



Master Seminar WS 18/19

Blockseminar

Patrick Schäfer

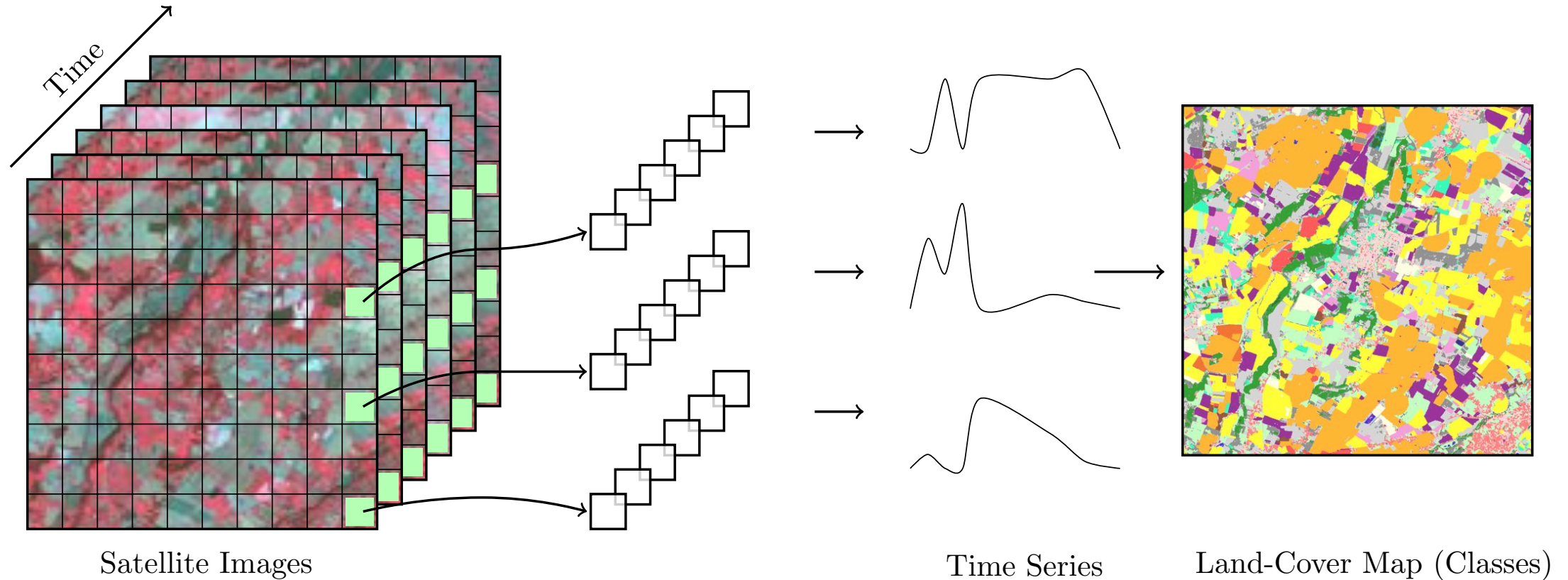
Friday, 01.02.2019

URL: <https://hu.berlin/landnutzung>

Agenda

- Today:
 - Blockseminar (**1.2.19 15-18 Uhr, RUD 25 4.410**)
 - Present your topic (30 min)
 - Some Dataset Information
 - Results of the Competition

From satellite images to pixel time series



From: Tan, Chang Wei, Geoffrey I. Webb, and François Petitjean. "Indexing and classifying gigabytes of time series under time warping." *Proceedings of the 2017 SIAM International Conference on Data Mining*. Society for Industrial and Applied Mathematics, 2017.

