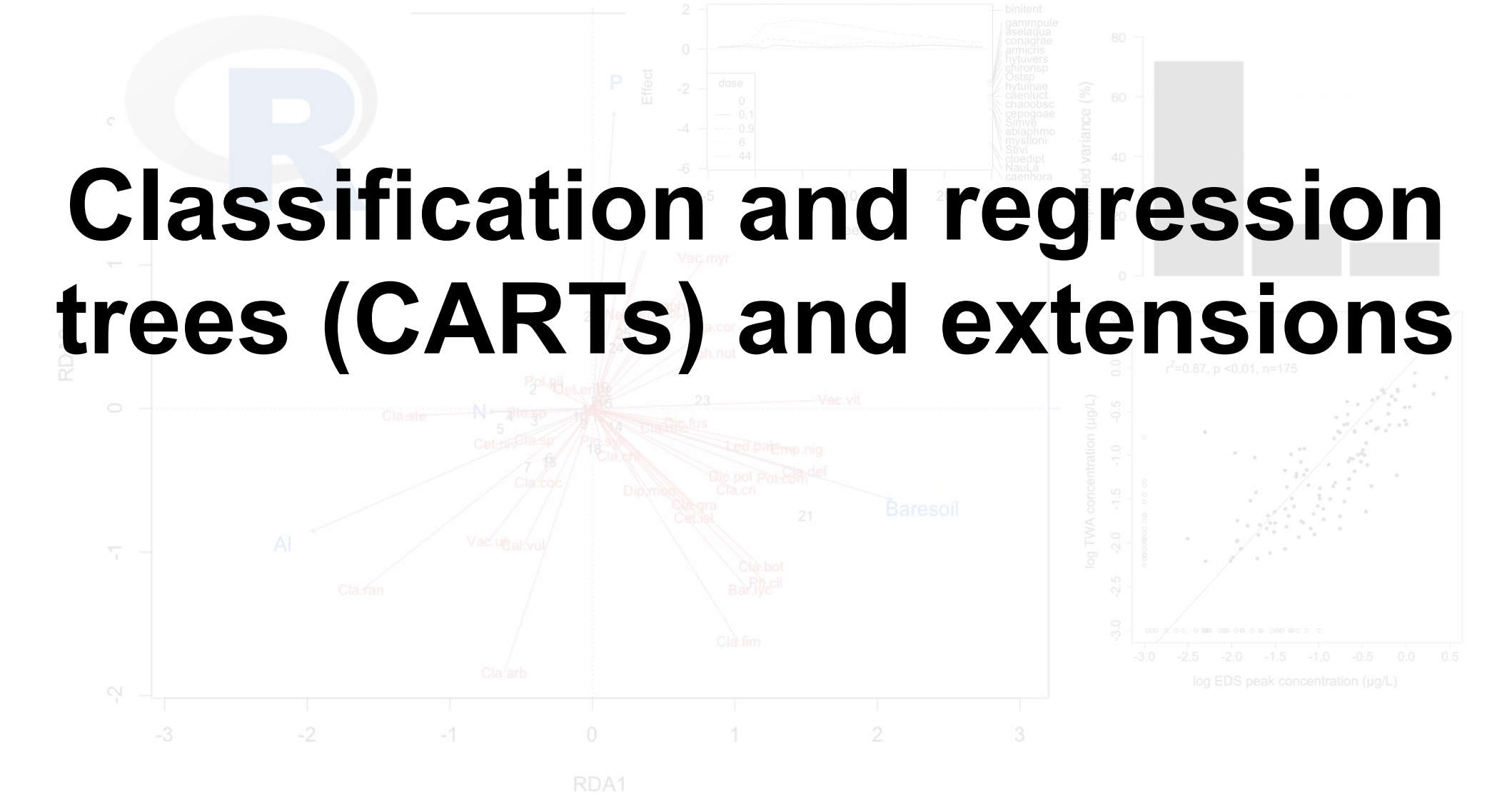
Tools for complex data analysis

University of Koblenz-Landau 2018/19



Classification and regression trees (CARTs) and extensions

Ralf B. Schäfer

Learning targets and study questions

- Knowledge on classification and terminology of machine learning methods.
- Comprehension of Classification and regression trees.
- Knowledge on extensions of CARTs.

Learning targets and study questions

- Knowledge on classification and terminology of machine learning methods
  - Describe the difference between supervised and unsupervised learning.
- Comprehension of Classification and regression trees.
  - Describe the differences between classification and regression trees.
  - Summarize the procedure for creating a tree.
  - Explain the role of impurity metrics and the cost-complexity parameter in this context.
  - Outline the different types of interactions in CARTs.
  - What is a surrogate variable?
  - What are major differences to (G)LMs and under which conditions should they be used?
  - Discuss advantages and disadvantages of CARTs.

Learning targets and study questions

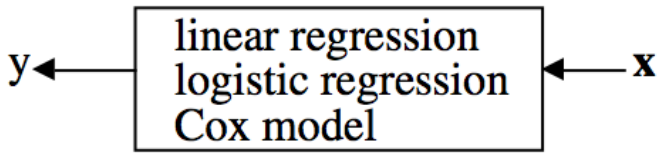
- Knowledge on extensions of CARTs.
  - Briefly discuss a few extensions of CARTS and their rationale.
  - Describe the application domain of multivariate regression trees.
  - Discuss major advantages and disadvantages of random forest compared to building a single tree.

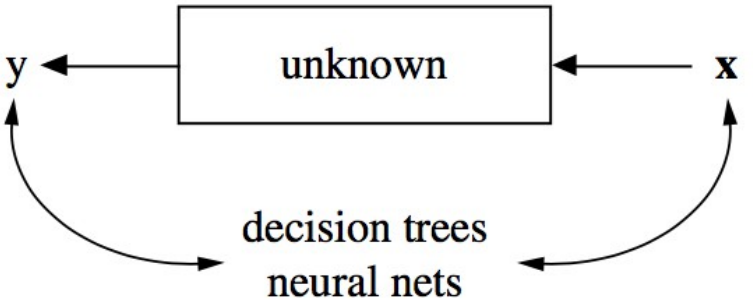
Classification and regression trees and extensions

Contents

Supervised and unsupervised learning

- Machine learning terminology
- Supervised learning:
  - Use statistical model or algorithm to map predictors x to response y

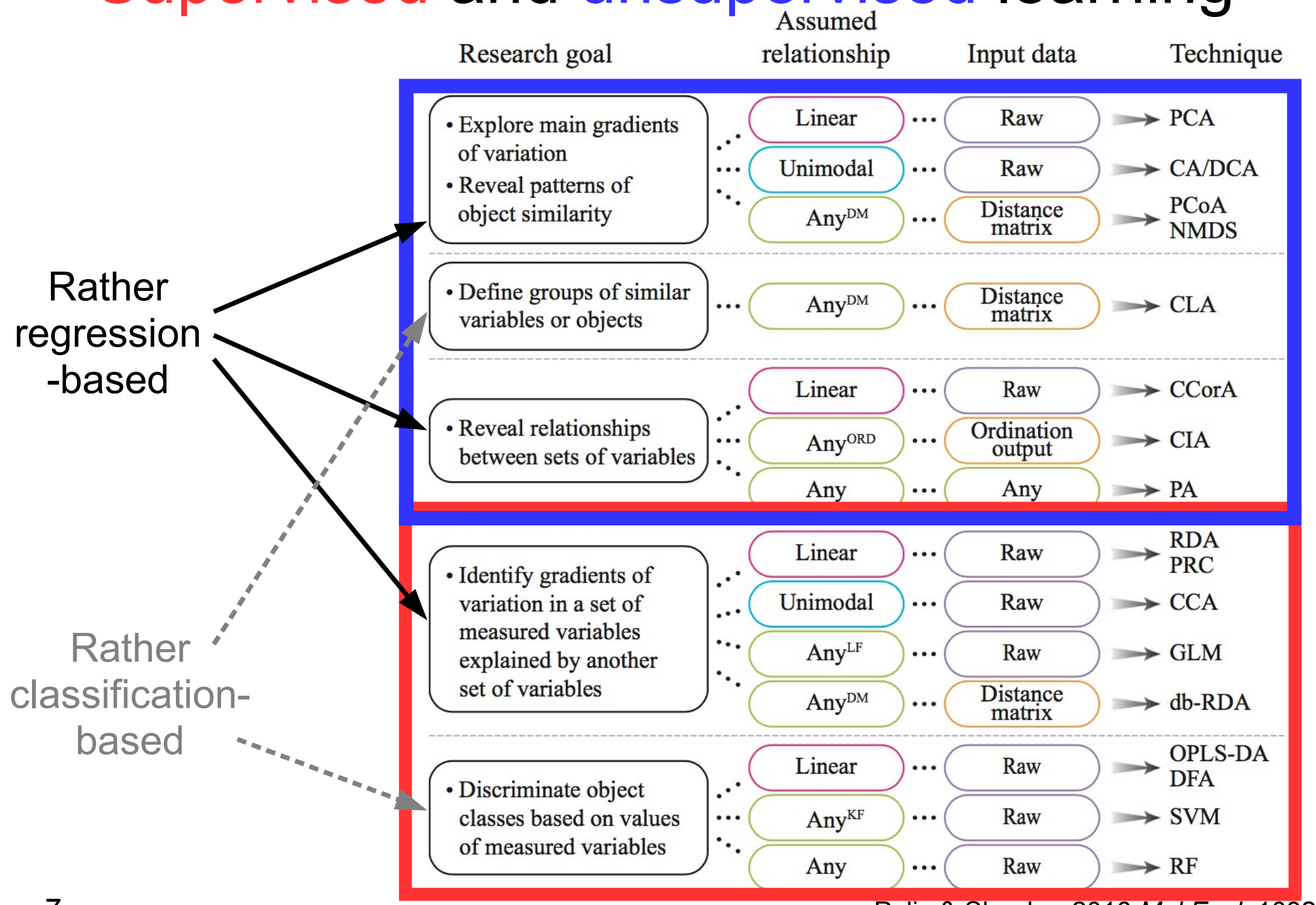




- Unsupervised learning:
  - Identify (hidden) patterns or associations within a data set without "ground truth"
- Unsupervised/supervised classification and regression

Breiman 2001 *Statistical Science* 16: 199

Supervised and unsupervised learning



Paliy & Shankar 2016 *Mol Ecol*: 1032

Supervised classification and regression

- Response (e.g. membership of observations to groups) known
- Aim: Identification of classification or regression rules (mainly for explanation or prediction)

Methods include classification and regression trees, discriminant analysis, artificial neural networks, but also our well known (generalized) linear regression models

Classification and regression trees (CART)

- Machine learning technique and method of recursive partitioning
- Used in prediction or explanation of a response variable through the construction of a decision tree

Breiman et al. 1984

Classification and regression trees and extensions

Contents

1. Supervised and unsupervised learning
2. **Intro: Classification and regression trees**
3. Goodness of split metrics, interactions, assumptions and predictor importance
4. Critical evaluation and comparison to (G)LMs
5. Extensions and alternatives 1: Conditional inference trees and multivariate trees
6. Extensions and alternatives 2: Random forest and boosted regression

Example: Which habitats are suitable for the turkeybeard *Xerophyllum asphodeloides*?

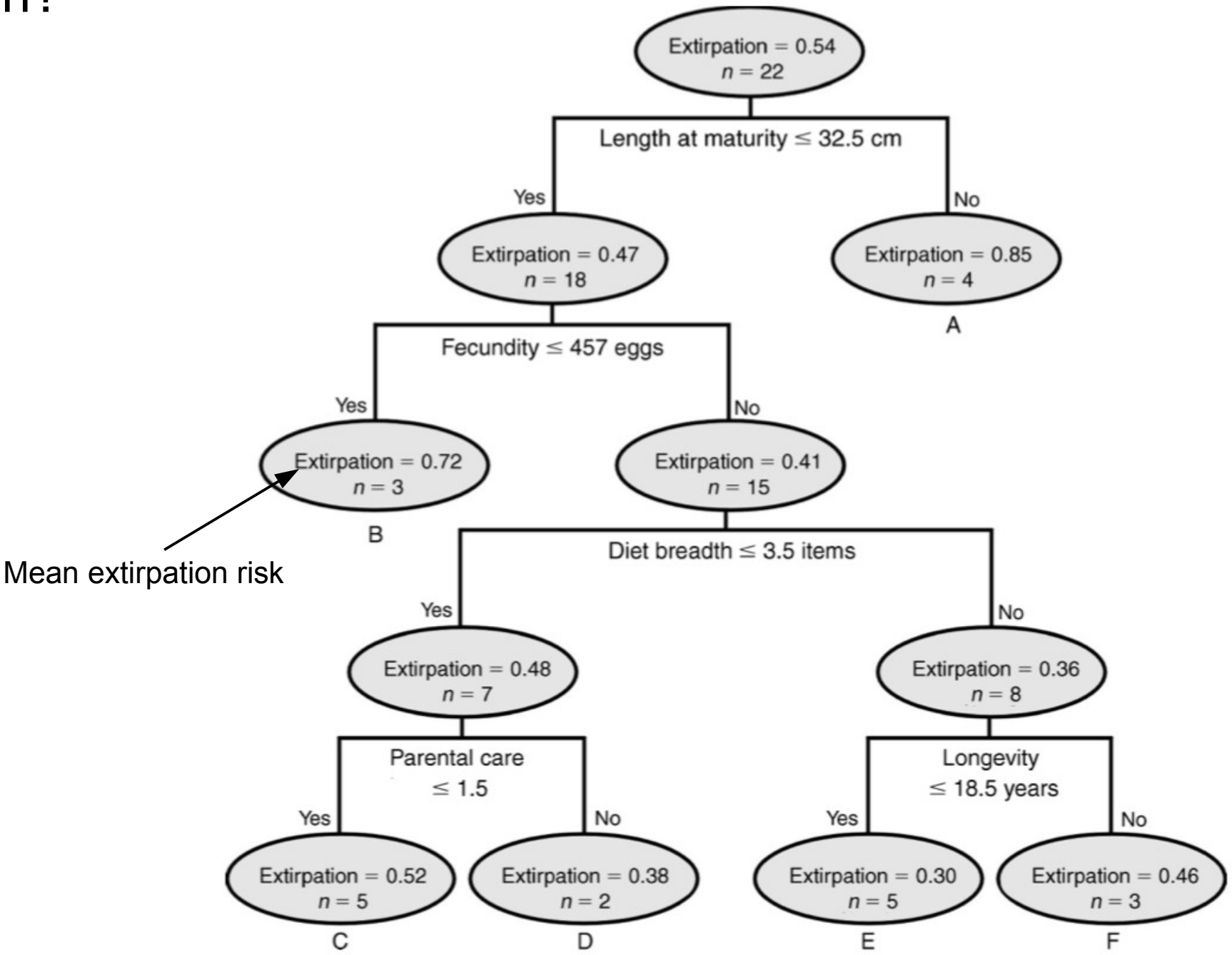
https://www.flickr.com/photos/blueridgekitties/5801875577

U = Unsuitable
Su = Suitable

139 U/ 14 Su

Bourg et al. 2005 *Ecology* 86: 2793

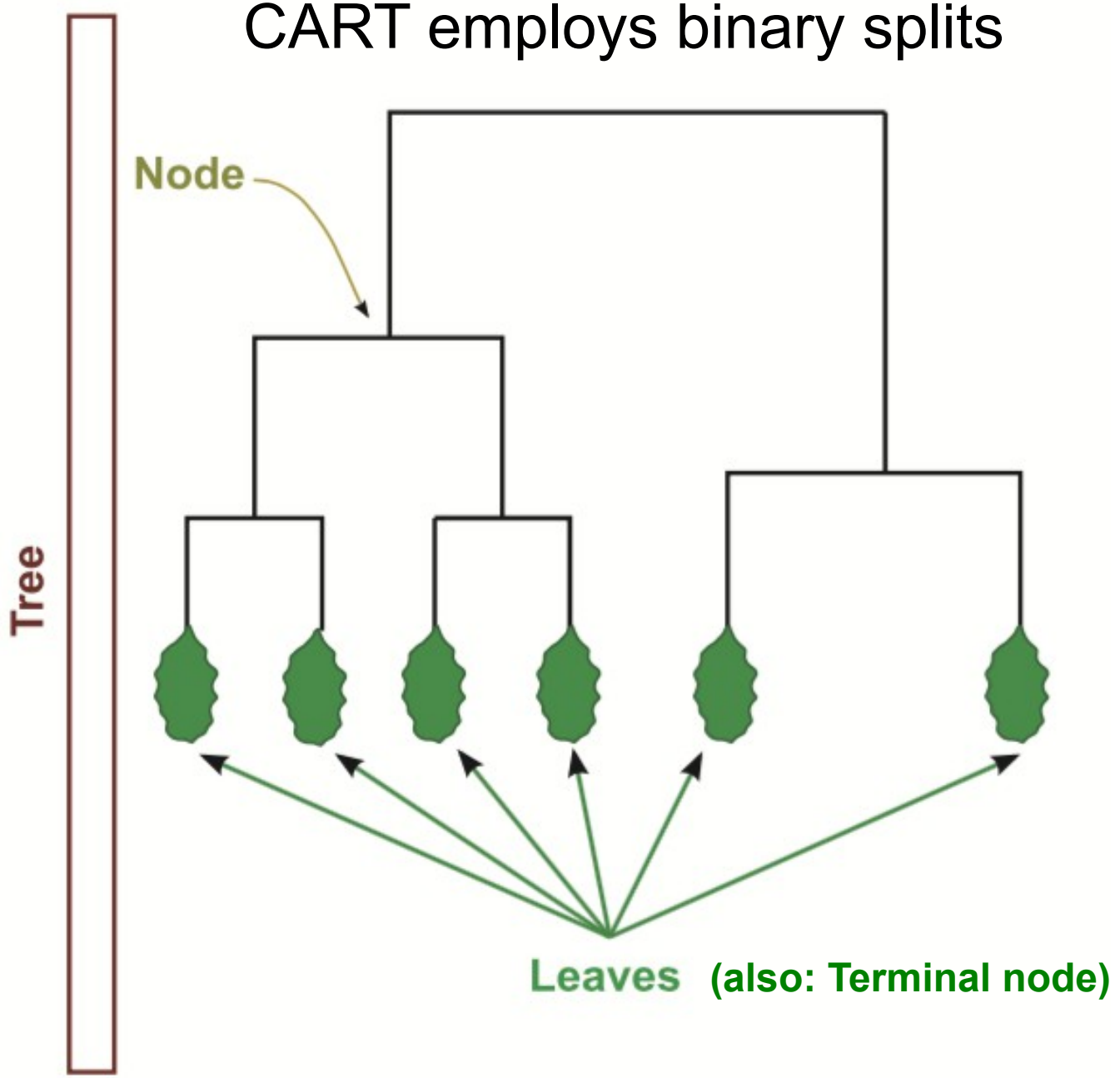
Example 2: Which traits determine the extinction risk of desert fish?



