STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®

Random Forest

Object-oriented mapping of urban trees using Random Forest classifiers

Daniel Kadyrov

CS559 – Machine Learning

Spring 2020

Presentation #3

Agenda

- Puissant, Anne & Rougier, Simon & Stumpf, Andre. (2014). Object-oriented mapping of urban trees using Random Forest classifiers. International Journal of Applied Earth Observation and Geoinformation. 26. 235–245. 10.1016/j.jag.2013.07.002.
  - https://www.researchgate.net/profile/Andre_Stumpf/publication/259124894_Object-oriented_mapping_of_urban_trees_using_Random_Forest_classifiers/links/5caf31ea299bf120975ddc2c/Object-oriented-mapping-of-urban-trees-using-Random-Forest-classifiers.pdf
- Introduction
- Study Site and Data
- Method
- Results and Discussion
- Conclusion

Introduction

- Many European cities enacted policies to track land cover use regarding vegetation
  - Topographical databases are rarely updated
  - Mapping of urban trees is difficult and expensive.
- Unsupervised and supervised per-pixel algorithms have been proposed to analyze very high resolution (VHR) satellite image.
  - Peterson et al., 2004; Tooke et al., 2009
  - Per-pixel classifications based on spectral characteristics alone are insufficient
    - Tuominen and Pekkarinen, 2005; Sheeren et al., 2009
- Object-based image analysis (OBIA) uses segmented images in homogenous regions and characterizes objects by spectral, spatial, and contextual properties.
  - Lang et al., 2006; Youjing and Hengtong, 2007; Mathieu et al., 2007; Pham and He, 2008; Tan and Wang, 2009; Vannier and Hubert-Moy, 2010

Introduction

- Random Forest classification has shown high accuracy, robustness against over-fitting training data, and integrates measures of variable importance.
  - Diaz-Uriarte and Alvarez de Andres, 2006
- RF is bias prone when the number of instances is distributed unequally among classes of interest (He and Garcia, 2009)
- Objective: investigate the application and performance of Random Forest learning algorithms with object-oriented approach.

Study Site and Data

Study site and dataset

- Strasbourg (North-east France)
- Figure 1: multispectral bands (red, green, blue, near infrared).
    - Urban district (city center and suburbs) area of 142 km2.
- Features wooded elements in public and private zones (hedges, group or isolated trees of different sizes, forest) and a river, the Rhine.
- According to the available vegetation inventory databases, 19.8% is wooded vegetation.

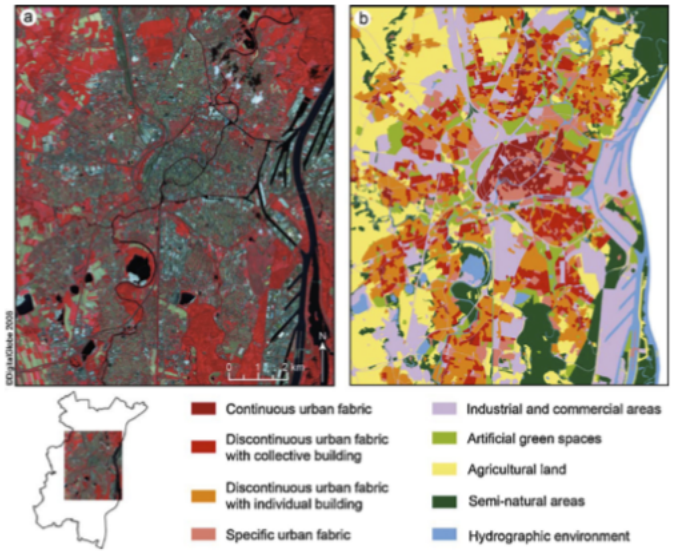


Fig. 1. (a) Study site and Quickbird image (DigitalGlobe 2008, P + MS, 60 cm) extension and (b) regional land cover/use database (BDOCCS©CIGAL, 2008).

Study Site and Data

Construction of the training and validation datasets

- The training set covers 5% of the image and is distributed over 15 stratified areas according to the urbanization index.
- Grid size is 700 m x 700 m.
- Urbanization index (I_u):
  - A_A: artificialized, A_NA: not artificialized, A_T: study area, 0.49 km2
  - I_u = -1 totally vegetated, I_u = 1 artificialized plots
- Index classified into 5 plots
- 15 validation plots were chosen, 3 for each classification.