NDVI time-series

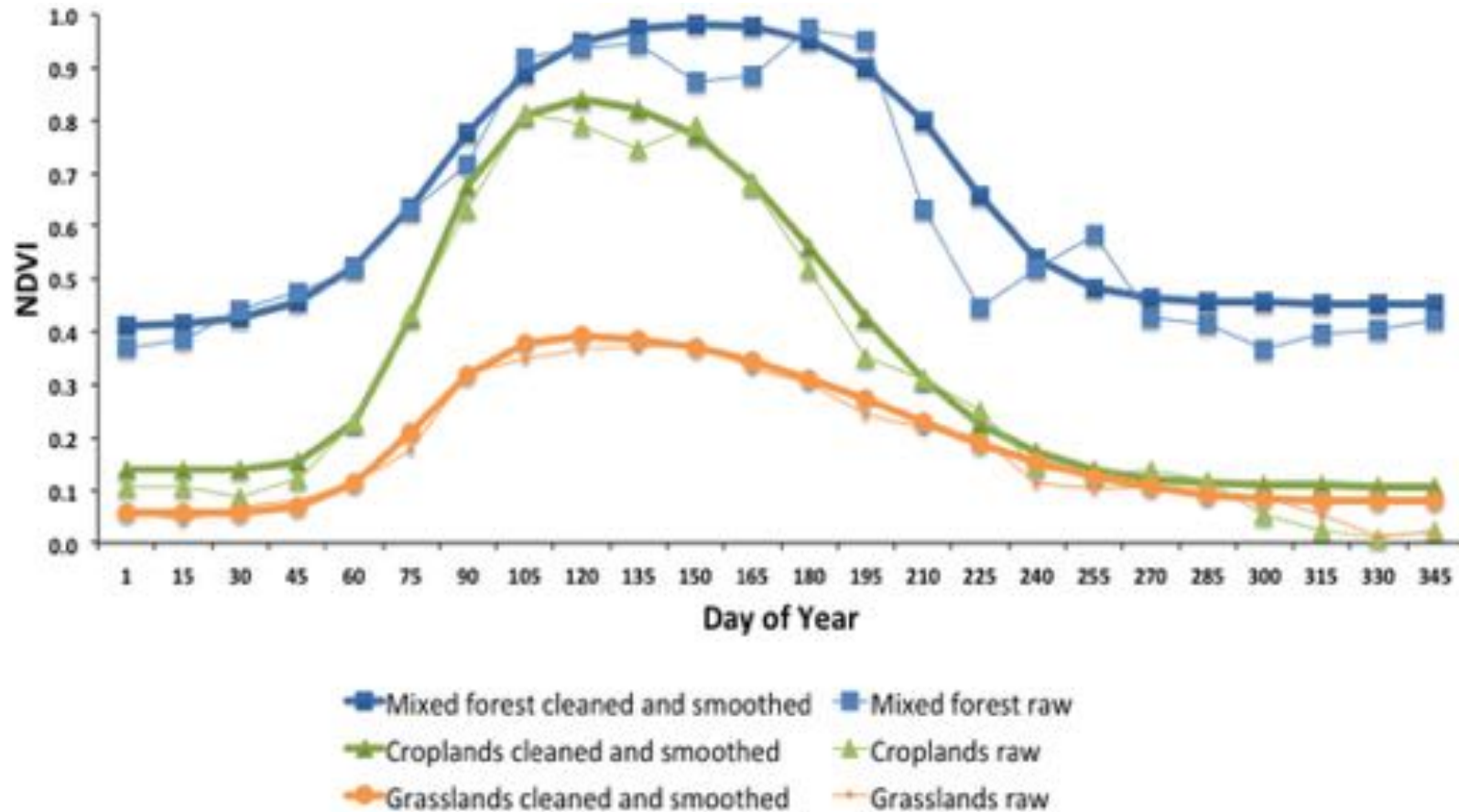


Fig. 3. Raw, and cleaned and smoothed NDVI time-series of mixed forest, croplands, and grasslands.

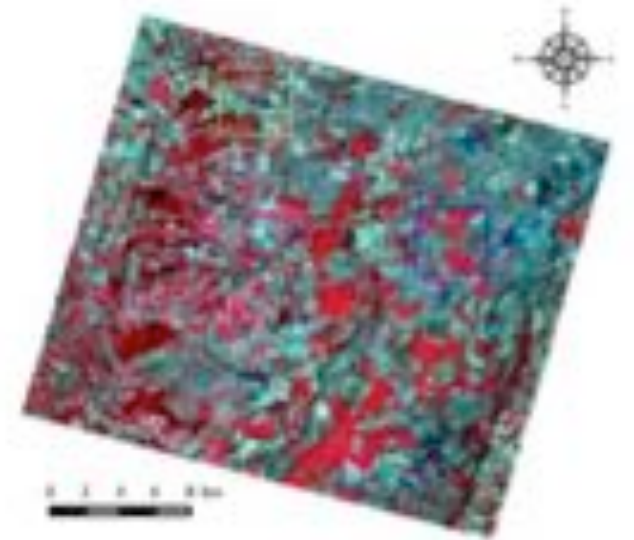
From: He, Yaqian, Eungul Lee, and Timothy A. Warner. "A time series of annual land use and land cover maps of China from 1982 to 2013 generated using AVHRR GIMMS NDVI3g data." *Remote Sensing of Environment* 199 (2017): 201-217.

Recap: Data Preparation

- Regarding feature normalization [1]:
 - [...] In machine learning, the input data are generally standardized by subtracting the mean and divided by the standard deviation for each feature where each time stamp is considered as a separate feature [...]
 - In machine learning, the input data are generally z-normalized by subtracting the mean and divided by the standard deviation for each time series. [...] z-normalization [...] leads to a loss of the significance of the magnitude that it is recognized as crucial for vegetation mapping, *e.g.* the corn will have higher NDVI values than other summer crops.

Recap: Train/Test dataset

- A **massive** land cover pixel time series (TS) dataset
 - 46 geometrically and radio-metrically corrected images taken by FORMOSAT-2
 - **Train data**: 6 mio pixels TS, 2,4GB
 - **Test data (hold-back, kaggle)**: 20.000 pixels
 - 46 time stamps between 06.2 and 29.11.2006
 - 3 surface reflectances: Near-Infra-Red, Red, Green
 - In total 3x46 values per pixel time series
- Contains **missing values** ,?’
- Overall, 24 land cover classes, labelled by experts
- **Note: This data is provided for the class only and it has to be deleted once the seminar is over**



<https://arxiv.org/pdf/1811.10166.pdf>

Recap: 24 Class Labels

prairie temporaire is mapped to #0

ble is mapped to #1

pre is mapped to #2

feuillus is mapped to #3

tournesol is mapped to #4

mais ensilage is mapped to #5

jachere is mapped to #6

bati dense is mapped to #7

bati diffus is mapped to #8

friche is mapped to #9

resineux is mapped to #10

sorgho is mapped to #11

pois is mapped to #12

orge is mapped to #13

bati indu is mapped to #14

soja is mapped to #15

eau is mapped to #16

eucalyptus is mapped to #17

colza is mapped to #18

lac is mapped to #19

peupliers is mapped to #20







mais is mapped to #21

graviere is mapped to #22


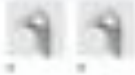




surface minerale is mapped to #23

Competition

Kaggle: Public Leaderboard

#	Team Name	Kernel	Team Members	Score @	Entries	Last
📍	Best Found Model (XGBoost)			0.77240		
1	Distributed Multivariate BoP ...			0.73750	13	5d
📍	Random Forest Benchmark			0.73200		
2	non-time-series-based			0.72860	9	5d
3	Univariate Dictionary-based			0.72780	18	2d
4	deep learning			0.67510	4	16d
5	UnivariateShapelets			0.62700	18	5d
📍	1-NN Manhattan Distance BFI...			0.62270		
6	DTW			0.59340	6	6d
📍	1-NN Manhattan Distance Na...			0.54810		

