Classification Models

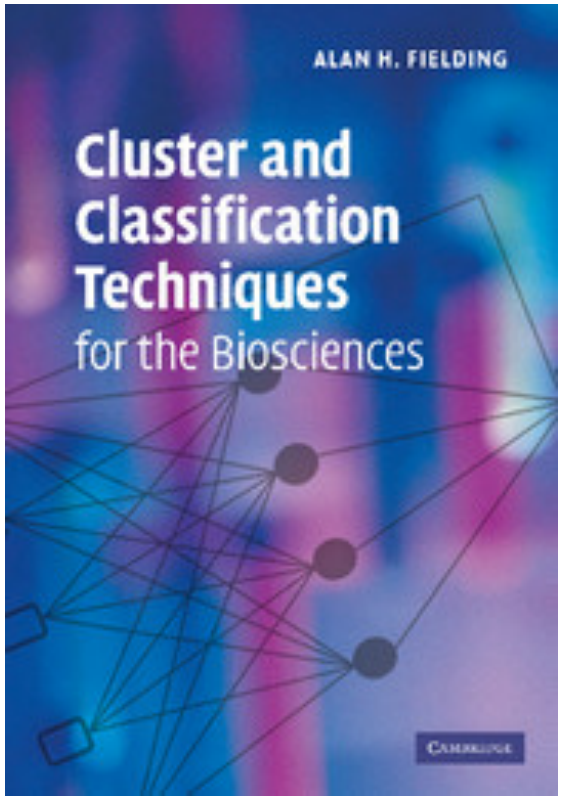
A Brief Overview

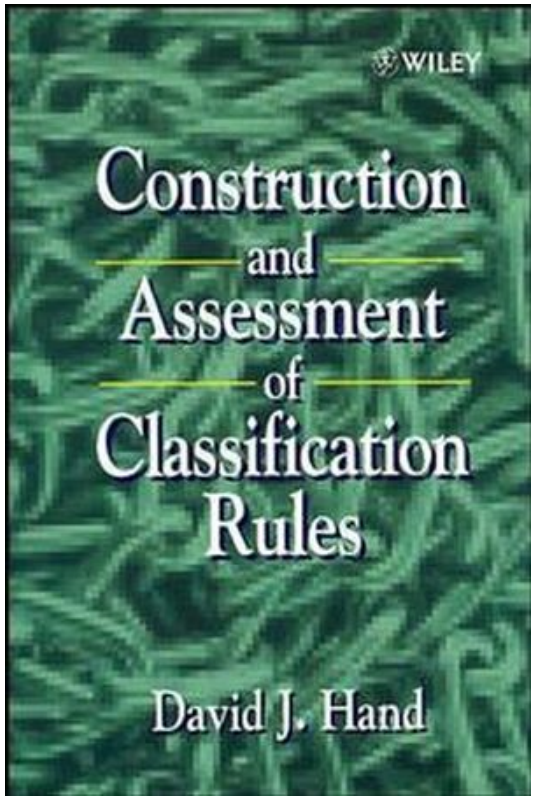
Alexander Brenning

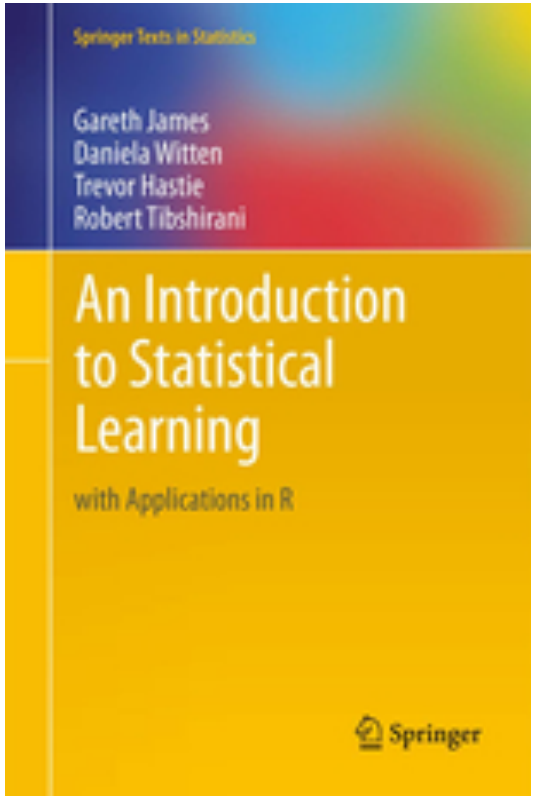
Department of Geography, Friedrich Schiller University Jena

Geo 408B

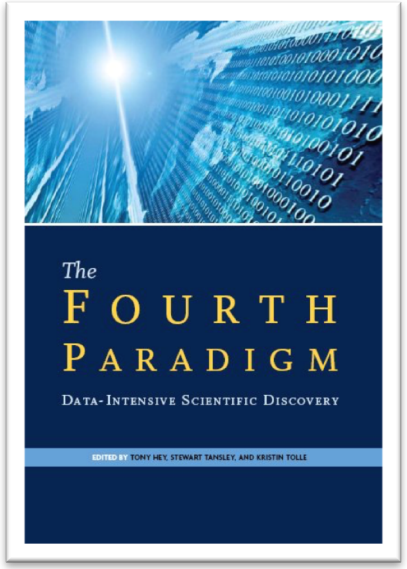
Recommended Textbooks
Statistical Learning / Classification







Data-Intensive Scientific Discovery: From Statistical *Analysis* to Statistical *Learning*



- **Jim Gray**: New Era of Data-Intensive Scientific Discovery
  - Exploding amounts of data from observations, experiments, simulations → shift from traditional inferential statistics to data-driven methods
- **Data Mining** or **Knowledge Discovery in Databases**:
  - Aims at making previously unknown, implicit patterns explicit
  - Data-driven, pragmatic, exploratory
  - Patterns are summarized by, for example, regression or classification rules
- **Machine learning** or **statistical learning**

  emphasizes the performance of predictive models
- In **Geocomputation**, added challenges

  related to scalability and interoperability

http://classes.engr.oregonstate.edu/eecs/fall2012/cs434/

Classification

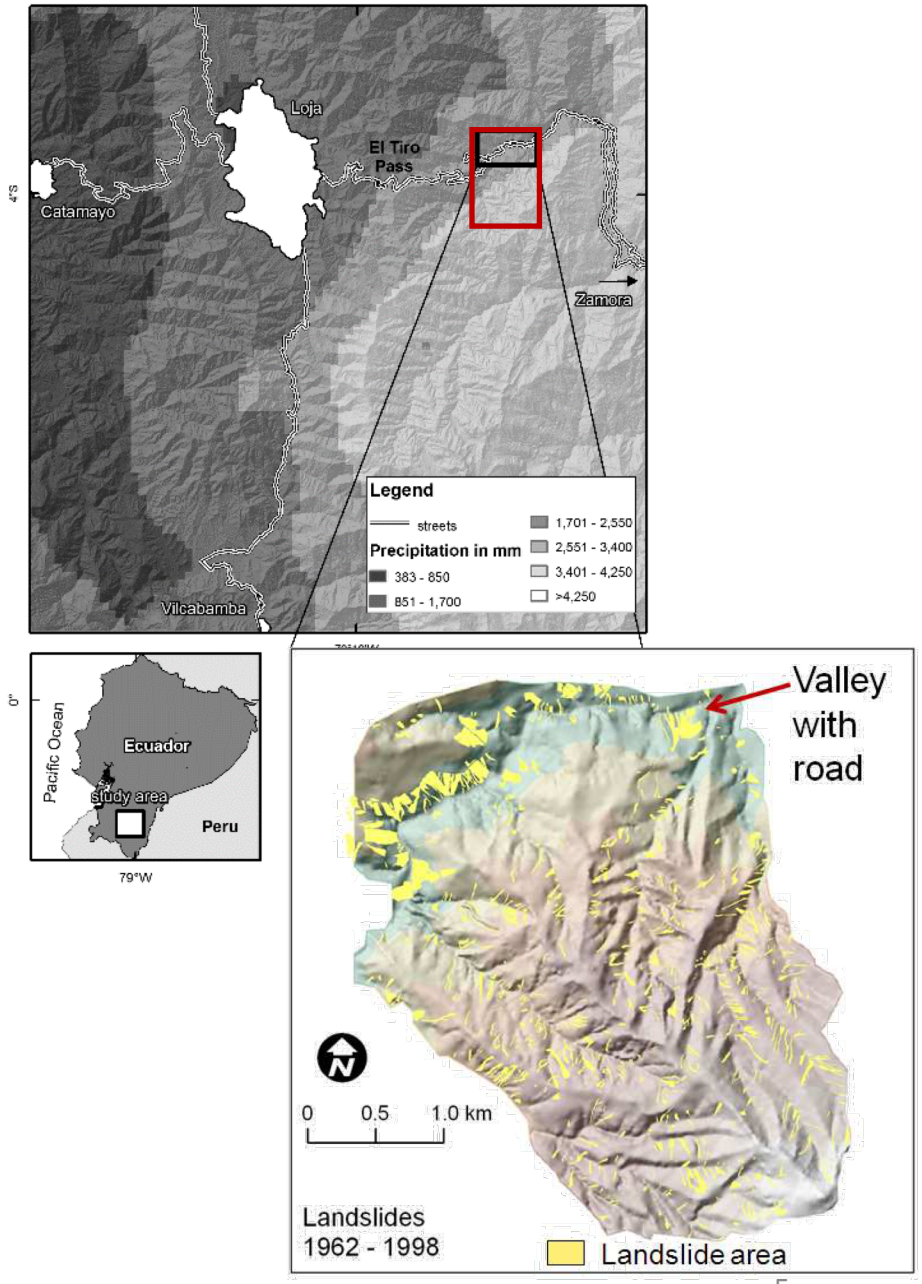
- General objective:
  - Identify a function C that predicts a categorical response variable y based on predictor variables $x^{(1)}, \ldots, x^{(p)}$:

$$\hat{y} = C\left(x^{(1)}, \ldots, x^{(p)}\right)$$

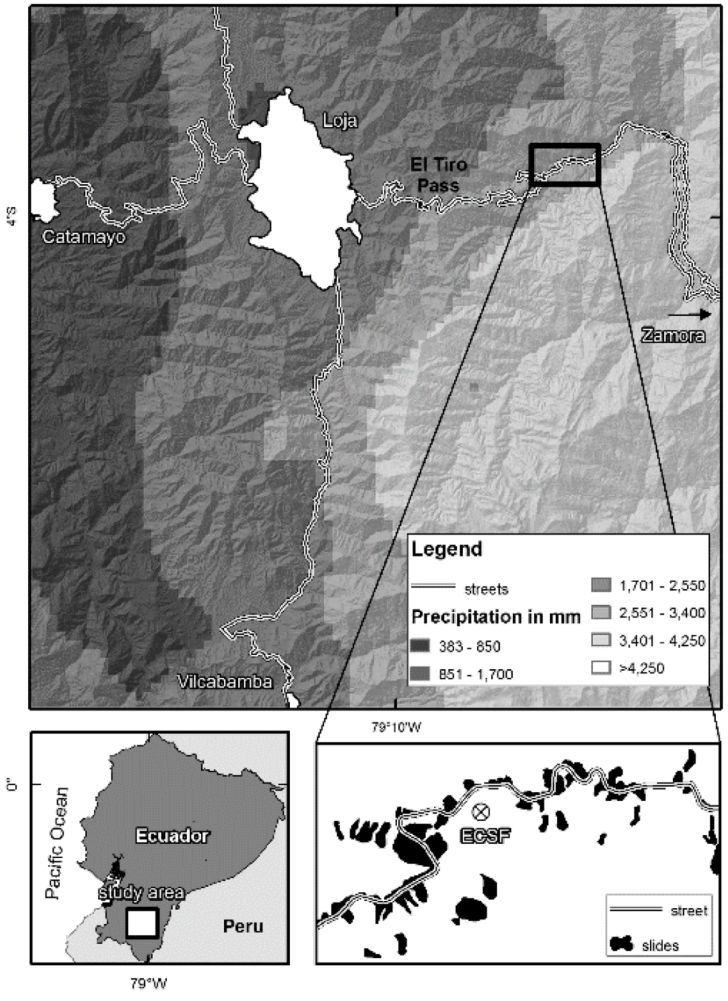
- **Supervised classification**:
  - The response is known for a sample that can be used to find C.
  - Like regression, but with a categorical response variable.
- **Unsupervised classification** (cluster analysis – *not covered here*):
  - Group the observations into previously unknown classes based on their features

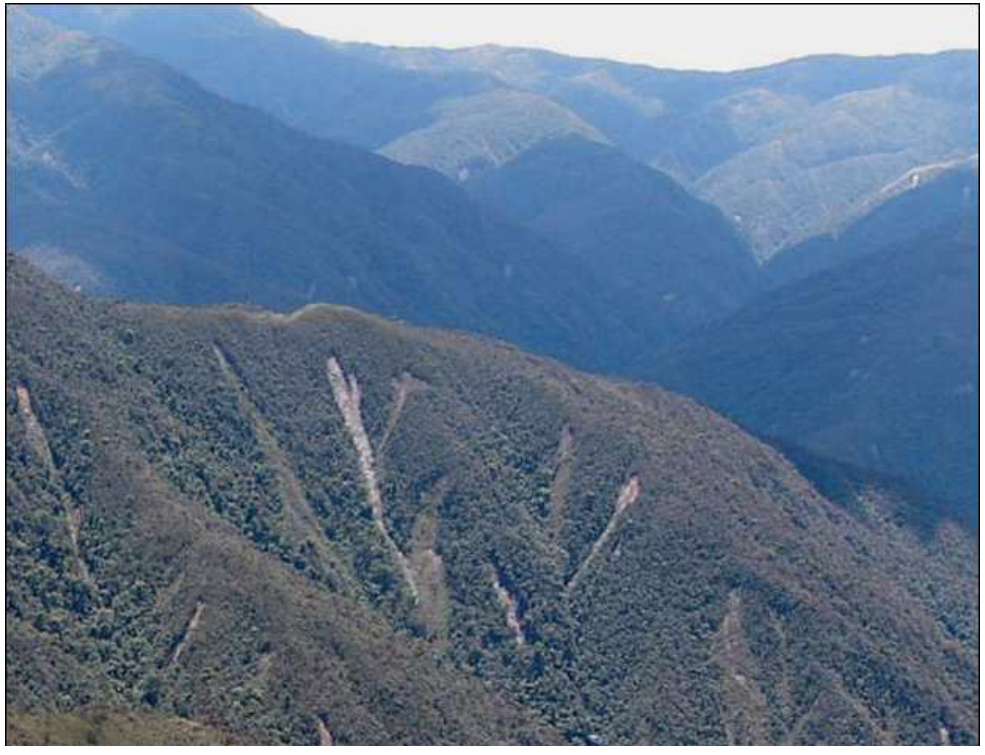
Case Study 1: Landslide Susceptibility

- Goal:
  - **Prediction**: Identify landslide-prone areas, and/or
  - **Analysis**: Identify preparatory factors
- Response:
  - Landslide presence / absence
- Predictors:
  - Terrain attributes (e.g., slope angle), land use, distance to road, rock type
- Two-class problem, „soft“ classification
- Step-by-step tutorial: RSAGA package vignette, Thursday‘s lab class