Olden et al. 2008 *Ecology* 89: 847

CART terminology

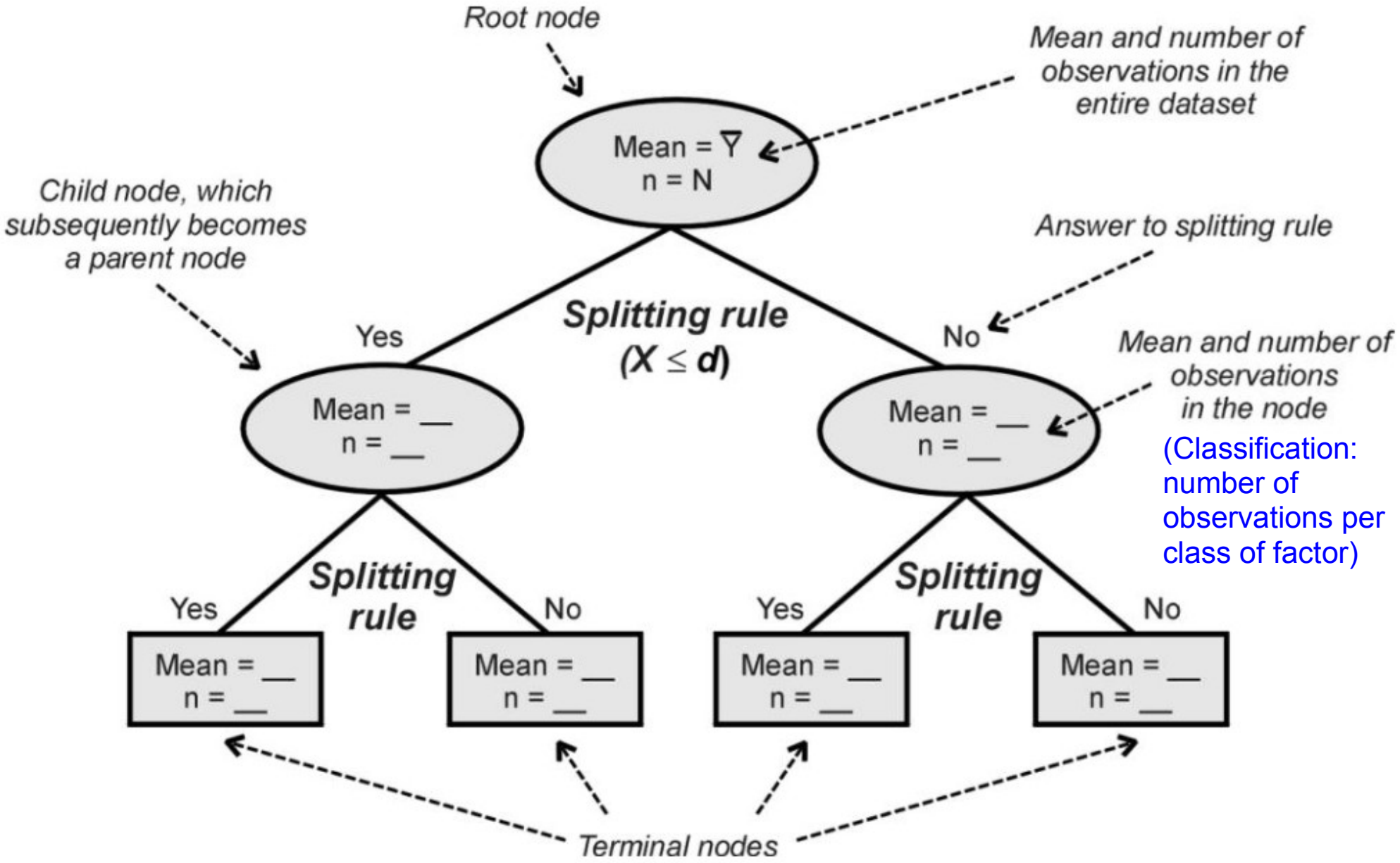
CART employs binary splits



Ouellette et al. 2012 *Methods Ecol. Evolut.* 3: 234

CART terminology and interpretation

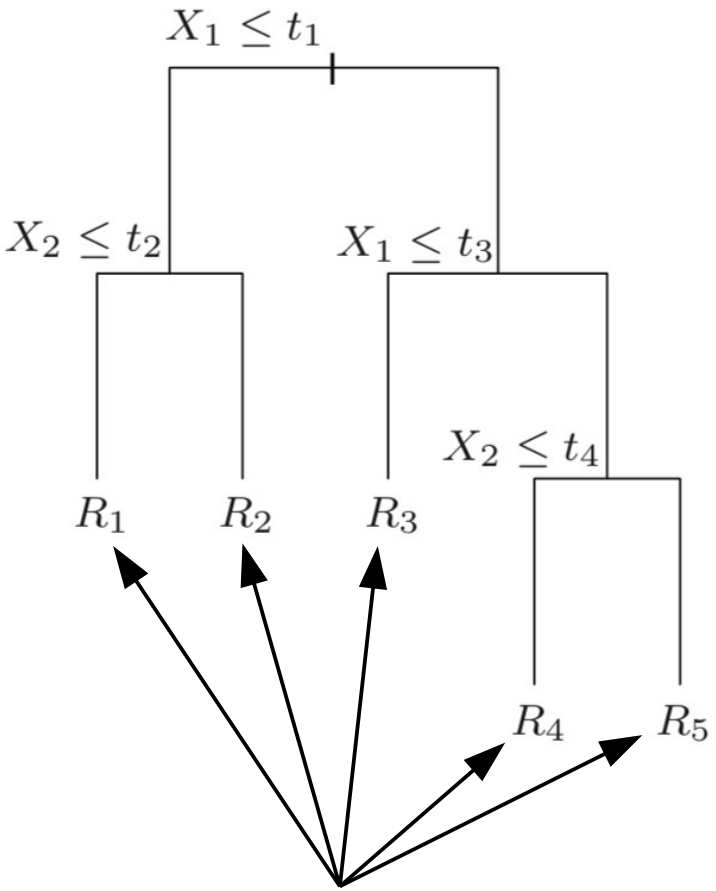
Described for regression tree



Olden et al. 2008 *Quart. Rev. Biol.* 83: 171

CART interpretation for two predictors

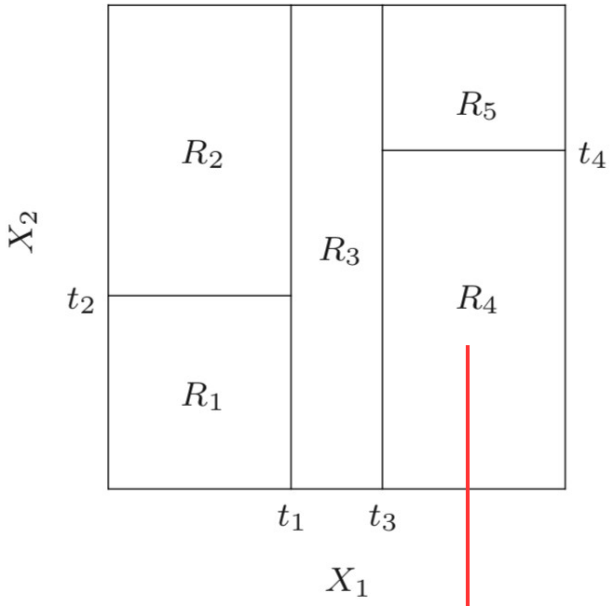
Splitting rule s is given at each node i and splits observations according to values t of predictors X_1 and X_2



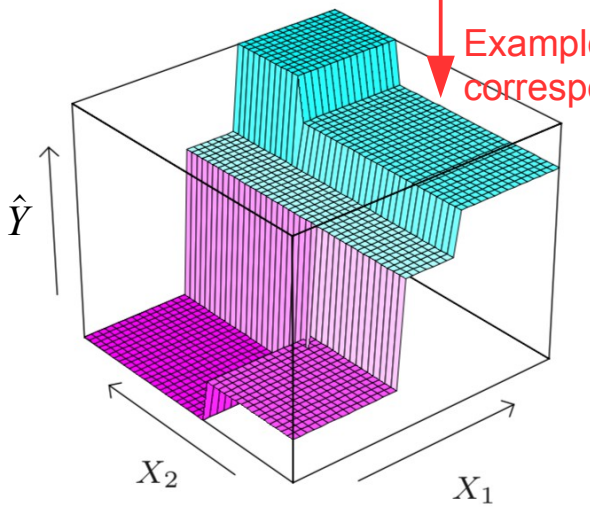
R_m: Region m of X_1 and X_2

Value of R_m: estimate of response $Y \rightarrow \hat{Y}$

Two-dimensional representation



Three-dimensional representation



Example for correspondence

Regions are assigned an estimate of $\hat{Y}$

Figures modified from Hastie et al. (2017: 306)

General procedure

1. Tree building

Sequential (binary) splitting into nested groups. Splitting is done to minimise impurity (e.g. Deviance)

→ Impurity or Goodness of split metrics

2. Stopping tree building

Implemented through three criteria:

a) Definition of a minimum number of observations in a node

→ No further splitting possible

b) All observations inside node have identical distribution

→ No further splitting possible

c) Defining a maximum number of splits

3. Pruning – Reducing size of tree

Based on complexity parameter that penalises increasing size (in conjunction with cross-validation error)

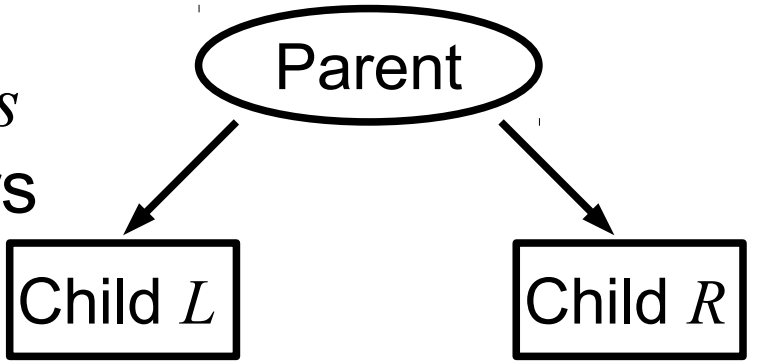
Classification and regression trees and extensions

Contents

1. Supervised and unsupervised learning
2. Intro: Classification and regression trees
3. **Goodness of split metrics, interactions, assumptions and predictor importance**
4. Critical evaluation and comparison to (G)LMs
5. Extensions and alternatives 1: Conditional inference trees and multivariate trees
6. Extensions and alternatives 2: Random forest and boosted regression

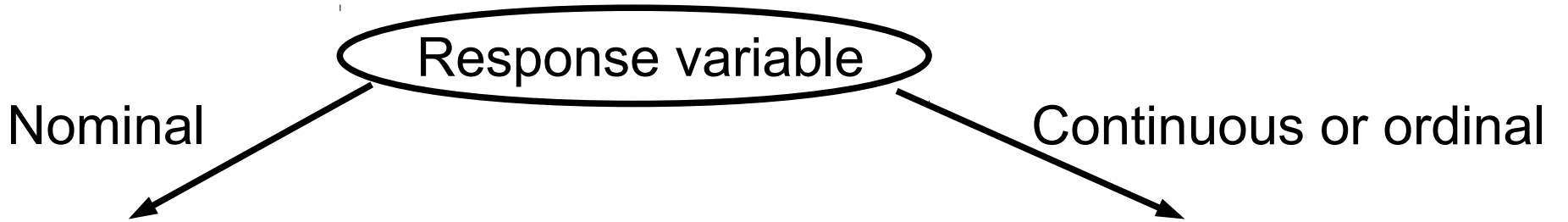
How to measure impurity?

- Deviance D
- During tree building, all possible splits s of node i are calculated for all predictors j and the purest is selected:



$$\arg\max_{s,j} \Delta D_i = D_{i,parent} - D_{i,child} \quad \text{with } D_{i,child} = D_{i,s,j,L} + D_{i,s,j,R}$$

Calculation of D



Classification tree: Gini index

$$D_i = \sum_{k=1}^{K} p_k(1-p_k)$$

K = number of classes of response

$\hat{p}_k$ = estimated proportion of observations in class k

Regression tree: MSE

$$D_i = \frac{1}{n_i}\sum_{k=1}^{n_i}(y_k - \hat{\mu}_i)^2$$

n_i = number of observations in parent or child node i

y_k = value of response of observation k

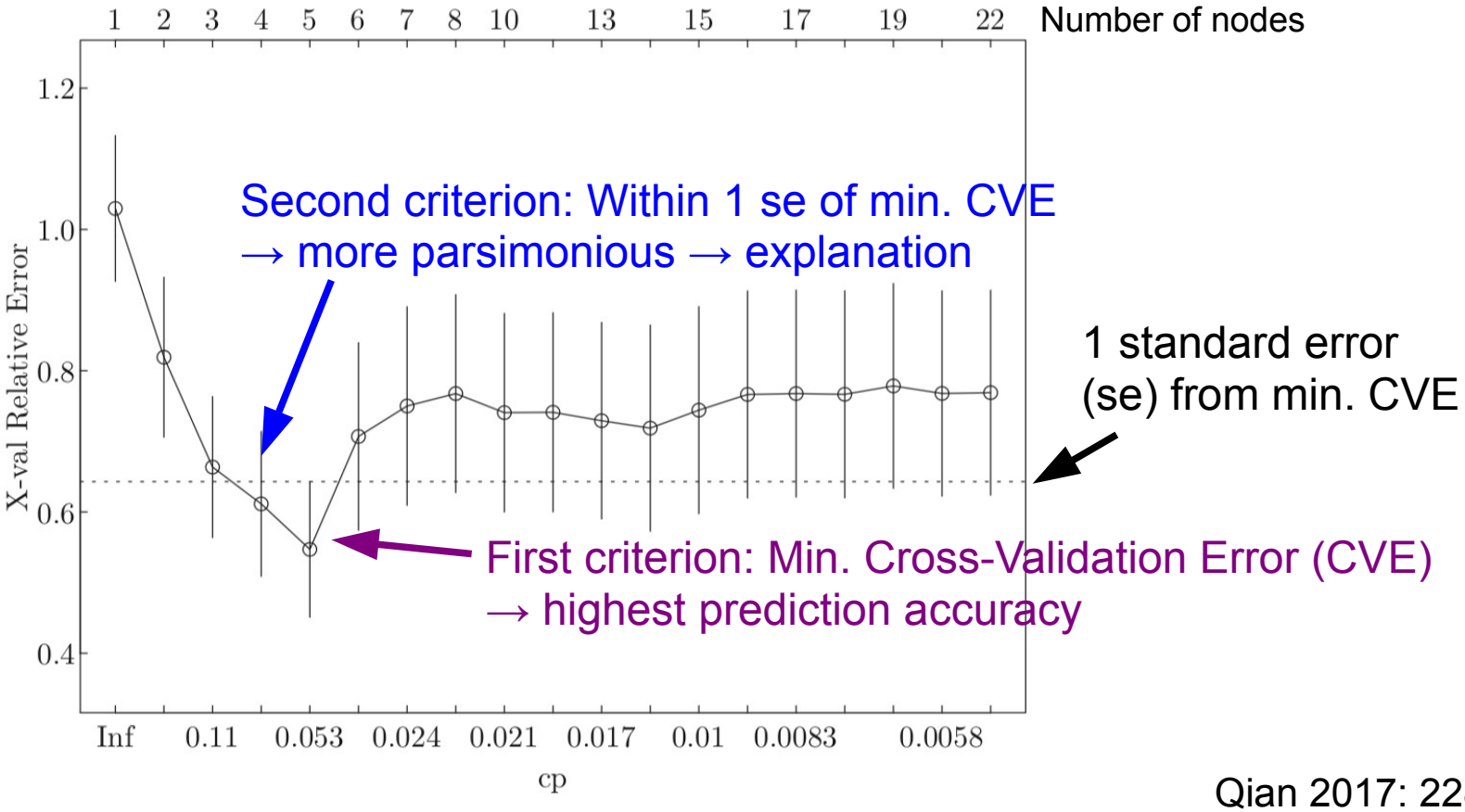
$\hat{\mu}$ = estimated mean of response for parent or child node i

How to select the optimal reduction in tree size?