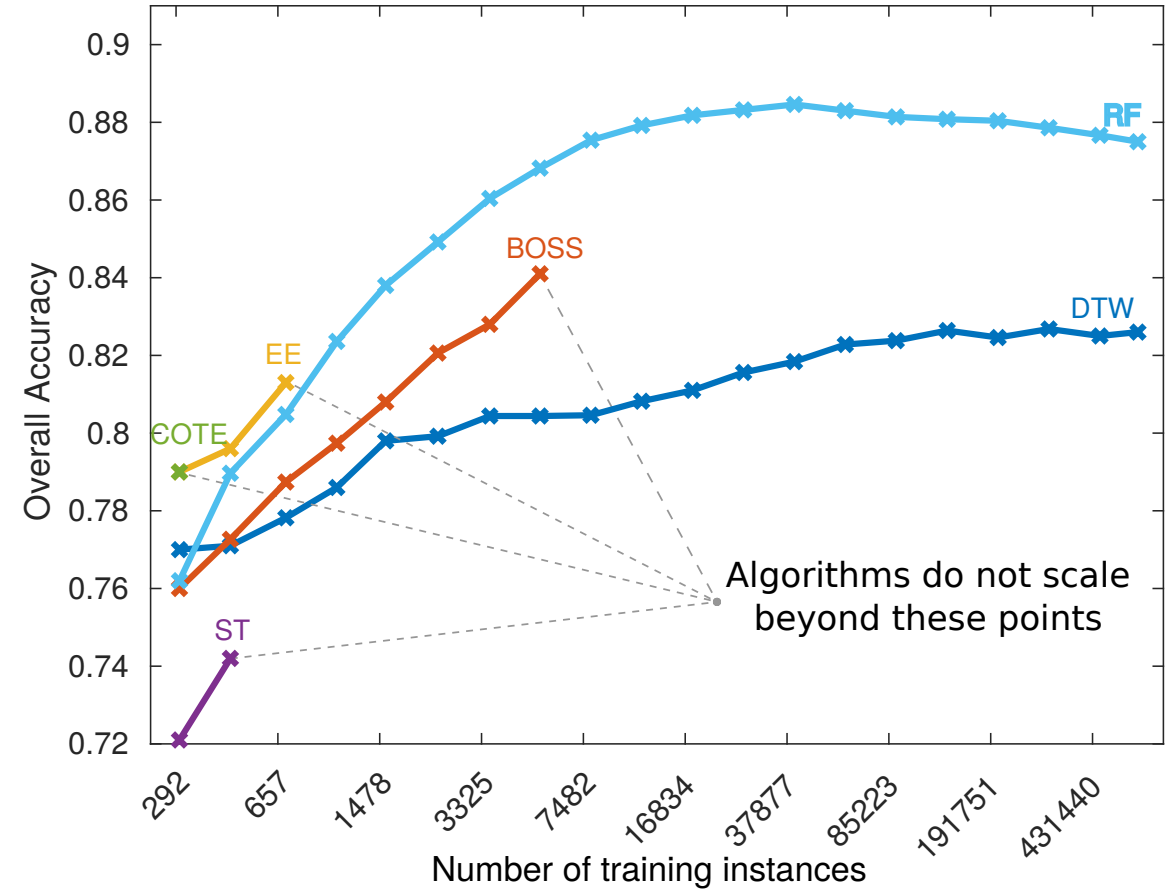
Kaggle: Private Leaderboard

#	Team Name	Kernel	Team Members	Score @	Entries	Last
📍	Best Found Model (XGBoost)			0.77350		
1	Distributed Multivariate BoP ...			0.73960	13	6d
📍	Random Forest Benchmark			0.73880		
2	Univariate Dictionary-based			0.73010	18	3d
3	non-time-series-based			0.72930	9	5d
4	deep learning			0.67870	4	16d
📍	1-NN Manhattan Distance BFI...			0.63060		
5	UnivariateShapelets			0.62890	18	6d
6	DTW			0.59419	6	7d
📍	1-NN Manhattan Distance Na...			0.55230		

Topic	Approach
(non-time series)	Feature Extraction/Selection: (a) Statistische Werte (TSFresh), NDVI, saisonale Features (Frühling, Sommer, Winter), Moving Averages, insgesamt 738 features => 700 Feature-Reduktion, jeder Fold gleiche Anzahl Samples/Klasse. (b) Autoencoder: supervised, Dense NN, kombiniert (70 Features), einzeln (40 Features) Classifier: Random Forests (200 trees)
Whole-Series	Features/Preprocessing: Imputation (linear), 3 bands: red, green, NIR Classifier: 1-NN, multivariates DTW / ED Probleme: Warping Window noch nicht getestet, NDVI
Univariate Shapelet	Features/Preprocessing: Interpolation (zero filling, linear, bfill), MinMax-Normierung auf -1 und 1 (teils problematisch) Classifier: Fast Shapelets (Shapelet Discovery) Probleme: z-normalisierung, 0-Filling
Univariate Dictionary	Features/Preprocessing: NDVI, Interpolation (0 filling best on train data) Classifier: SFA, WEASEL Probleme: Bebauung schwer zu unterscheiden (Green-Index im NDVI fehlt)
Multivariate Dictionary	Features/Preprocessing: NDVI, Range Normalization, Time-Synchronization (2 Tages-Intervalle, gleiche Intervalllängen vergleichen), Backward-Fill Classifier: SAX (keine Mean-Bildung) und Bag-of-Pattern pro Channel, Random Forests classifier, Concurrent implementation Probleme: Skalierbarkeit (Memory),
Deep Learning	Feature/Processing: TinyDNN (days) / Keras (Minutes), Red, Green, NIR, NDVI Classifier: TimeCNN, ResNet, FCN