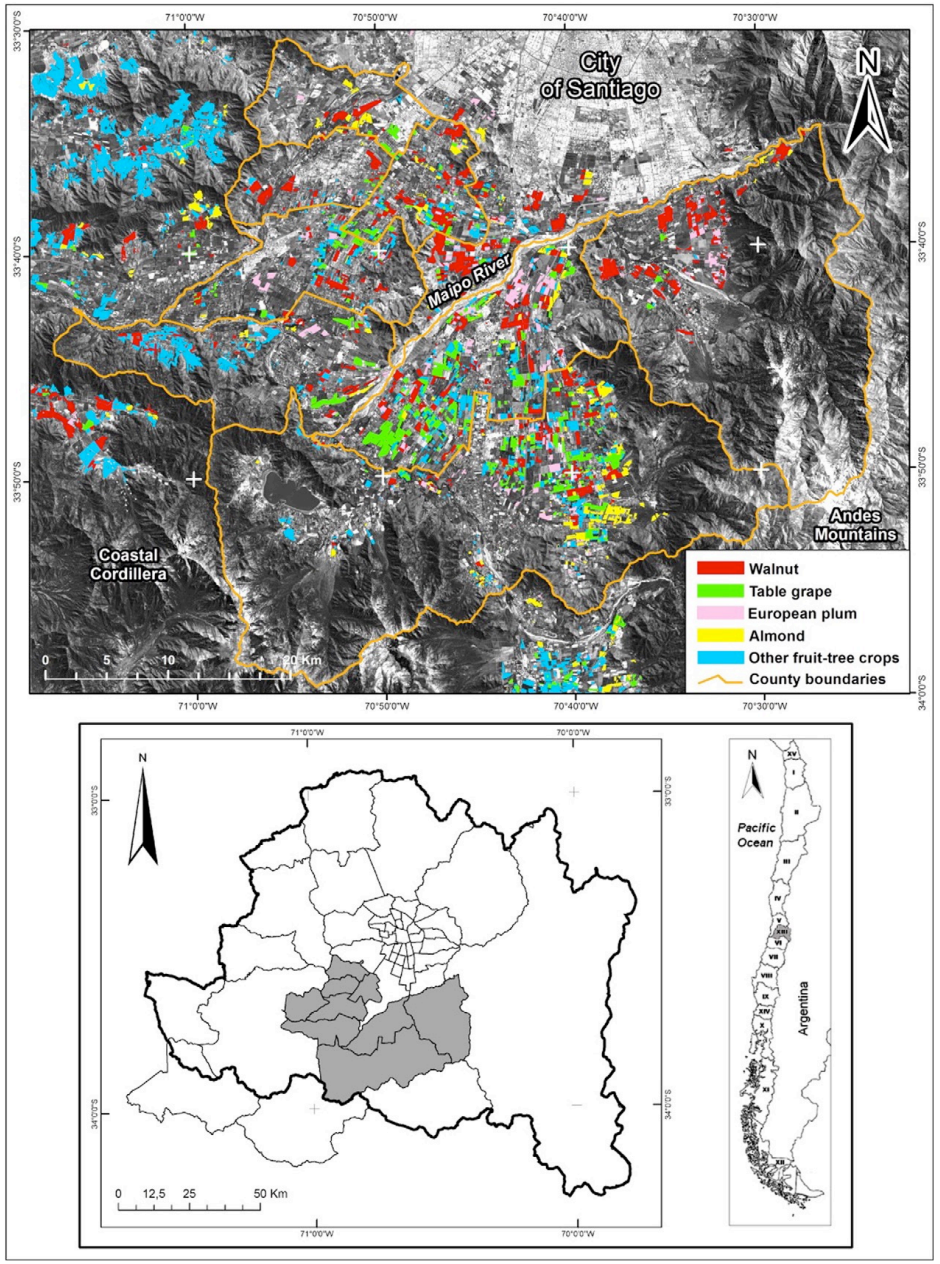
Landslides in Southern Ecuador





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

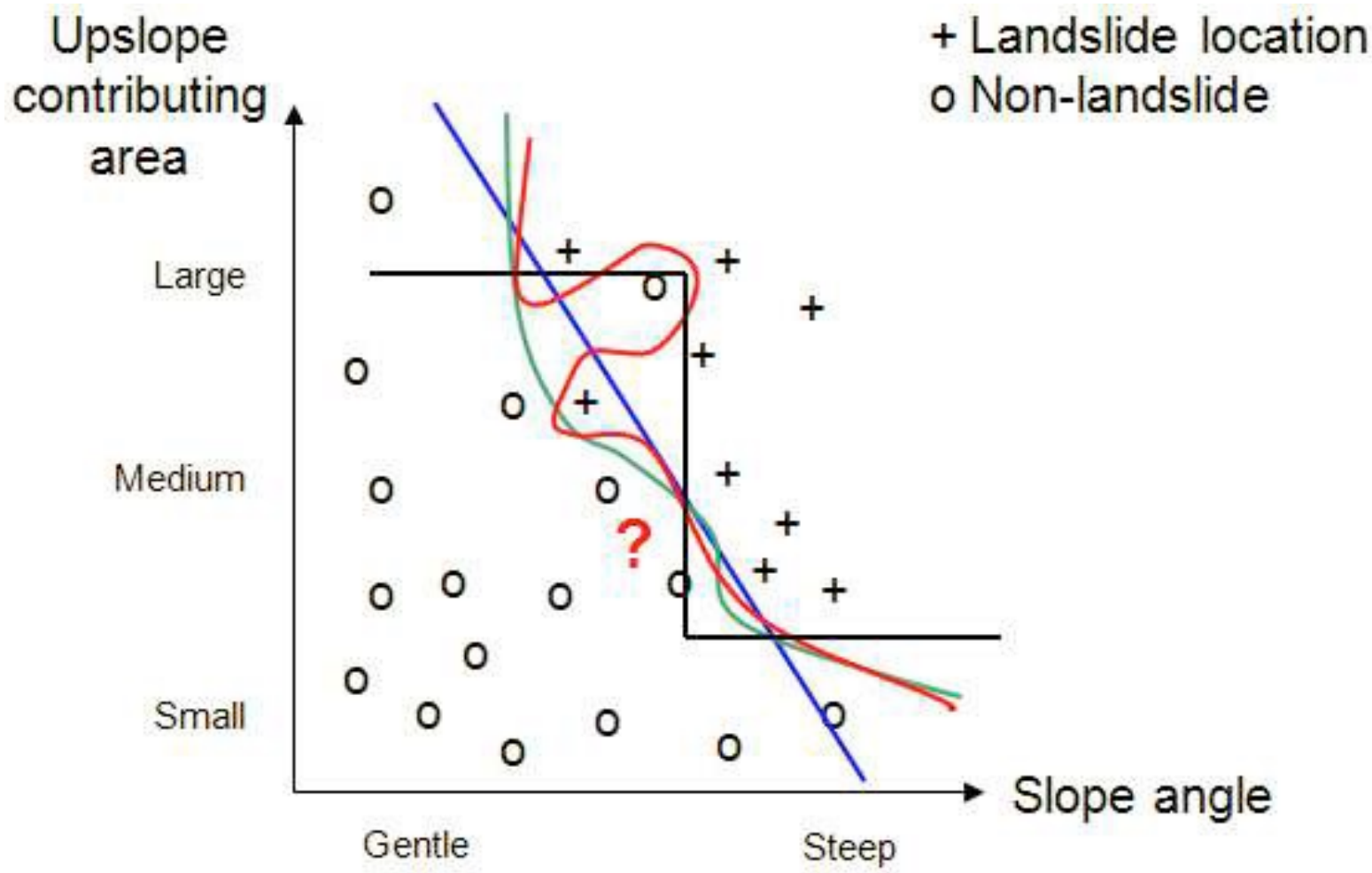


Classification Techniques

- *k* nearest neighbours classification
- Generalized linear model: logistic regression
- Generalized additive model
- Linear discriminant analysis (and extensions)
- Classification trees
- Ensemble methods: bagging, random forest
- Boosting
- Support vector machine
- Artificial neural networks
- …

Fancy names, but at the end of the day…

How Classifiers Work (Basically…)



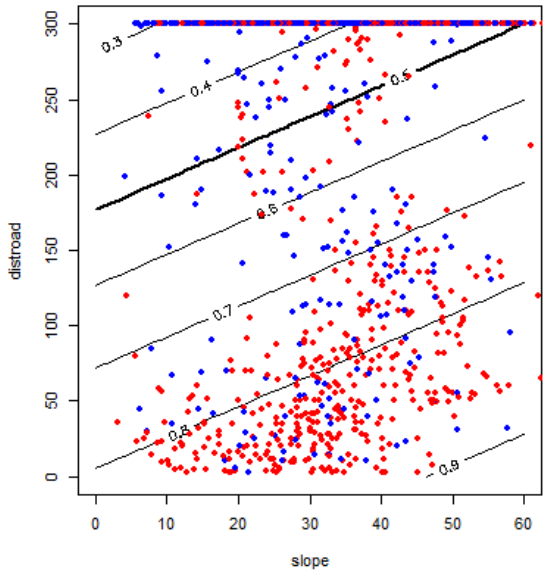
Model Predictions in Feature Space: Comparison

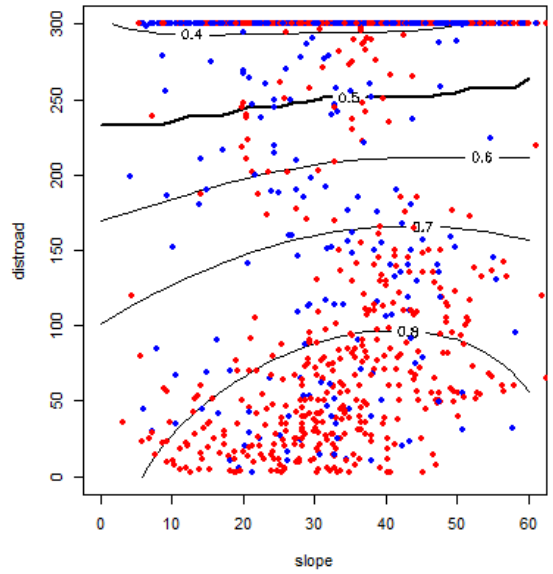
For illustration only:

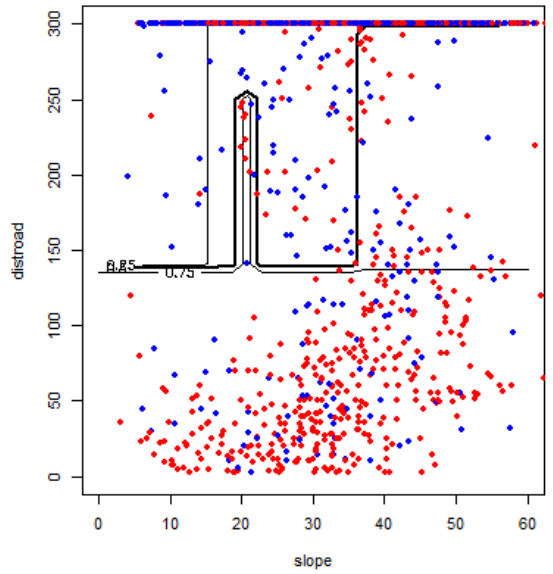
Using only slope and distroad as predictors

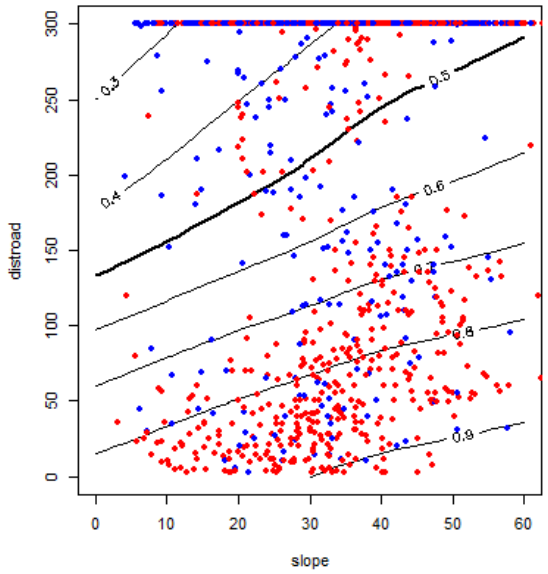
Contours are lines of equal predicted "probability"

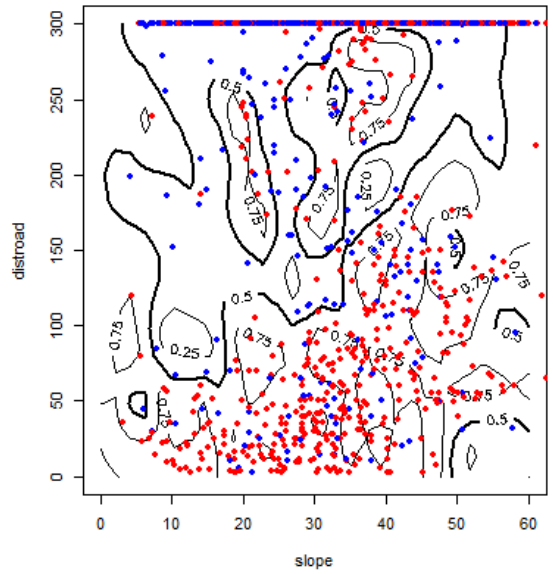
Points are landslide (red) and non-landslide (blue) observations

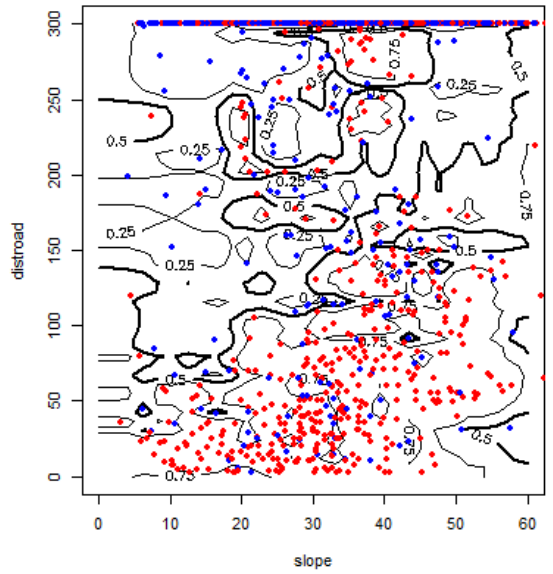












A Formal Model

- There are N objects for which we observe
  - the class membership y,
  - the corresponding values of p predictor variables (or features) $\mathbf{x} = \left(x^{(1)}, \ldots, x^{(p)}\right)^T$.
- The set

$$L = \{(y_i, \mathbf{x}_i) : i \in 1, \ldots, N\}$$

  of the N known class memberships and their corresponding predictor values is referred to as the **learning sample** or **training set**.
- This data is available for constructing a classifier.
- Based on L, we wish to find a **classifier** C_L that predicts the class membership y (of a possible new object) only based on the predictors:

$$\hat{y} = C_L(\mathbf{x})$$

k Nearest Neighbours Classification

- Consider a learning sample L and a "new" object with predictor values $\boldsymbol{x}$ and unknown class membership y.
- Which observations are the most similar to the new object?
  - Use a distance measure (in feature space), $d(\mathbf{x}, \tilde{\mathbf{x}})$,
  - to identify the k nearest neighbours of $\mathbf{x}$ among $\mathbf{x}_1, \ldots, \mathbf{x}_N$.
  - Denote their indices by $i_1, \ldots, i_k$.
- Predict the conditional probability for class j by

$$P(Y = j | X = \mathbf{x}) = \frac{1}{k} \sum_{l=1}^{k} I(y_{i_l} = j)$$

- Classify the "new" object to the class with the largest probability.
- *How will different values for k influence the shape of C_L?*

A Formal Model

- The classifier C_L is conditional on the learning sample, L!
- A classification technique is a procedure used for constructing C_L from a learning sample L:

$$C: L \rightarrow C_L$$

- **Parametric techniques**: C_L is derived by estimating coefficients $\beta_1, \dots, \beta_p$ from the learning sample
- **Non-parametric techniques**: Often algorithmic, cannot be written as a simple mathematical formula, or do not involve parameters
- Classifiers may also depend on **hyperparameters** that control their general behaviour:
  - E.g. in k-NN: $\theta = k$

Classification Techniques

Logistic Regression

- Generalized linear models (GLM) for binary response variables.
- We model $p = P(Y = 1|\mathbf{X} = \mathbf{x})$.
- Instead of modelling it linearly, we transform the probabilities to **logits**, the logarithm of the **odds**:

$$logit(p) = \log\big(odds(p)\big) = \log\left({}^{p}/_{(1-p)}\right)$$

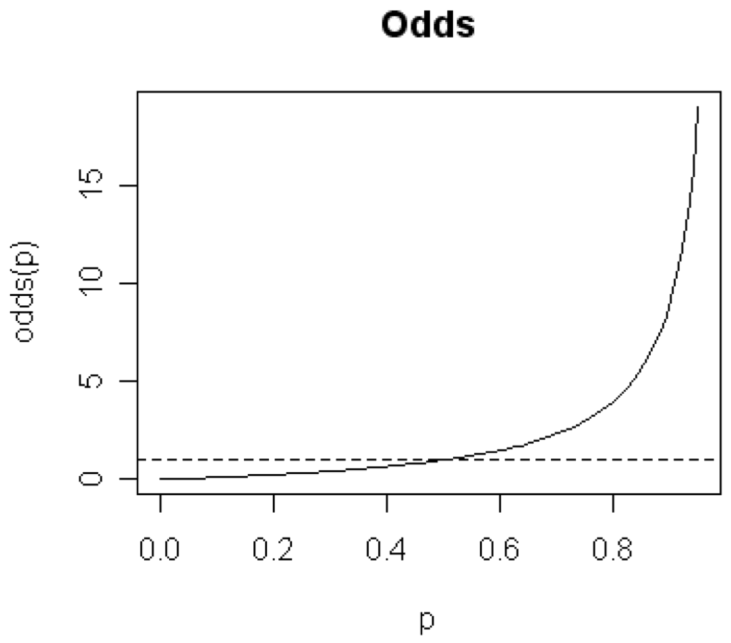
- This quantity is modelled linearly:

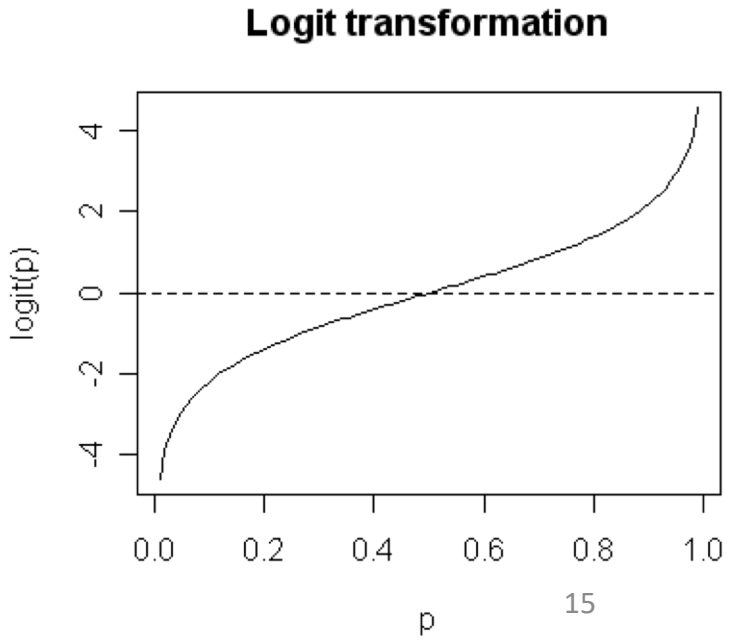
$$logit\big(P(Y = 1|\mathbf{X} = \mathbf{x})\big) = \alpha + \sum_{i=1}^{p} \beta_i x^{(i)}$$

- Logits can be back-transformed into probabilities:

$$p = \frac{\exp(logit)}{1 + \exp(logit)}$$

- Hosmer & Lemeshow (2000) is a great reference.