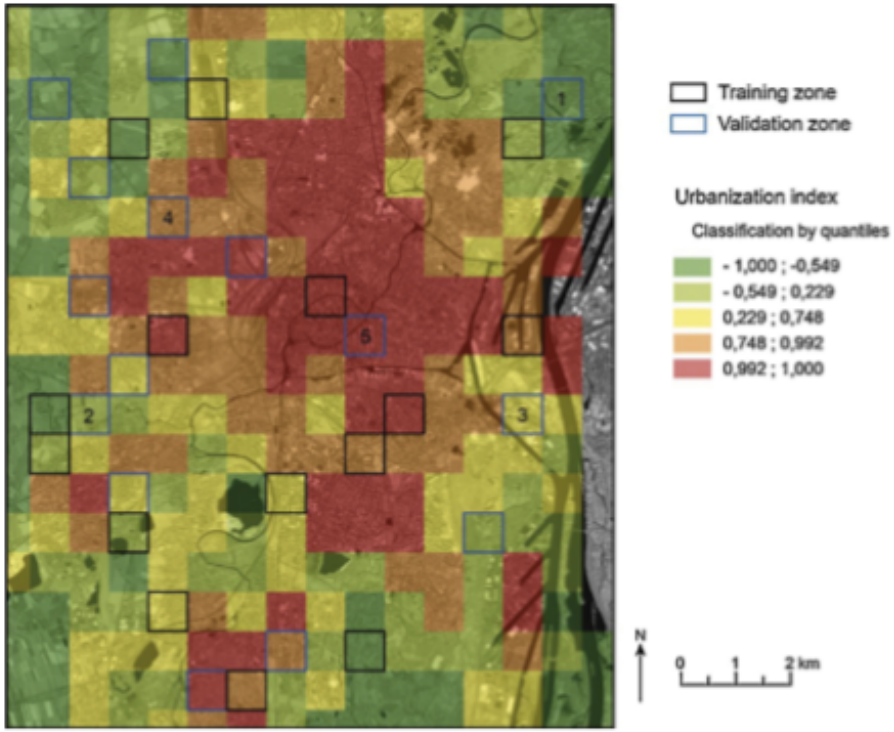
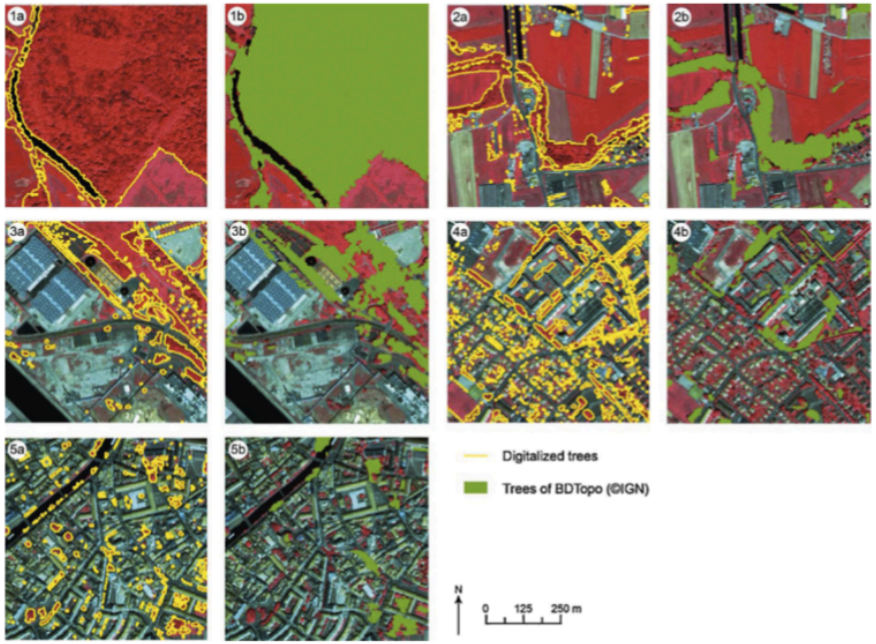


Fig. 2. Urbanization index in 5 classes with the localization of training and validation zones.

$$I_U = \frac{A_A - A_{NA}}{A_T} \quad (1)$$

Method

- Three Steps:
  1. Segmentation and computation of object attributes
     - Algorithms are applied using the 4 multispectral bands and MSAVI index. Objective to improve and optimize classification procedure.
       - Test A: correct bias.
       - Test B: select more suitable metrics
       - Test C: modifies number of variables chosen for nodes.
  2. Classification
  3. Validation



Method

Image Segmentation

- Image segmentation generates regions
- Delineation quality of the target images directly influences the accuracy of image classifier.
- Algorithms have been developed to delineate homogenous and meaningful segments.
- Multi-resolution image segmentation (MRIS): region growing algorithm that merges adjacent pixels or regions based on a criterion (object size, color, weight).
    - Over segmentation is preferable in the context of supervised classification since it enables small objects (Duro et al., 2012a; Smith, 2010; Stumpf and Kerle, 2011).
- Spectral Difference Algorithm (SDA): merges neighboring objects based on the mean layer intensity values and user-defined maximum thresholds.
- Each segment, 100 objects features for classification of wooden elements are computed including spectral, shape, texture, and attributes.