- Deviance D is penalised by cost-complexity (cp) parameter

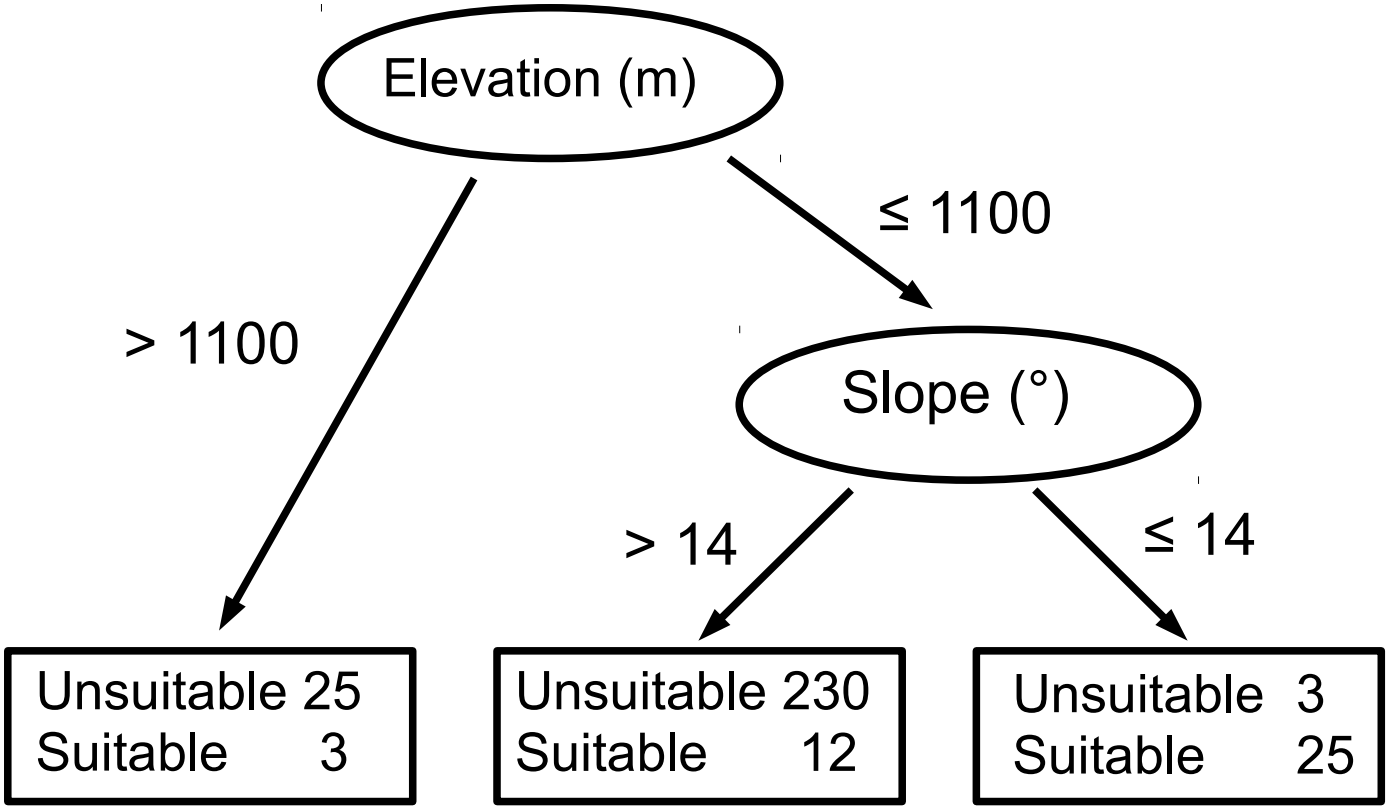
$$D_{cp} = D + cp \times \text{size-of-tree}$$

- Same rationale as information-theoretic criteria (e.g. AIC)
- Optimal tree size determined by cp. How to set cp?
  → Cross-validation!



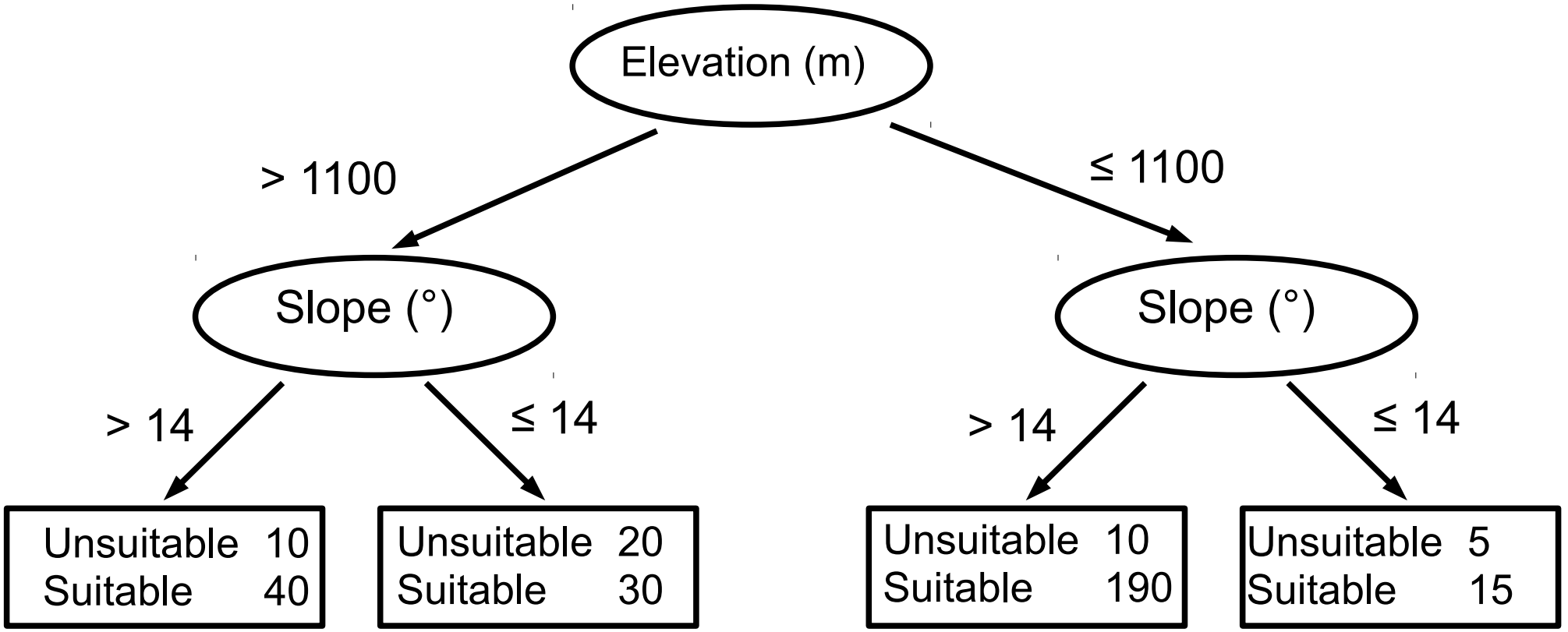
Qian 2017: 228

Interpreting interactions in decision trees



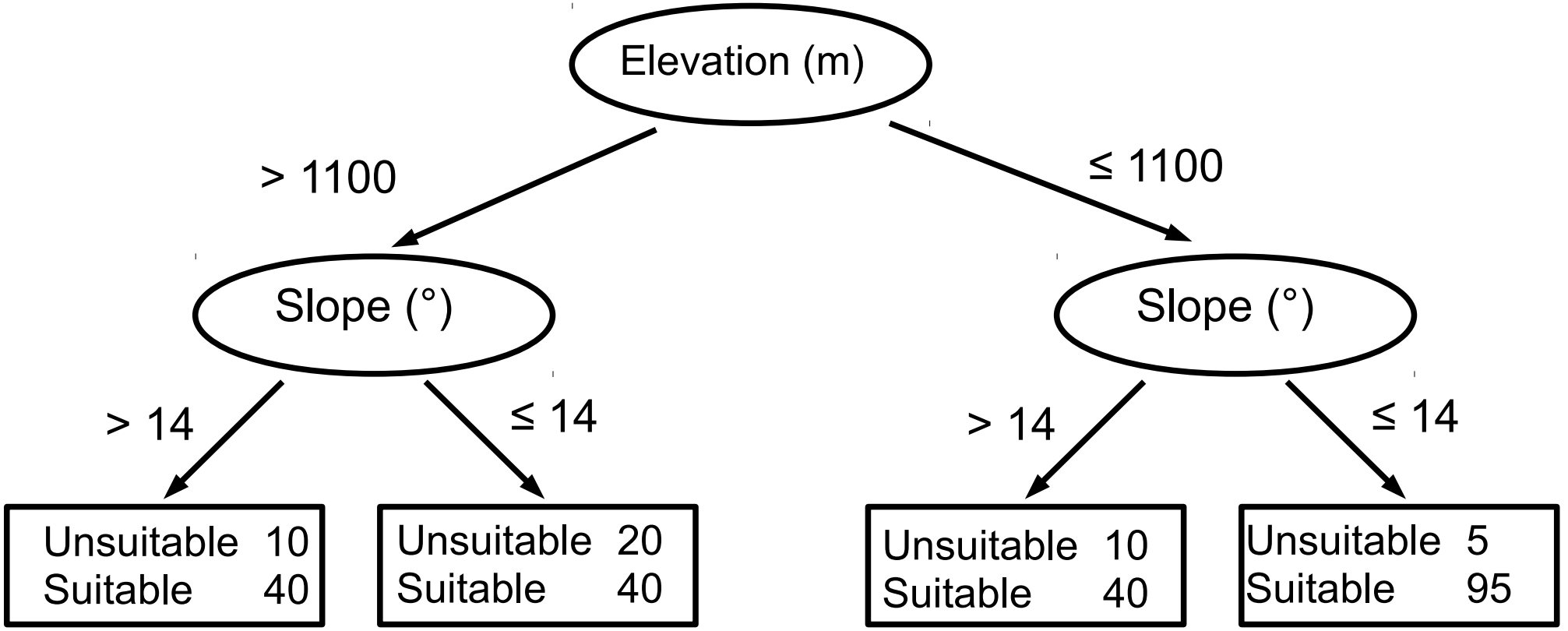
- Asymmetric interaction: Split variable only in one branch
- Very common in trees

Interpreting interactions in decision trees



- No interaction: Same split variable in both branches with same cutpoint and same effect (Slope effect in each branch: fraction of suitable habitat decreases by 20%)
- Very rare → unlikely to have same split variables and cutpoints, even if data originates from population with main effects only

Interpreting interactions in decision trees



- Symmetric interaction: Same split variable in both branches at same cutpoint, but different effects
- Equivalent to classical interaction model: A + B + A × B
- Very rare in trees → unlikely to have same split variables and cutpoints, even if data originates from population with classical interaction

Assumptions and predictor importance

- Trees do not make distributional assumptions
- Scaling of predictors required to avoid undue influence of predictors with higher (absolute) variance
- Predictor importance based on sum of improvement of goodness of split over all nodes a predictor is considered

Why are non-split variables listed? → **Surrogate variable:**

- Alternative split variable with lower improvement in purity
- Replaces split variable in case of missing values in node
- Considered in variable importance because may be more important than some split variables and helps to identify collinearity

Classification and regression trees and extensions

Contents

1. Supervised and unsupervised learning
2. Intro: Classification and regression trees
3. Goodness of split metrics, interactions, assumptions and predictor importance
4. **Critical evaluation and comparison to (G)LMs**
5. Extensions and alternatives 1: Conditional inference trees and multivariate trees
6. Extensions and alternatives 2: Random forest and boosted regression

Advantages and limitations of CART

- Relatively easy to interpret with respect to groups with similar response and predictor importance
- Very powerful to deal with non-linearity and interactions
- Sparse assumptions: Largely unaffected by outliers and no data transformation required
- Can easily handle missing values

- Collinearity and large trees complicate interpretation
- Non-interacting predictors often represented as interacting
- Bias for predictors with higher number of distinct values
- Many observations required
- Continuous variables treated as discrete
- Single tree not very robust and not necessarily optimal

Olden et al. 2008 *Quart. Rev. Biol.* 83: 171

(G)LM vs. CART

Characteristic	GLM	CART
Data Requirements		
Accommodate "mixed" data types	Low	High
Accommodate missing values of predictors	Low	High
Insensitive to monotonic transformations of predictors	Low	High
Robust to outliers in predictors	Low	Moderate
Insensitive to irrelevant predictors	Low	High
Modeling Process		
Automation (i.e., low degree of user involvement)	High	Moderate
Transparency of the modeling process	High	Moderate
Ability to model nonlinear relationships	Low	Moderate
Accommodate interactions among predictors	Low	Moderate
Model Output		
Explanatory insight and variable interpretability	High	Moderate
Predictive power	Low	Moderate
Software Availability and Ease-of-Use	High	Moderate

Additionally, CART requires larger data sets than (G)LM for meaningful results

Olden et al. 2008 *Quart. Rev. Biol.* 83: 171

(G)LM vs. regression trees

Relationship between predictors and response linear in population

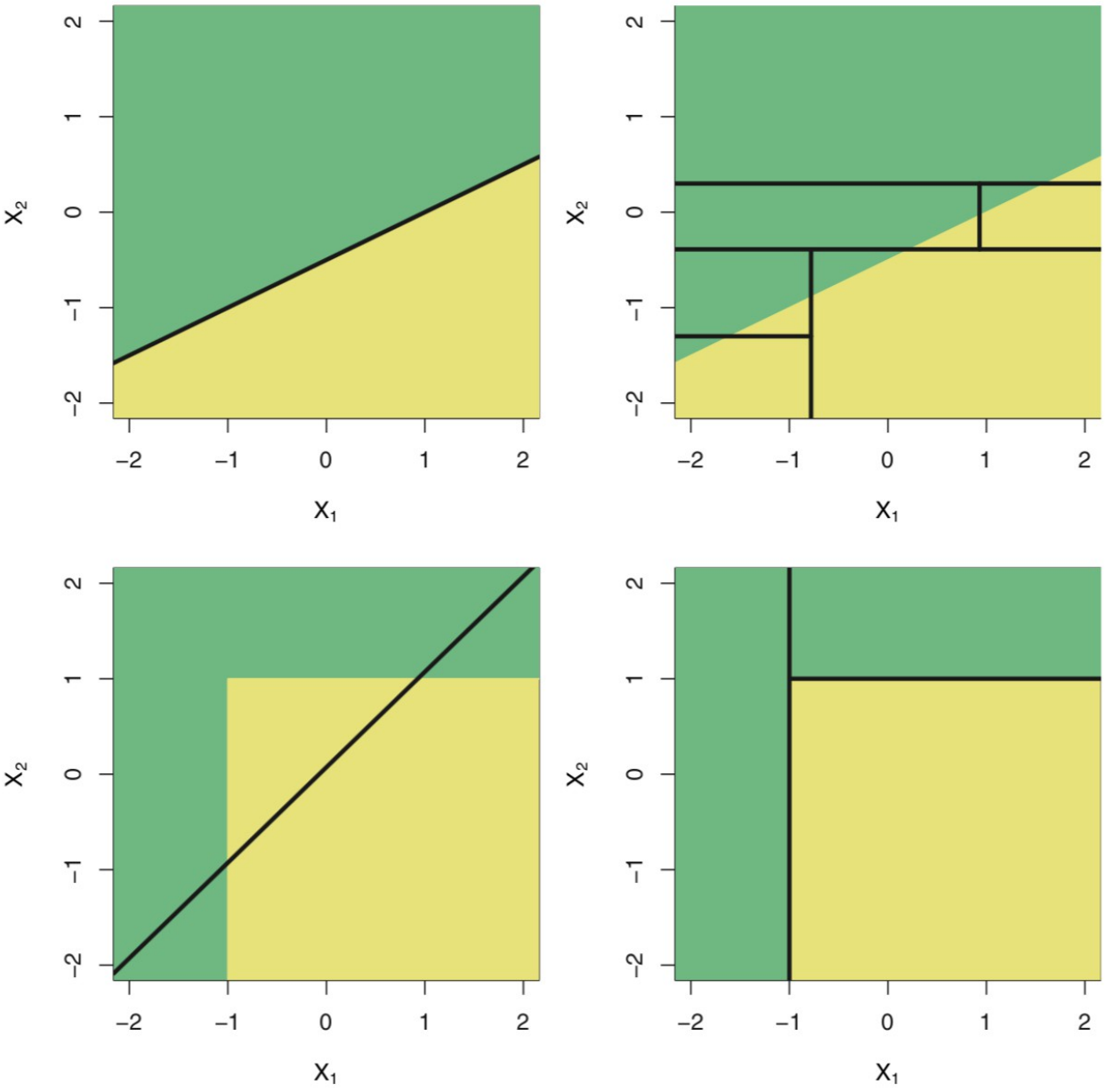
→ (G)LMs outperform regression trees

Relationship between predictors and response non-linear and complex in population

