Impact de l'autocorrélation spatiale sur la qualité des modèles d'apprentissage automatique

étude dans le cadre de la classification d'essences forestières à partir de données satellitaires

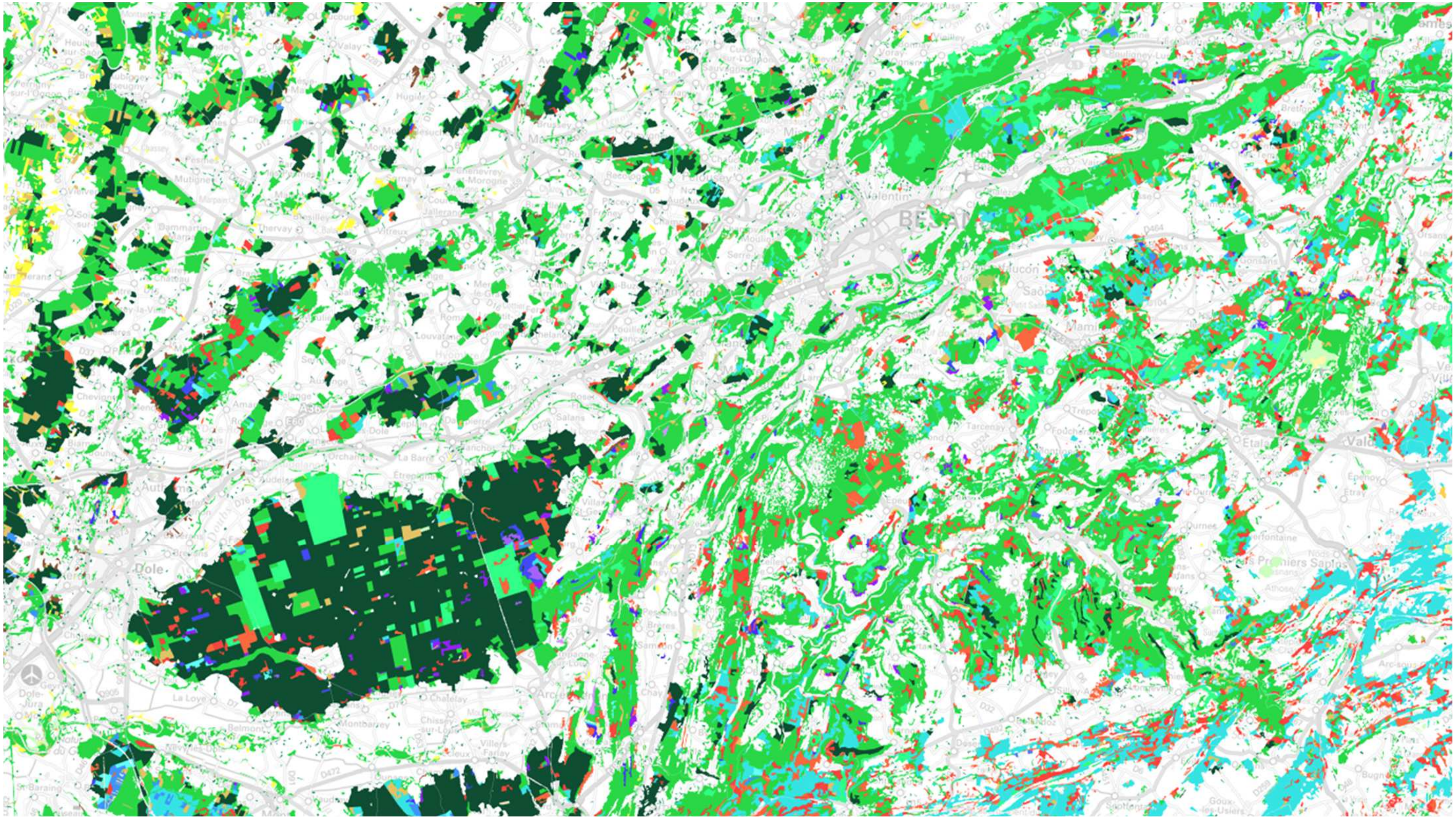
N. Karasiak[1], D. Sheeren[1], J-F Dejoux[2], C. Monteil[1].

