Comparing CART and Random Forest for the monitoring of wetland vegetation with multispectral data.

Julie Campagna, phD student, Angers France

Aurélie Davranche, Associate professor Angers France

IBS-DR Biometry Workshop
Würzburg University 09 october 2015

SUMMARY

- **Decisions trees : Generalities**
- **CART and Random Forest presentation**
- **CART functionment**
- **Random Forest functionment**
- **Exemple of application : Remote sensing**

DECISION TREES

- **Method of classification (or regression)**
- **Non parametric method**
- **Can deal with a lot of data**
- **Separate each sample to obtain the most homogeneous classes as possible**
- **Separability criterions existing :**
  - Gini Index : CART
  - Chi square automatic interaction detection : CHAID
  - Shannon Entropy :C5.0

COMPARAISON CART ET RANDOM FOREST

Two decision tree methods developped essentially by Breiman et al.

- **Cart was the first in 1984**
- **Random Forest 2001**
- **Different applications: biology, medecine, remote sensing,…**
- **Deal with a lot of data sample and variables**
- **Not perturbated with extrems data or variables not required**

CART : FUNDAMENTALS

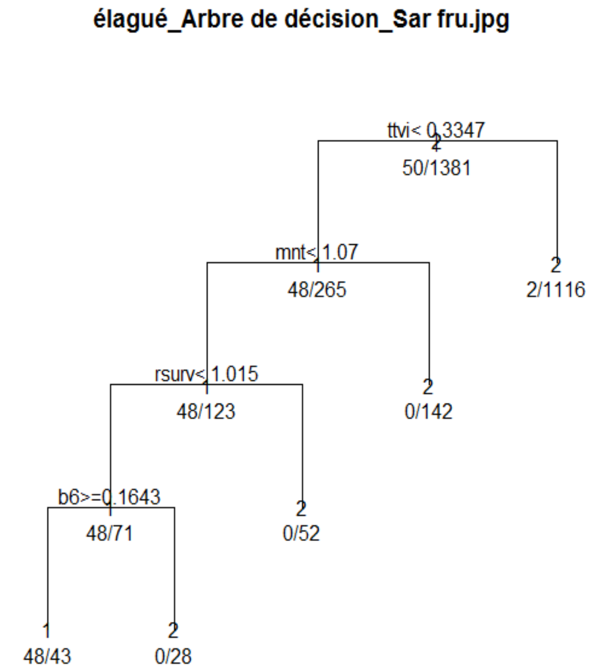
- **Cart use Ginny criterion to separate a training sample**

$$\boxed{I = 1 - \sum_{i}^{n} f_i^{\,2}}$$

n : Number of class to predict

Fi : Class frequencie in the node

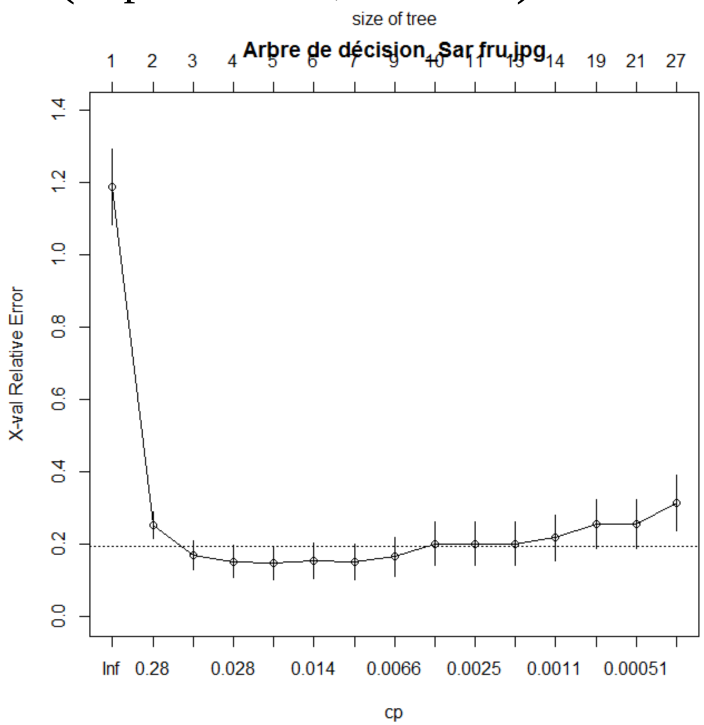
- **Dichotomous partionning**
- **Decision rule appears**