→ Regression trees outperform (G)LMs

(G)LM

CART



James et al. 2017: 315

Classification and regression trees and extensions

Contents

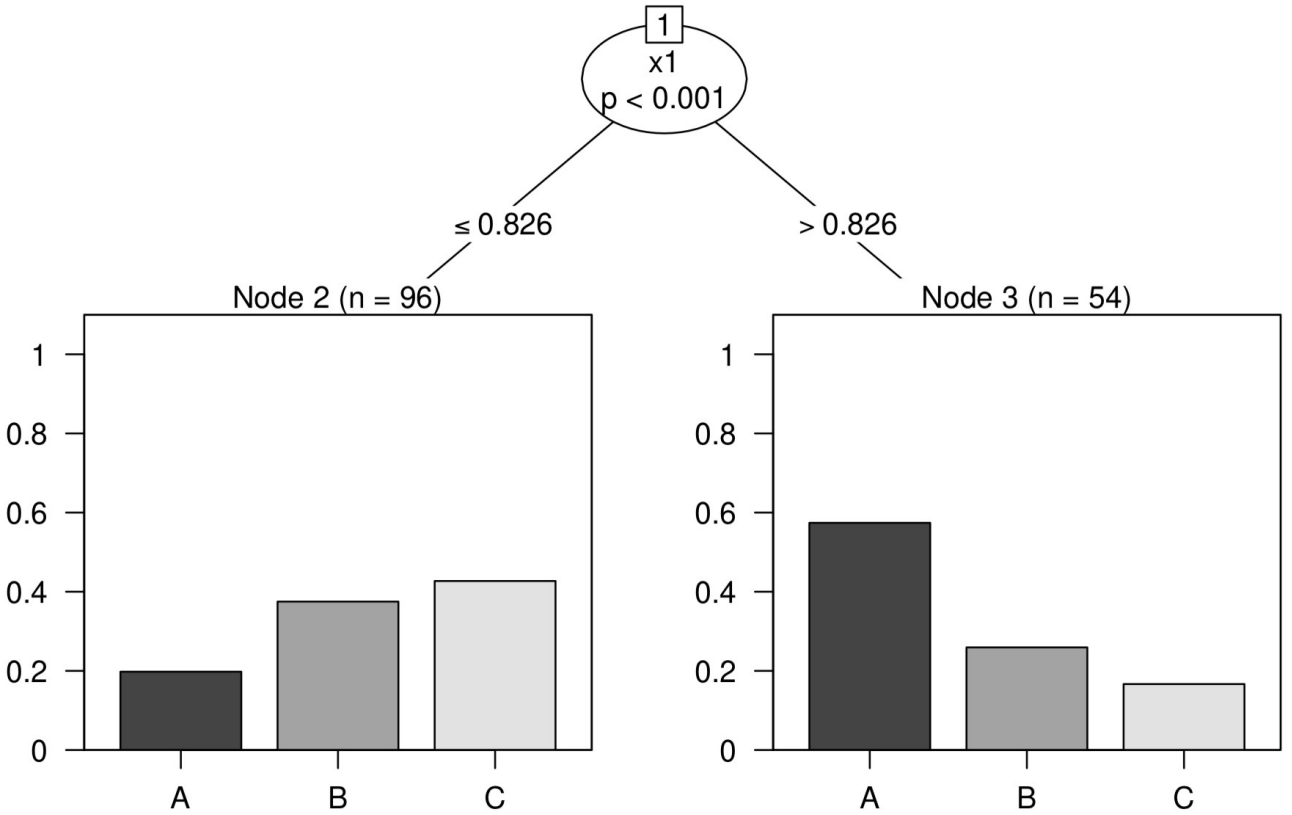
1. Supervised and unsupervised learning
2. Intro: Classification and regression trees
3. Goodness of split metrics, interactions, assumptions and predictor importance
4. Critical evaluation and comparison to (G)LMs
5. **Extensions and alternatives 1: Conditional inference trees and multivariate trees**
6. Extensions and alternatives 2: Random forest and boosted regression

Extensions and alternatives to CART

- CART lacks a probabilistic foundation (e.g. no information on uncertainty from confidence intervals, no p-values) → Conditional inference trees
- CART for inspecting parametric models (i.e. does the slope differ → are additional variables relevant?) → model-based recursive partitioning (Zeileis et al. 2008)
- Hierarchical CART for set of nested variables (e.g. different ecological scales) → Cascade regression trees (Ouellette et al. 2012)
- Multivariate response → Multivariate regression trees
- More robust trees → Random forest and Boosted regression trees

Conditional inference trees

- Provides (frequentist) statistical framework for variable selection in decision trees via permutation tests
- Separates variable selection and splitting → no bias towards predictors with higher number of distinct values
- Comparison to CART: Estimation accuracy generally higher, prediction accuracy dependent on data set



Hothorn et al. 2006 *J. Comp. Graph.* 15: 651–674

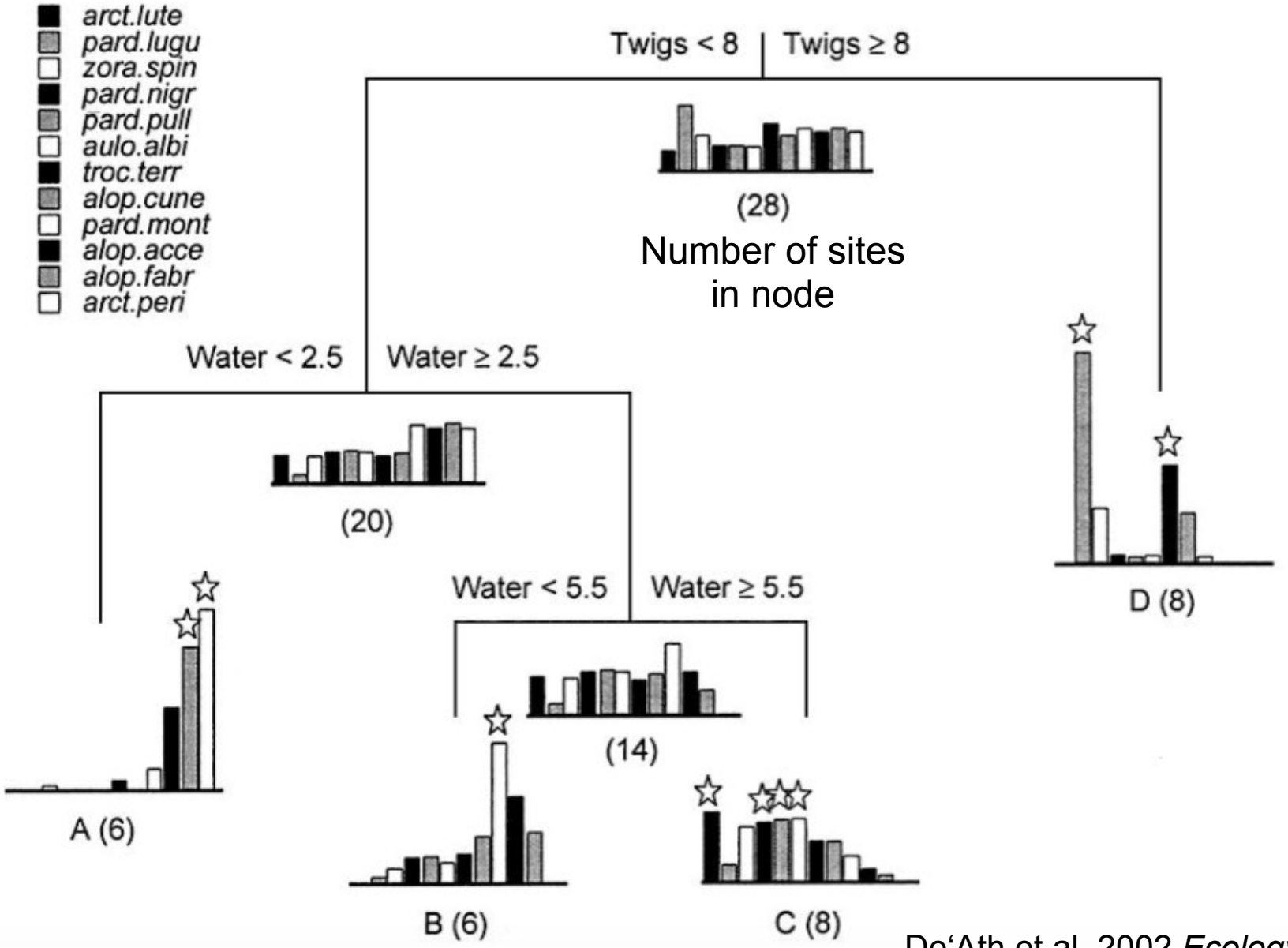
Multivariate regression trees

- Extension to multivariate case: multiple response variables
- Impurity measure summed over responses: Sum of squares to multivariate mean or median
- For community data, method allows to: (1) identify species that drive splits, (2) create biplots displaying sites and species, (3) identify representative species for nodes
- Comparison with unconstrained cluster analysis (discussed later) can inform whether relevant explanatory variables have been captured
- Alternative to ordination (discussed later), especially for non-linear and complex species-environment relationships

De'Ath et al. 2002 *Ecology* 83: 1105

Multivariate regression trees: Results

Bars display means of species, stars denote indicator species



De'Ath et al. 2002 *Ecology* 83: 1105

Multivariate regression trees: Results

Analysis provides information on variance explained by tree and its splits.

TABLE 1. Tabulation of species variance for the tree analysis of the hunting spider data.

Species	Species variance (%) explained by tree splits and whole tree				
	Twigs < 8	Water < 2.5	Water < 5.5	Tree total	Species total
Arctosa perita	1.90	15.54	0.00	17.45	20.85
Alopecosa fabrilis	2.33	6.72	0.79	9.83	13.44
Pardosa monticola	1.75	1.34	5.06	8.16	10.82
Alopecosa accentuata	1.93	0.70	2.15	4.78	5.77
Arctosa lutetiana	0.47	0.71	1.78	2.96	4.37
Aulonia albimana	0.35	0.90	0.70	1.96	3.28
Pardosa pullata	0.47	0.99	0.48	1.94	2.53
Pardosa nigriceps	0.20	0.88	0.41	1.49	2.23
Zora spinimana	0.80	0.64	0.64	2.08	4.04
Alopecosa cuneata	0.25	0.84	0.02	1.11	2.89
Pardosa lugubris	23.22	0.02	0.04	23.28	24.41
Trochosa terricola	3.29	0.40	0.05	3.74	5.38
Total species variance	36.97	29.69	12.12	78.78	100.00

Notes: The total species variance is partitioned by species, the whole tree, and the three splits of the tree. The first split is dominated by *Pardosa lugubris* (23 percentage points of 37% of the total species variance explained by that split), the second by *Arctosa perita* (15 percentage points of 30%), and the third by *Pardosa monticola* (5 percentage points of 12%).

De'Ath et al. 2002 *Ecology* 83: 1105

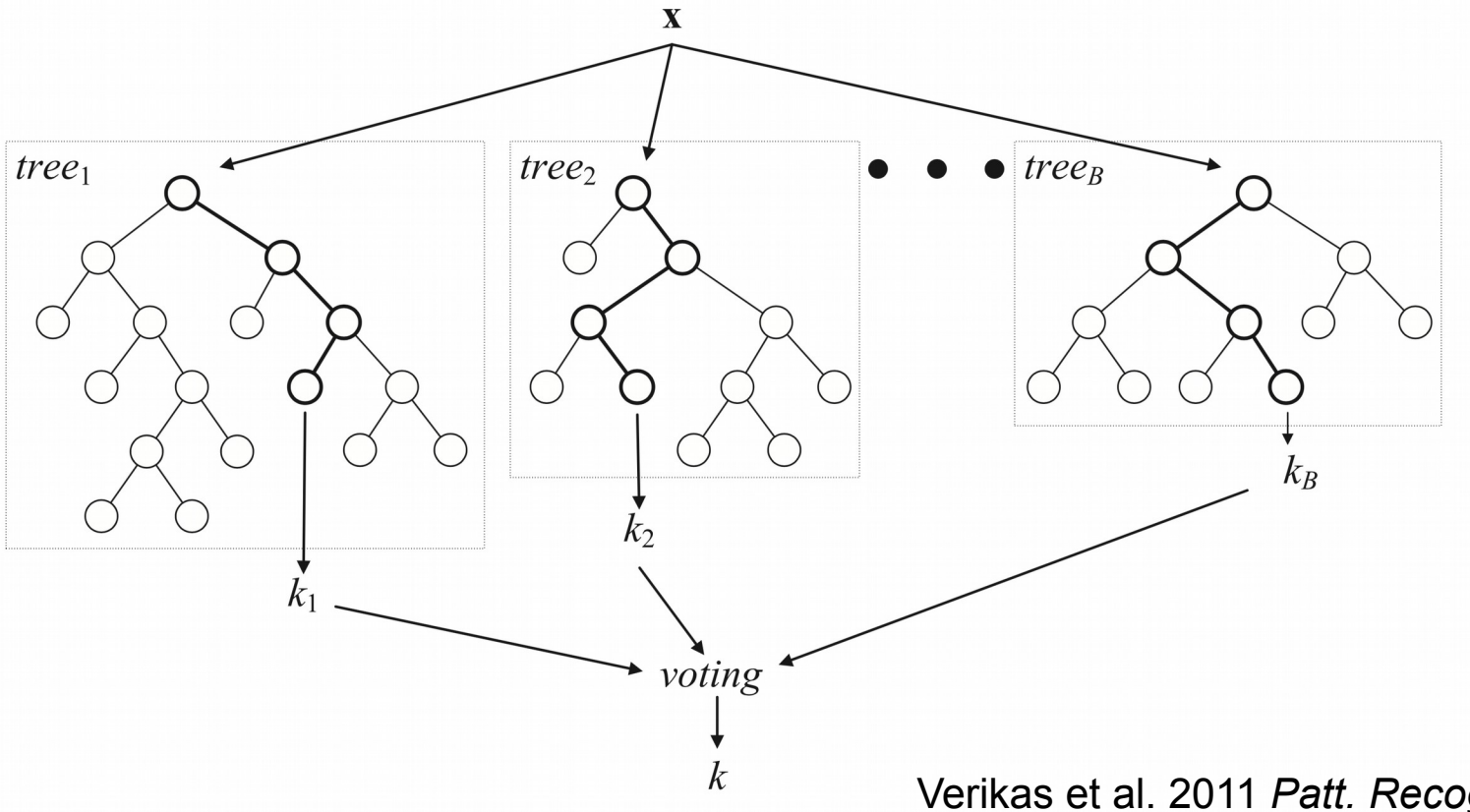
Classification and regression trees and extensions

Contents

1. Supervised and unsupervised learning
2. Intro: Classification and regression trees
3. Goodness of split metrics, interactions, assumptions and predictor importance
4. Critical evaluation and comparison to (G)LMs
5. Extensions and alternatives 1: Conditional inference trees and multivariate trees
6. **Extensions and alternatives 2: Random forest and boosted regression**

Random forest (RF)

- Single tree unstable, but average of multiple trees reliable and provides unbiased estimation or prediction
  → Grow many trees (forest) on bootstrapped samples
- At each split a predictor subset is randomly selected: Variable masked by other predictors can be identified and may improve global performance



Verikas et al. 2011 *Patt. Recogn.* 44: 330

Random forest (RF)

- Ensemble method: Prediction requires aggregation (via mean (regression) or majority vote (classification))
- Evaluation of prediction accuracy: Out-of-bag (OOB) error (observations not used in tree building)
- Variable importance:
  - Average improvement of node impurity over all trees
  - Better: Difference in prediction accuracy before and after permutation
- RFs typically exhibit a prediction accuracy similar or higher than other methods (see Verikas et al. 2011), but results are less easy to interpret: There is no average tree!

Boosted regression trees (BRT)

- Related to *Boosting* (Ensemble method)
- Sequential growing of trees, focus on minimising residuals from previous trees to improve accuracy
- Typically (slightly) higher prediction accuracy than RF, but:
  - Larger number of trees required → computationally more costly
  - More tuning parameters → more complex

Journal of Animal Ecology

British Ecological Society

Journal of Animal Ecology 2008, 77, 802–813 doi: 10.1111/j.1365-2656.2008.01390.x

A working guide to boosted regression trees

J. Elith[1]*, J. R. Leathwick[2] and T. Hastie[3]

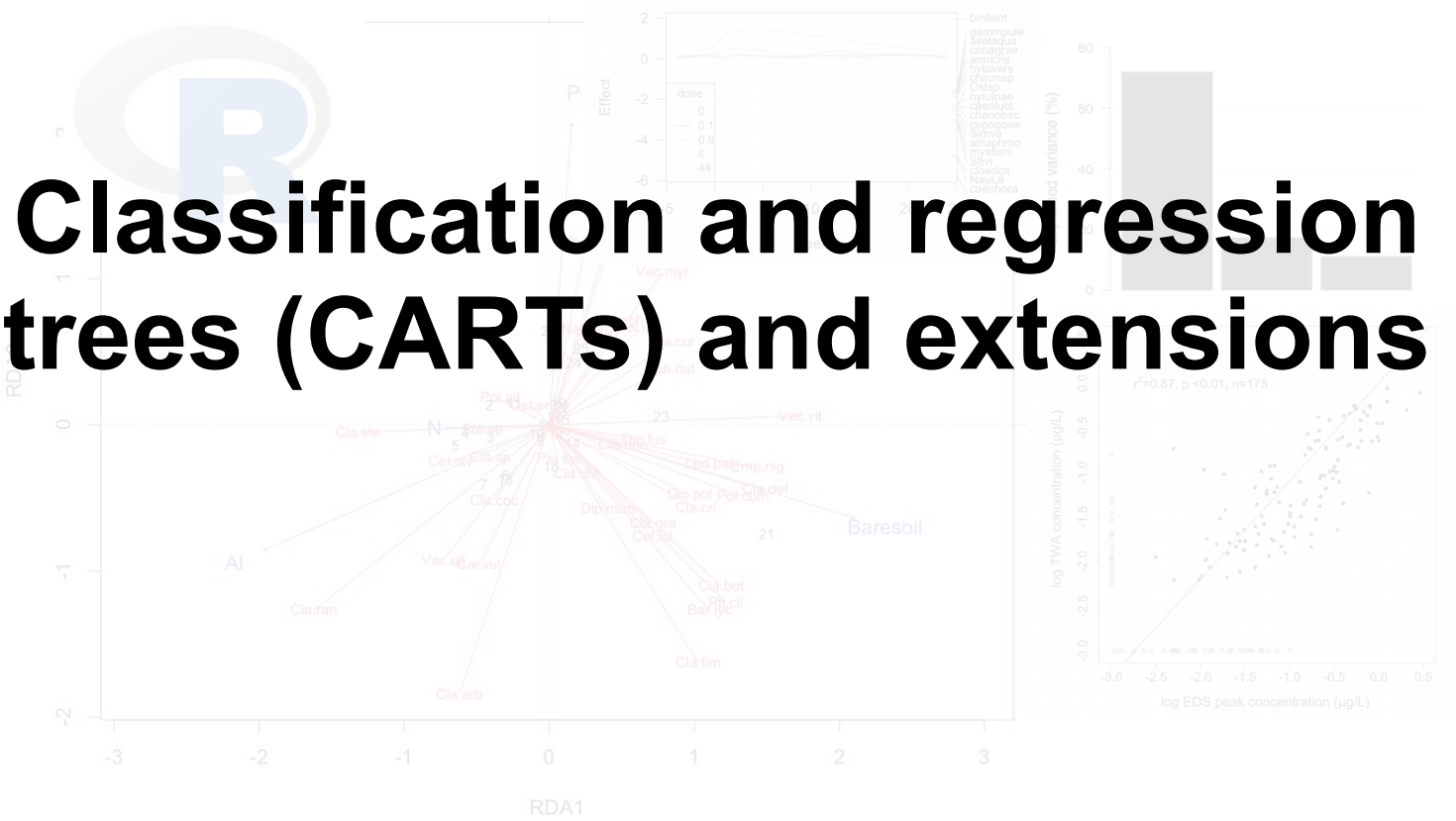
[1]*School of Botany, The University of Melbourne, Parkville, Victoria, Australia 3010;* [2]*National Institute of Water and Atmospheric Research, PO Box 11115, Hamilton, New Zealand; and* [3]*Department of Statistics, Stanford University, CA, USA*

Summary

1. Ecologists use statistical models for both explanation and prediction, and need techniques that

Tools for complex data analysis

University of Koblenz-Landau 2018/19



Classification and regression trees (CARTs) and extensions

Ralf B. Schäfer

These slides and notes complement the lecture with exercises “Tools for complex data analysis” for ecotoxicologists and environmental scientists. Do not hesitate to contact me if you have any comments or you find any errors (slides, slide notes, or code): schaefer-ralf@uni-landau.de

While I made notes below the slides, some aspects are only mentioned in the R demonstration associated with the lecture.