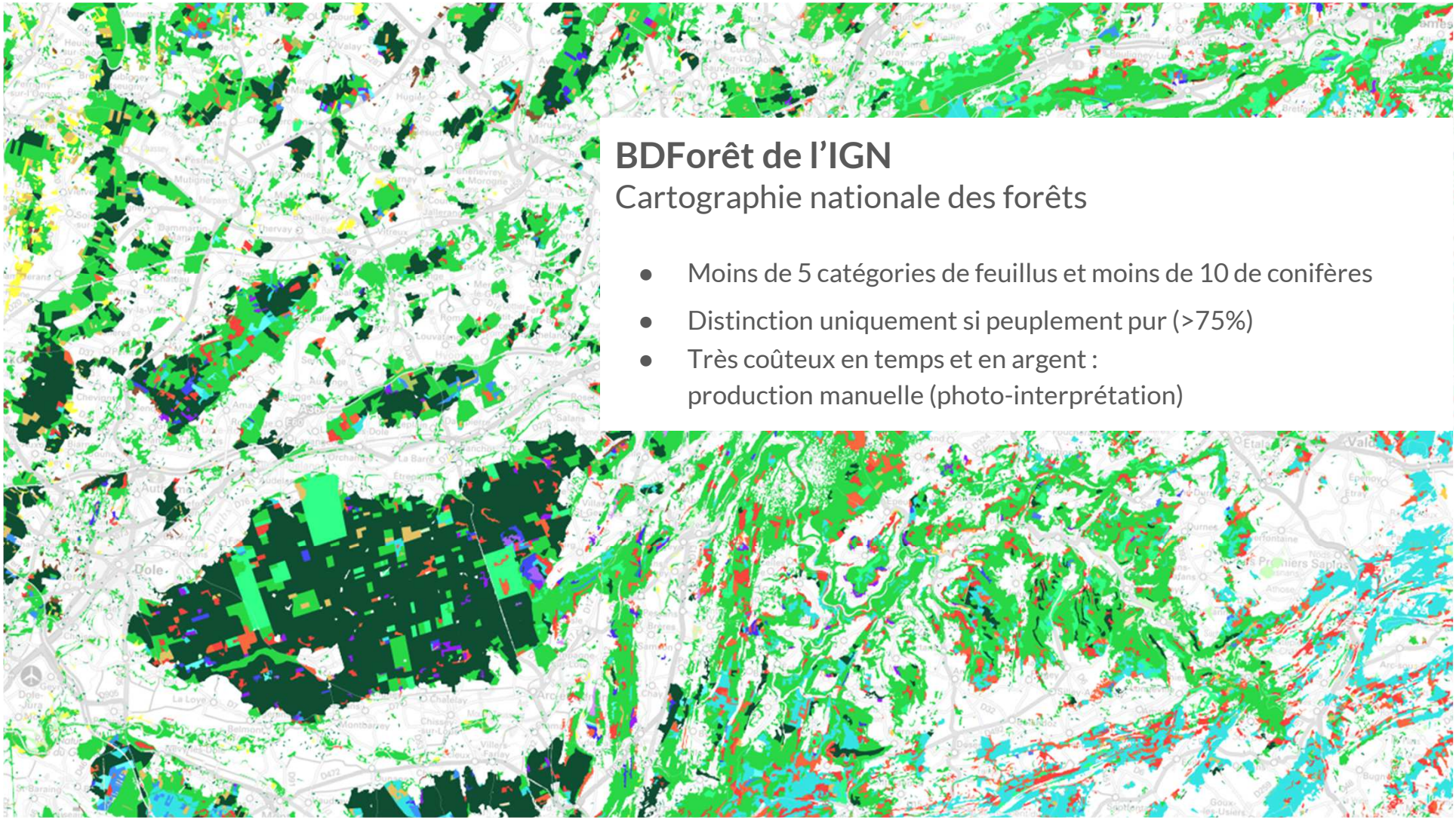
[1]DYNAFOR, Université de Toulouse, INRA, Castanet-Tolosan, France
[2]CESBIO, Université de Toulouse, CNES, CNRS, IRD, INRA, Toulouse, France.