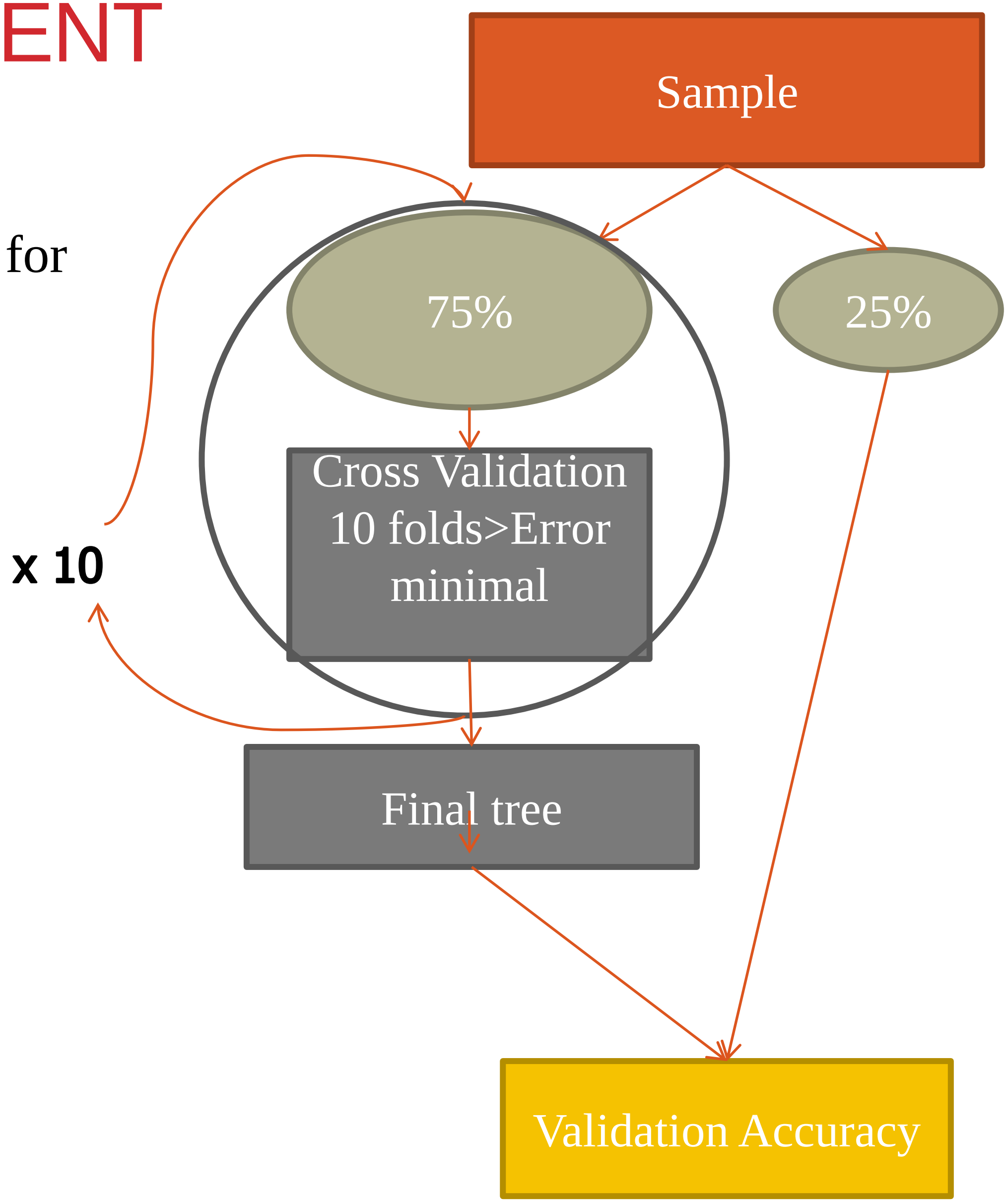


CART : IMPROVEMENT

Choose the result tree

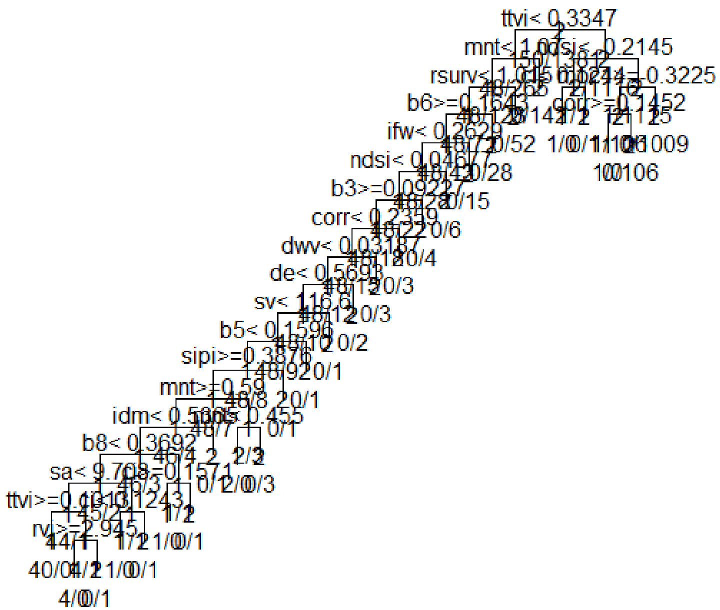
- 75% for training sample and 25% for validation
- 10 cross validation
- (Esposito et al, CV-1SE)