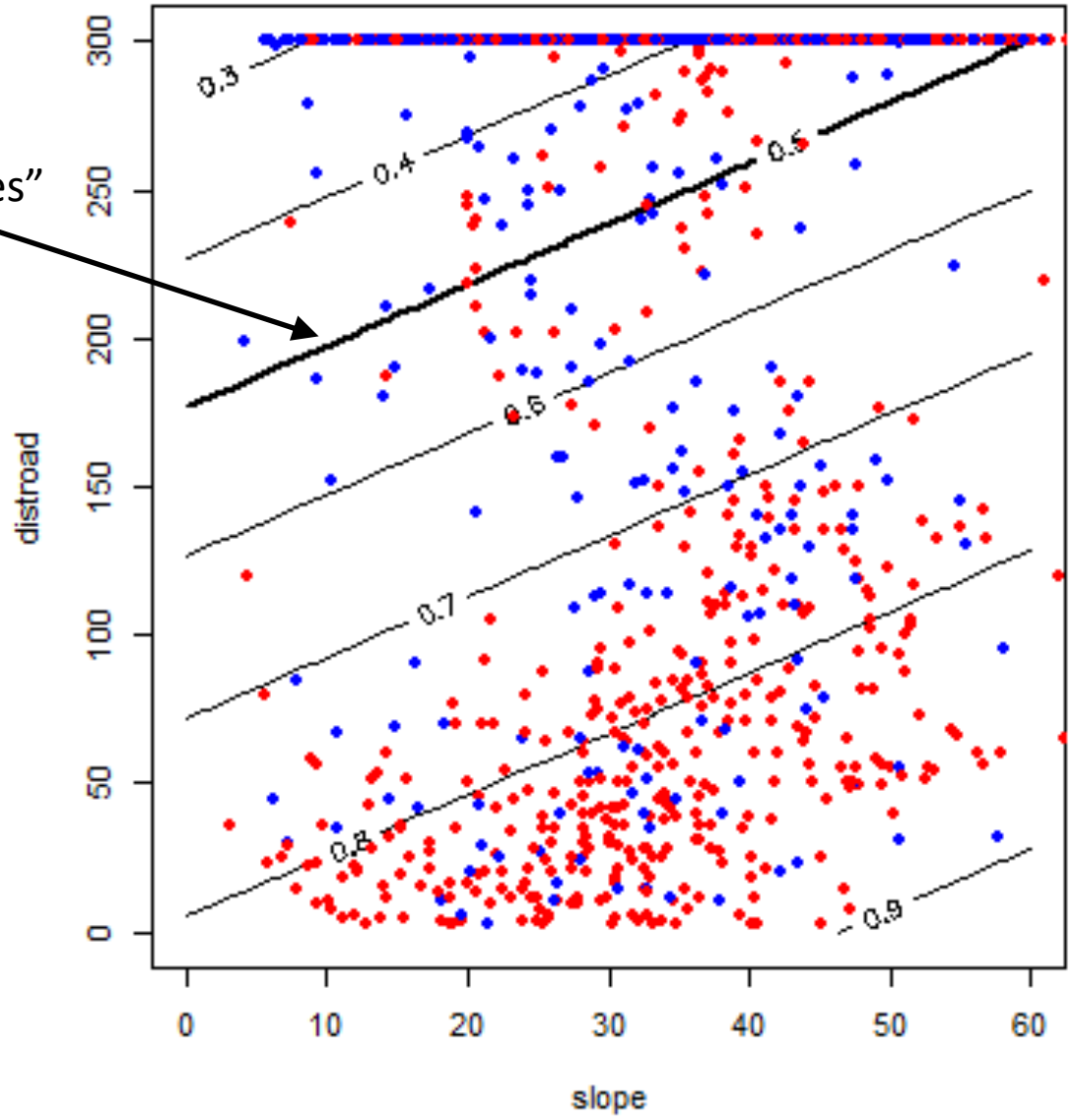
Logistic Regression

- Parametric, (generalized) linear model → fairly easy to interpret in terms of odds ratios
- Usually only applied to two-class problems
  - Multinomial logistic regression, equivalent to the maximum entropy (MaxEnt) classifier
- AIC / BIC, tests for coefficients, …
- Mixed-models extensions
- Spatial logistic models with residual autocorrelation e.g. `glmmPQL` in R package `MASS`.
  - Computationally expensive

Model Predictions in Feature Space: Logistic Regression

Simple model with two predictors used for illustration only

Contour lines of predicted “probabilities”




```
Call:
glm(formula = slides89 ~ slope + distroad, family = binomial,

    data = d)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1164  -1.0021  -0.1789   1.2662   1.5757

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.4286950  0.2004692   7.127 1.03e-12 ***
slope        0.0165778  0.0043754   3.789 0.000151 ***
distroad    -0.0080696  0.0005515 -14.633  < 2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2695.0  on 1943  degrees of freedom
Residual deviance: 2417.5  on 1941  degrees of freedom
AIC: 2423.5
```

Generalized Additive Model (GAM)

- Semi-parametric extension of the GAM
- E.g. logistic additive model:

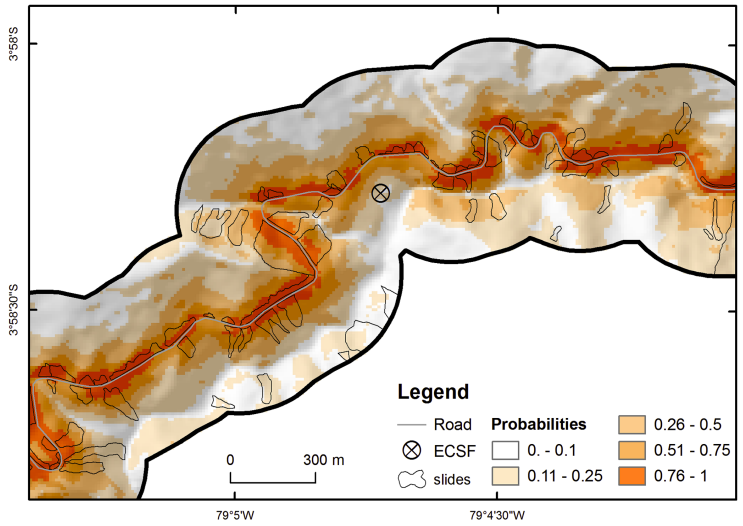
$$logit(P(Y = 1|\mathbf{X} = \mathbf{x})) = \alpha + \sum_{i=1}^{p} \beta_i s_i(x^{(i)})$$

  where the s_i are transformation functions:
  - Usually non-parametric smoothing splines
  - Mix linear and non-linear terms
  - Bivariate smoothers represent interaction terms…
- Degree of smoothing determined by their degrees of freedom
  - Tuned e.g. by generalized cross-validation (GCV)
- GAM with random effects and residual spatial autocorrelation available in R package `mgcv` (Wood, 2006)
- Can be computationally expensive, may have convergence problems

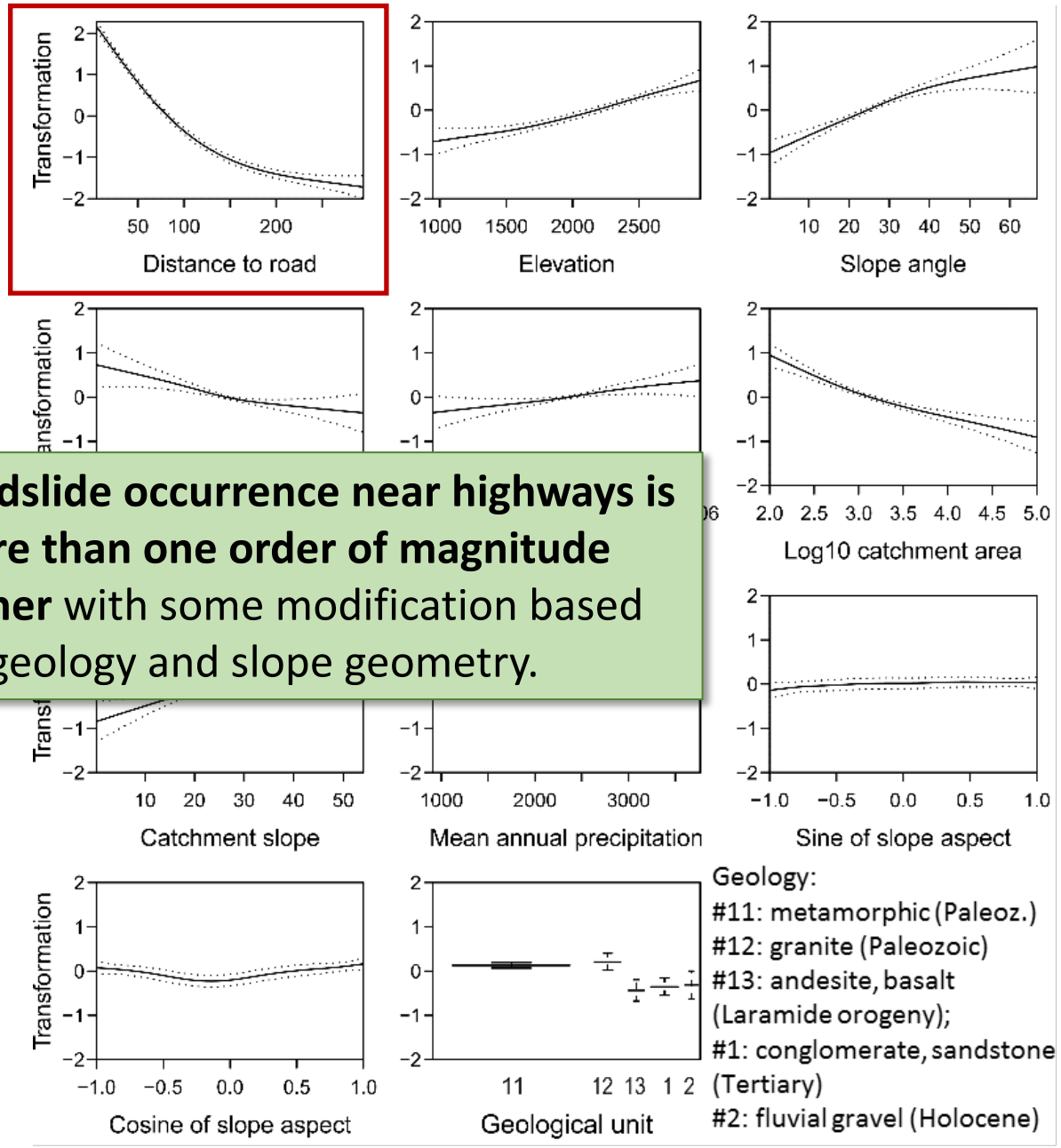
GAM Example: Landslides & Highways

	Parametric estimation	Spatial block bootstrap
GAM	21.2*	19.6 [15.5–25.3]
LM	18.4 [15.5–22.0]	18.9 [13.7–26.6]

* Parametric confidence intervals not available for the GAM.



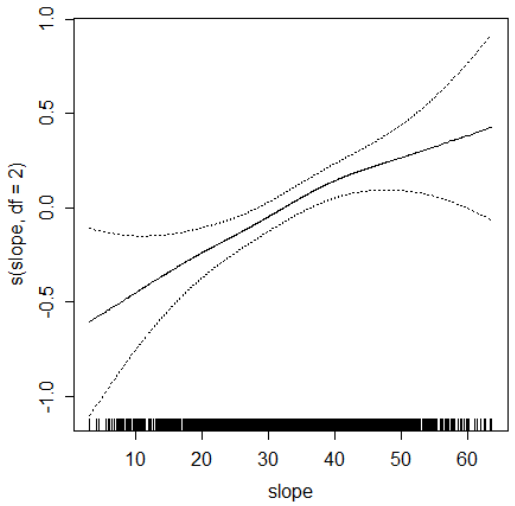
Brenning et al. (2015)

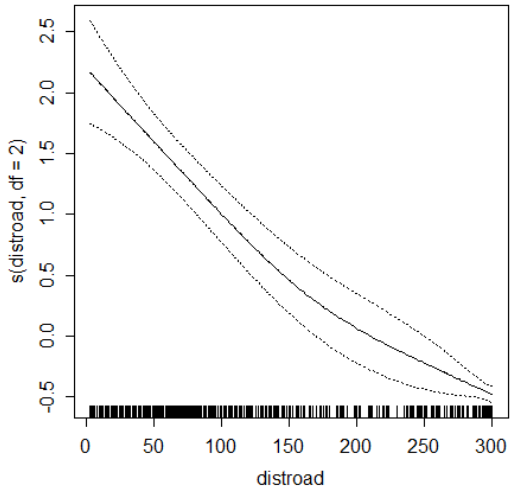


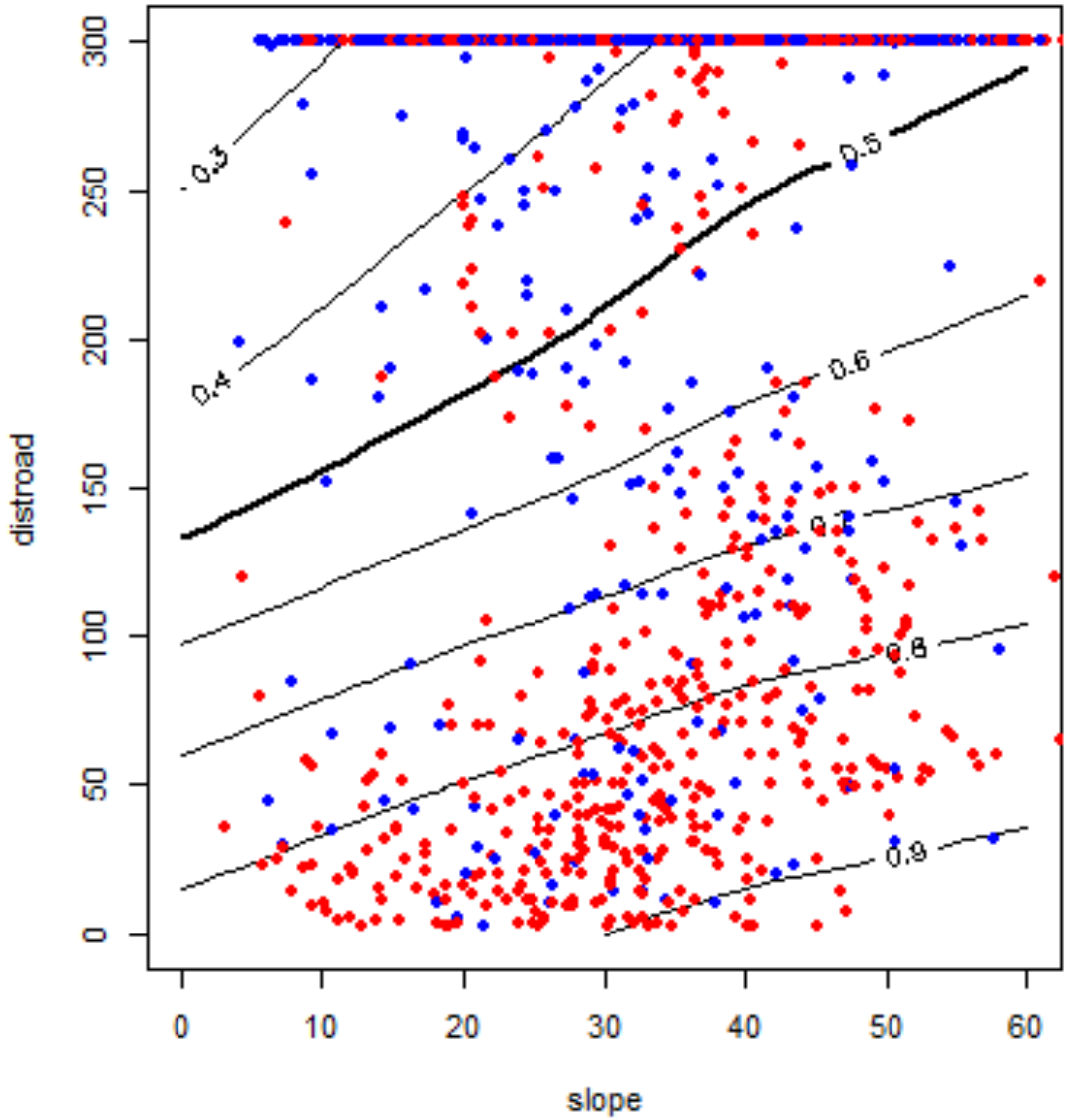
Landslide occurrence near highways is more than one order of magnitude higher with some modification based on geology and slope geometry.