INRA SCIENCE & IMPACT

DYNAFOR UMR 1201 | INRA - INP TOULOUSE

INP TOULOUSE ENSAT

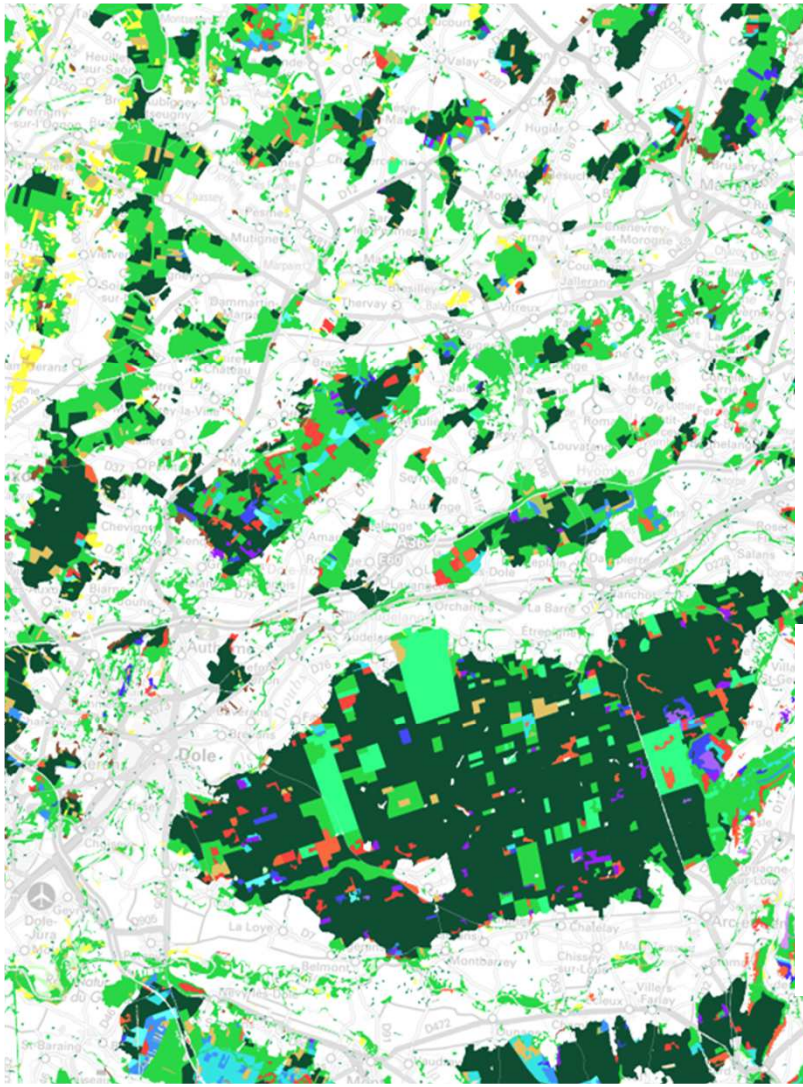

Dole
La Loye
Montbarrey
Salans
La Barre
Tarcenay
Etalans
Epenoy
Etray
Bugny
Belmont



BDForêt de l'IGN

Cartographie nationale des forêts

- Moins de 5 catégories de feuillus et moins de 10 de conifères
- Distinction uniquement si peuplement pur (>75%)
- Très coûteux en temps et en argent :
  production manuelle (photo-interprétation)