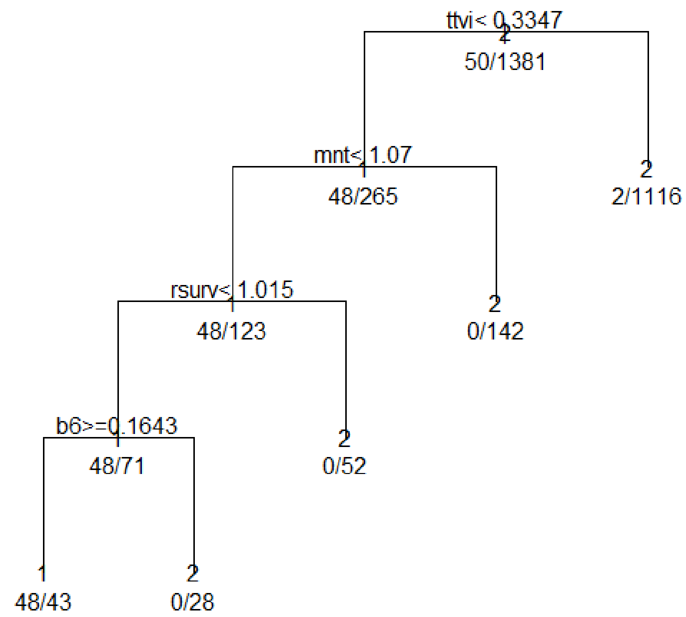


CART : PRUNNING RESULT

Arbre de décision_Sar fru.jpg



élagué_Arbre de décision_Sar fru.jpg



CART : PARAMETERS

- **Cart was implement in R using the package Rpart**

- **Presence = « 1 » ; absence = « 2 »**
- **Unbalanced sample**

  Optimal « Prior » parameter : iterative runs of the algorithm

RANDOM FOREST : GENERAL OPERATION

- **RF grows many classification trees**

- **To classify, each variable goes down each of the trees in the forest.**

- **Each tree gives a classification: we say the tree “votes” for that class.**

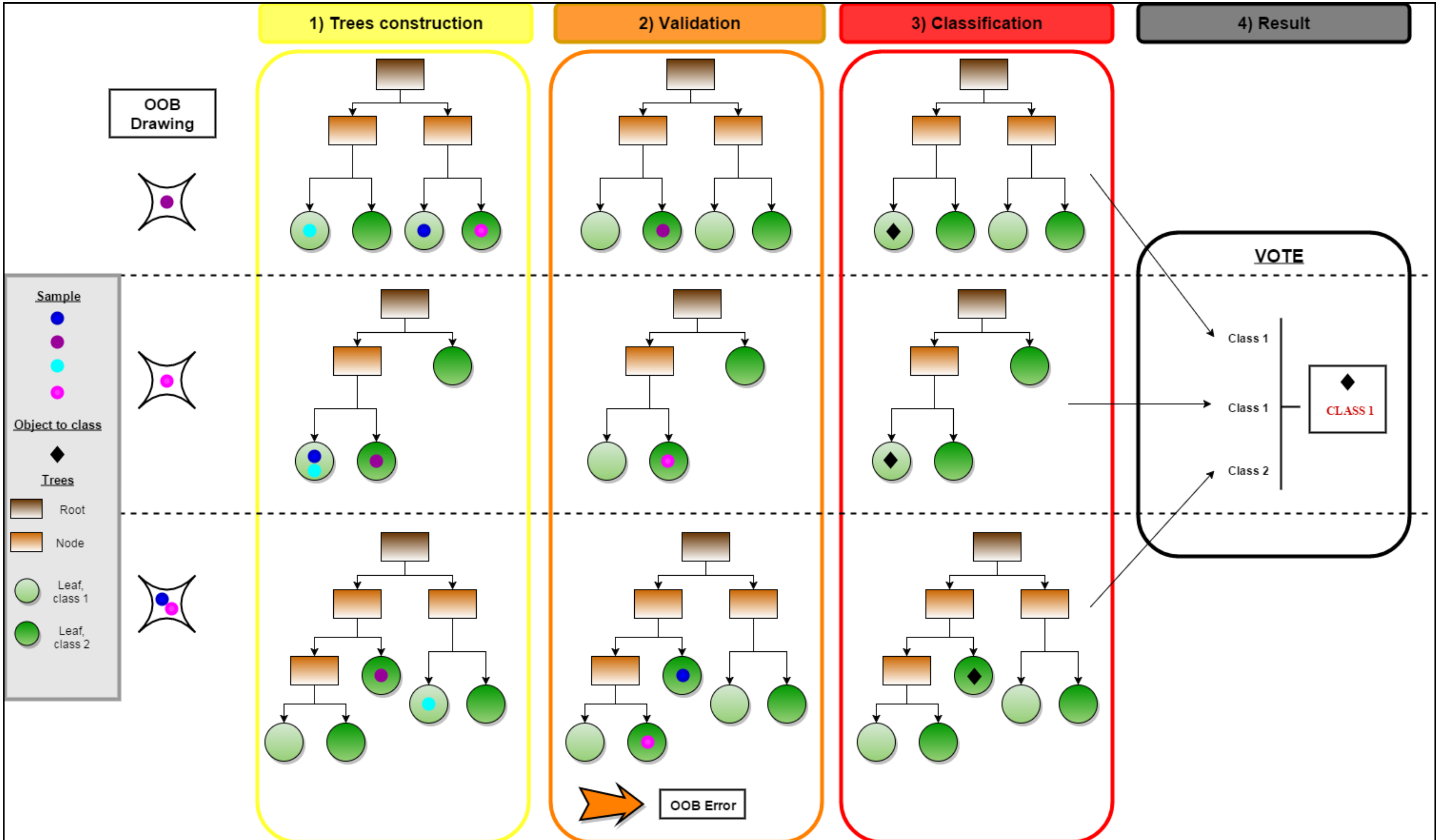
- **The forest choses the classification having the most vote (over all the trees in the forest).**

RANDOM FOREST : STEP ONE

- **For each tree it selects randomly 2/3 of the sample for training set and 1/3 for validation (Out Of Bag, OOB)**

- **Variables are chosen randomly (generally sqrt(variables)) at each node with replacement**

RANDOM FOREST : STEP TWO FOREST CONSTRUCTION



RANDOM FOREST : PARAMETERS

- **Can not deal with unbalanced samples**
- **Two ways to ajust datas :**
  - **Up-sampling** based on the size of the largest class
  - **Down-sampling** based on the size of the smallest class