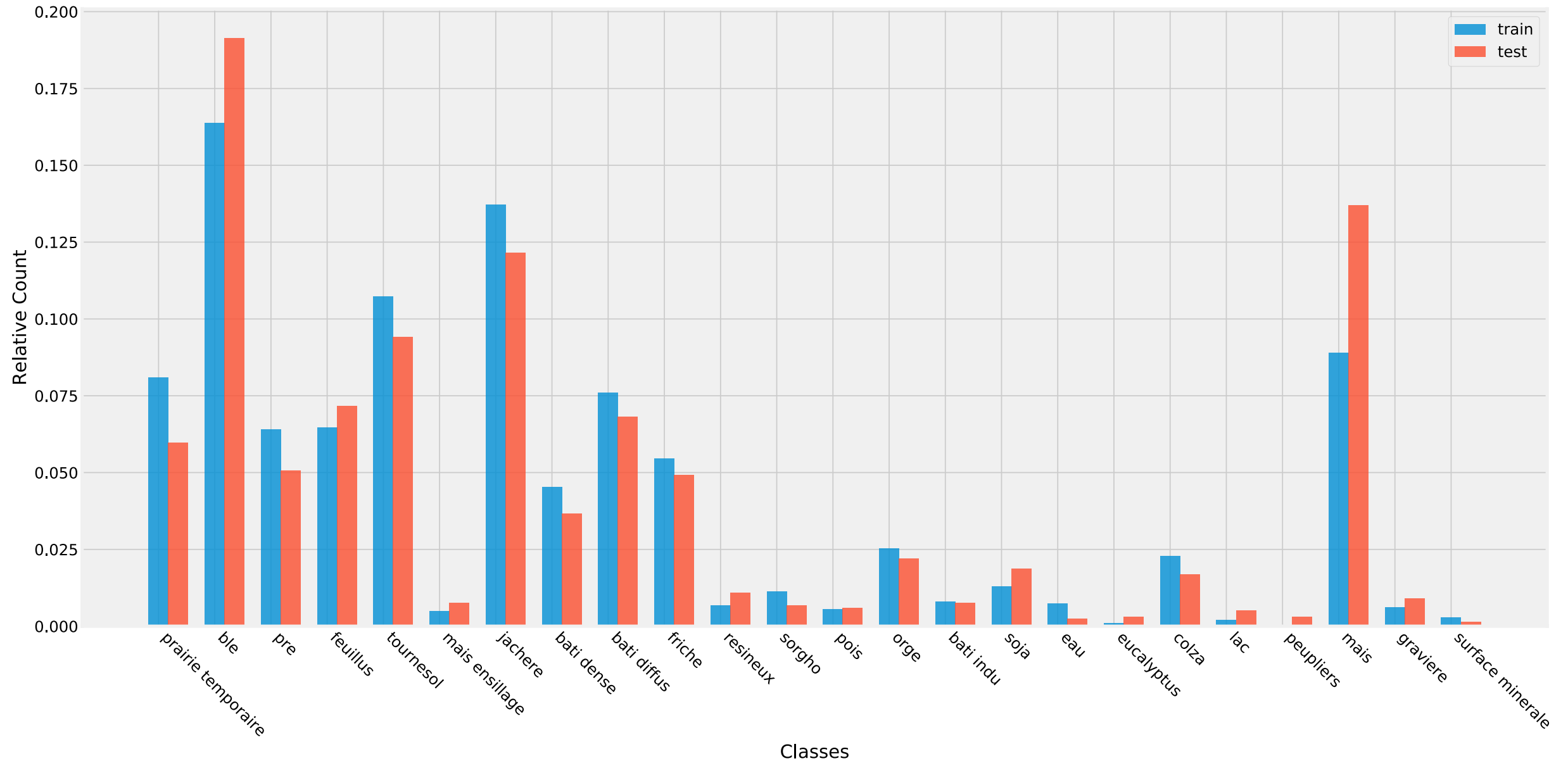
Recap: Accuracy...

- Competitors
 - DTW (warping window size is fixed at 25%)
 - Elastic Ensemble
 - BOSS
 - Shapelet Transform
 - COTE
- Only NDVI features
- Only 1000 test samples
- Limit at 24 hours single core runtime
- Normalization?
- Using (inefficient) codes from www.timeseriesclassification.com



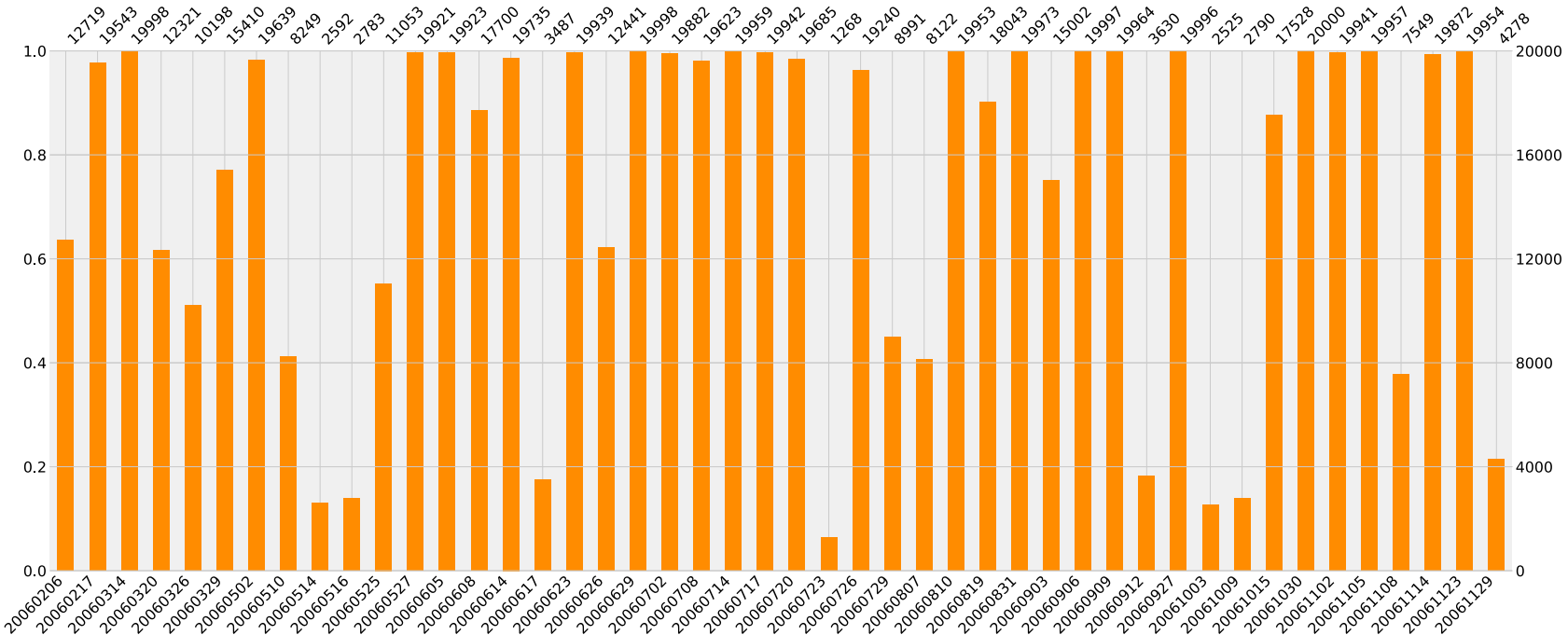
[1] <https://arxiv.org/pdf/1811.10166.pdf>