BDForêt de l'IGN

Cartographie nationale des forêts

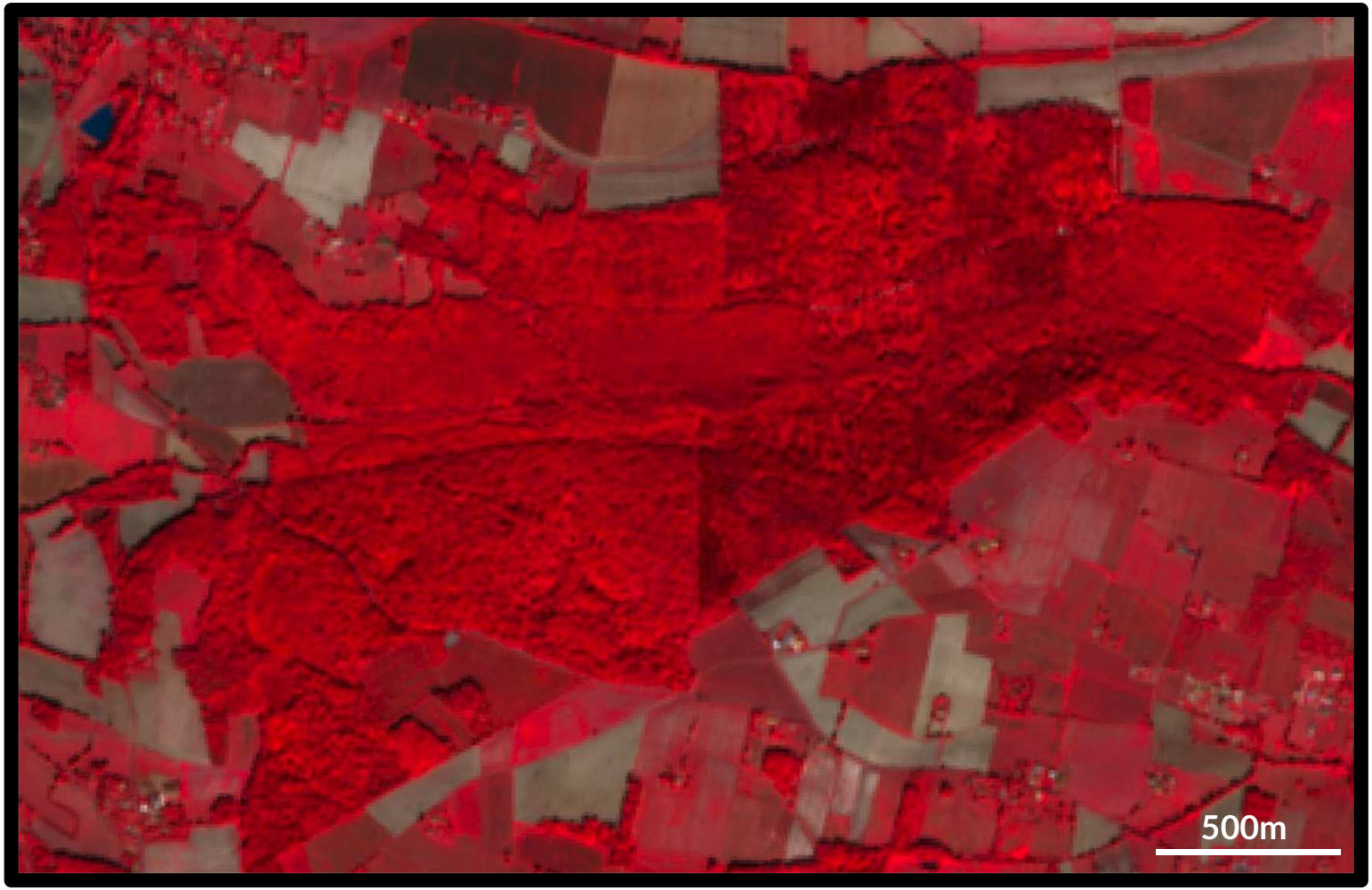
- Moins de 5 catégories de feuillus et moins de 10 de conifères
- Distinction uniquement si peuplement pur (>75%)
- Très coûteux en temps et en argent :
  production manuelle (photo-interprétation)