Learning targets and study questions

- Knowledge on classification and terminology of machine learning methods.
- Comprehension of Classification and regression trees.
- Knowledge on extensions of CARTs.

2

Learning targets and study questions

- Knowledge on classification and terminology of machine learning methods
  - Describe the difference between supervised and unsupervised learning.
- Comprehension of Classification and regression trees.
  - Describe the differences between classification and regression trees.
  - Summarize the procedure for creating a tree.
  - Explain the role of impurity metrics and the cost-complexity parameter in this context.
  - Outline the different types of interactions in CARTs.
  - What is a surrogate variable?
  - What are major differences to (G)LMs and under which conditions should they be used?
  - Discuss advantages and disadvantages of CARTs.

Learning targets and study questions

- Knowledge on extensions of CARTs.
  - Briefly discuss a few extensions of CARTS and their rationale.
  - Describe the application domain of multivariate regression trees.
  - Discuss major advantages and disadvantages of random forest compared to building a single tree.

4

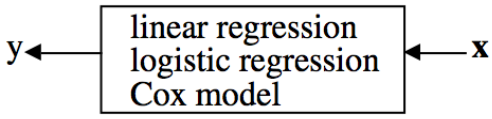
Classification and regression trees and extensions

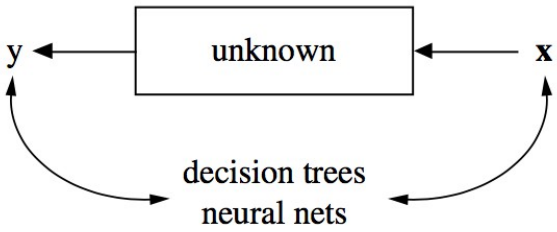
Contents

5

Supervised and unsupervised learning

- Machine learning terminology
- Supervised learning:
  - Use statistical model or algorithm to map predictors x to response y



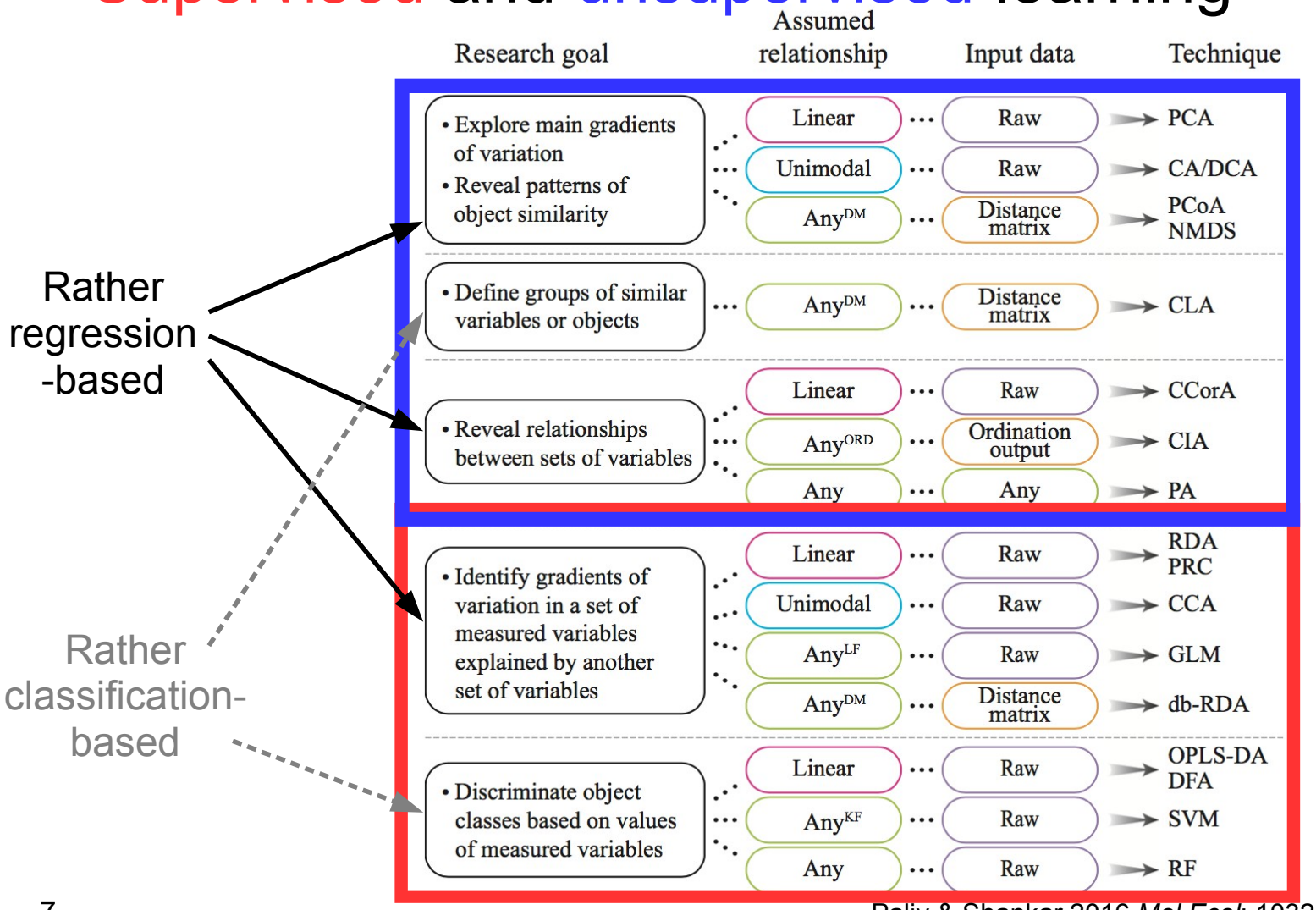


- Unsupervised learning:
  - Identify (hidden) patterns or associations within a data set without "ground truth"
- Unsupervised/supervised classification and regression

Breiman 2001 *Statistical Science* 16: 199

Super- and unsupervised learning methods can be further subdivided into regression and classification, where regression focuses on continuous phenomena and classification on groups.

Supervised and unsupervised learning



Paliy & Shankar 2016 *Mol Ecol*: 1032

7

Supervised classification and regression

- Response (e.g. membership of observations to groups) known
- Aim: Identification of classification or regression rules (mainly for explanation or prediction)

Methods include classification and regression trees, discriminant analysis, artificial neural networks, but also our well known (generalized) linear regression models

Classification and regression trees (CART)

- Machine learning technique and method of recursive partitioning
- Used in prediction or explanation of a response variable through the construction of a decision tree

Breiman et al. 1984

Supervised classification and regression are mainly used for explanation and prediction. They can also be used to assess hypotheses and determine probabilities, though this requires the use of a specific method, i.e. conditional inference trees (discussed later), that provides p-values from permutation.

For a very readable overview of the different methods in the context of CARTs that have been developed see Loh (2014). We will focus on classical CARTs and briefly touch on a few extensions.

Note that Qian (2017) differs in his evaluation in that he considers CART rather as an exploratory tool to identify variables for (parametric) model building. This should only be done, if an independent data set is available. By no means, the data used for fitting the tree (or random forest, see later) should be re-used (Strobl et al. 2009).

Breiman L., Friedman J., Stone C.J. & Olshen R.A. (1984) Classification and regression trees, Repr. Chapman & Hall, Boca Raton.

Loh W.-Y. (2014). Fifty Years of Classification and Regression Trees. *International Statistical Review* 82, 329–348. Free to download at: http://www.stat.wisc.edu/~loh/treeprogs/guide/LohISI14.pdf

Qian SS (2017) Environmental and ecological statistics with R, 2nd ed. Chapman & Hall/CRC, Boca Raton, Fla.

Strobl C., Malley J. & Tutz G. (2009) An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods* 14, 323–348.

Classification and regression trees and extensions

Contents

9

Example: Which habitats are suitable for the turkeybeard *Xerophyllum asphodeloides*?

https://www.flickr.com/photos/blueridgekitties/5801875577

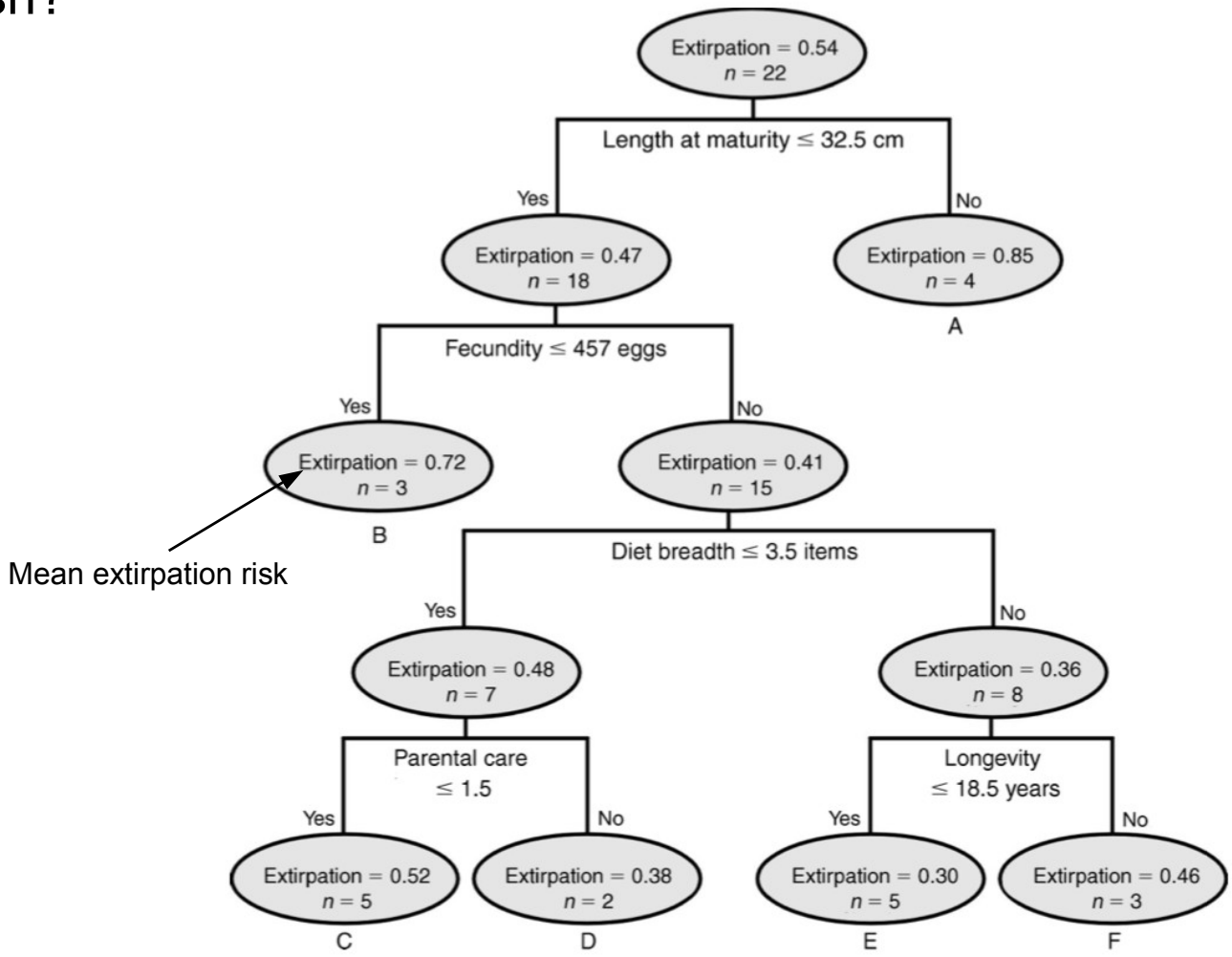
Bourg et al. 2005 *Ecology* 86: 2793

CART is used in different environmental contexts, for example to identify the most important habitat variables for conservation and restoration of habitats for organisms or to derive classification rules for land cover types from remote sensing data. We will later discuss its application in the context of multivariate response variables (e.g. community data).

The example shows a classification tree. Suitable habitats in bold.

Bourg N.A., McShea W.J. & Gill D.E. (2005) Putting a CART before the search: Successful habitat prediction for a rare forest herb. *Ecology* 86, 2793–2804.

Example 2: Which traits determine the extinction risk of desert fish?



11

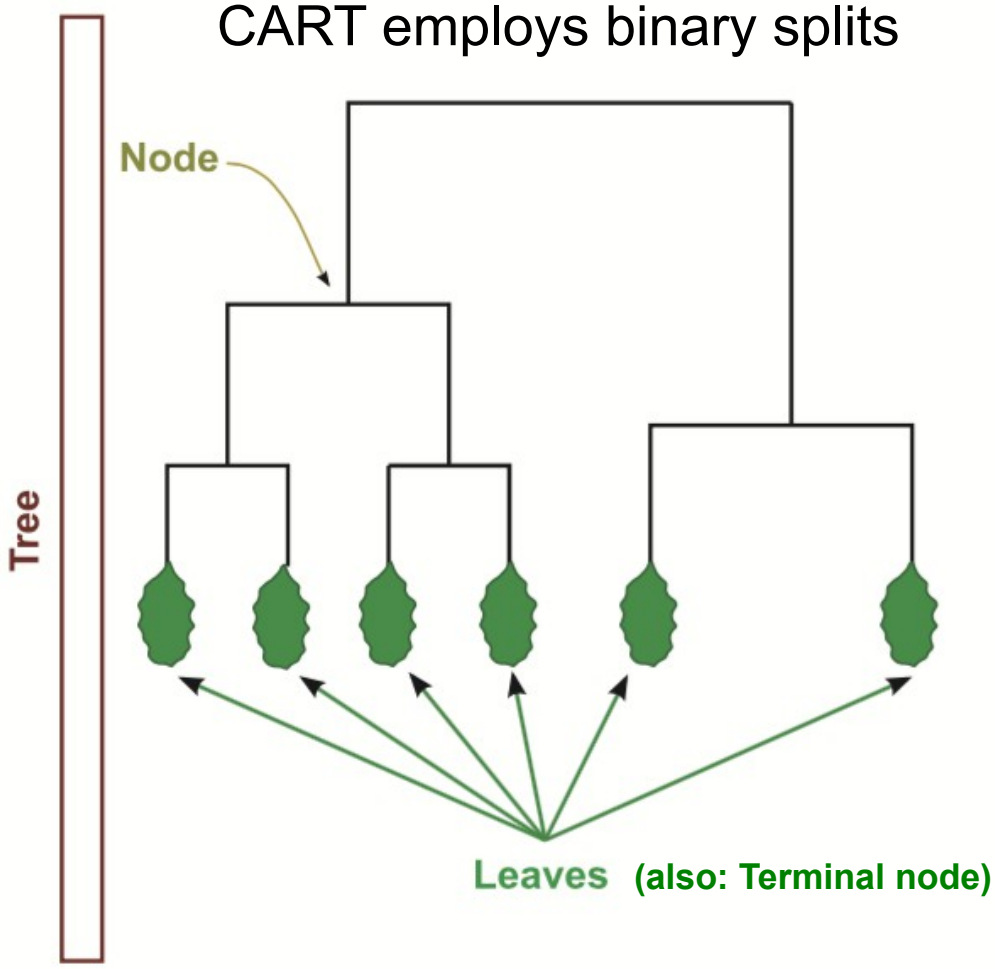
Olden et al. 2008 *Ecology* 89: 847

Example for a regression tree.

Olden J.D., Poff N.L. & Bestgen K.R. (2008) Trait synergisms and the rarity, extirpation, and extinction risk of desert fishes. *Ecology* 89, 847–856.

CART terminology

CART employs binary splits



12

Ouellette et al. 2012 *Methods Ecol. Evolut.* 3: 234

Splitting can also be termed partitioning and CART represents a form of recursive partitioning. The formed groups in CART are mutually exclusive, i.e. each observation can only belong to one terminal.