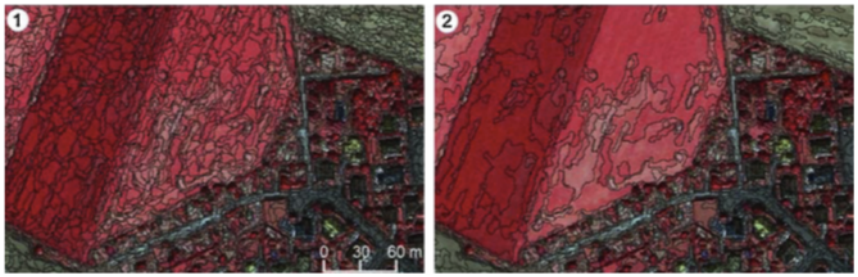
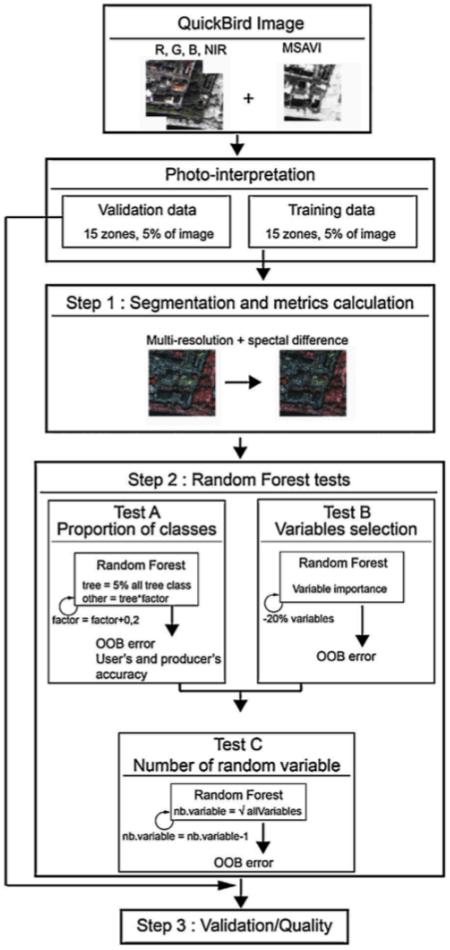


Fig. 5. Segmentation results with (1) MRIS and (2) MRIS + SDA.

Method

Image Segmentation



Overview of computed features for each region.

Type	Features	Band or index used
Spectral	Mean of spectral bands	R, G, B, NIR, MSAVI
	Brightness	/
	Maximum difference	/
	Standard deviation of pixel values	R, G, B, NIR, MSAVI
	Minimum pixel value	R, G, B, NIR, MSAVI
	Maximum pixel value	R, G, B, NIR, MSAVI
	Ratio (band/Brightness)	R, G, B, NIR
	Skewness of pixel value	R, G, B, NIR, MSAVI
	Mean difference to neighbors	R, G, B, NIR, MSAVI
	Mean difference to brighter neighbors	R, G, B, NIR, MSAVI
	Mean difference to darker neighbors	R, G, B, NIR, MSAVI
Geometry	Area, perimeter, length, width, length/width, asymmetry, border index, compactness, density, radius of largest enclosed ellipse, radius of smallest enclosed ellipse, rectangular fit, roundness, shape index	/
Textural	GLCM (all bands) (all directions, 0°, 45°, 90°, 135°)	Homogeneity, contrast, dissimilarity, entropy, angular 2nd moment, mean, standard deviation, correlation

Method

Image Classification: Random Forest (RF)

$$cT\sqrt{MN}\log N$$

- RF is a supervised machine learning algorithm for regression and classification.
  - Multiple decision tree classified based on classification and regression tree (CART)
  - Performs bootstrap sampling and enables the calculation of an error estimate on the instances remaining out of the bag (OOB) on each tree.
  - Does not consider all variables at each node to determine the best split threshold but a random subset of the original set of features.
  - Number of variable nodes is set to the square root of the total number of variables but can be adjusted by user.
- The number of trees must be sufficiently large to capture the full variability of the data and yield classification accuracy.
- Final class is assigned to an object based on the majority vote of all trees in the forest.
- The RF package in R was used.
- Calculation time for training depends on constant complexity, number of trees, number of variables, and number of instances.
- Each tree was built from a stratified bootstrap sample comprising only 5% of the training set.
  - Reduced computation time and memory requirements