Model Predictions in Feature Space: GAM

Transformation functions:







(Or Would You Rather Use a Physically Based Model?)

SHALSTAB model
for shallow rainfall-induced landslides
Montgomery & Dietrich (1994)

- For given material properties, stability is only determined by slope angle and contributing area
- Empirical models will normally perform better than such simple physically based models (Goetz et al. 2011 in *Geomorphology*)

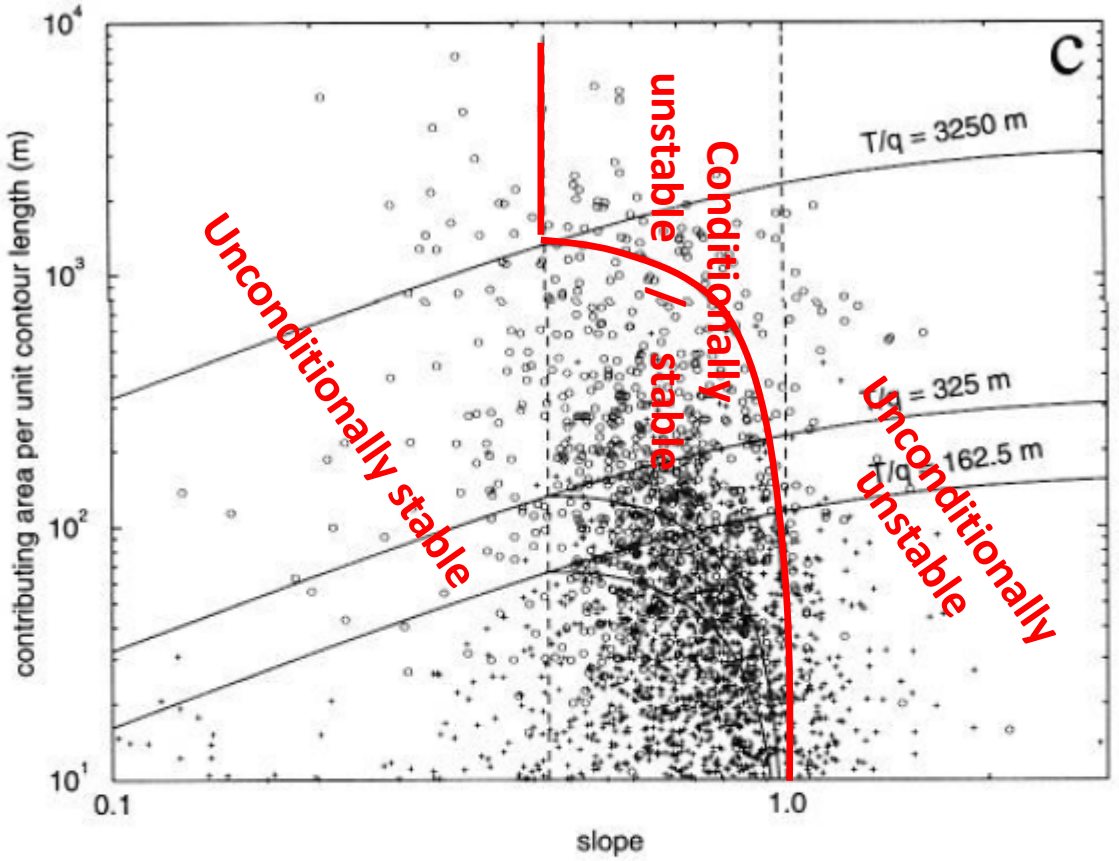
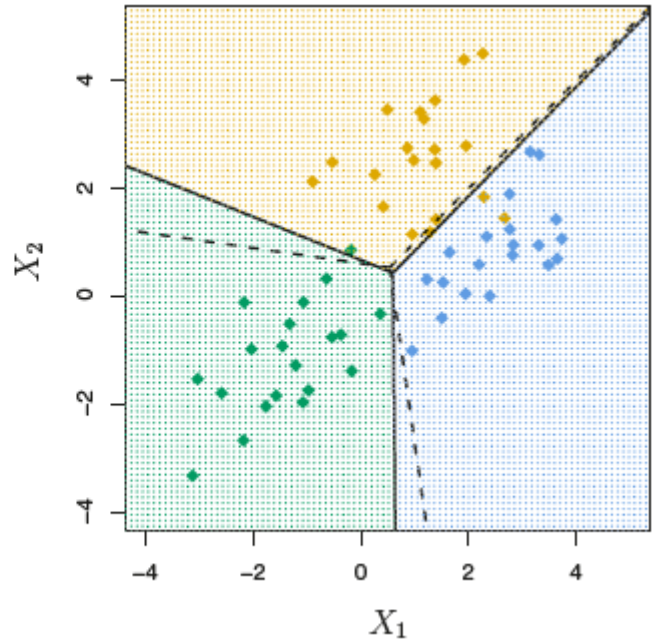


Figure 6. Plots of contributing area per unit contour length versus slope (tan θ) for convergent (circles) and divergent (crosses) topographic elements in the (a) Tennessee Valley, (b) Mettman Ridge, and (c) Split Creek study catchments showing the effect of varying T/q on the topographic thresholds for soil saturation and slope stability (solid lines). Dashed lines indicate limits of the slope stability model.

Linear Discriminant Analysis

- Uses linear combinations of predictors – the **discriminant functions**
- This results in a **separating hyperplane** in feature space



James et al. (2013)

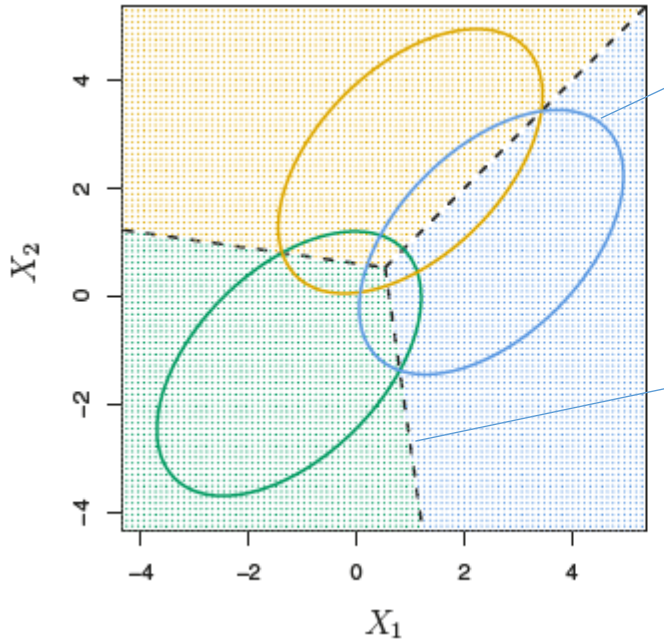
- LDA models the conditional probability distributions

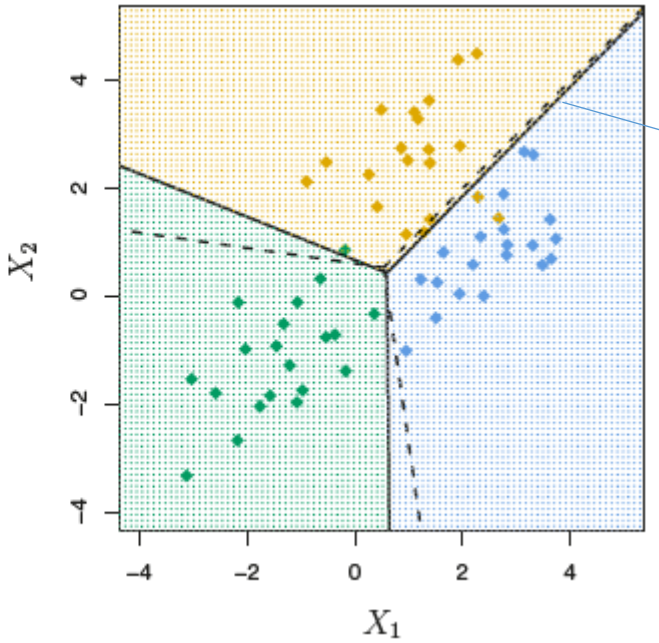
  $P(\mathbf{X}|Y = i)$ for each of the classes *i*

  and uses Bayes‘ theorem to translate this into $P(Y = 1|\mathbf{X} = \mathbf{x})$

Linear Discriminant Analysis

- Assume that $\mathbf{X}|Y = i$ follows a multivariate normal distribution with *common* covariance matrices, Σ, but centered at *different* mean values, μ_i.
- Σ and μ_i (and the prior class probabilities $P(Y = i)$) are estimated from the data.
  - $p(p+1)/2$ covariances to estimate
  - mp mean values
  - m probabilities
- Assumption of multivariate normal distribution requires, in particular, quantitative predictor variables.

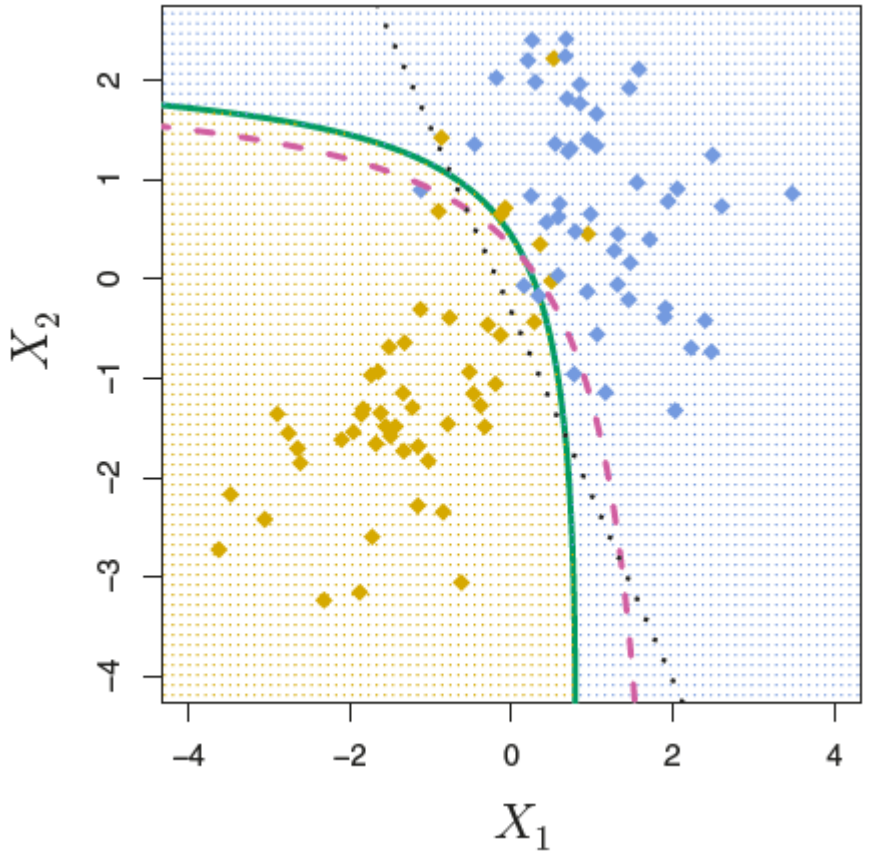




James et al. (2013)

Quadratic Discriminant Analysis

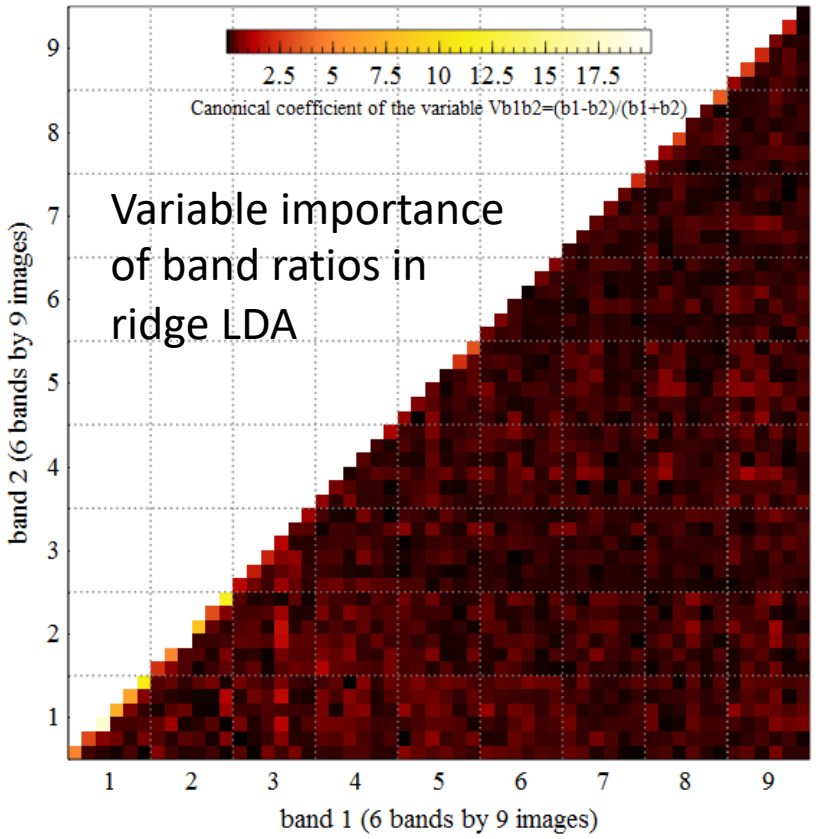
- Each probability distribution $P(\mathbf{X} = \mathbf{x}|Y = j)$ has a different covariance matrix, Σ_j.
  - Estimate *mp*(*p*+1)/2 covariances
  - E.g. crop classification (*m*=4, *p*=64): 8320 covar.
- This results in a discriminant function that is quadratic in **x**,
- and a parabolic separating hypersurface in feature space
  - Much more flexible than LDA, may overfit
- Known as maximum likelihood classification in the remote sensing literature.