Class Distribution: Test / Train

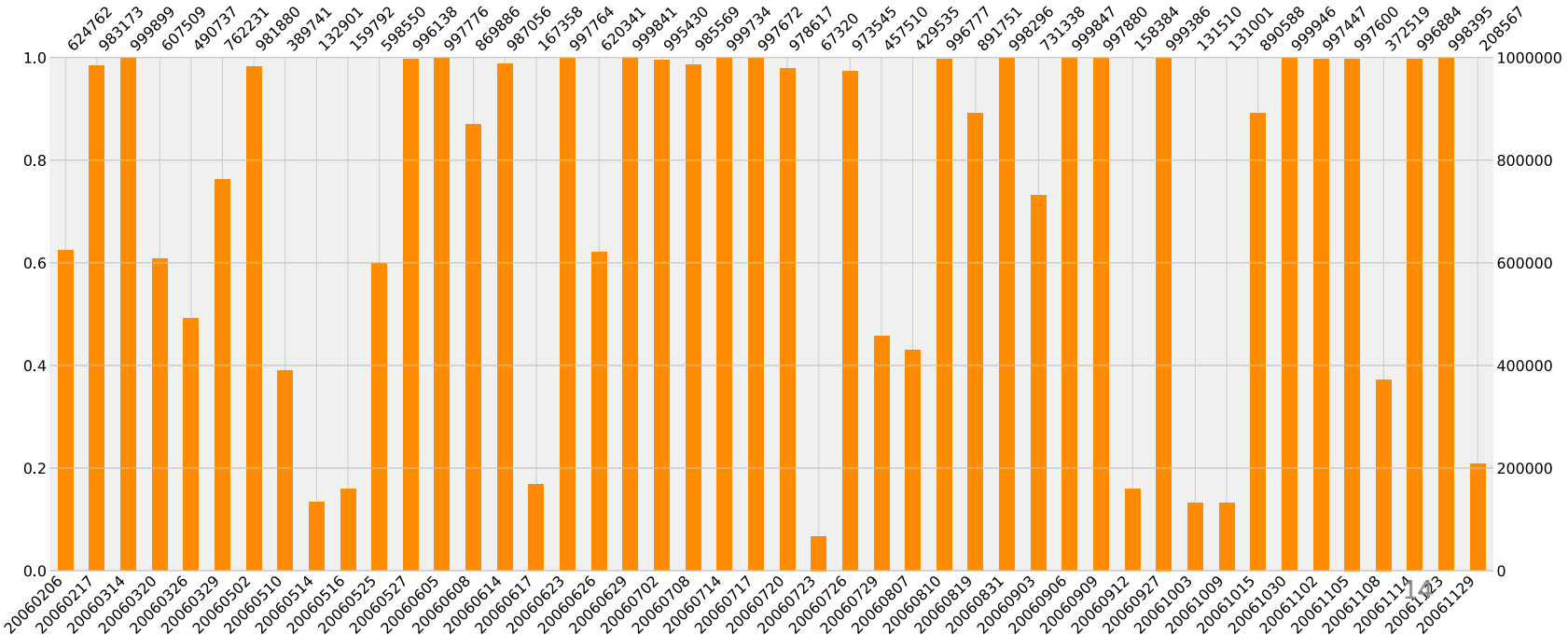


Missing Values Train / Test

Test data
20.000 samples



Train data
Subset 10^6 samples



My best found model

- Pre-processing / imputation method used:

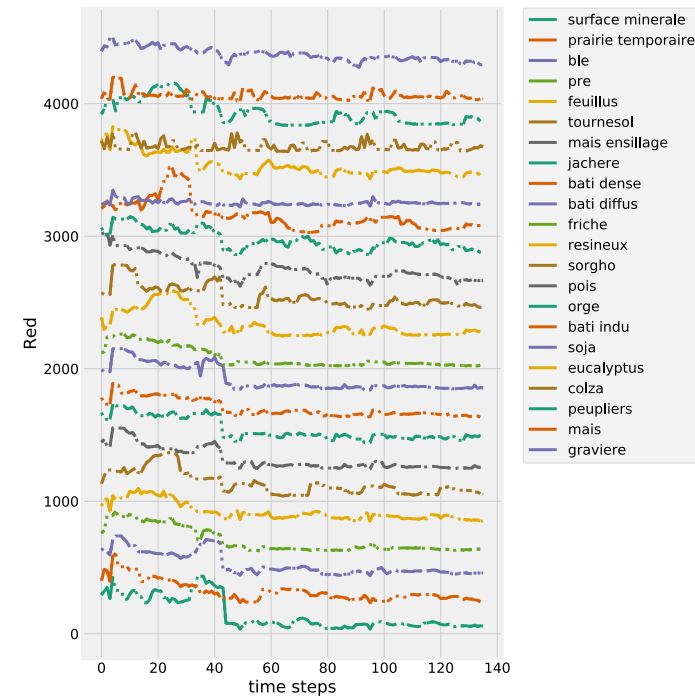
- Backward-fill using last value
- Forward-fill using last value
- No normalization applied

- Feature Engineering [1]:

- 3 Spectral Bands: Red, Green, NIR
- NDVI (with red): $(\text{NIR} - \text{red}) / (\text{NIR} + \text{red})$
- NDWI (with green): $(\text{NIR} - \text{green}) / (\text{NIR} + \text{green})$
- Chlorophyll Vegetation Index: $(\text{NIR} * \text{red}) / (\text{green}^2)$

- Classifiers:

- Random Forests with 1000 trees, all samples: **75.4% on Kaggle**
- Gradient Boosting with 1000 trees, 10^6 samples: **76.8% on Kaggle**