Method

Random Forest Tuning

- Test A: Proportion of Classes
  - # Wooded elements < # Non-wooden elements
  - Iterative training and predicting on subsamples can avoid bias against the minority class.
  - Bootstrap samples corresponding to 5% of the wooded segments were multiplied by a factor. At first by 1, and the increased until reaching original class-ratio.
- Test B: Variable Selection
  - First step, large number of trees (n=5000) were generated to measure variable importance.
  - New RFs are constructed by disregarding 20% of the least important variables.
  - OOB error is calculated and the model with the lowest error is selected.
- Test C: Number of Random Variables
  - Square root of total number of variables by default.
  - Smaller number of random variables with a sufficiently large number of trees yields higher classification accuracy on a remote sensing dataset. (Gislason et al. (2006))
  - Final accuracy test is carried out comparing classification results at the 15 areas against mappings resulting from visual image interpretation.
    - Three accuracy measures are used, user accuracy (UA), producer accuracy (PA) and the F measure which allows to measure classification performance independence class imbalance.

$$F\text{-measure} = \frac{2 * UA * PA}{UA + PA}$$

Results and Discussion

Segmentation Results

Parameters used for segmentation framework and results.

		MRIS	SDA
Segmentation parameters	Scale parameter	15	20
	Color/shape	0.7/0.3	/
	Compactness/smoothness	0.3/0.7	/
	Weight of spectral bands (B, G, R, NIR, MSAVI)	1, 1, 1, 1, 4	
Number of segments (%)	Total	792,614	736,583
	Wooded elements	/	105,756 (13%)
	Non wooded elements	/	688,651 (87%)
Size of segments (in pixels)	Mean	26.44	28.45
	Minimum	1	1
	Maximum	2012	142,986

- MRIS segmentation yielded less than 800,000 segments.
- SDA reduced the segment number to 55,000 with increase mean size of segments and maximum size of segments.
- Classes were assigned after the segmentation in a GIS (Geographic Information System). Number of segments does not correspond to sum of segments from both classes
- Calculation of figures was intersected with the ground truth. Segmentation boundaries pixels followed the correctly generalized objects. Wooded elements represented 13%.

Results and Conclusion

Random Forest

Result of Proportion of Classes (Test A):

- Showed bias between wooded and non-wooded elements. Non-wooded elements overrepresented.
  - Wooded elements 105,756 segments, Non-wooded elements 688,651 segments.
- Proportion of wooded and non-wooded elements equalized by multiplying segment size by a factor until a balance is reached.

Result of Variable Selection (Test B):

- 7/8 MSAVI-based features/20 important variables
- NIR bands were the lowest ranked attributes
- 80 features is the best possible solution for high accuracy prediction (but high computational costs.
- 33 attributes were selected.
- Both a 33 and an 80-feature model were used.

Result of Number of Randomly Selected Variables (Test C):

- For the construction of each 1000 trees, 5% of the wooded class segments are used.
- 33 and 80 feature models had similar results
- Error reduction was 0.15
  - One variable per split yields a stronger decorrelation of individual trees in the forest and enhances the classification results.
- Only used training data

Results and Conclusion

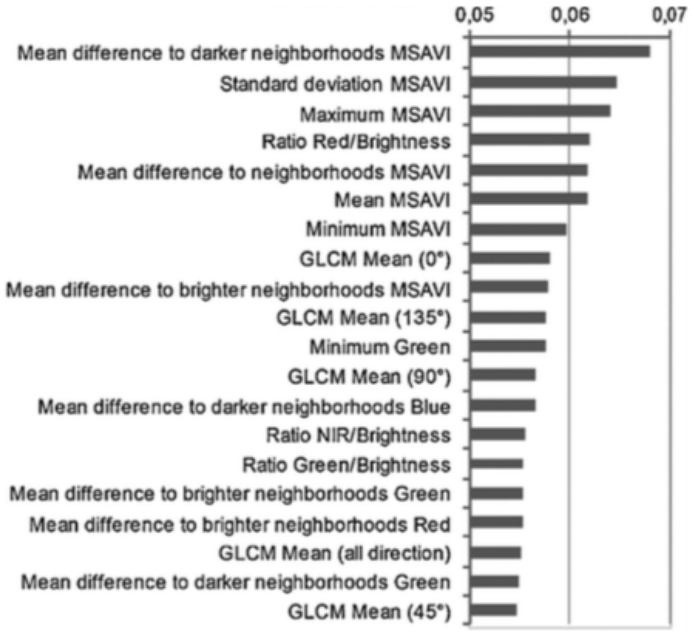


Fig. 7. Features importance.