James et al. (2013)

High-Dimensional Problems in Remote Sensing

- Parameters usually not estimable when *# par.* > *N*
- Hyperspectral remote sensing:
  - E.g. Hyperion imager onboard EO-1 satellite with 220 spectral bands, 10 nm bandwidth each
  - Applications in mineral potential or plant disease mapping
- Texture filters:
  - Filters with varying moving window sizes and other settings
  - Applications in very-high-resolution remote sensing
- Multitemporal vegetation indices
  - Normalized differences indices built from spectral bands across multiple image dates
- What is our "honest" sample size when adjacent image pixels are strongly correlated?
  - E.g. our crop data: 7713 observations, but only 400 fields



Peña et al. (in prep.)

Penalized LDA Techniques

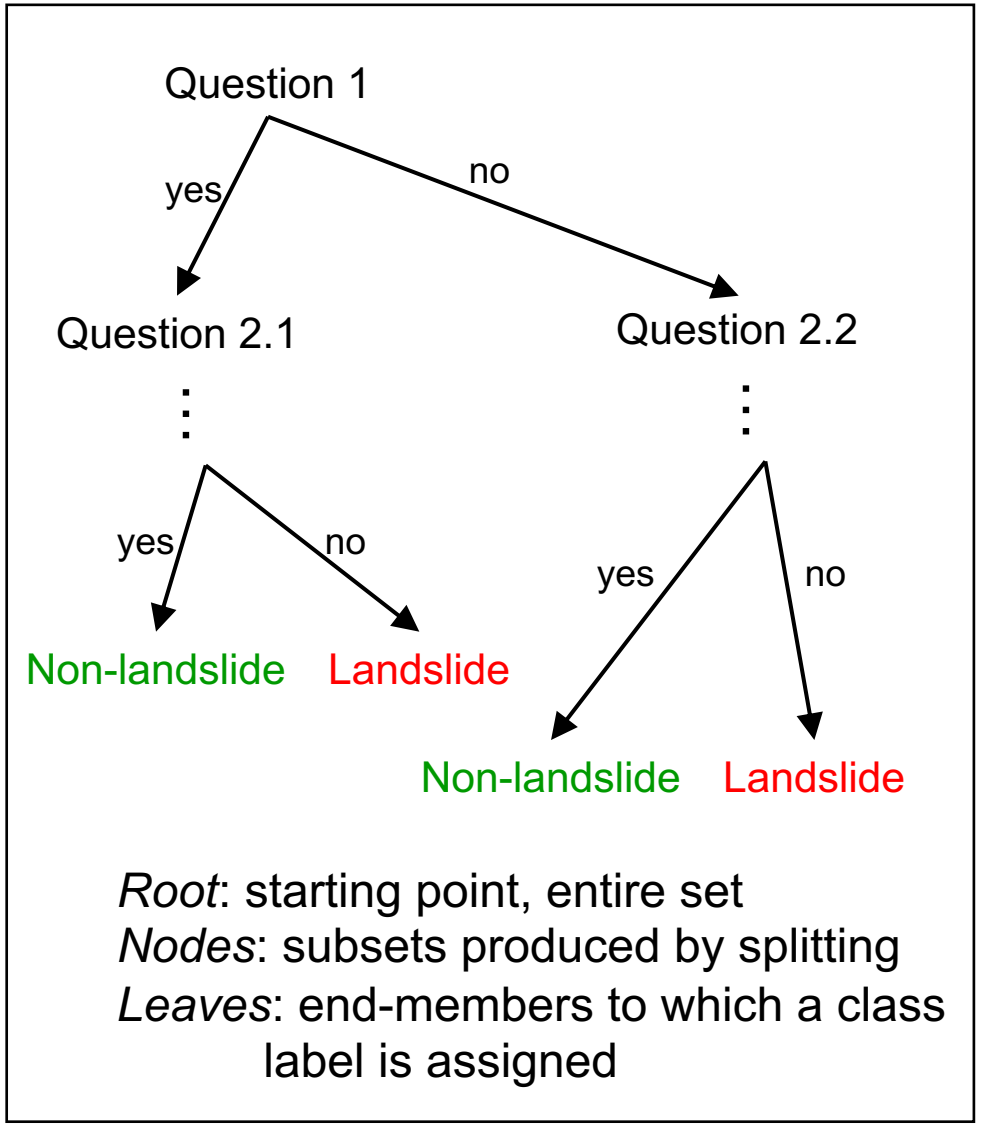
- Impose penalties on model coefficients
  - Hyperparameter controls penalty
- **Lasso penalty**:
  - Variable selection by shrinking some/most of the coefficients to (exactly) 0
  - Suitable when looking for / expecting that there is a small subset of relevant features while all others are irrelevant (*needle in the haystack*)
  - R package `penalizedLDA`
- **Ridge penalty**:
  - Shrink coefficients towards 0, without reaching 0
  - R package `mda`
- **Principal component LDA**:
  - Dimension reduction by picking only the first principal components


MONTY PYTHON'S
AND NOW FOR SOMETHING COMPLETELY DIFFERENT
モンティ・パイソン
アンド・ナウ
THE BEST OF MONTY PYTHON'S FLYING CIRCUS

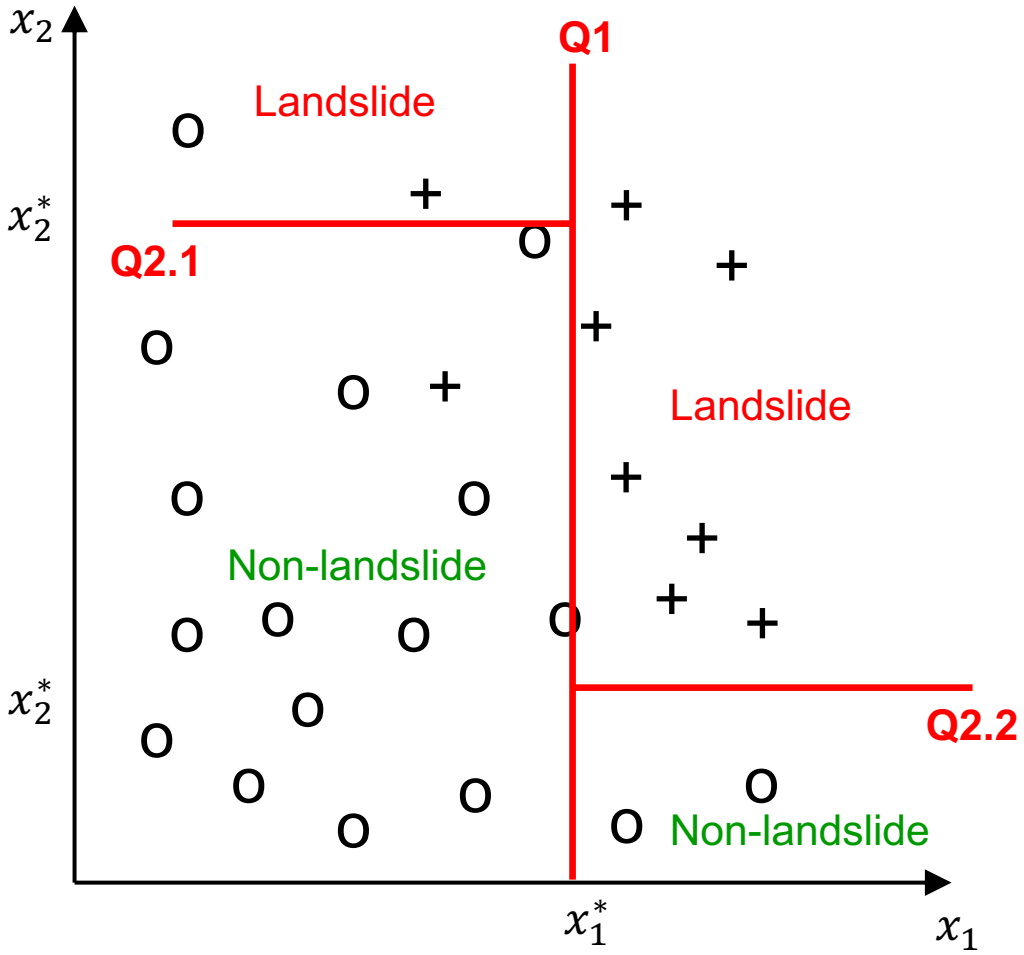
Blu-ray.com

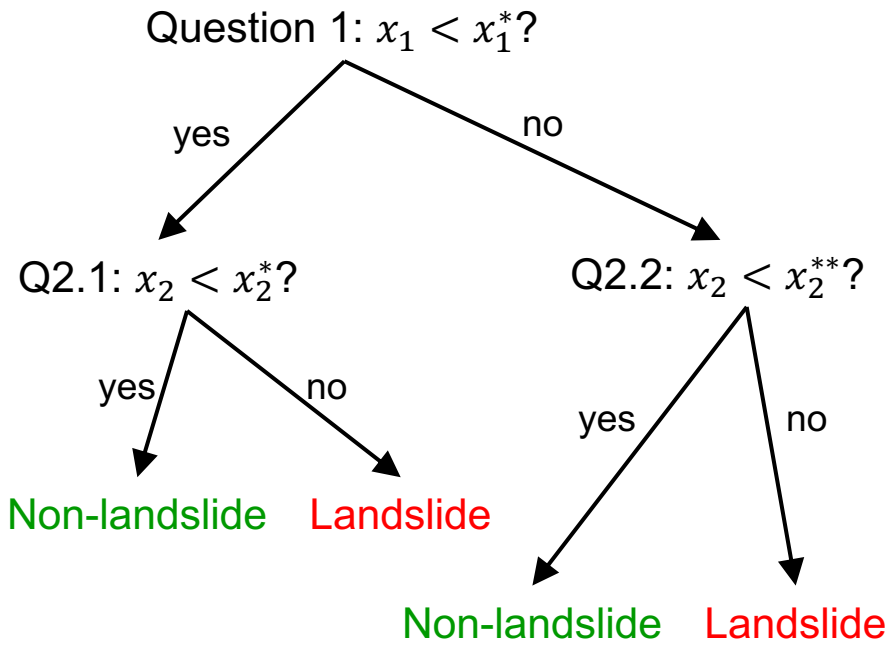
Classification Trees

- Consist of a series of yes/no questions
- Can be represented by a tree
- Equivalent to a binary partitioning of feature space
- Constructed algorithmically based on splitting and stopping rules
  - The splits are only locally optimal.
- Complete search for best tree among all possible trees is computationally not feasible.



Classification Tree

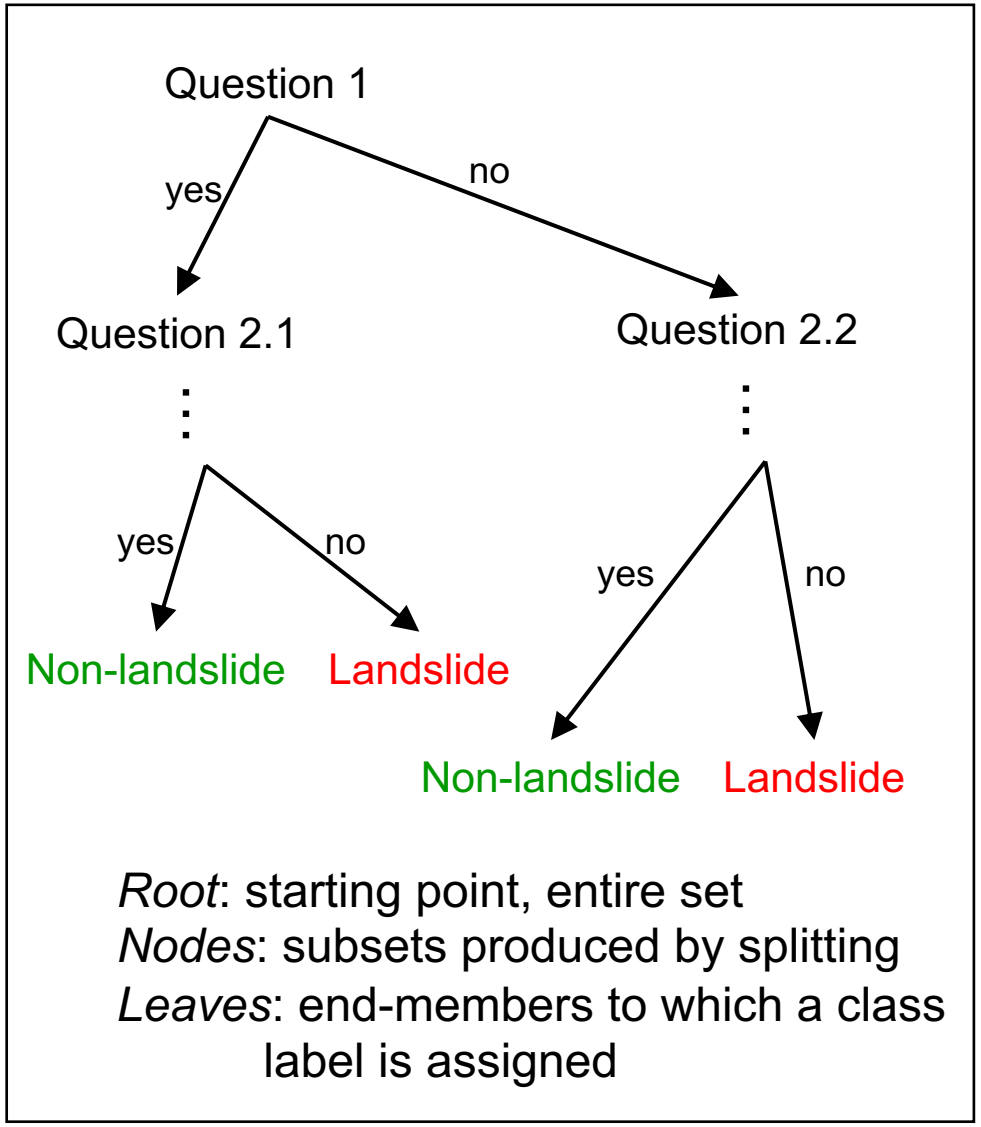




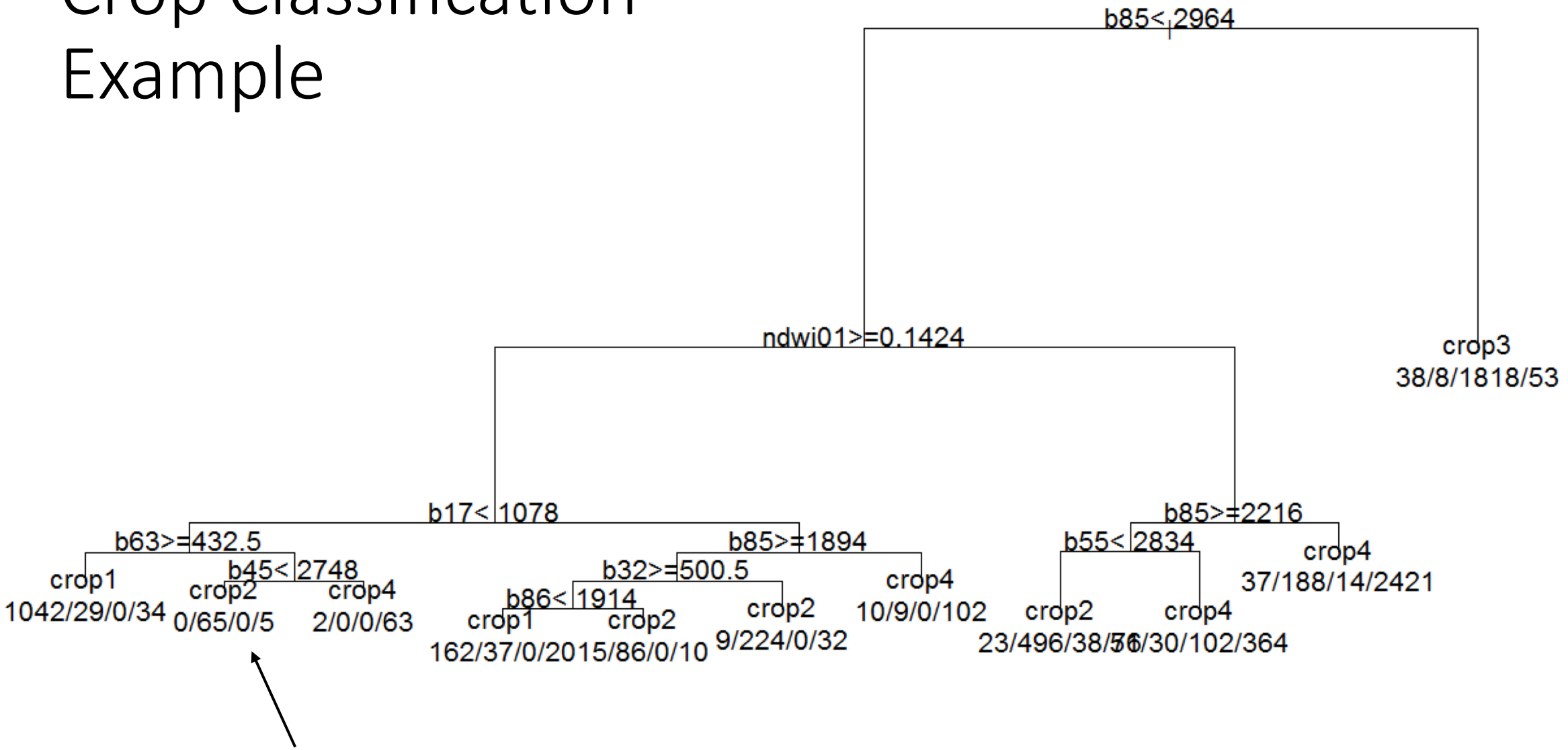
Note:
-Q2.1 and Q2.2 may use different variables
-The same variable may be used several times within a tree structure

Growing Trees

- Compare all possible splits in each variable
- Select candidate split that minimizes impurity of the nodes it generates
  - Typically measured by the Gini impurity
- Repeat this procedure recursively within each subset generated in the previous step
- Stop as soon as a node is “pure”, too small or does not improve model fit.
- **Pruning** cuts tree down to size based on its predictive performance
- Each step is independent any subsequent steps!