Besoin de cette cartographie des essences forestières pour

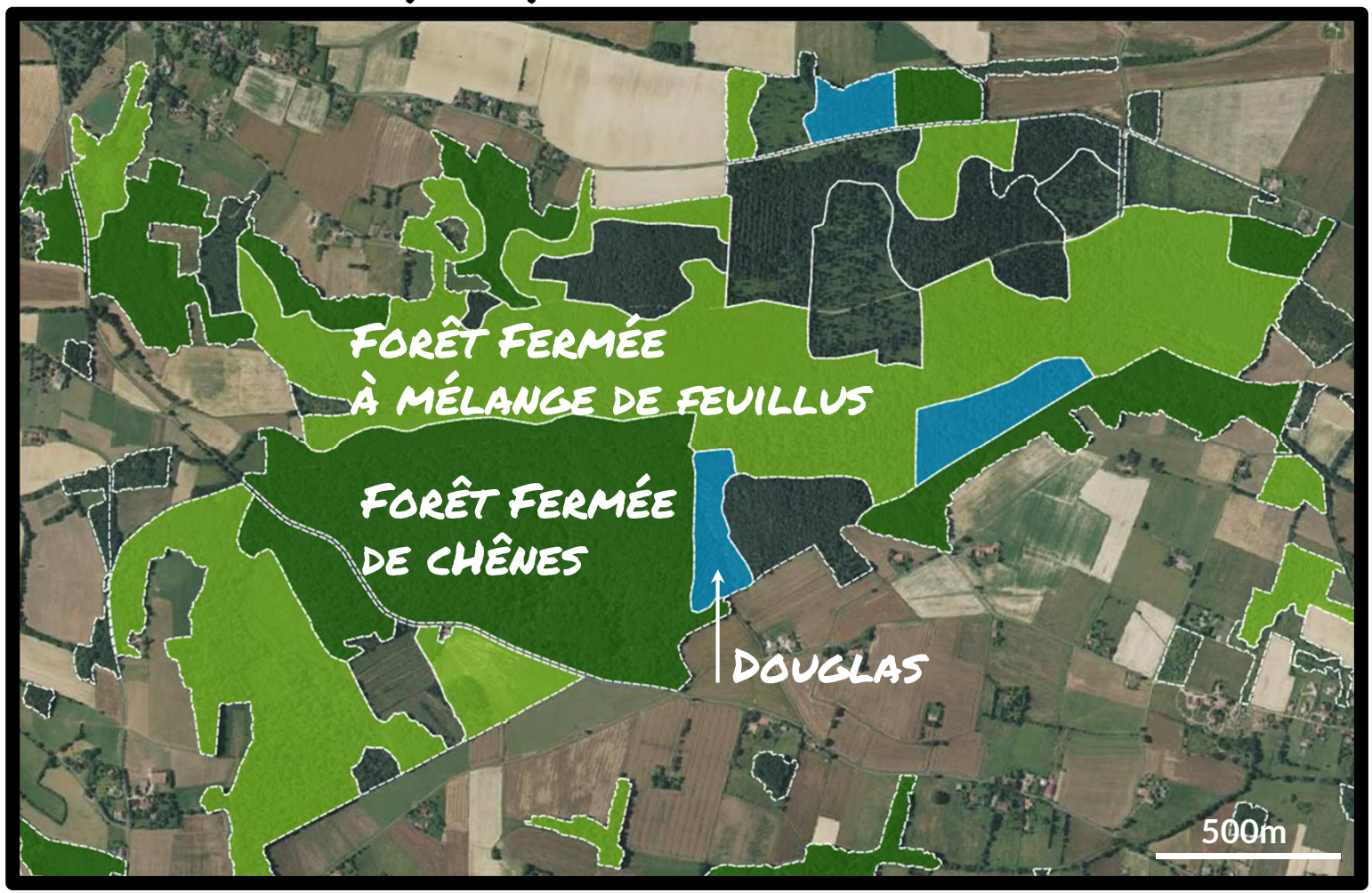
- Gérer et suivre la ressource, détecter les situations à risque
  (dépérissement, maladies...)
- Évaluer l'état de la biodiversité (caractériser les habitats, mesurer les services écosystémiques, évaluer la résilience...)