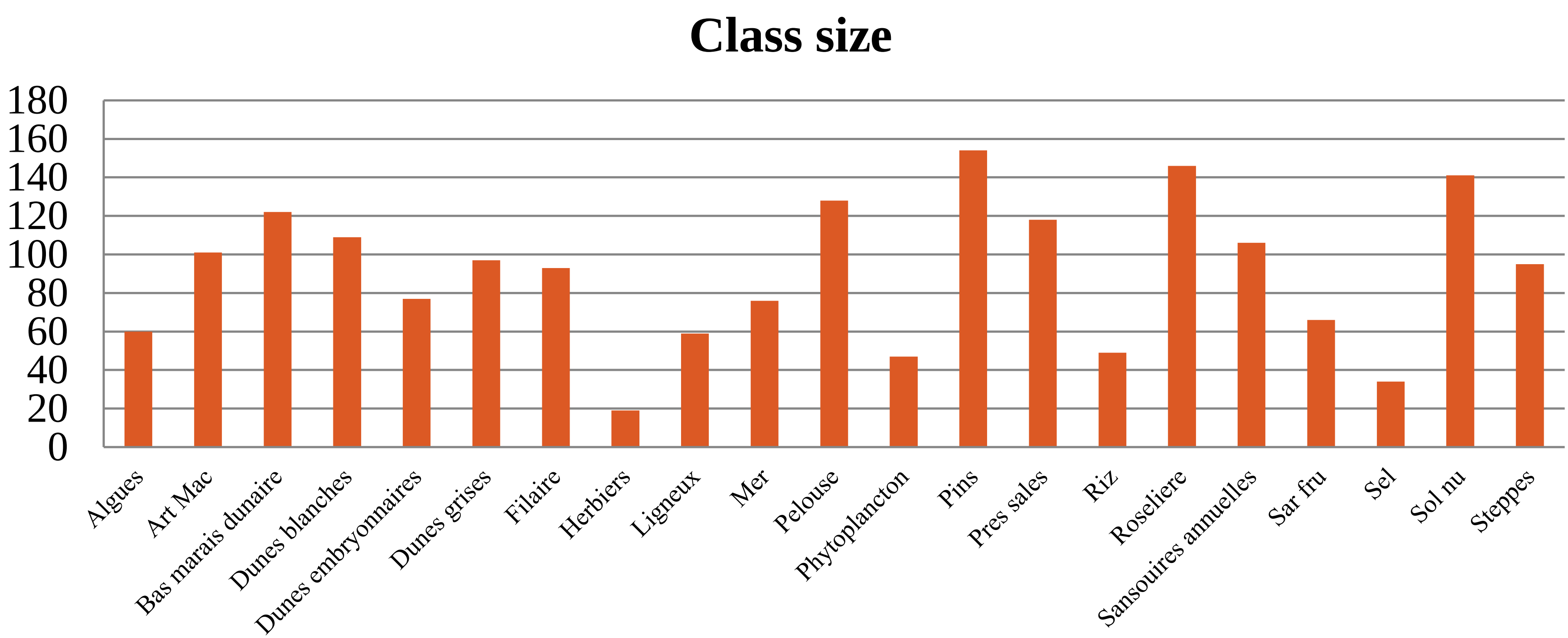
EXEMPLE OF APPLICATION : REMOTE SENSING

- **Satellite images usefull for monitoring of wetland environments**
- **In this case we used a high spatial resolution image (World View 2) on Camargue in South of France.**
- **Needs :**
  - Mapping the vegetation
  - Create a method easy to apply without knowledge in remote sensing and R programmation

[image removed]

SAMPLE

- **21 landcover classes from field data**
- **49 descriptive variables : reflectance values from bands spectral data and multispectral indices**