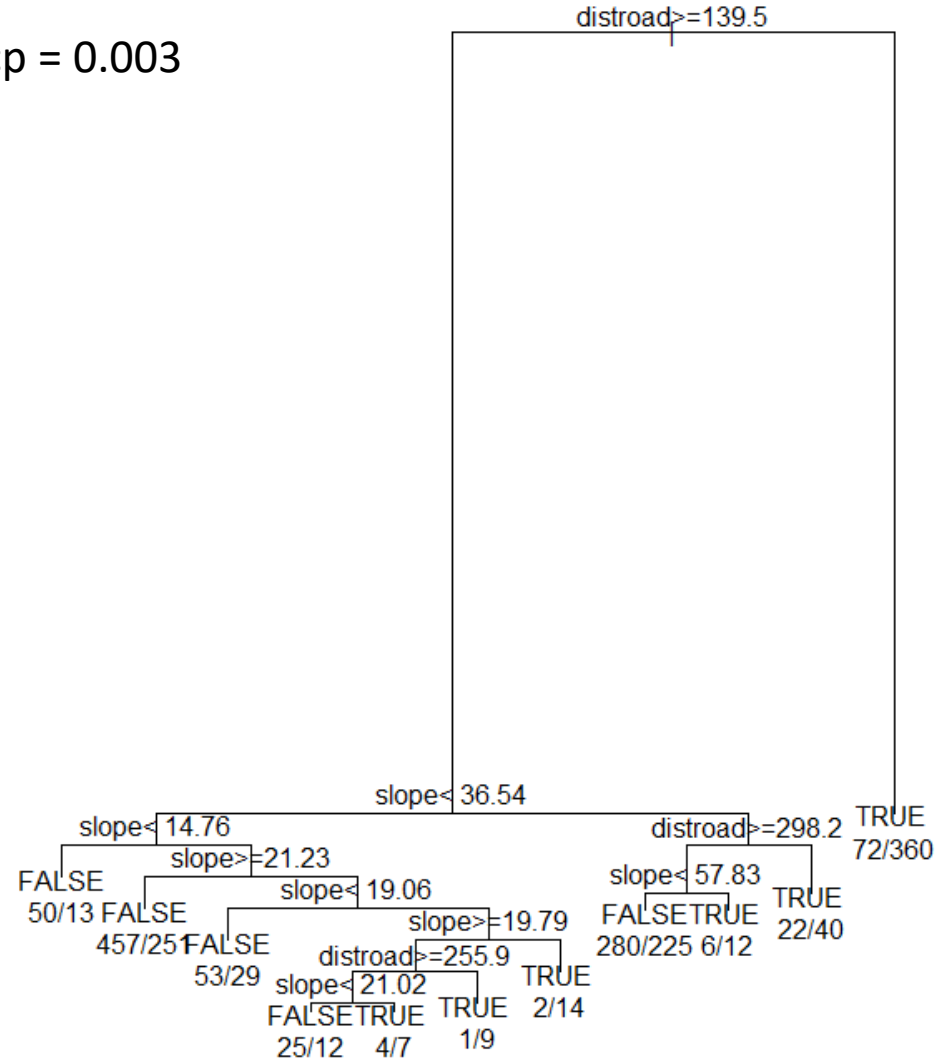
Crop Classification Example

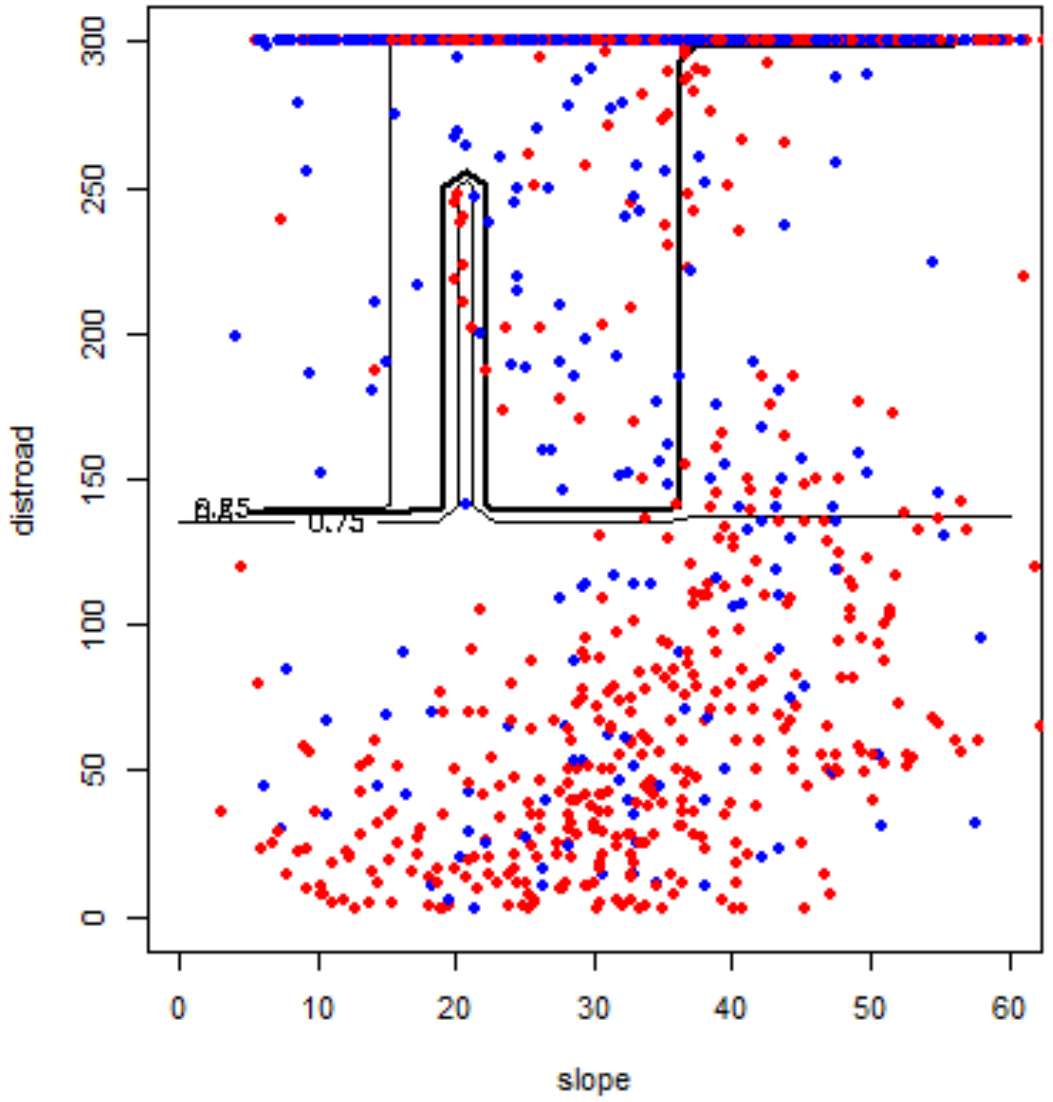


This terminal node contains 70 observations. All of them are predicted to be crop2. However, only 65 of them are in fact crop2 while 5 are crop4.

Model Predictions in Feature Space: Classification Tree

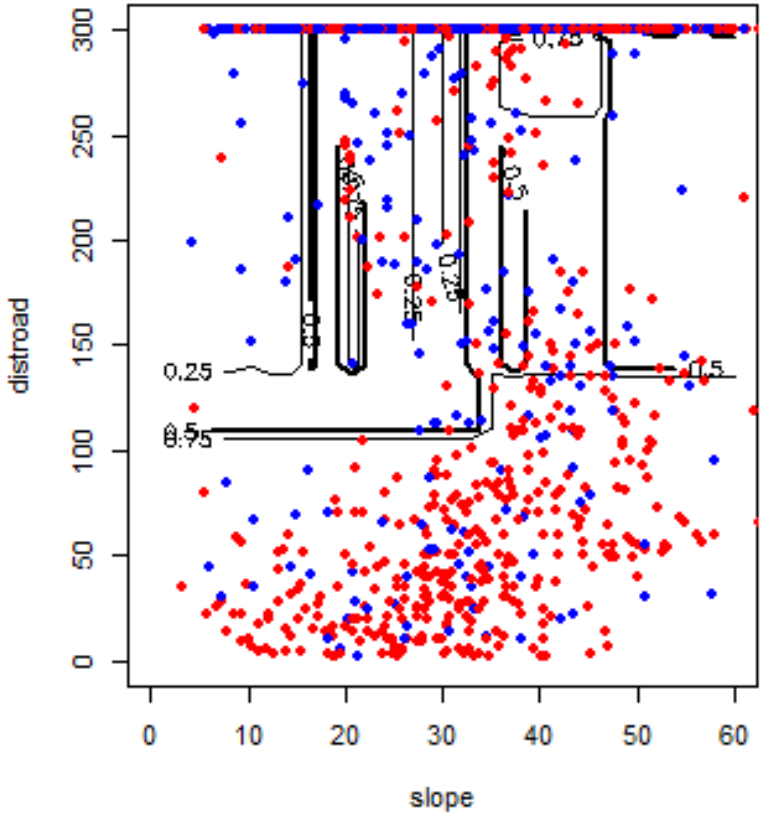
Using cp = 0.003



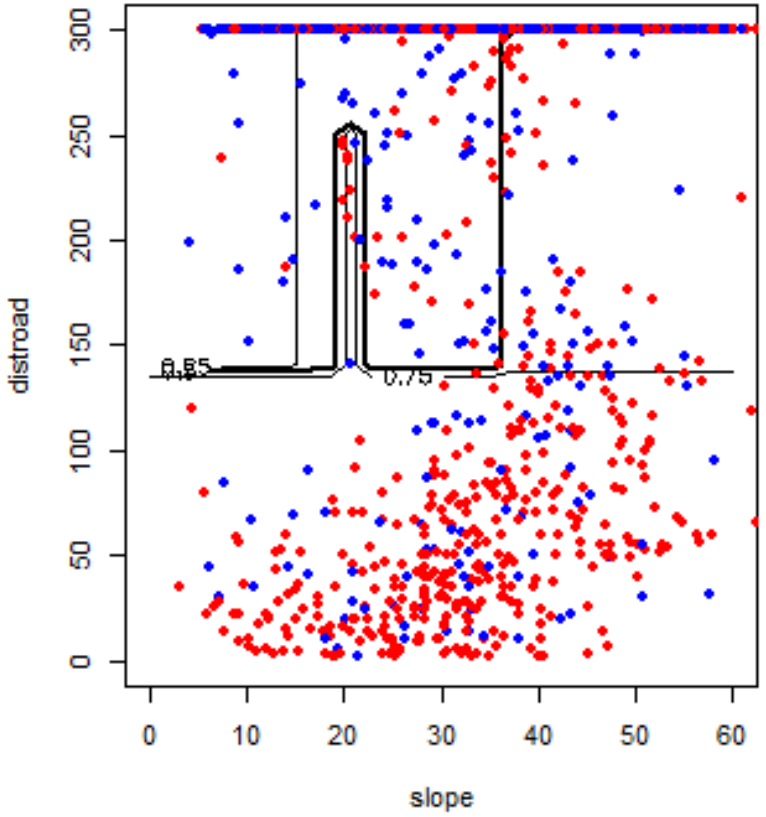


Model Predictions in Feature Space: Classification Tree

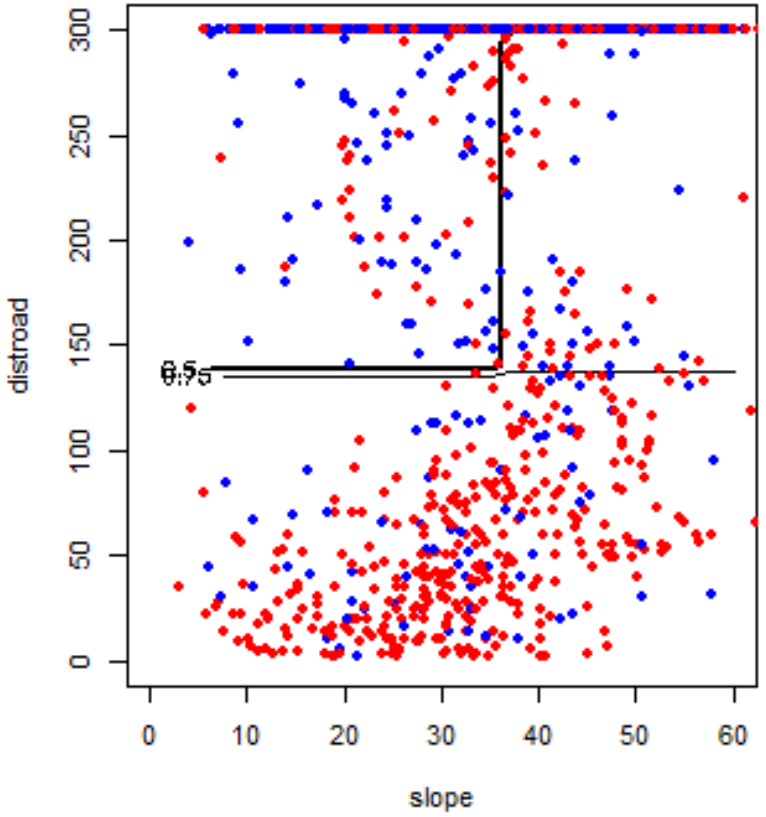
cp = 0.000



cp = 0.003



cp = 0.006



Characteristics of Classification Trees

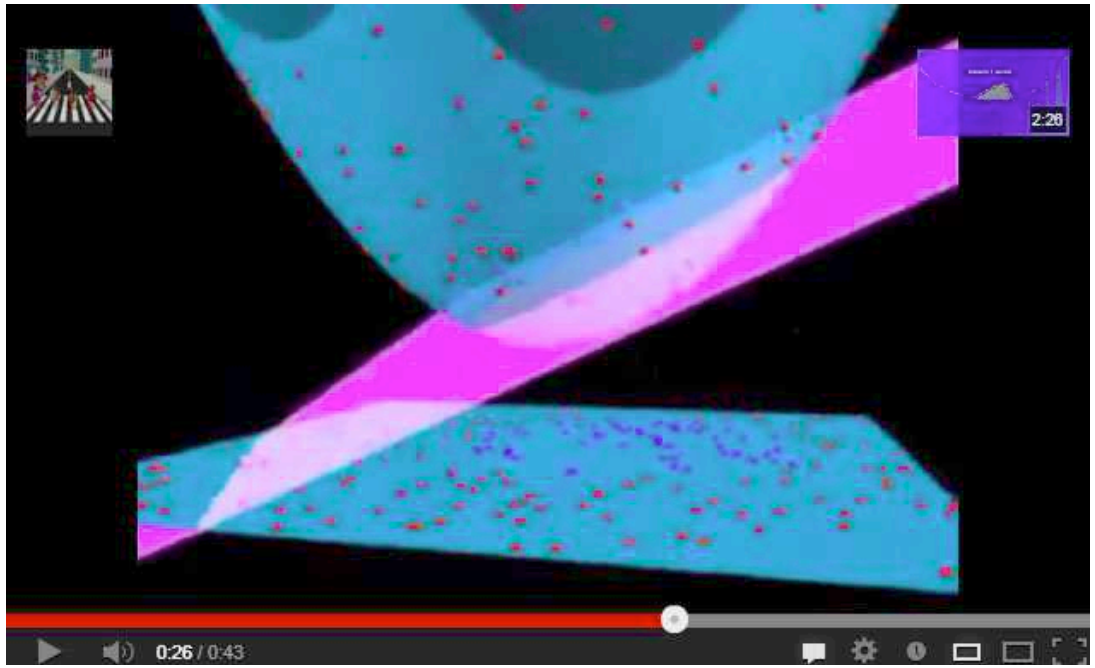
- Easy to interpret (and use) by end-user
- Can be adapted to handle missing values in predictors
- Inherently able to model nonlinearities and interactions
- Insensitive to outliers
- Suitable for quantitative, ordinal and nominal predictors