Forêt de rieumes, IGN, 2013

Forêt de rieumes, IGN, 2013



Forêt de rieumes, IGN, 2013

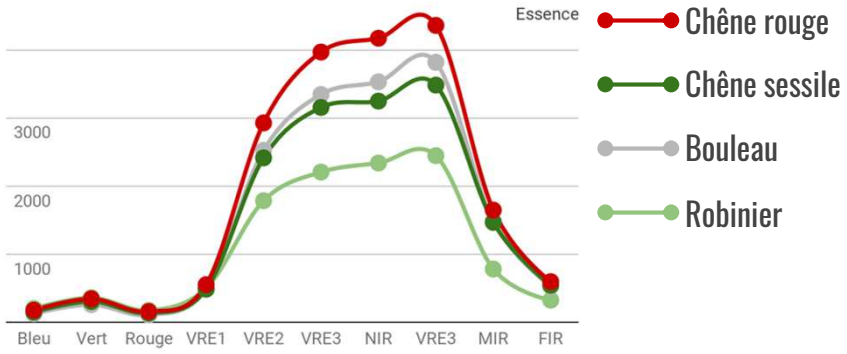


La télédétection

1km

25 juillet 2018

11 novembre 2018

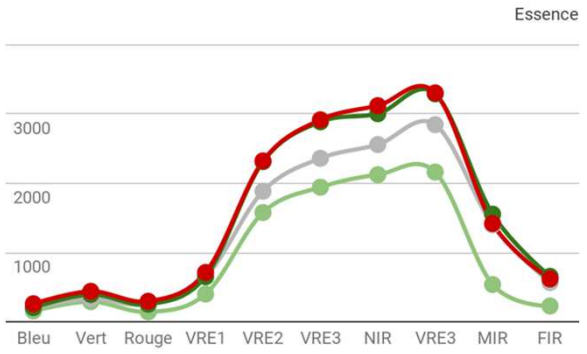


Chêne rouge

Chêne sessile

Bouleau

Robinier

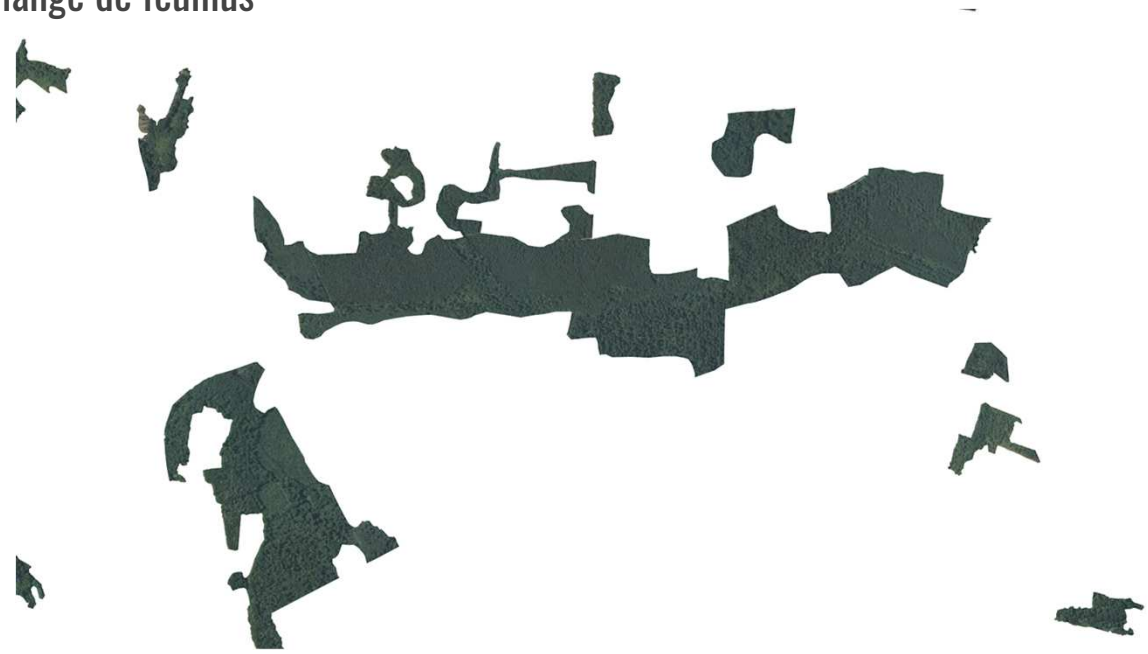


Théo Quant
Nouvelles approches en Géographie Théorique & Quantitative

La télédétection

La télédétection

Forêt fermée à mélange de feuillus



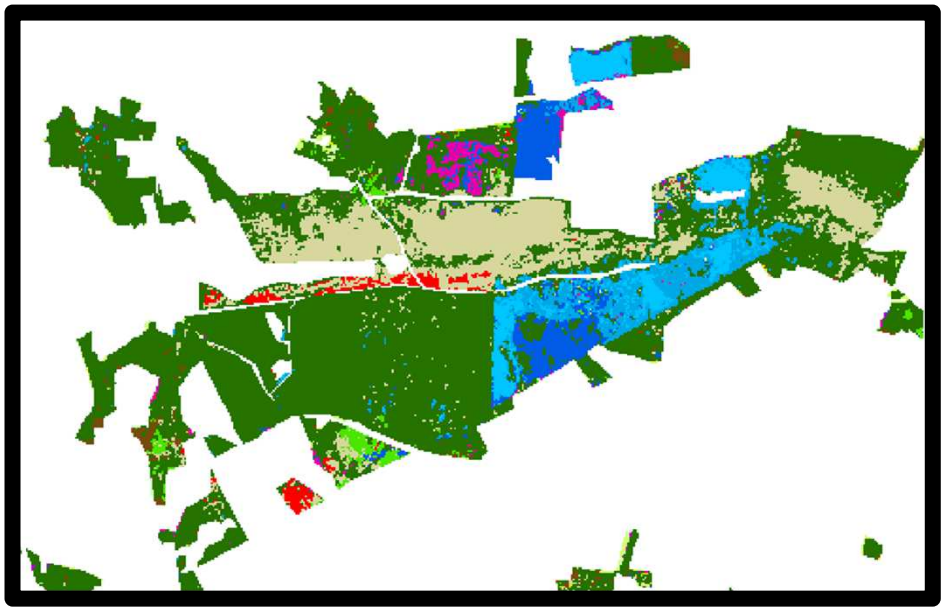
La télédétection

Forêt fermée à mélange de feuillus

Forêt de rieumes, IGN, 2013
BD Forêt V2

À chaque forêt de plus de 0.5ha
un type de forêt

Cartographie
par apprentissage automatique



À chaque pixel une essence

Problématique

Discriminer les essences forestières par télédétection : une question délicate

Réponse spectrale des espèces influencée par de nombreux facteurs

- Propriétés biochimiques des feuilles, structure interne
- État sanitaire
- Structure du peuplement, âge, densité, pratiques sylvicoles
- Conditions stationnelles et environnementales
- Phénologie
- Conditions d'éclairement et date de prise de vue
- Ombre portée

Problème ?