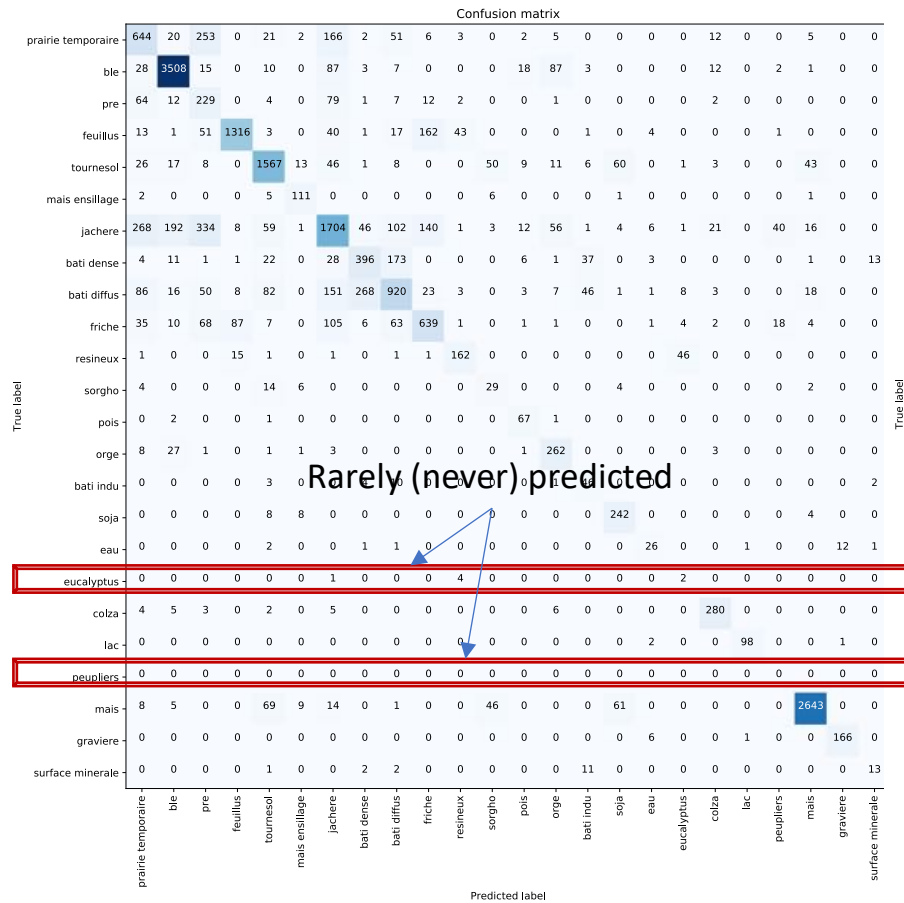
Classification-Report: Sensitivity

- GBT much better at representing underrepresented (low support) classes
- Most prominent class: peupliers with 0% for RF and 33% for GB

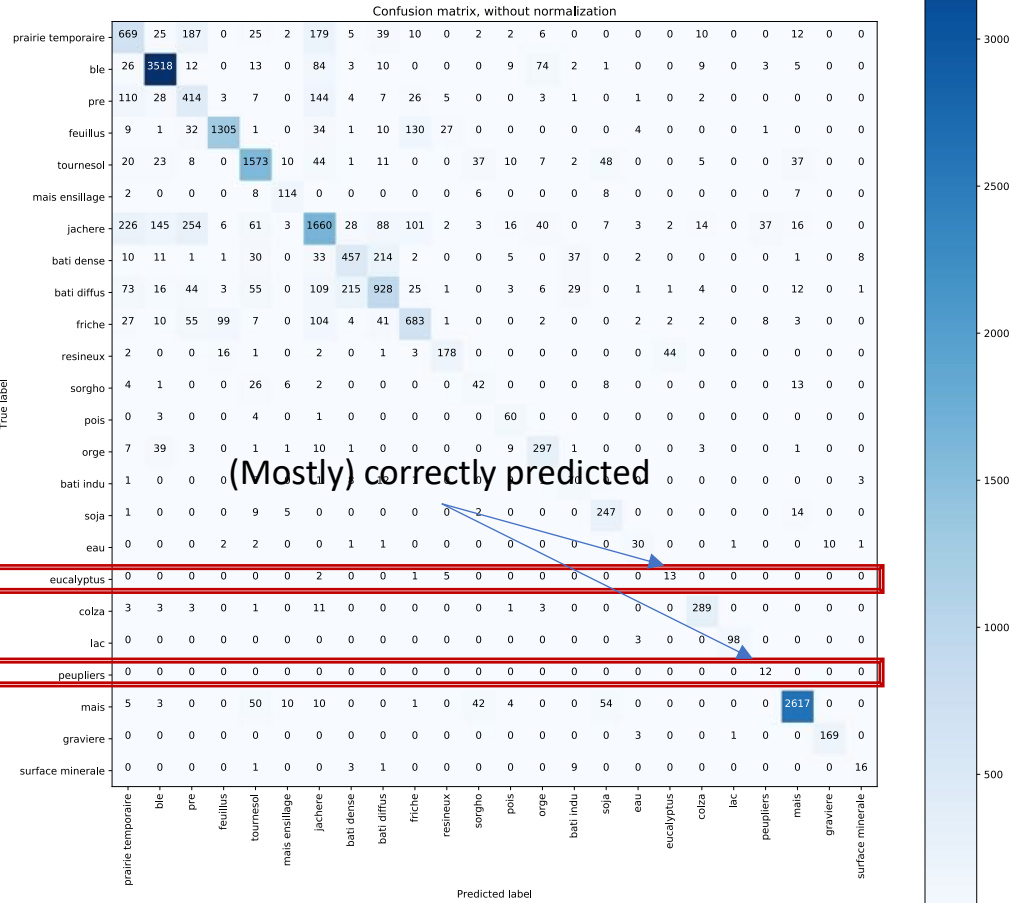
	Random Forests (RF)			Gradient Boosting Trees (GBT)			
Class Labels	Precision	Recall	F1-score	Precision	Recall	F1-score	Class Support
prairietemporaire	54%	52%	53%	57%	56%	57%	1195
ble	93%	91%	92%	93%	92%	93%	3826
pre	55%	21%	30%	55%	41%	47%	1013
feuillus	79%	93%	85%	84%	91%	87%	1435
tournesol	82%	83%	83%	86%	84%	85%	1882
maisensillage	89%	73%	80%	79%	75%	77%	151
jachere	55%	70%	62%	61%	68%	65%	2430
batidense	56%	54%	55%	56%	63%	59%	731
batidiffus	54%	66%	60%	61%	68%	64%	1363
friche	60%	64%	62%	65%	69%	67%	983
resineux	74%	76%	75%	72%	81%	76%	219
sorgho	38%	15%	22%	41%	31%	36%	134
pois	100%	37%	54%	88%	50%	64%	119
orge	82%	58%	68%	80%	68%	73%	439
batiindu	68%	30%	42%	67%	46%	55%	151
soja	94%	64%	76%	89%	66%	76%	373
eau	53%	53%	53%	62%	61%	62%	49
eucalyptus	50%	6%	11%	62%	21%	31%	62
colza	91%	82%	86%	92%	86%	89%	338
lac	98%	94%	96%	97%	98%	98%	100
peupliers	0%	0%	0%	100%	20%	33%	61
mais	93%	96%	95%	94%	96%	95%	2738
graviere	96%	93%	94%	98%	94%	96%	179
surfacemineral	43%	41%	42%	53%	55%	54%	29
avg/total	75%	75%	74%	78%	77%	77%	20000