But…

- While each split is locally optimal, the splits and the whole tree are not globally optimal
- Depending on sample size and number of classes, nominal predictors may be preferred or neglected by tree algorithm
- Small changes in the training sample may change the tree completely
- No statistical significance of splits

Support Vector Machine

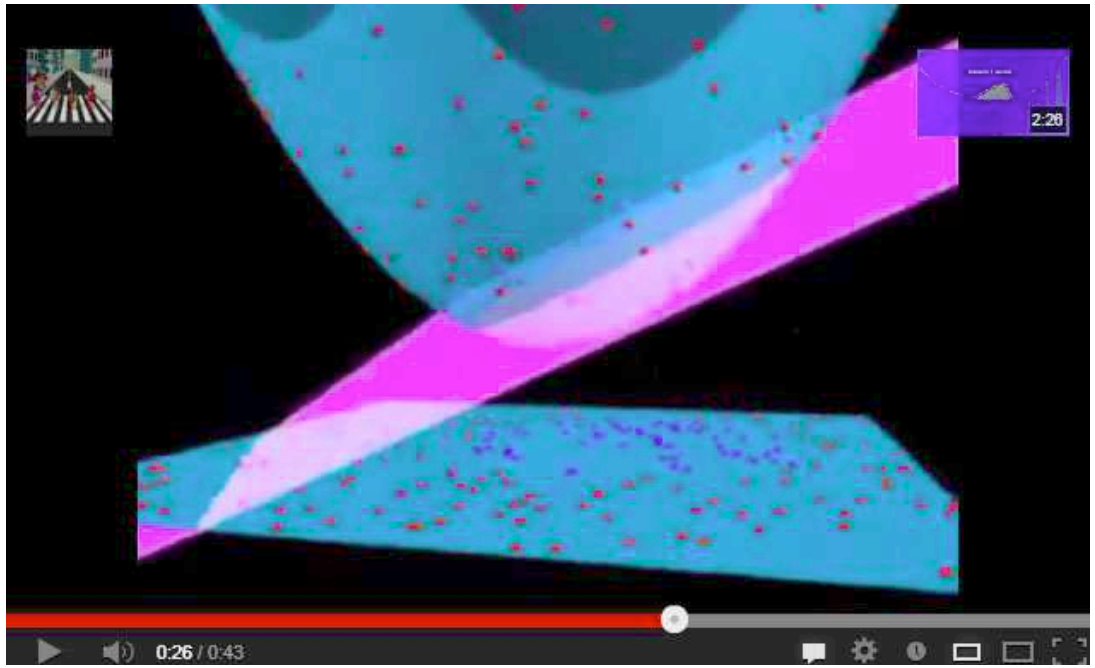
- General mathematical framework for classification and regression modelling
- Nonlinear transformations map the predictors into a higher dimensional feature space
- A separating hyperplane exists in this high-dimensional space
- Flexibility of SVM is controlled by hyperparameters



SVM with polynomial kernel provides a perfect separation of the red and blue dots. Watch the video at http://youtu.be/3liCbRZPrZA

SVM Hyperparameters

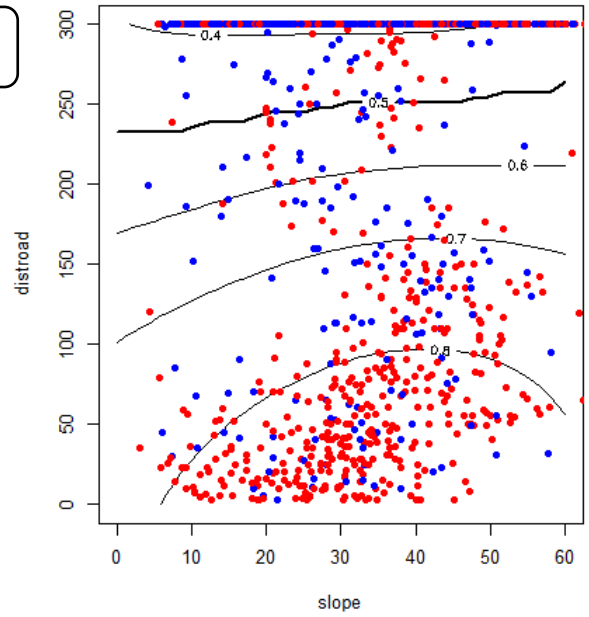
- **Kernel type**: linear, polynomial, radial basis function
- **Kernel bandwidth**, γ
- **In C-classification**, a **cost parameter** C defines the cost of misclassification: low cost → simpler model → avoids overfitting but tends to overgeneralize
- **In ν-classification**, ν defines an upper bound for the training error rate: large value → stronger generalization
- Hyperparameters should be tuned using a *nested* cross-validation

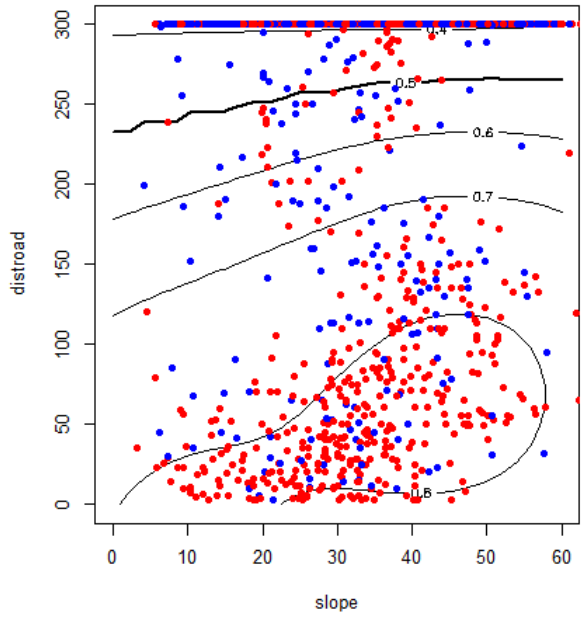


SVM with polynomial kernel provides a perfect separation of the red and blue dots. Watch the video at http://youtu.be/3liCbRZPrZA

Model Predictions in Feature Space: SVM (1)

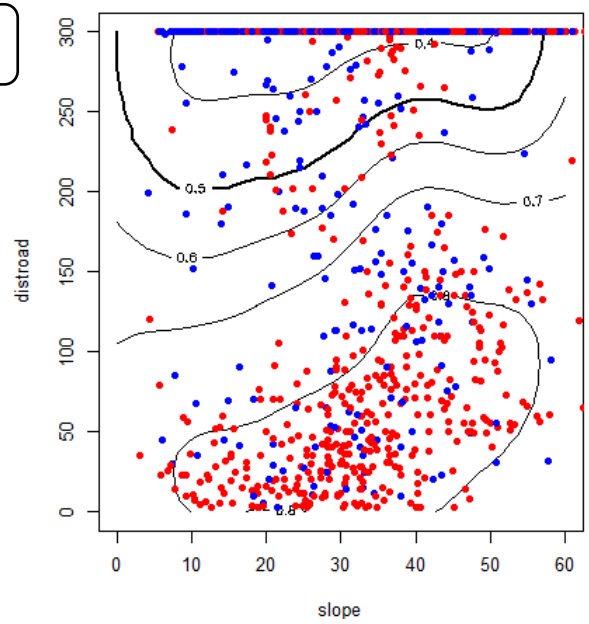
SVM (C=.1, γ=.1)

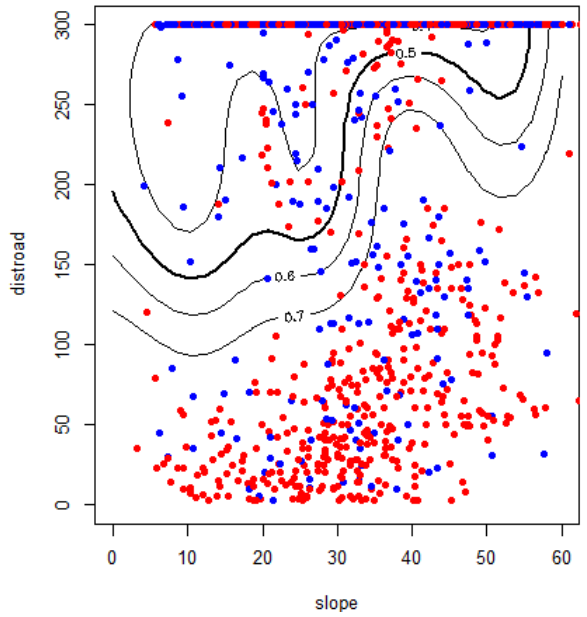




SVM (C=1, γ=.1)

SVM (C=.1, γ=1)



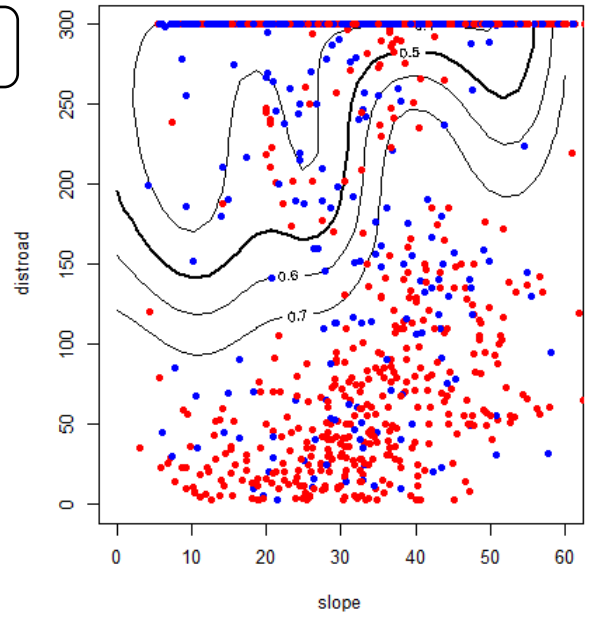


SVM (C=1, γ=1)

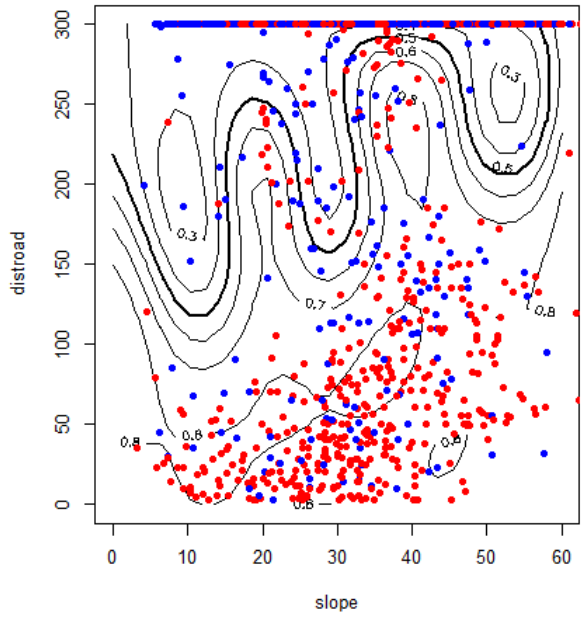
C-classification,
radial basis function kernel

Model Predictions in Feature Space: SVM (2)

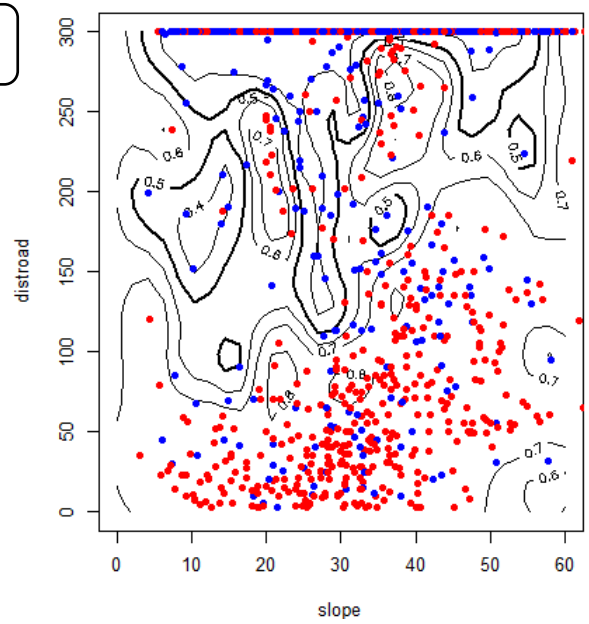
SVM (C=1, γ=1)



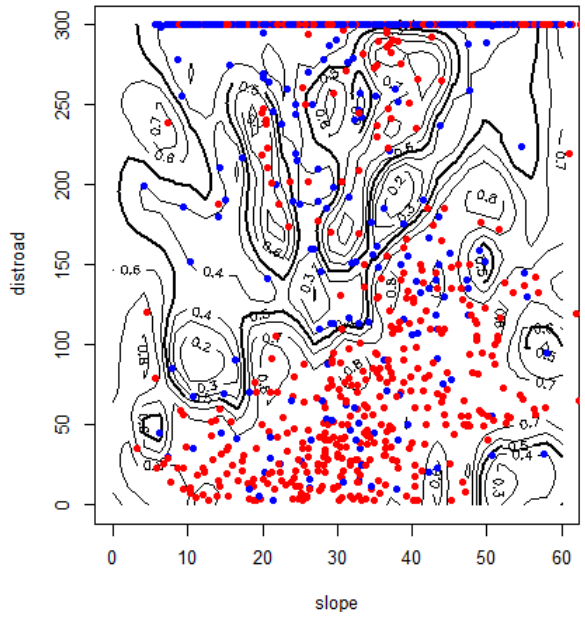
SVM (C=10, γ=1)



SVM (C=1, γ=10)



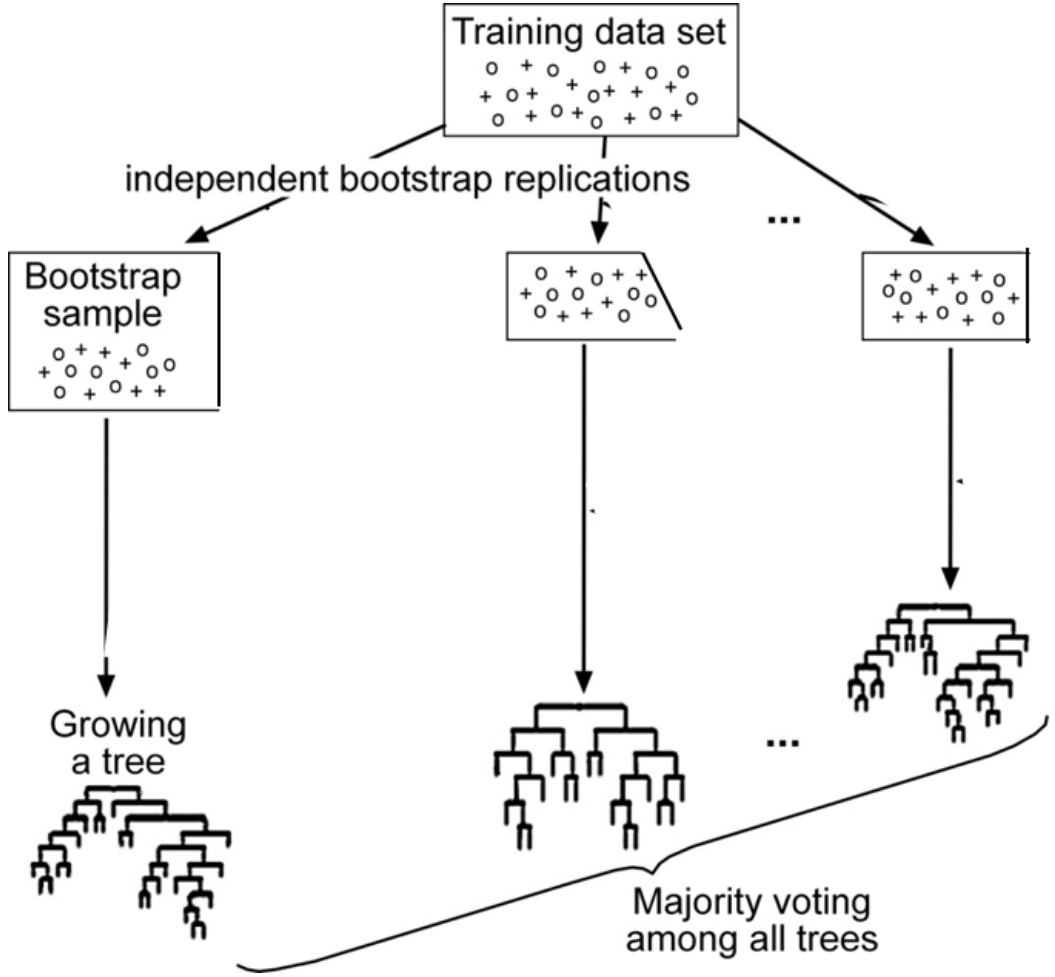
SVM (C=10, γ=10)