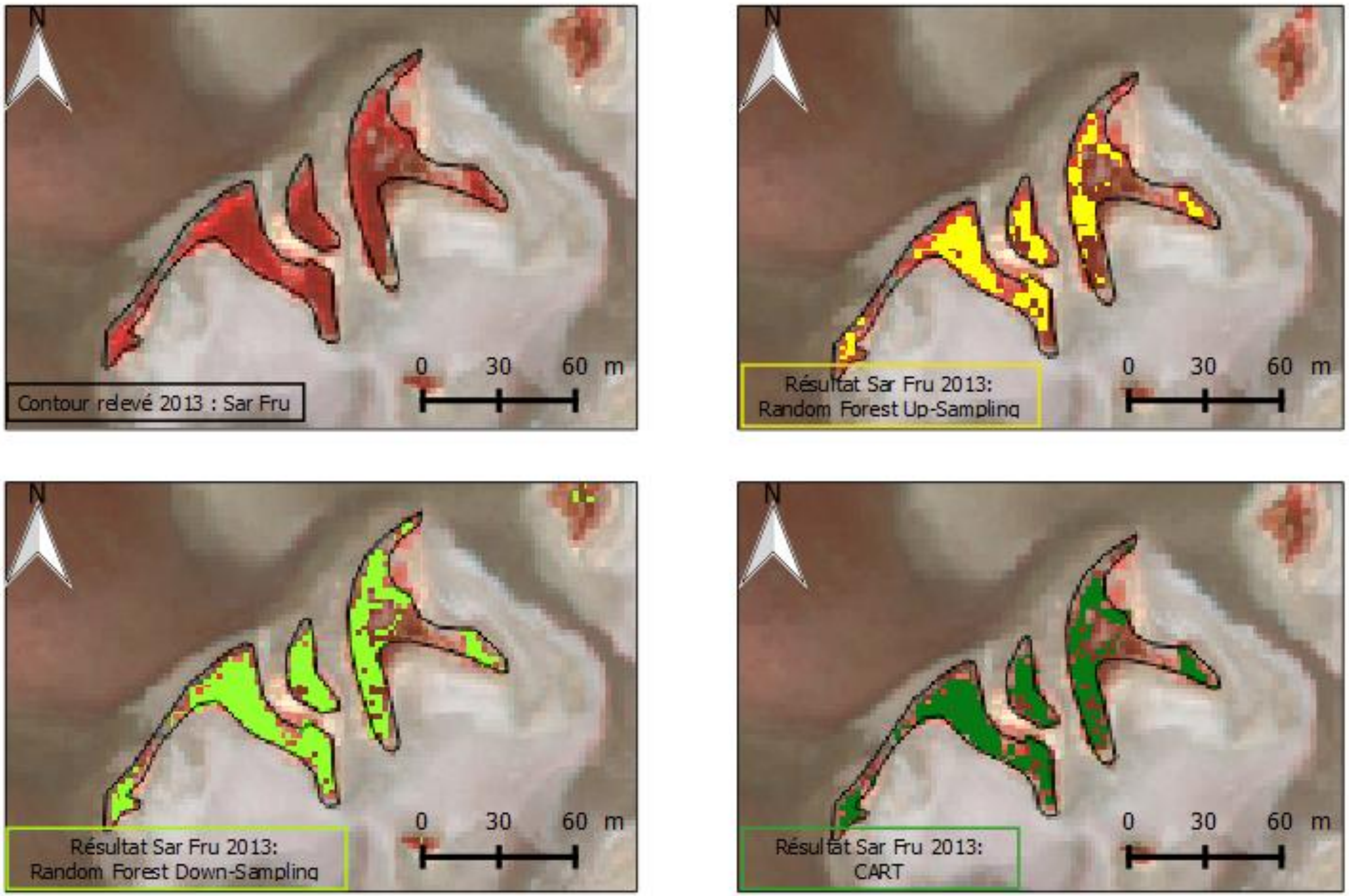
EXEMPLE OF APPLICATION : REMOTE SENSING

Classification of *Salicornia Fruticosa*

[image removed]

EXEMPLE OF APPLICATION : REMOTE SENSING

Cartography Results : *Sarcocornia Fruticosa*



EXEMPLE OF APPLICATION : REMOTE SENSING

Confusion matricies :

Cart		Carte de référence				
				Classe 1	Classe 2	Précision Globale
	Carte produite, classification	Entraînement	Classe 1	49	65	
			Classe2	1	1316	
				50	1381	**0,953878407**
		Erreur d'omission		0,02	0,047067	
		Validation	Classe 1	16	22	
			Classe2	0	428	
				16	450	**0,9527897**
		Erreur d'omission		0	0,048889	
		Total	Classe 1	65	87	
			Classe2	1	1744	
				66	1831	**0,953610965**
		Erreur d'omission		**0,015**	**0,047**	

RF		Carte de référence					
				Classe 1	Classe 2	Précision Globale	Erreur OOB
	Carte produite, classification	RF_Up	Classe 1	858	9		
			Classe2	0	1822		
				858	1831	**0,991**	0,26%
				0	0,00494		
		RF_Down	Classe 1	59	50		
			Classe2	7	1781		
				66	1831	**0,97**	3%
		Erreur d'omission		**0,11**	**0,027**		

➢ **Close classification accuracy values**

EXEMPLE OF APPLICATION : REMOTE SENSING

- The difference between global accuracy is really low between CART and Random forest (around 1,5%) and both results are good.
- CART provides an explicit model, the one of Random Forest is implicit
- An explicit model can be used again on a new dataset or an other image of the same date without repeat all the steps of modeling : more easy to use without specific knowledge

CONCLUSION AND DISCUSSION

- **On a same dataset and with all parameters suitable to CART we obtain results not significantly different from Random Forest**
- **This two models need some parameters to be capable to deal with unbalanced samples**
- **CART can generate an explicit model as Random Forest can’t**
- **This two algorithms also permit to identify important variables**

THANKS FOR YOUR ATTENTION !