Confusion Matrix

Random Forests



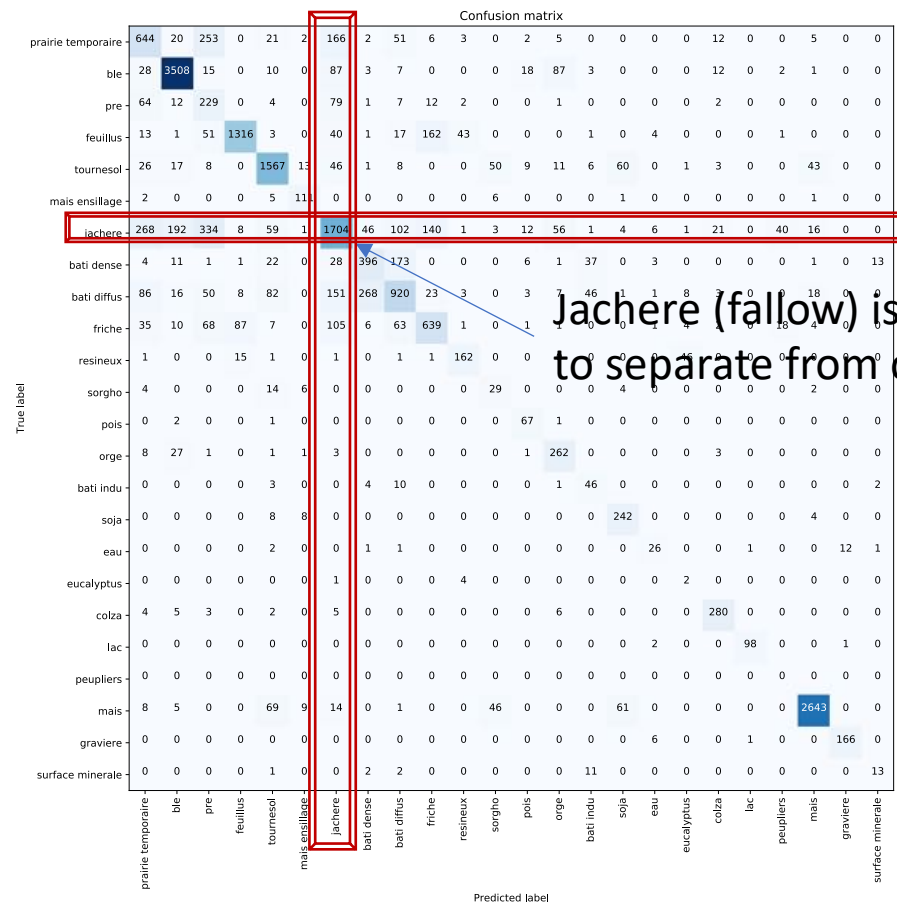
Gradient Boosting Trees



Classes with low support are often correctly predicted in the GBT but never predicted in the RF