La variabilité intra-spécifique est souvent équivalente voire supérieure à la variabilité inter-spécifique (Boureau 2008)

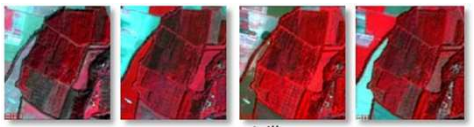
Différentes approches en télédétection

SPOT-5 (10m)

(Foody et Hill 1996, Kayitakire et al. 2002, Carleer et Wolff 2004, Immitzer et al. 2012)

résolution spatiale

résolution temporelle



Février Mai Juillet Août

(Key et al. 2001, Hill et al. 2010, Cano et al. 2014)

résolution spectrale

VIS PIR SWIR

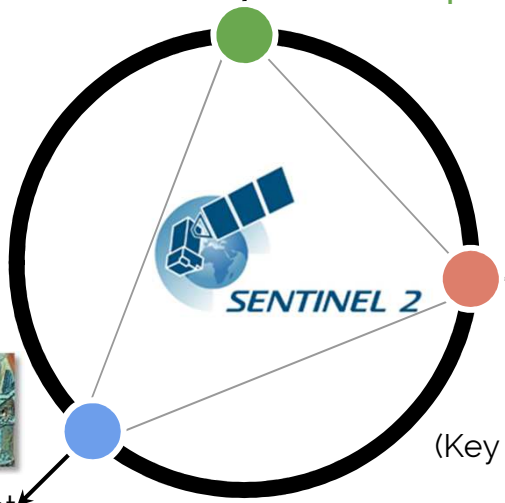
(Boschetti et al. 2007, Ghiyamat et Shafri 2010, Dalponte et al. 2012)

Différentes approches en télédétection

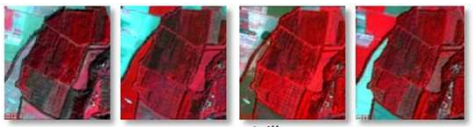
SPOT-5 (10m)

(Foody et Hill 1996, Kayitakire et al. 2002, Carleer et Wolff 2004, Immitzer et al. 2012)

résolution spatiale



résolution temporelle



Février Mai Juillet Août

(Key et al. 2001, Hill et al. 2010, Cano et al. 2014)

résolution spectrale

VIS PIR SWIR

(Boschetti et al. 2007, Ghiyamat et Shafri 2010, Dalponte et al. 2012)

L'autocorrélation spatiale ?

Corrélation d'une variable avec elle-même dans l'espace.

Concrètement ? Plus on s'éloigne d'un endroit, moins on a de chance de tomber sur une endroit qui lui ressemble.

Plus concrètement ? Si on est adossé à un chêne, on a de fortes chances d'être à côté d'un autre chêne.

Quel lien entre autocorrélation spatiale et apprentissage automatique ?

L'autocorrélation spatiale, un problème pas si récent...

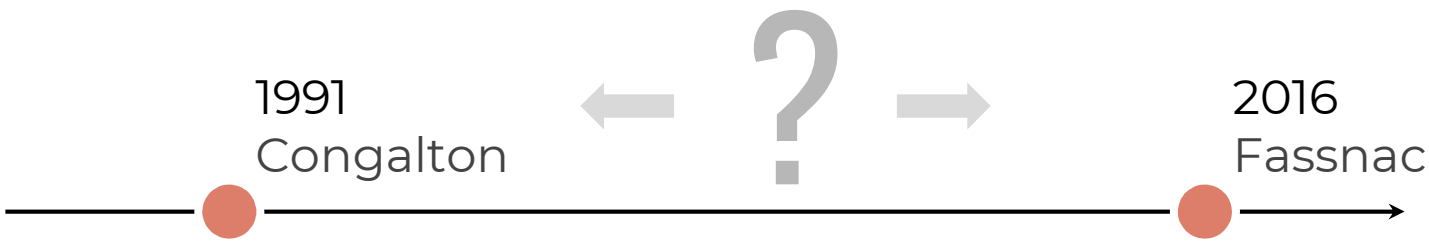
Quelles conséquences ?

L'effectif donné à l'algorithme est plus faible que l'effectif réel :

- les échantillons se ressemblent trop
- surestimation de la qualité car validation avec des pixels trop semblables

Avec le progrès des algorithmes... la qualité grimpe (parfois 100%).

