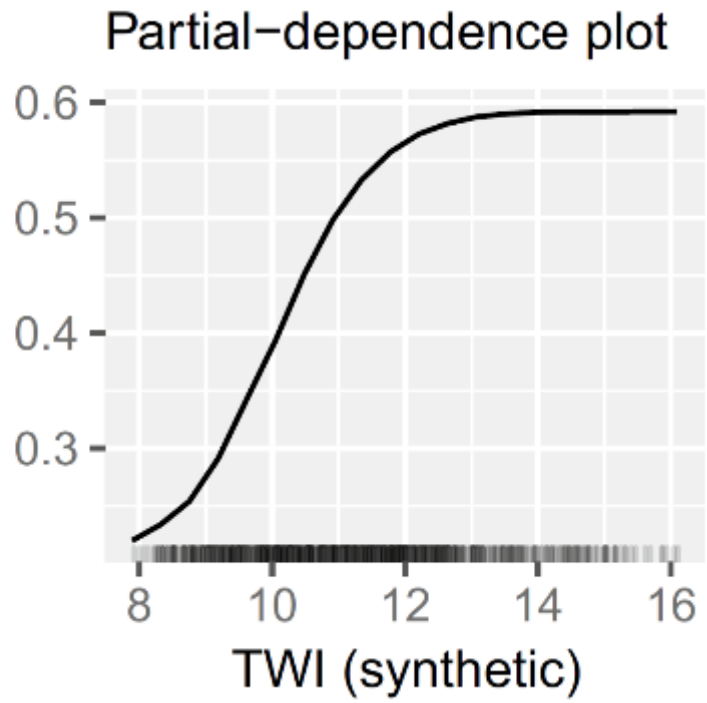
ALE plots



Brenning (2023) in *Machine Learning*

Case study: Rock glaciers
Interpretation with *synthetic* features





Brenning (2023) in *Machine Learning*

Lessons Learned

- To interpret complex machine-learning models, we have to represent them in a simplified way, in a one- or two-dimensional subspace of the features space.
  - This way, we usually ignore interactions among the predictors.
- Permutation-based feature importance provides a very rough measure of a variable's contribution to the predictive performance.
- SHAP feature importance requires less extrapolation.
- A simpler approach is to compare performances achieved with different feature sets.
- We can assess feature importances in different predictive settings.
  - E.g. different prediction distances, different performance measures.
- We can interpret models in lower-dimensional, transformed feature spaces.

What Can Go Wrong?

- For complex ML model, interpretation tools never tell you the whole story.
  - They are models of models…
- If a key objective is to interpret your data, use an interpretable model.
  - Additive models are great.
- Related features will always “steal” importance from each other.
  - Importance means “importance while accounting for the other variables in the model”.
- Different features can be important in different predictive settings.

Words of Wisdom (Coombs 1964, *Theory of Data*)

"we buy information with assumptions"

Clyde Hamilton Coombs (1912-1988)

Links & References

@geobrenning

- R code & data related to this class: https://github.com/alexanderbrenning/ogh23_ml
- 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. [link] [package]
- Brenning, A. (2023). Spatial machine-learning model diagnostics: a model-agnostic distance-based approach. *International Journal of Geographical Information Science*. [link] [github]
  - Blog article: https://geods.netlify.app/post/spatial-ml-model-diagnostics/
- Brenning, A. (2023). Interpreting machine-learning models in transformed feature space with an application to remote-sensing classification. *Machine Learning*. [link] [package]
  - Blog article: https://geods.netlify.app/post/interpretable-ml-with-a-twist/
- Molnar, C. 2023. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. [link]
- Schratz, P., Becker, M., Lang, M., Brenning, A. (2022). Mlr3spatiotempcv: Spatiotemporal resampling methods for machine learning in R. *arXiv preprint*. [link]
- Wadoux, A., Heuvelink, G.B.M., de Bruin, S., Brus, D.J. (2021). Spatial cross-validation is not the right way to evaluate map accuracy. *Ecological Modelling*. [link]