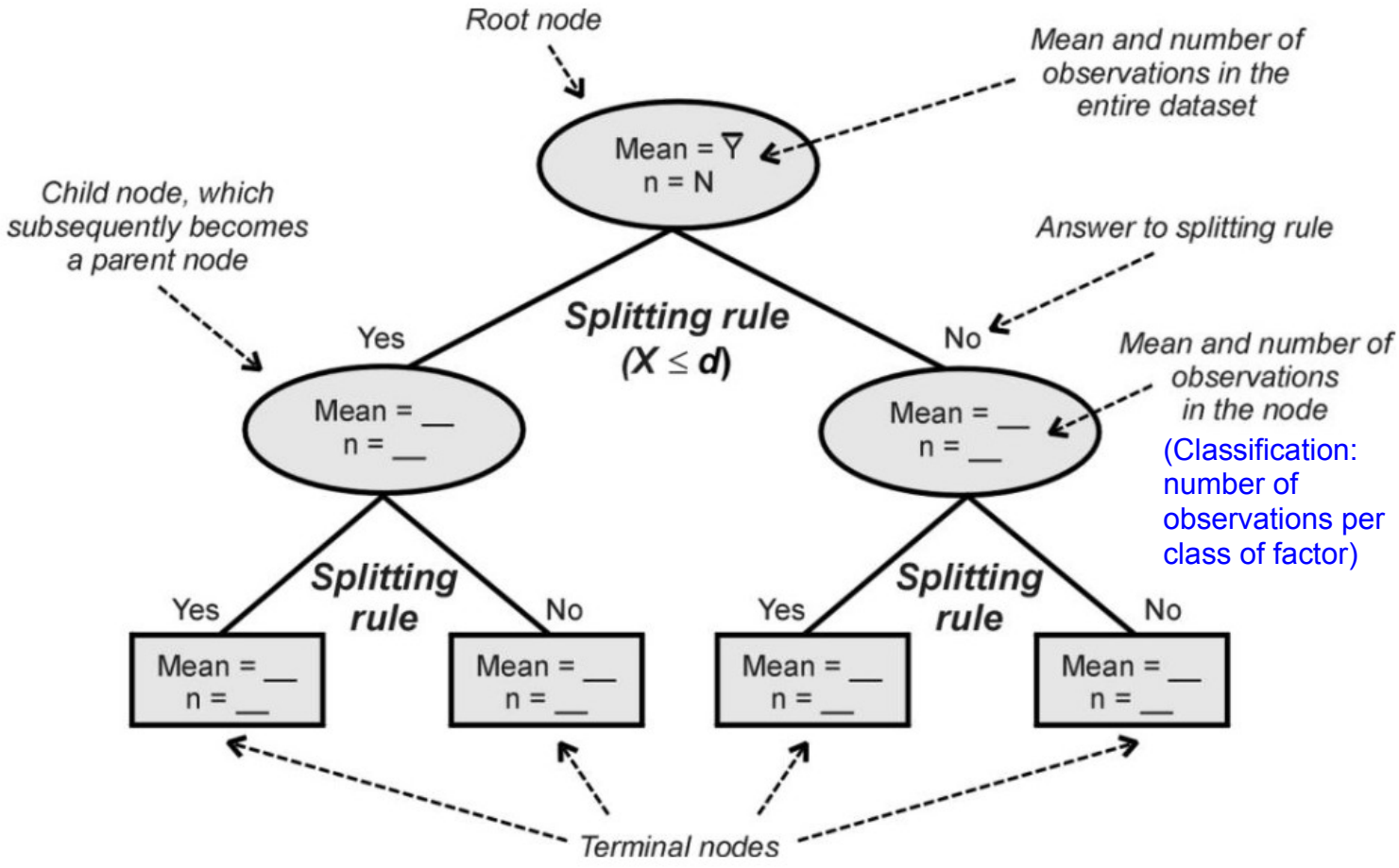
If no further splitting is possible, the node becomes a terminal node.

Note that CARTs grow upside-down, with the root node at the top.

Ouellette M.-H., Legendre P. & Borcard D. (2012) Cascade multivariate regression tree: a novel approach for modelling nested explanatory sets. *Methods in Ecology and Evolution* 3, 234–244.

CART terminology and interpretation

Described for regression tree



13

Olden et al. 2008 *Quart. Rev. Biol.* 83: 171

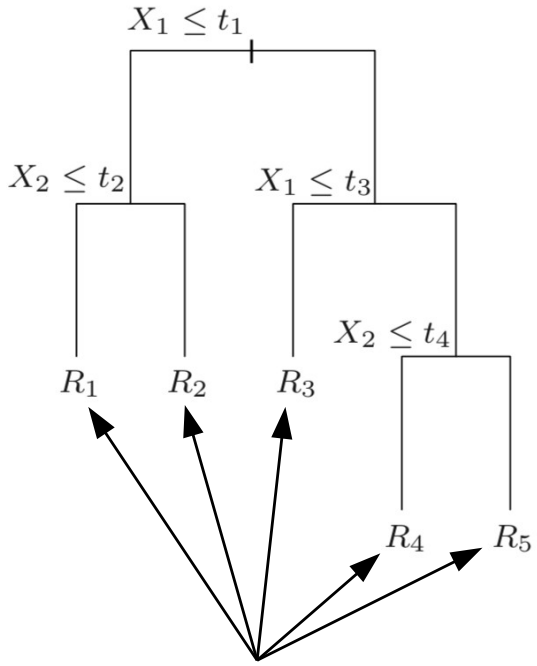
The splitting rules are conditions for continuous or ordinal variables (e.g. precipitation > 10 mm) and discrete values (e.g. dispersal capacity is high) for categorical variables. The variable related to the splitting rule is called split variable.

For ordinal and continuous response variables (as shown in the figure), each node is assigned the mean (or median) of the response variable of all observations in the node. For nominal response variables, each node is assigned the probabilities for all classes of the response variable, though often only the class with the greatest probability is visualised in the decision tree.

Olden J.D., Lawler J.J. & Poff N.L. (2008) Machine Learning Methods Without Tears: A Primer for Ecologists. *The Quarterly Review of Biology* 83, 171–193.

CART interpretation for two predictors

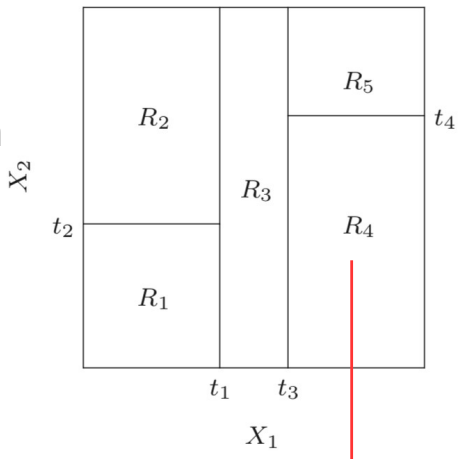
Splitting rule s is given at each node i and splits observations according to values t of predictors X_1 and X_2



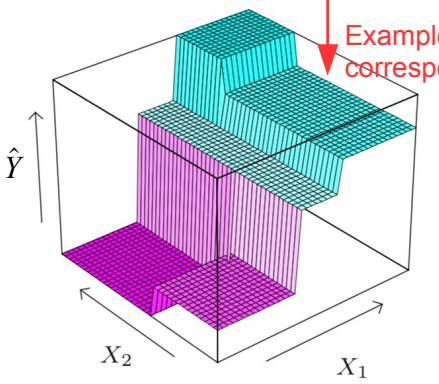
R_m: Region m of X_1 and X_2

Value of R_m: estimate of response $Y \rightarrow \hat{Y}$

Two-dimensional representation



Three-dimensional representation



Regions are assigned an estimate of $\hat{Y}$

14 Figures modified from Hastie et al. (2017: 306)

General procedure

1. Tree building

Sequential (binary) splitting into nested groups. Splitting is done to minimise impurity (e.g. Deviance)

→ Impurity or Goodness of split metrics

2. Stopping tree building

Implemented through three criteria:

a) Definition of a minimum number of observations in a node

→ No further splitting possible

b) All observations inside node have identical distribution

→ No further splitting possible

c) Defining a maximum number of splits

3. Pruning – Reducing size of tree

Based on complexity parameter that penalises increasing size (in conjunction with cross-validation error)

15

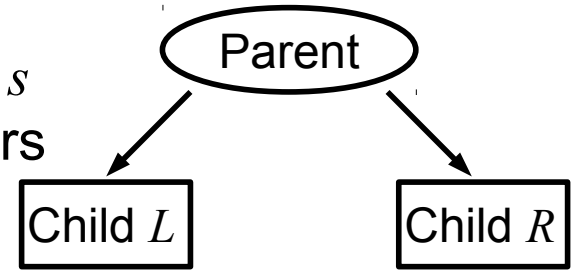
Classification and regression trees and extensions

Contents

16

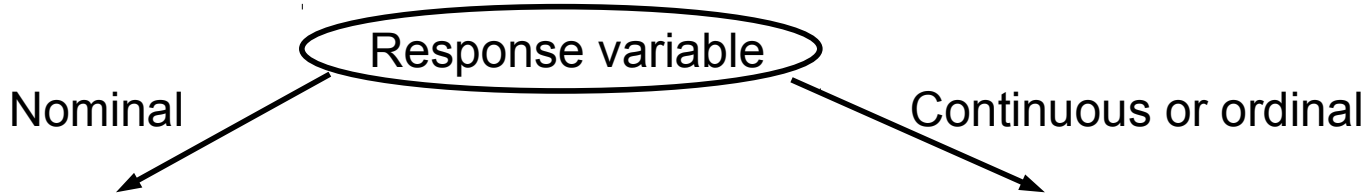
How to measure impurity?

- Deviance D
- During tree building, all possible splits s of node i are calculated for all predictors j and the purest is selected:



$$\arg\max_{s,j} \Delta D_i = D_{i,parent} - D_{i,child} \quad \text{with } D_{i,child} = D_{i,s,j,L} + D_{i,s,j,R}$$

Calculation of D



Classification tree: Gini index

$$D_i = \sum_{k=1}^{K} p_k(1-p_k)$$

K = number of classes of response
$\hat{p}_k$ = estimated proportion of observations in class k

Regression tree: MSE

$$D_i = \frac{1}{n_i}\sum_{k=1}^{n_i}(y_k - \hat{\mu}_i)^2$$

n_i = number of observations in parent or child node i
y_k = value of response of observation k
$\hat{\mu}$ = estimated mean of response for parent or child node i

You should be familiar with the Mean Squared Error (MSE) from the linear model. During tree building the observations are split between two child nodes as to maximise the difference in MSE to the parent (which is equivalent to minimising the MSE).

Note that the calculation of the Deviance measure in the regression tree relies on the number of observations in the respective parent or child node, whereas in the classification tree, the total number of classes of the response K can be used, which simplifies the formula (if p for a class k is 0, this does not matter for the calculation. Hence, it is irrelevant whether all K classes of the response are considered in the calculation or only those of the respective node (i.e. all $p_k > 0$)).

For the relationship between the deviance as defined here and as defined in the context of the GLM see Qian (2017: 297f).

The different measures of Deviance are also termed “goodness of split” metrics. For the classification tree, the two most commonly employed metrics to calculate *D* are the Gini index (default in R, see slide) and the Information index (also Entropy index). Both indices yield to their minimum if all observations are of the same class, and their maximum if the proportions of all classes are the same. The information index is given as:

$$D_i = -\sum_{k=1}^{K} p_k \log(p_k)$$

For an example-based calculation of Deviance metrics see Zuur et al. (2007): 143ff and for a more formal mathematical notation and additional details see Izenman (2008): 282ff.

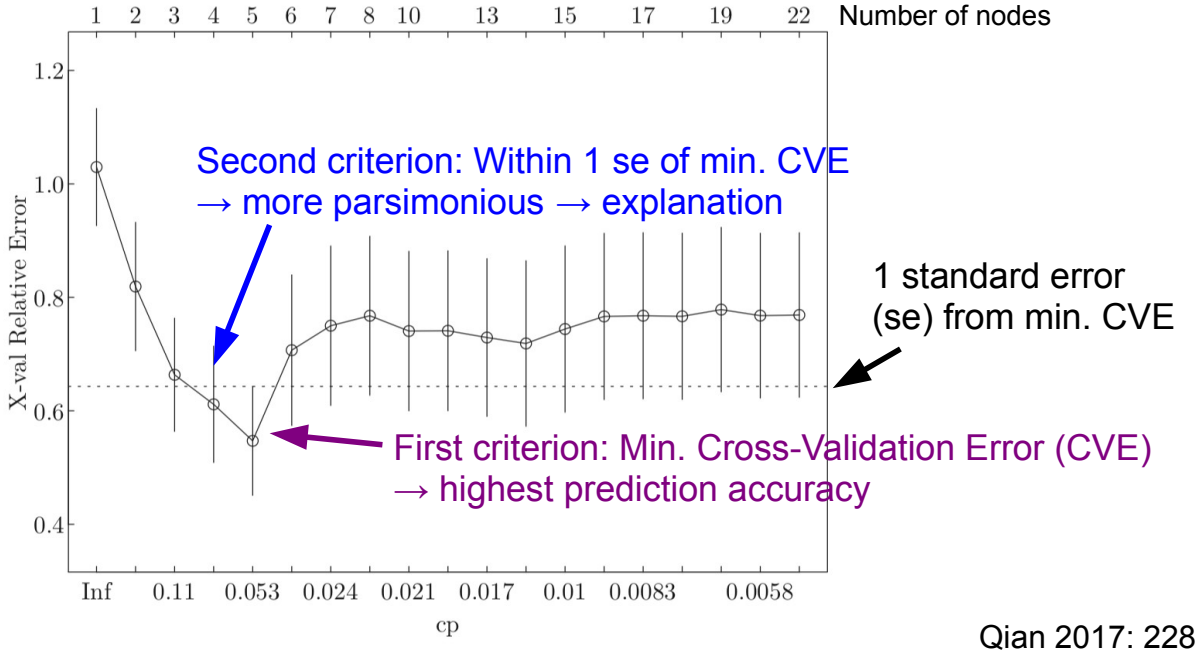
See literature list for most references. Additional reference:
Izenman A.J. (2008) Modern multivariate statistical techniques: regression, classification, and manifold learning. Springer, New York, NY.
Qian SS (2017) Environmental and ecological statistics with R, 2nd ed. Chapman & Hall/CRC, Boca Raton, Fla.

How to select the optimal reduction in tree size?

- Deviance D is penalised by cost-complexity (cp) parameter

$$D_{cp} = D + cp \times \text{size-of-tree}$$

- Same rationale as information-theoretic criteria (e.g. AIC)
- Optimal tree size determined by cp. How to set cp?
  → Cross-validation!



Qian 2017: 228

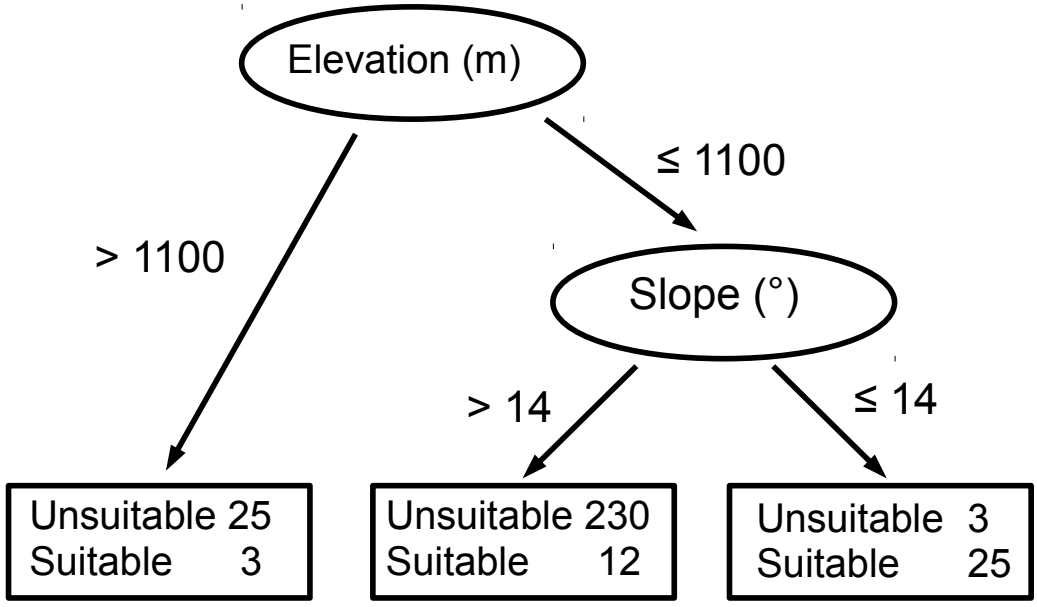
When the size of the tree increases, D typically becomes smaller whereas the term $cp \times$ size-of-tree increases and thus penalises for a larger tree size (as AIC and BIC contain a term consisting of a constant multiplied with the number of parameters in the model).

For a detailed mathematical treatment on how to find the optimal tree size (i.e. pruning), see Izenmann (2008): 295 ff.

Izenman A.J. 2008 Modern multivariate statistical techniques: regression, classification, and manifold learning. Springer, New York, NY.

Qian SS (2017) Environmental and ecological statistics with R, 2nd ed. Chapman & Hall/CRC, Boca Raton, Fla.