C-classification,
radial basis function kernel

Bagging and Random Forest

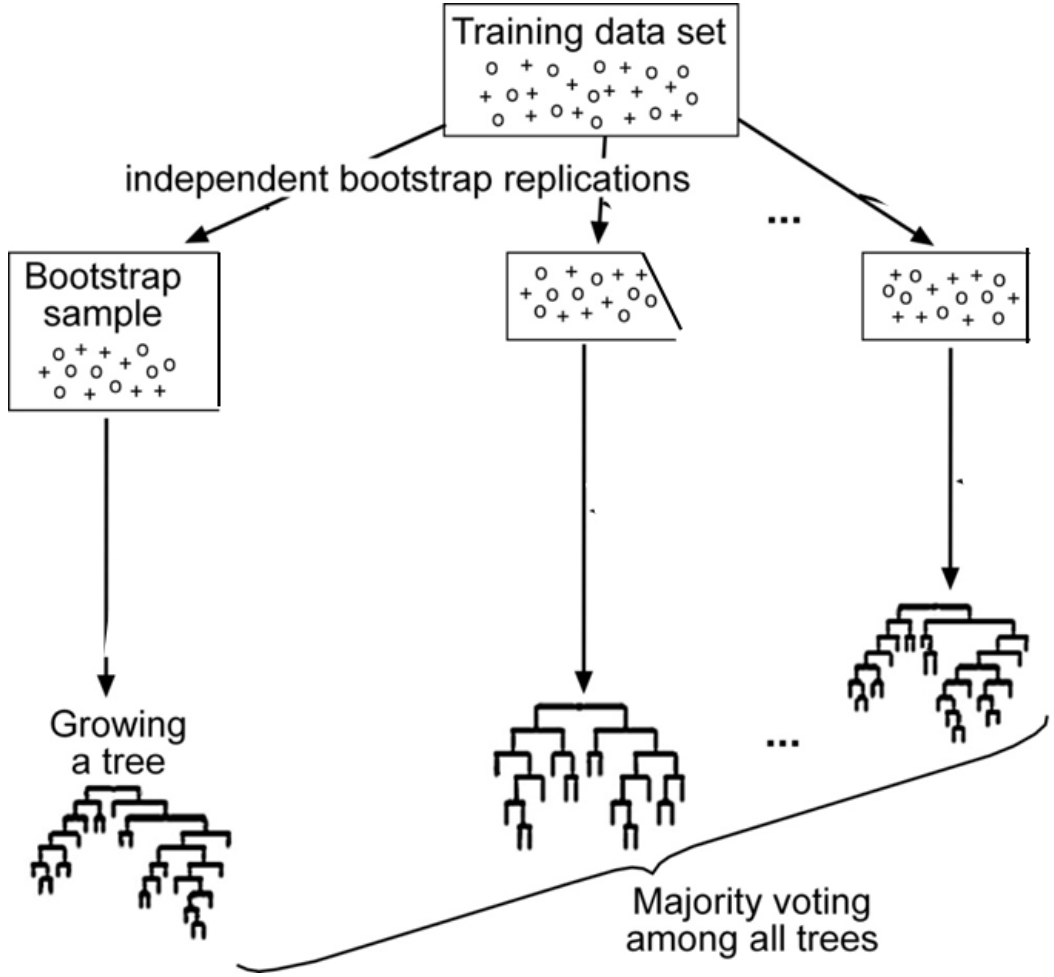
- Individual trees are unstable → combine many trees!
- Generate (e.g.) 100 **bootstrap samples** by drawing N out of N observations with replacement.
- Grow a tree on each bootstrap sample.
- To make bagging predictions for a new object, drop the object down each tree to obtain 100 predictions, and combine these predictions by majority voting.
  - **Bagging** = bootstrap aggregating
  - Alternatively, use fraction of votes as "soft" prediction.



Bagging and Random Forest

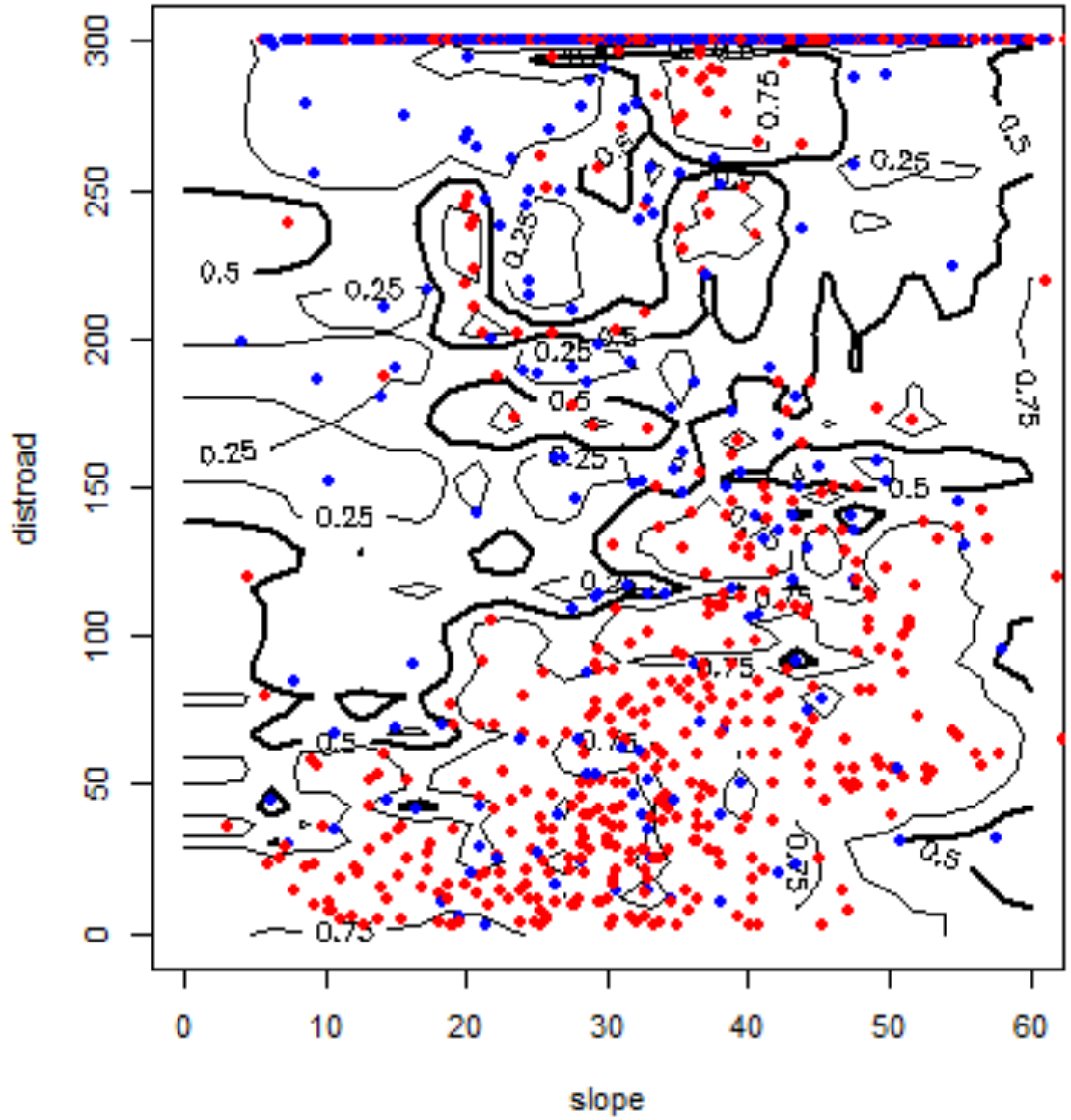
- Breiman (1999): “Bagging goes a way toward making a silk purse out of a sow’s ear, especially if the sow’s ear is twitchy.”
  - Improves poor (unstable) methods, but may make good methods worse

- **Random forest** (Breiman, 2001) modifies bagging to also consider only a random subset of predictors.
  - Avoids over-reliance on dominant predictors



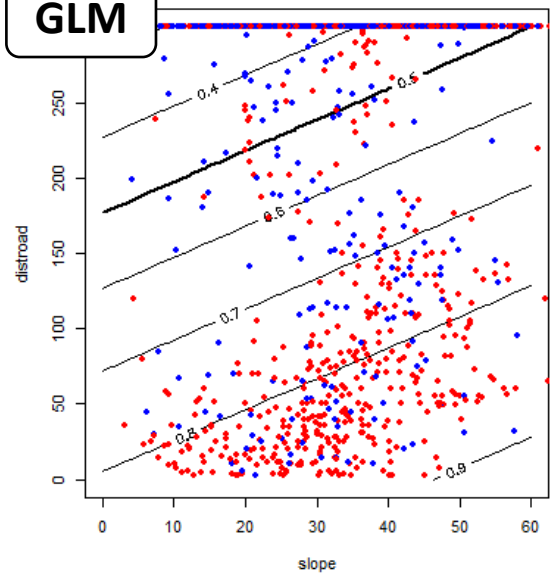
Model Predictions in Feature Space: Random Forest

Using ntree = 5000 to reduce random variability – but still pretty noisy…

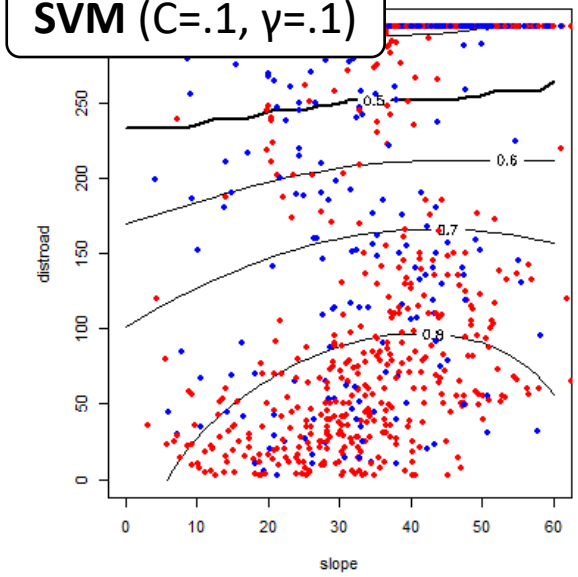


Model Predictions in Feature Space: Comparison

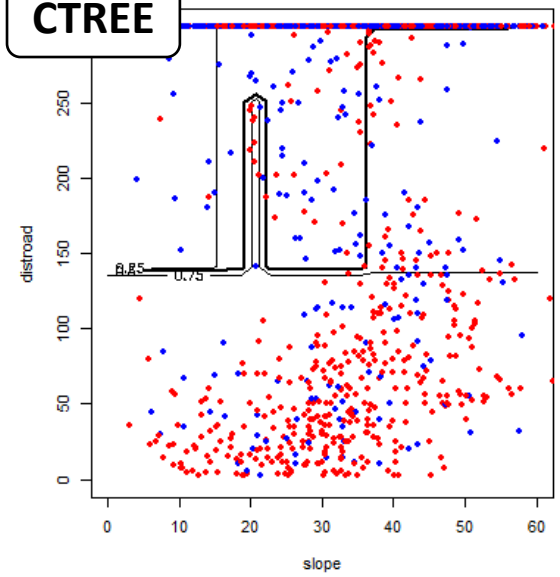
GLM



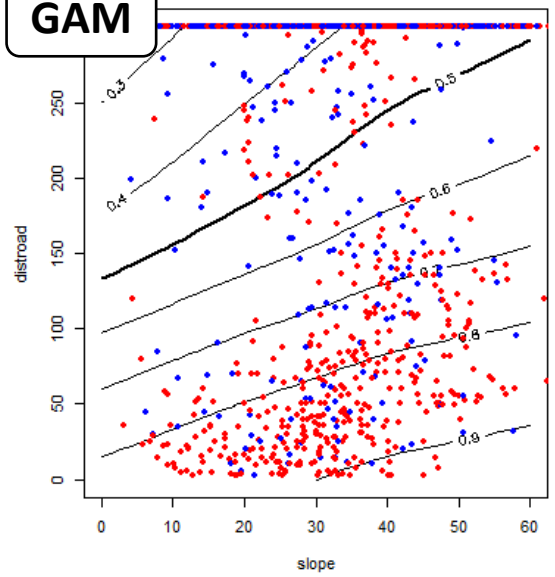
SVM (C=.1, γ=.1)



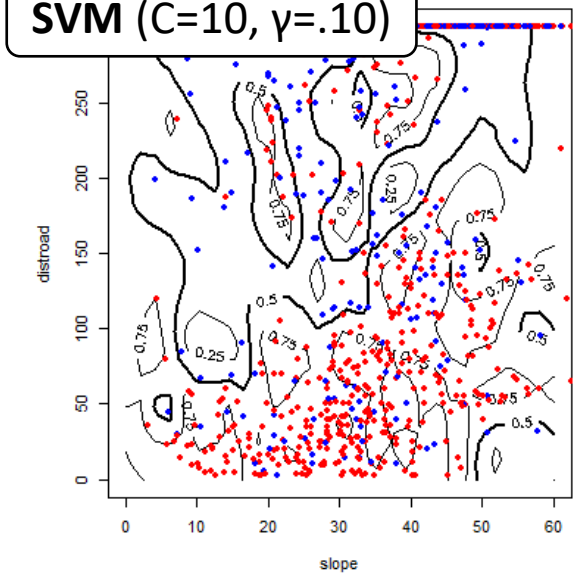
CTREE



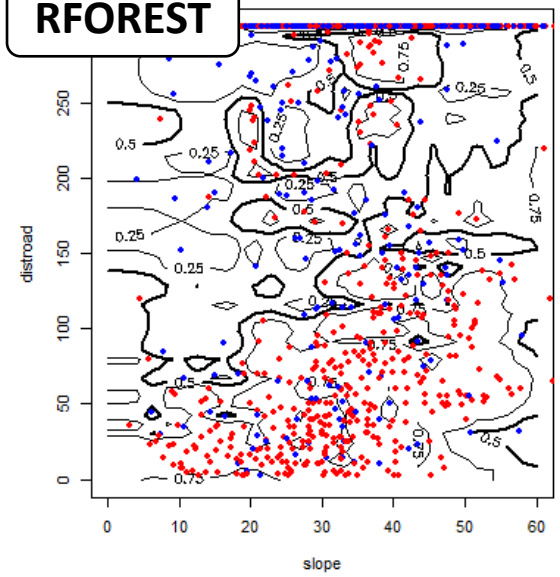
GAM



SVM (C=10, γ=.10)



RFOREST



Lessons Learned

- In **predictive modelling**, we can be pragmatic about the type of model used – as long as it provides good predictions.
  - Prediction models are often poorly interpretable.
  - More flexible models tend to overfit to the training data.
  - We need to learn more about model assessment and tuning and the interpretation of black-box models.

- In **spatial analysis**, use “transparent” models that allow you to…
  - tell a story about the data
  - quantify contrasts
  - perform statistical inference
  - incorporate spatial (or other) autocorrelation

And Remember…

All models are wrong, but some are useful

George E. P. Box (1919-2013)

References

- Brenning, A., Schwinn, M., Ruíz-Páez, A.P., Muenchow, J. (2015). Landslide susceptibility near highways is increased by 1 order of magnitude in the Andes of southern Ecuador, Loja province. *Natural Hazards and Earth System Sciences*, 15: 45-57.
- 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.