Results and Conclusion

Validation Step

- Producer's accuracy for "wooded" class higher than user's accuracy by 5% for both classifications (33 and 80 feature)
  - F-measures were equivalent
  - Modeling 33 features presents better results and requires less computation time.
- Accuracy performance of RF depends on urban fabric.
  - Small trees or little wedges cause omission errors.
  - Collective buildings like schools, hospitals also cause problems.
  - Shrubby vegetation, small gardens.

Table 3
Results of classifications calculated with segments.

	User accuracy (%)	Producer accuracy (%)	*F*-measure (%)
80 variables classification	62.20	67.42	67.70
33 variables classification	62.53	67.32	64.84

Results and Conclusion

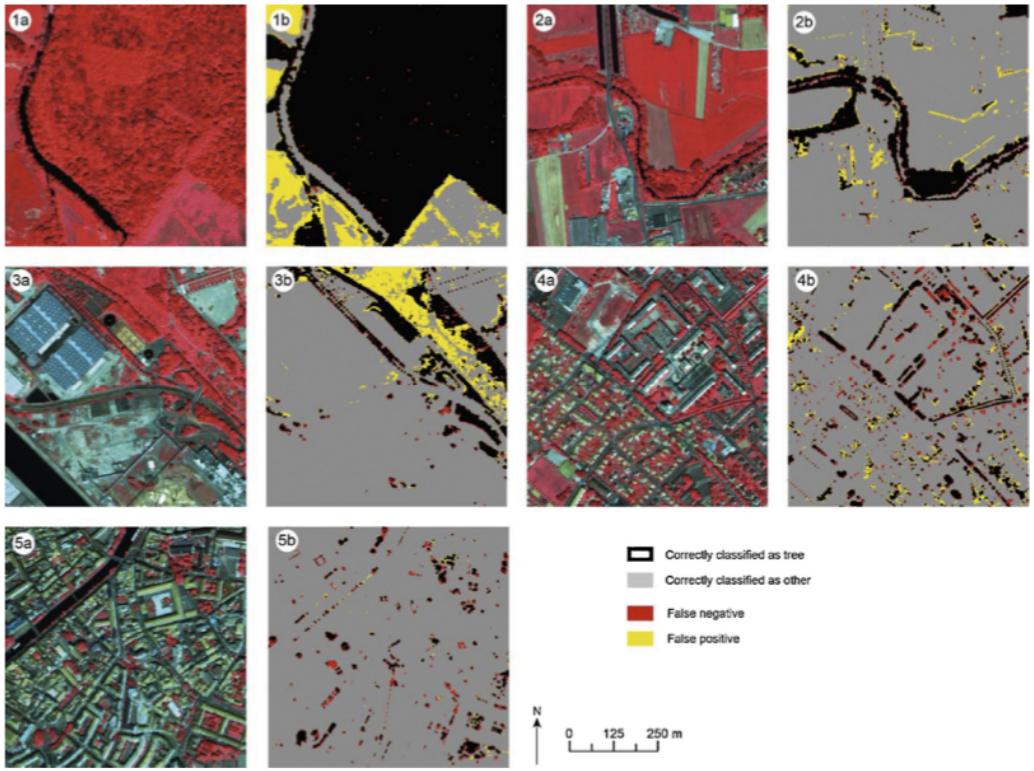


Fig. 10. (a) Quickbird image and (b) results with the SDA segmentation algorithm and 33 variables.

Conclusion

- Indexes from MSAVI were crucial for the classification of the images.
- Random Forest classification proved to be robust to extract wooded vegetation.
  - Accuracy did not considerably change when changing class balance or included features.
  - Model was not sensitive to over fitting and all features could have been used.
    - Features were discarded for computational time; these features could have revealed their importance for classification
  - Does not require tuning of many parameters so it can identify urban vegetation efficiently.
- Future Tests:
  - Active Learning: classifiers recommends iteratively the most valuable samples or for user to label areas.
  - Transfer Learning: reuse the clasifier and ground truth for classification of other images by adjusting the classifier to the changes.

STEVENS

INSTITUTE *of* TECHNOLOGY

THE INNOVATION UNIVERSITY®

stevens.edu