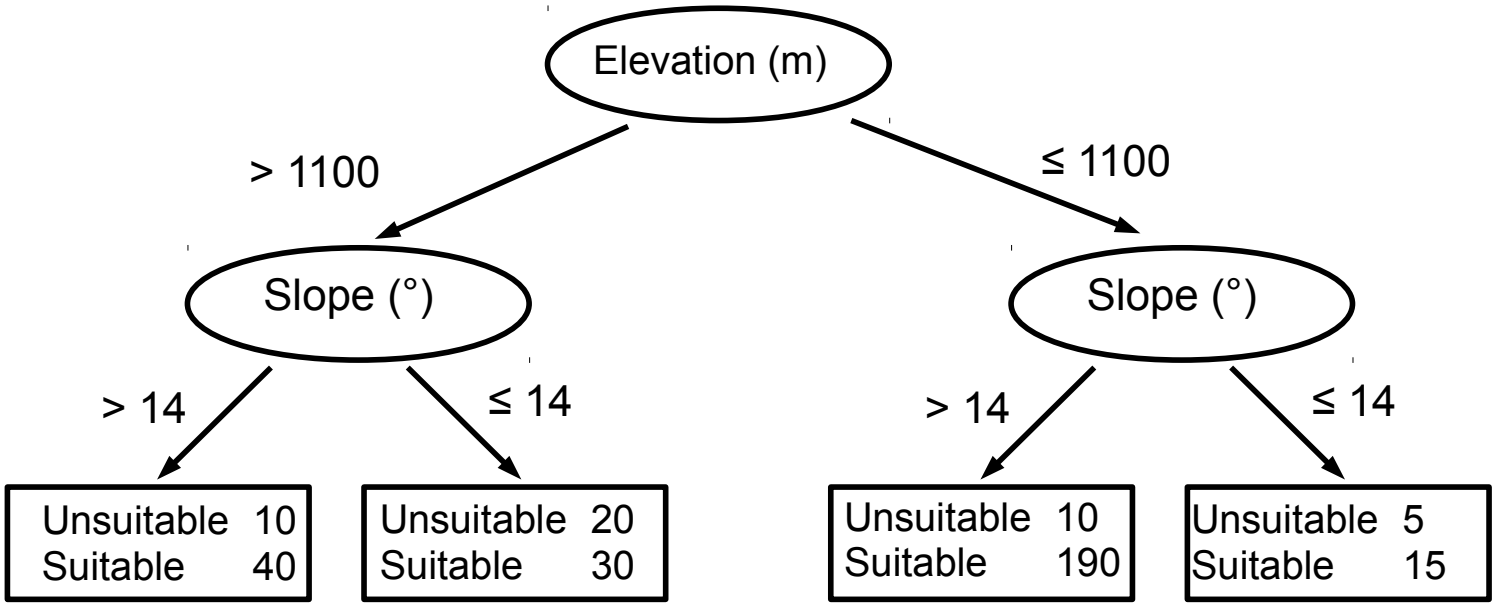
Interpreting interactions in decision trees



- Asymmetric interaction: Split variable only in one branch
- Very common in trees

19

Interpreting interactions in decision trees



- No interaction: Same split variable in both branches with same cutpoint and same effect (Slope effect in each branch: fraction of suitable habitat decreases by 20%)
- Very rare → unlikely to have same split variables and cutpoints, even if data originates from population with main effects only

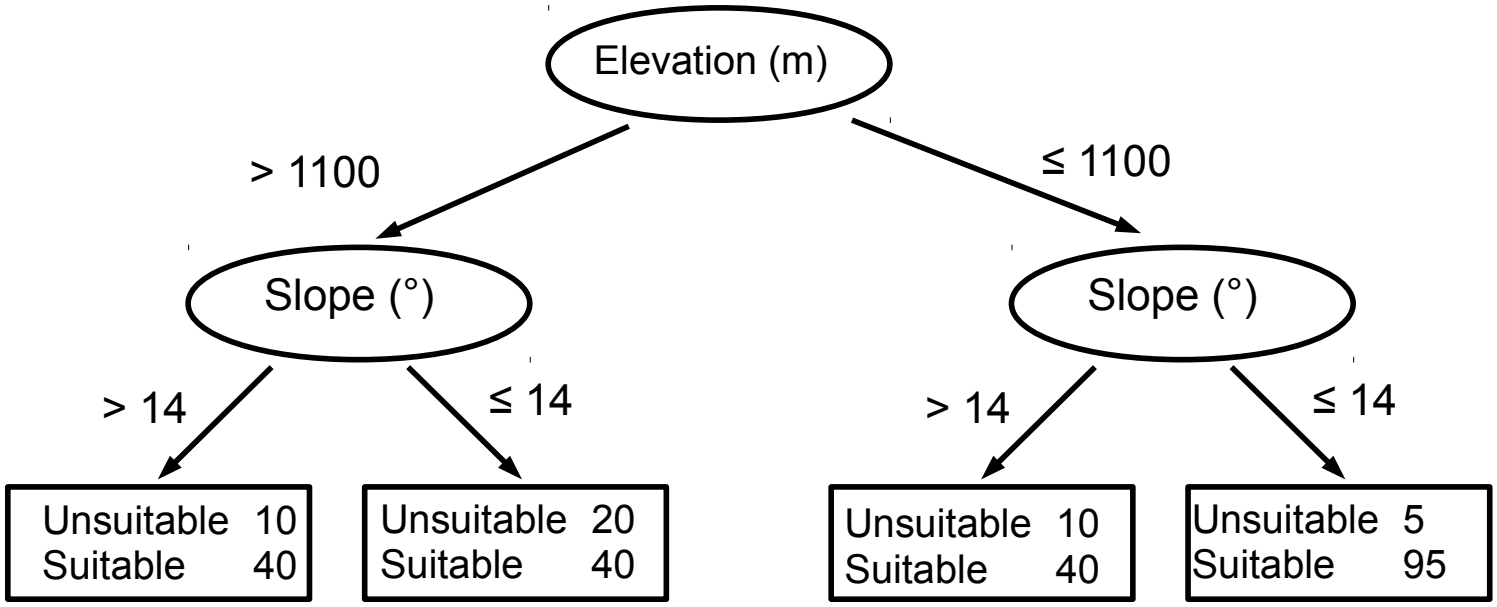
20

In both branches the fraction of unsuitable habitat increases by 0.2 (left: 0.2 to 0.4, right: 0.05 to 0.25).

For details on interactions and why it is unlikely to have trees that only represent main effects, see Strobl et al. (2009).

Strobl C., Malley J. & Tutz G. (2009) An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods* 14, 323–348.

Interpreting interactions in decision trees



- Symmetric interaction: Same split variable in both branches at same cutpoint, but different effects
- Equivalent to classical interaction model: A + B + A × B
- Very rare in trees → unlikely to have same split variables and cutpoints, even if data originates from population with classical interaction

21

In the left branch, the fraction of suitable habitat decreases with a lower slope, whereas in the right branch it increases.

For details on interactions and why it is unlikely to have trees that only represent main effects, see Strobl et al. (2009).

Strobl C., Malley J. & Tutz G. (2009) An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods* 14, 323–348.

Assumptions and predictor importance

- Trees do not make distributional assumptions
- Scaling of predictors required to avoid undue influence of predictors with higher (absolute) variance
- Predictor importance based on sum of improvement of goodness of split over all nodes a predictor is considered

Why are non-split variables listed? → **Surrogate variable:**

- Alternative split variable with lower improvement in purity
- Replaces split variable in case of missing values in node
- Considered in variable importance because may be more important than some split variables and helps to identify collinearity

22

Classification and regression trees and extensions

Contents

1. Supervised and unsupervised learning
2. Intro: Classification and regression trees
3. Goodness of split metrics, interactions, assumptions and predictor importance
4. **Critical evaluation and comparison to (G)LMs**
5. Extensions and alternatives 1: Conditional inference trees and multivariate trees
6. Extensions and alternatives 2: Random forest and boosted regression

23

Advantages and limitations of CART

- Relatively easy to interpret with respect to groups with similar response and predictor importance
- Very powerful to deal with non-linearity and interactions
- Sparse assumptions: Largely unaffected by outliers and no data transformation required
- Can easily handle missing values

- Collinearity and large trees complicate interpretation
- Non-interacting predictors often represented as interacting
- Bias for predictors with higher number of distinct values
- Many observations required
- Continuous variables treated as discrete
- Single tree not very robust and not necessarily optimal

Olden et al. 2008 *Quart. Rev. Biol.* 83: 171

Collinearity may lead to several situations where the split and surrogate variable differ only slightly in their explanatory power, rendering interpretation somewhat arbitrary.

As discussed before, trees would likely not appropriately represent predictors that exhibit only main effects and that are independent in their explanatory power from all other predictors.

Furthermore, CART exhibits a “selection bias” towards predictors with a higher number of distinct values.

For further issues of tree-based methods including adjustments of CART and alternative methods see Hastie et al. (2017): 310-313. We will discuss a few alternatives in the context of extensions.

(G)LM vs. CART

Characteristic	GLM	CART
Data Requirements		
Accommodate "mixed" data types	Low	High
Accommodate missing values of predictors	Low	High
Insensitive to monotonic transformations of predictors	Low	High
Robust to outliers in predictors	Low	Moderate
Insensitive to irrelevant predictors	Low	High
Modeling Process		
Automation (i.e., low degree of user involvement)	High	Moderate
Transparency of the modeling process	High	Moderate
Ability to model nonlinear relationships	Low	Moderate
Accommodate interactions among predictors	Low	Moderate
Model Output		
Explanatory insight and variable interpretability	High	Moderate
Predictive power	Low	Moderate
Software Availability and Ease-of-Use	High	Moderate

Additionally, CART requires larger data sets than (G)LM for meaningful results

Olden et al. 2008 *Quart. Rev. Biol.* 83: 171

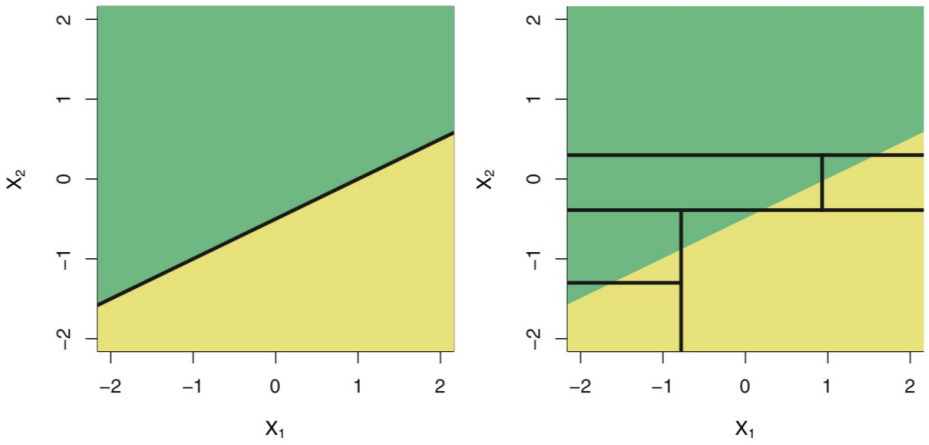
25

(G)LM vs. regression trees

(G)LM CART

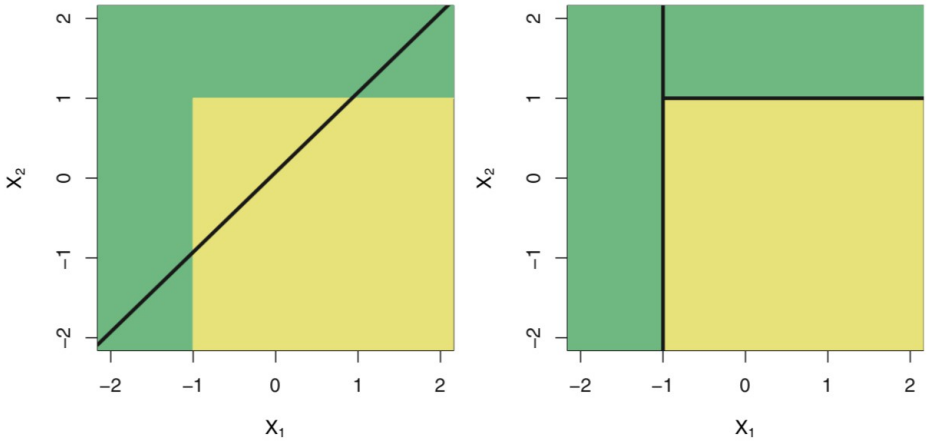
Relationship between predictors and response linear in population

→ (G)LMs outperform regression trees



Relationship between predictors and response non-linear and complex in population

→ Regression trees outperform (G)LMs



James et al. 2017: 315

In several studies, CARTs outperformed GLMs (e.g. Greenacre & Primicerio 2013, Naghibi & Pourghasemi 2015, Razi & Athappilly 2005, Munoz & Felicismo 2004, Felicismo et al. 2013), though the differences were often minor and occasionally the opposite result was observed (e.g. Kurt et al. 2008).

Felicísimo Á.M., Cuartero A., Remondo J. & Quirós E. (2013). Mapping landslide susceptibility with logistic regression, multiple adaptive regression splines, classification and regression trees, and maximum entropy methods: a comparative study. *Landslides* 10, 175–189.

Greenacre M.J. & Primicerio R. (2013) Multivariate analysis of ecological data. Fundación BBVA, Bilbao.

James G., Witten D., Hastie T. & Tibshirani R. (2017). An introduction to statistical learning: with applications in R. Springer, New York.

Kurt I., Ture M. & Kurum A.T. (2008). Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. *Expert Systems with Applications* 34, 366–374.

Muñoz J. & Felicísimo Á.M. (2004). Comparison of statistical methods commonly used in predictive modelling. *Journal of Vegetation Science* 15, 285–292.

Naghibi S.A. & Pourghasemi H.R. (2015). A Comparative Assessment Between Three Machine Learning Models and Their Performance Comparison by Bivariate and Multivariate Statistical Methods in Groundwater Potential Mapping. *Water Resources Management* 29, 5217–5236.

Razi M.A. & Athappilly K. (2005). A comparative predictive analysis of neural networks (NNs), nonlinear regression and classification and regression tree (CART) models. *Expert Systems with Applications* 29, 65–74.

Classification and regression trees and extensions

Contents

1. Supervised and unsupervised learning
2. Intro: Classification and regression trees
3. Goodness of split metrics, interactions, assumptions and predictor importance
4. Critical evaluation and comparison to (G)LMs
5. **Extensions and alternatives 1: Conditional inference trees and multivariate trees**
6. Extensions and alternatives 2: Random forest and boosted regression

27

Extensions and alternatives to CART

- CART lacks a probabilistic foundation (e.g. no information on uncertainty from confidence intervals, no p-values) → Conditional inference trees
- CART for inspecting parametric models (i.e. does the slope differ → are additional variables relevant?) → model-based recursive partitioning (Zeileis et al. 2008)
- Hierarchical CART for set of nested variables (e.g. different ecological scales) → Cascade regression trees (Ouellette et al. 2012)
- Multivariate response → Multivariate regression trees
- More robust trees → Random forest and Boosted regression trees

Model-based partitioning allows to investigate whether the relationship of a model e.g. GLM differs when considering further variables. Hence, this can be used to examine whether crucial additional variables are missing in a model and whether parameter estimates are robust.

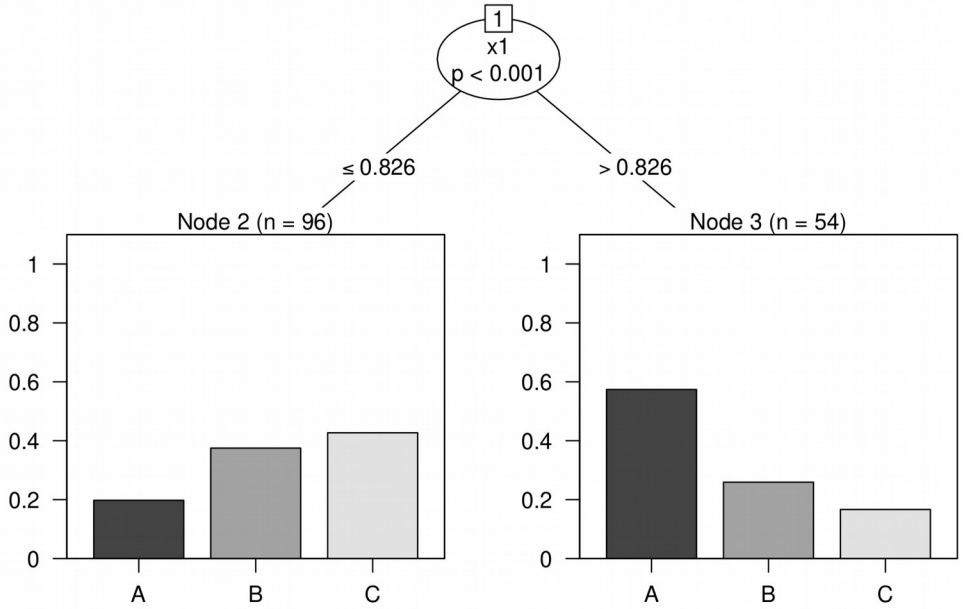
Cascade regression trees can be used for nested designs. For example, they allow to study the influence of anthropogenic stressors on species assemblages after the influence of well-known natural drivers of assemblages such as climate or geology have been accounted for.

Ouellette M.-H., Legendre P. & Borcard D. (2012). Cascade multivariate regression tree: a novel approach for modelling nested explanatory sets. *Methods in Ecology and Evolution* 3, 234–244. https://doi.org/10.1111/j.2041-210X.2011.00171.x

Zeileis A., Hothorn T. & Hornik K. (2008). Model-Based Recursive Partitioning. *Journal of Computational and Graphical Statistics* 17, 492–514. https://doi.org/10.1198/106186008X319331

Conditional inference trees

- Provides (frequentist) statistical framework for variable selection in decision trees via permutation tests
- Separates variable selection and splitting → no bias towards predictors with higher number of distinct values
- Comparison to CART: Estimation accuracy generally higher, prediction accuracy dependent on data set



29

Hothorn et al. 2006 *J. Comp. Graph.* 15: 651–674

Figure taken from https://cran.r-project.org/web/packages/partykit/vignettes/ctree.pdf

For details on why conditional inference trees (CIT) avoid the selection bias see Strobl et al. (2007).

Parameter estimation of CIT was not influenced by missing values or the number of distinct values, whereas CART exhibited a related bias (Hothorn et al. 2006). Similarly, the estimated tree more frequently matched the original data for CIT than for CART (79% vs. 63% see Hothorn et al. 2006). However, in the same study, the prediction accuracy for new observations was relatively similar (equivalent for 8 of 12 cases, in 3 cases CIT was better and in 1 case CART was better).

For an application of CIT in the area of freshwater ecology and pollution see Bunzel et al. (2013).

Bunzel K., Liess M. & Kattwinkel M. (2013). Landscape parameters driving aquatic pesticide exposure and effects. *Environmental Pollution* 186, 90–97.