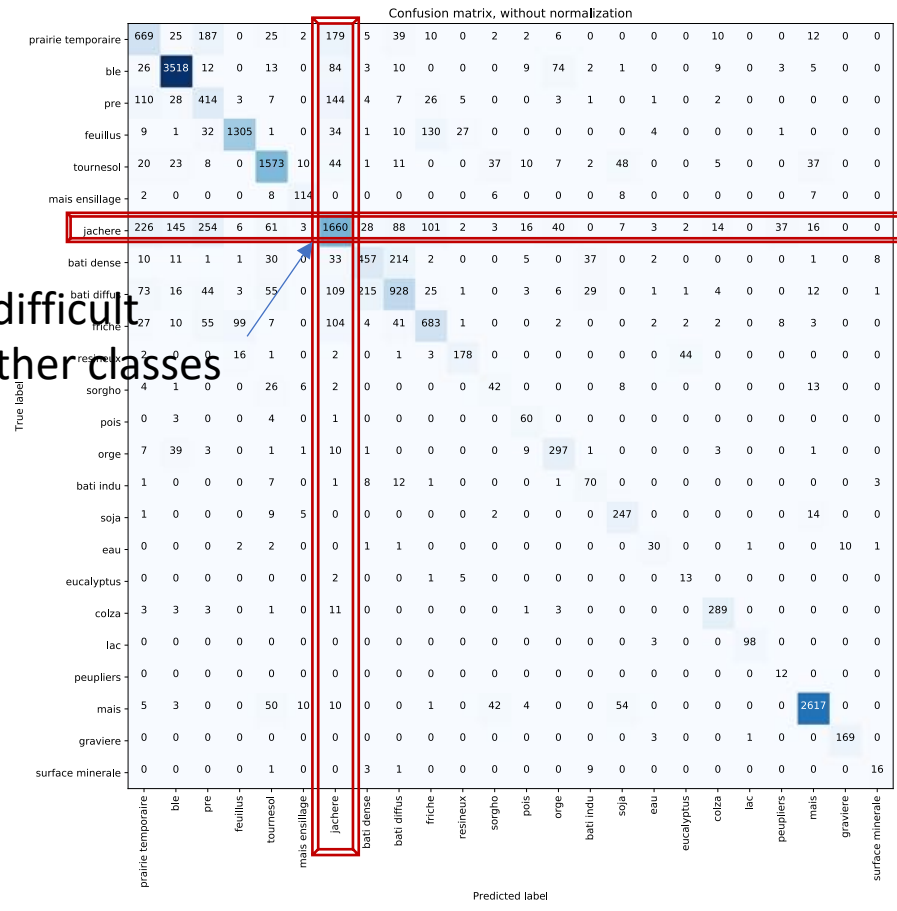
Confusion Matrix

Random Forests



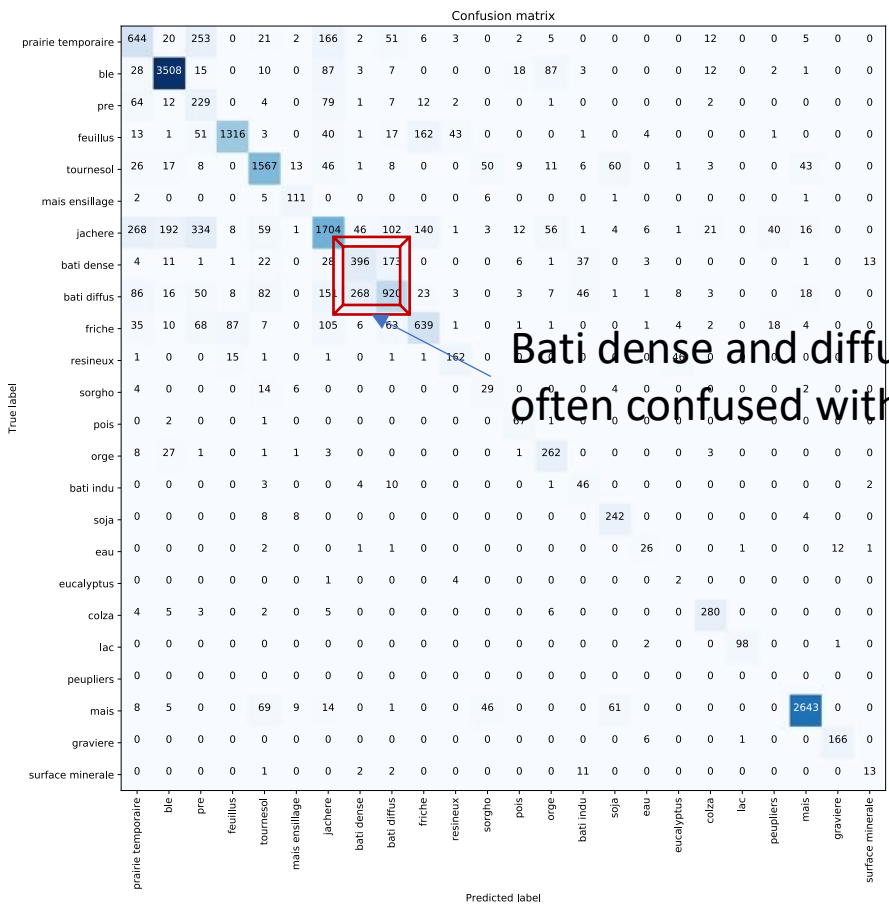
Jachere (fallow) is difficult to separate from other classes

Gradient Boosting Trees



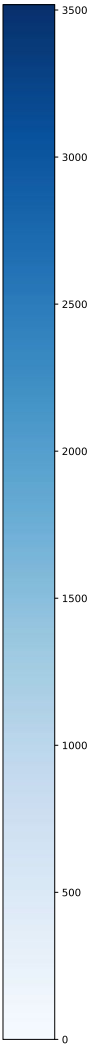
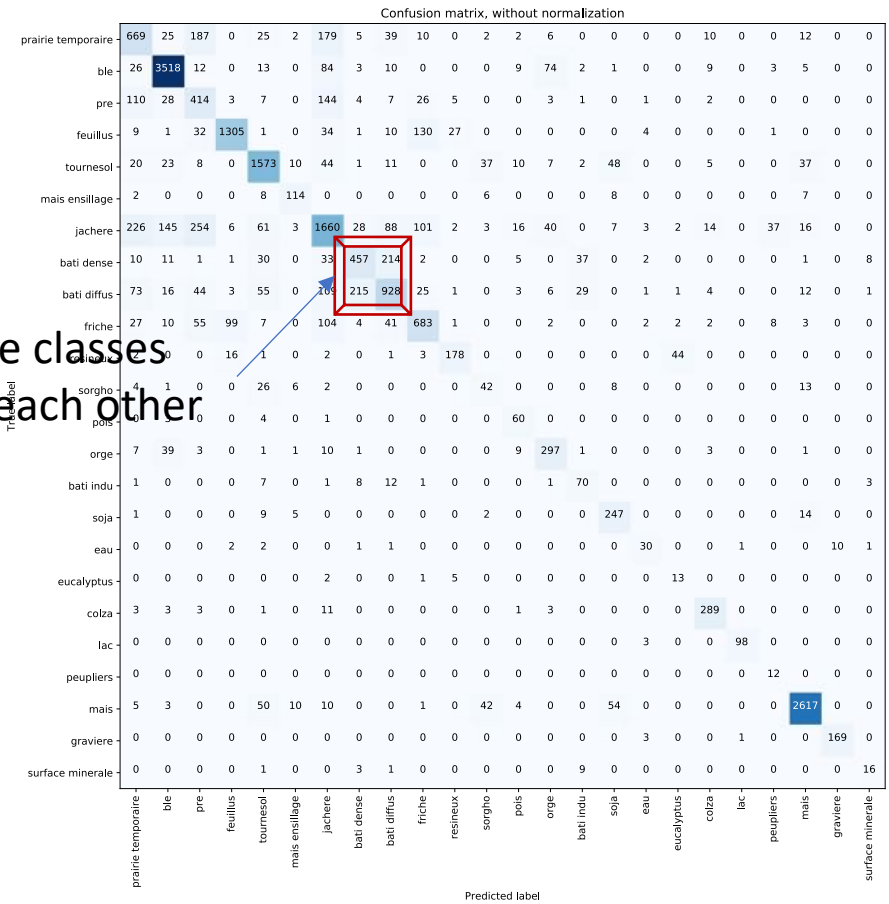
Confusion Matrix

Random Forests



Bati dense and diffuse classes often confused with each other

Gradient Boosting Trees



Next steps...

- Please, send me your presentation slides
- Seminar Thesis before **31.03.2019**
 - **write seminar thesis (~20 pages)**
https://hu.berlin/checkliste_seminar

Hinweise zur Ausarbeitung

- Eine elektronische Version schicken (± 20 Seiten)
 - Selbstständigkeitserklärung (einscannen oder abgeben) unterschreiben
- Referenzen:
 - Im Text referenzieren, Liste am Schluss
- Korrekt zitieren
 - Vorsicht vor Übernahme von kompletten Textpassagen oder Abbildungen; wenn, dann deutlich kennzeichnen
 - Aussagen mit Evidenz oder Verweis auf Literatur versehen
- Siehe: https://hu.berlin/checkliste_seminar

Questions?