Hothorn T., Hornik K. & Zeileis A. (2006) Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics* 15, 651–674.

Strobl C., Boulesteix A.-L. & Augustin T. (2007). Unbiased split selection for classification trees based on the Gini Index. *Computational Statistics & Data Analysis* 52, 483–501.

Multivariate regression trees

- Extension to multivariate case: multiple response variables
- Impurity measure summed over responses: Sum of squares to multivariate mean or median
- For community data, method allows to: (1) identify species that drive splits, (2) create biplots displaying sites and species, (3) identify representative species for nodes
- Comparison with unconstrained cluster analysis (discussed later) can inform whether relevant explanatory variables have been captured
- Alternative to ordination (discussed later), especially for non-linear and complex species-environment relationships

De‘Ath et al. 2002 *Ecology* 83: 1105

A general introduction into multivariate analysis will follow in the next session. For the time being, it is sufficient to note that multiple response variables make for a very different situation compared to a single (univariate) response.

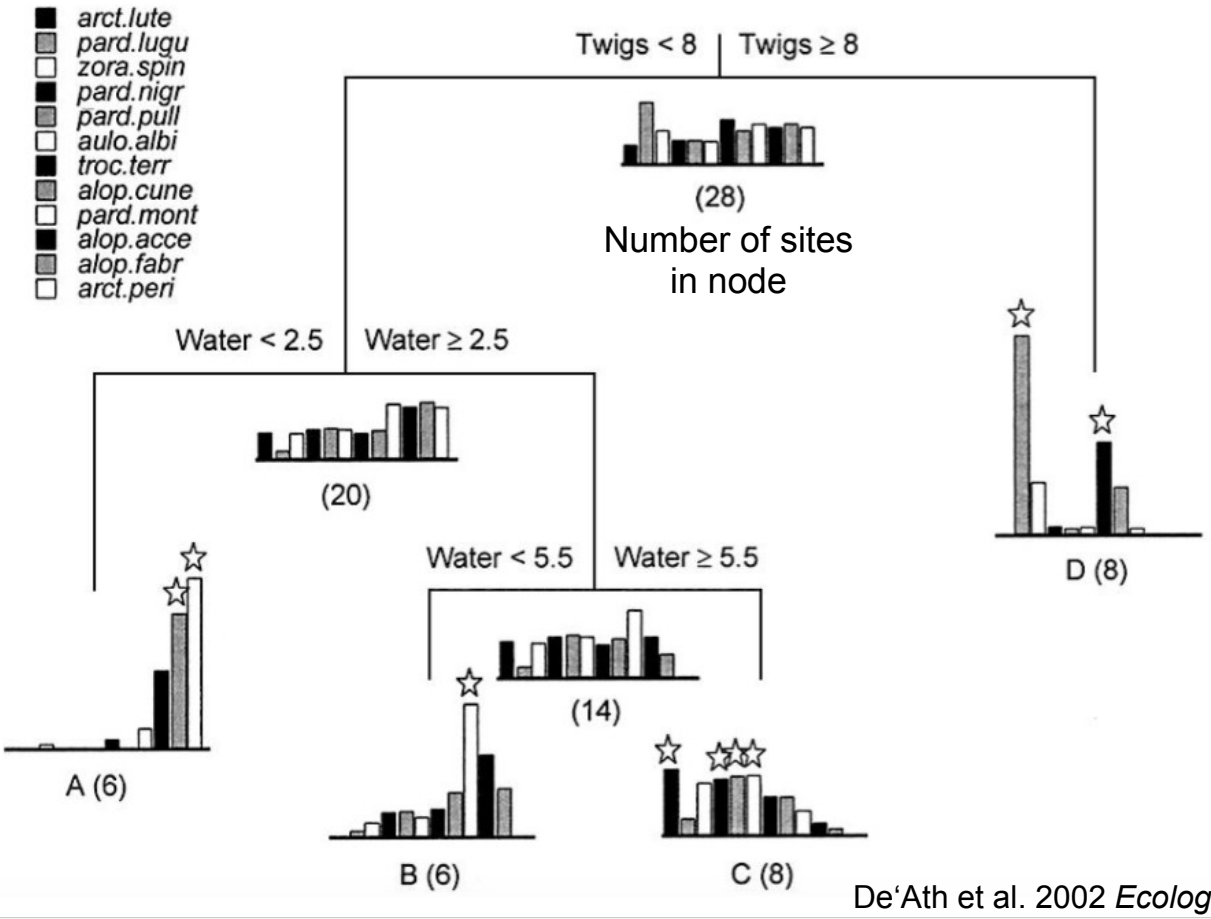
In an ecological context, the multivariate response is typically an assemblage of species, i.e. the abundances of different species across sites.

Cluster analysis and ordination are discussed later in the course. Unconstrained means that potential explanatory variables (in a second data set) are not considered in the analysis. If the consideration and non-consideration of explanatory variables leads to fairly different grouping structures, then the different grouping structure in the unconstrained case points to the omission of relevant explanatory variables (e.g. environmental variables or ecological processes).

De’Ath G. (2002) Multivariate regression trees: a new technique for modeling species-environment relationships. *Ecology* 83, 1105–1117.

Multivariate regression trees: Results

Bars display means of species, stars denote indicator species



De'Ath et al. 2002 *Ecology* 83: 1105

Indicator species are species that indicate certain environmental conditions and therefore can be used to establish groups of sites.

Multivariate regression trees: Results

Analysis provides information on variance explained by tree and its splits.

TABLE 1. Tabulation of species variance for the tree analysis of the hunting spider data.

Species	Species variance (%) explained by tree splits and whole tree				
	Twigs < 8	Water < 2.5	Water < 5.5	Tree total	Species total
Arctosa perita	1.90	15.54	0.00	17.45	20.85
Alopecosa fabrilis	2.33	6.72	0.79	9.83	13.44
Pardosa monticola	1.75	1.34	5.06	8.16	10.82
Alopecosa accentuata	1.93	0.70	2.15	4.78	5.77
Arctosa lutetiana	0.47	0.71	1.78	2.96	4.37
Aulonia albimana	0.35	0.90	0.70	1.96	3.28
Pardosa pullata	0.47	0.99	0.48	1.94	2.53
Pardosa nigriceps	0.20	0.88	0.41	1.49	2.23
Zora spinimana	0.80	0.64	0.64	2.08	4.04
Alopecosa cuneata	0.25	0.84	0.02	1.11	2.89
Pardosa lugubris	23.22	0.02	0.04	23.28	24.41
Trochosa terricola	3.29	0.40	0.05	3.74	5.38
Total species variance	36.97	29.69	12.12	78.78	100.00

Notes: The total species variance is partitioned by species, the whole tree, and the three splits of the tree. The first split is dominated by *Pardosa lugubris* (23 percentage points of 37% of the total species variance explained by that split), the second by *Arctosa perita* (15 percentage points of 30%), and the third by *Pardosa monticola* (5 percentage points of 12%).

De'Ath et al. 2002 *Ecology* 83: 1105

32

Classification and regression trees and extensions

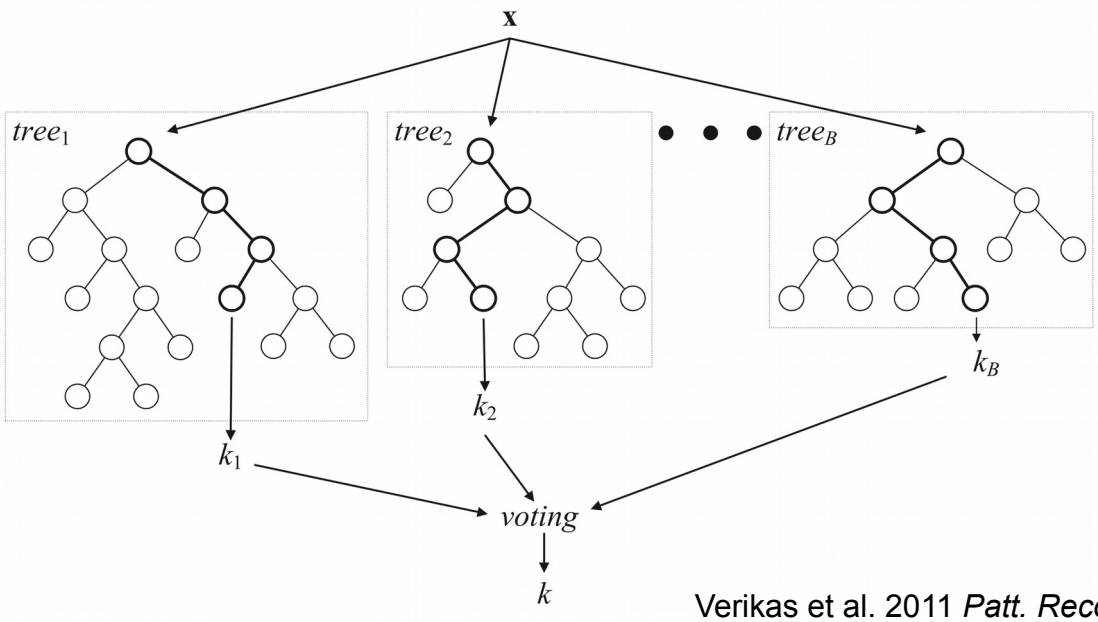
Contents

1. Supervised and unsupervised learning
2. Intro: Classification and regression trees
3. Goodness of split metrics, interactions, assumptions and predictor importance
4. Critical evaluation and comparison to (G)LMs
5. Extensions and alternatives 1: Conditional inference trees and multivariate trees
6. **Extensions and alternatives 2: Random forest and boosted regression**

33

Random forest (RF)

- Single tree unstable, but average of multiple trees reliable and provides unbiased estimation or prediction → Grow many trees (forest) on bootstrapped samples
- At each split a predictor subset is randomly selected: Variable masked by other predictors can be identified and may improve global performance



34

Verikas et al. 2011 *Patt. Recogn.* 44: 330

The seminal paper on Random Forest (Breiman et al. 2011) has received more than 40,000 citations (google scholar, 18.2.2018). The method is applied in many disciplines and is among the most popular machine learning methods. Strobl et al. (2009) provide a very readable introduction to RF. For a more technical introduction see Verikas et al. (2011) and Biau & Scornet (2016). Verikas et al (2011) also provide an overview of many studies comparing RF to other methods of data analysis. Biau & Scornet (2016) feature the theoretical background of RF including recent developments.

Random forests are build on the premise, supported by many studies (see Strobl et al. 2009), that an individual tree is unstable (e.g. tree can vary substantially for minor changes in the data), but that overall the method is unbiased and on average yields to a reliable estimation and prediction.

Since splitting is evaluated locally (i.e. best predictor for a given parent node) ignoring the consequence for the whole tree, a variable can be locally more relevant, but lead to a globally lower estimation or prediction accuracy (the same applies to other sequential methods such as stepwise variable selection for the linear regression model). The random forest method randomly restricts the number of predictors available per split, which allows to identify predictors that lead to a lower improvement in the local goodness-of-split metric, but to a higher global prediction accuracy.

Biau G. & Scornet E. (2016). A random forest guided tour. *TEST* 25, 197–227.

Breiman L. (2001) Random Forests. *Machine Learning* 45, 5–32.

Strobl C., Malley J. & Tutz G. (2009) An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods* 14, 323–348.

Verikas A., Gelzinis A. & Bacauskiene M. (2011) Mining data with random forests: A survey and results of new tests. *Pattern Recognition* 44, 330 – 349.

Random forest (RF)

- Ensemble method: Prediction requires aggregation (via mean (regression) or majority vote (classification))
- Evaluation of prediction accuracy: Out-of-bag (OOB) error (observations not used in tree building)
- Variable importance:
  - Average improvement of node impurity over all trees
  - Better: Difference in prediction accuracy before and after permutation
- RFs typically exhibit a prediction accuracy similar or higher than other methods (see Verikas et al. 2011), but results are less easy to interpret: There is no average tree!

Bagging is another extension of CART to enhance predictive power, but is not discussed in this course, because it has been superseded by random forest and boosted regression trees (See James et al. 2017: 316ff and Strobl et al. 2009). Moreover, bagging is a special case of RF, where the number of selected predictors is set to the total number of predictors.

Some observations, the so-called out-of-bag observations, are not included in the learning process and can be used for evaluating prediction accuracy. What is the difference to cross-validation? A bootstrapped sample is used (typically of the same size as the initial sample size), instead of holding back parts of the original observations and fitting the model to a smaller sample.

Both variable importance measures are provided in the RandomForest R package. The permutation method should be preferred. It calculates the average prediction accuracy (e.g. for classification: the number of correctly classified out-of-bag observations) over all trees before and after permuting a variable. For details see Strobl et al. (2009).

Note that the importance scores should not be compared across studies, as they depend on the choice of tuning parameters (e.g. tree size, number of predictors selected per split). Furthermore, Strobl et al. (2009) caution against the use of statistical tests on importance scores and rather suggest to exclude variables where the importance score is close to zero.

For an overview on recent developments regarding variable importance in the context of random forests, in particular with respect to genetic analysis, see Brieuc et a. (2018). This publication also features R code.

Brieuc M.S.O., Waters C.D., Drinan D.P. & Naish K.A. (2018). A practical introduction to Random Forest for genetic association studies in ecology and evolution. *Molecular Ecology Resources* 18, 755–766.

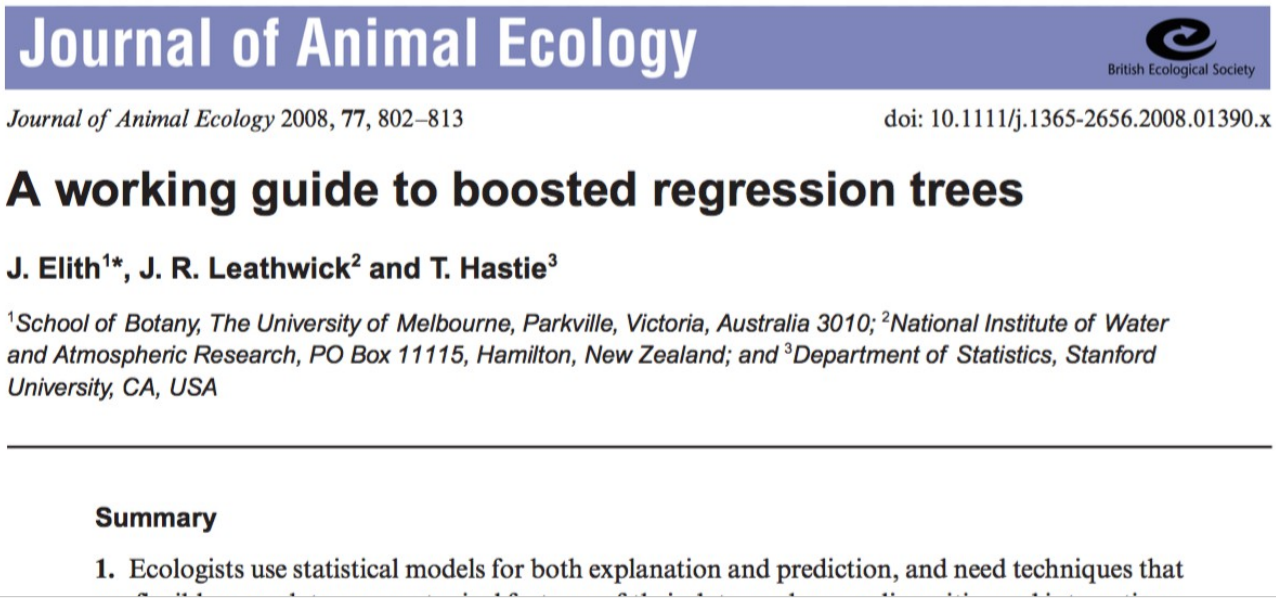
James G., Witten D., Hastie T. and Tibshirani R. 2017. An introduction to statistical learning: with applications in R. Springer: NY.

Strobl C., Malley J. & Tutz G. (2009) An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods* 14, 323–348.

Verikas A., Gelzinis A. & Bacauskiene M. (2011) Mining data with random forests: A survey and results of new tests. *Pattern Recognition* 44, 330 – 349.

Boosted regression trees (BRT)

- Related to *Boosting* (Ensemble method)
- Sequential growing of trees, focus on minimising residuals from previous trees to improve accuracy
- Typically (slightly) higher prediction accuracy than RF, but:
  - Larger number of trees required → computationally more costly
  - More tuning parameters → more complex



Journal of Animal Ecology

British Ecological Society

Journal of Animal Ecology 2008, 77, 802–813 doi: 10.1111/j.1365-2656.2008.01390.x

A working guide to boosted regression trees

J. Elith[1]*, J. R. Leathwick[2] and T. Hastie[3]

[1]School of Botany, The University of Melbourne, Parkville, Victoria, Australia 3010; [2]National Institute of Water and Atmospheric Research, PO Box 11115, Hamilton, New Zealand; and [3]Department of Statistics, Stanford University, CA, USA

Summary

1. Ecologists use statistical models for both explanation and prediction, and need techniques that

Boosting is beyond the level of this course, for an introduction see Hastie et al. (2017: Chapter 10). They praise it as one of the most powerful learning method of the last two decades.

According to Strobl et al. (2009), in some studies RF outperformed BRT, whereas in others these outperformed RFs. However, in most cases in Hastie et al. (2017: Chapter 15), BRT outperform RF. Moreover, they show that RF is more susceptible to overfitting, but this only became problematic when the ratio of noise to relevant variables was a factor 10 or higher Hastie et al. (2017: 597).

When should you, in general, select a RF or BRT over LASSO and related methods? If it cannot be assumed that the problem at hand is sparse, linear and interactions play no major role.

Hastie, T.; Tibshirani, R.; Friedman, J. 2017. The elements of statistical learning: data mining, inference, and prediction. 12th printing with corrections; Springer: New York, NY.

Strobl C., Malley J. & Tutz G. (2009) An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods* 14, 323–348.