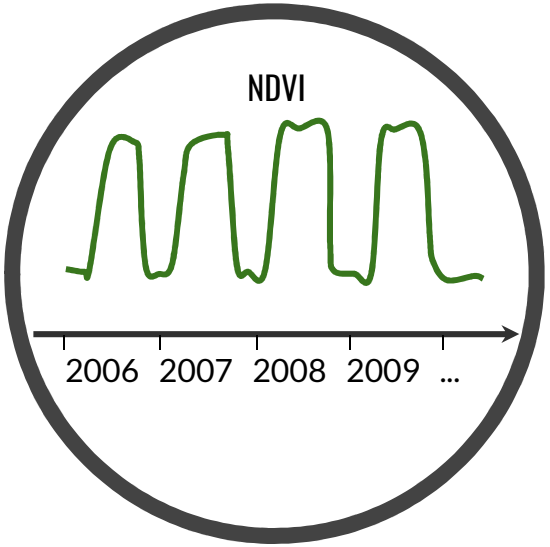
Deux reviews alertent :

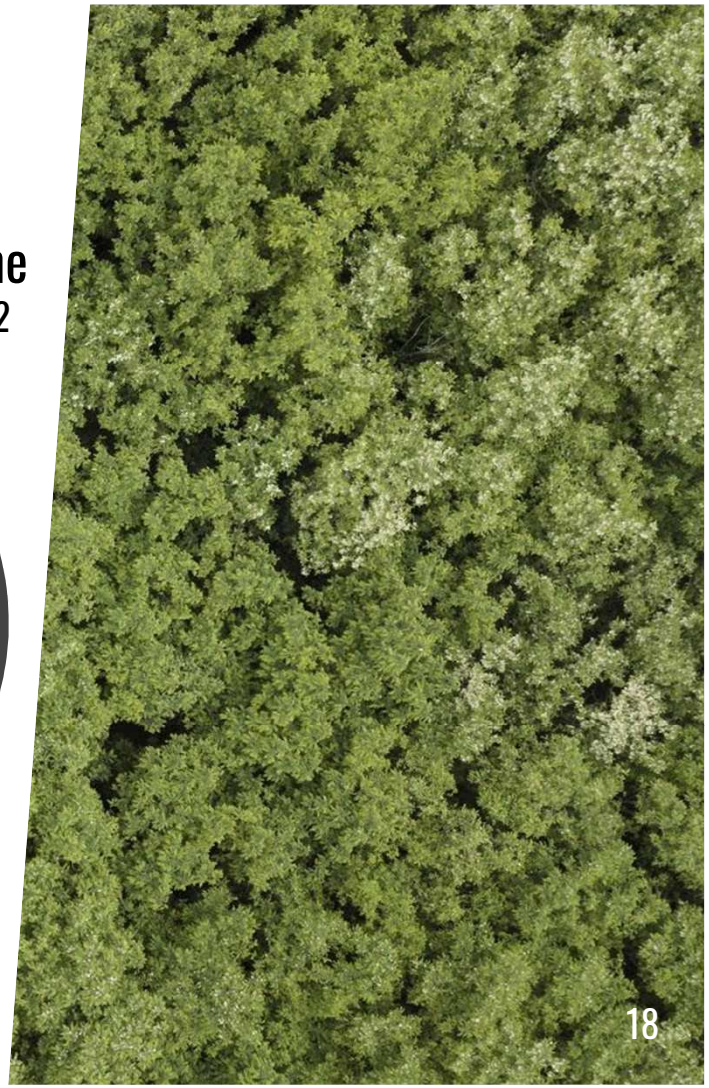


1.

L’identification du problème

9 années avec le satellite Formosat-2





Présentation

Apprentissage automatique

Cartographie avec l'algorithme SVM (Support vector Machine) des essences pour chaque année.
Validation des modèles en comparant deux types de validations :
- Monte Carlo
- Spatial Leave-One-Out

13 essences forestières
8 feuillus / 5 résineux

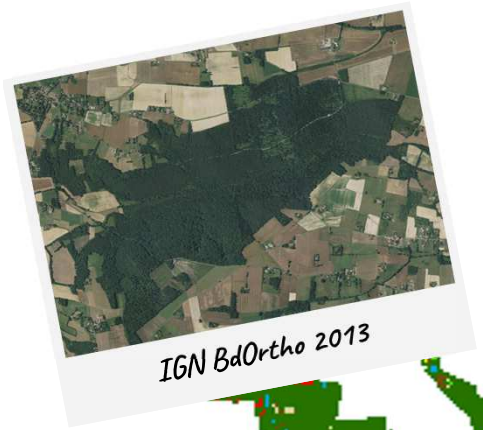
1262 points terrain localisant au pixel chaque espèce.

9 années de séries temporelles

De 2006 à 2014, mais échantillonnage temporel irrégulier (à cause des nuages essentiellement).

Hypothèse :

La dimension temporelle augmente la séparabilité des essences.



Validation : MC-CV 50%

kappa

94%

BOULEAU

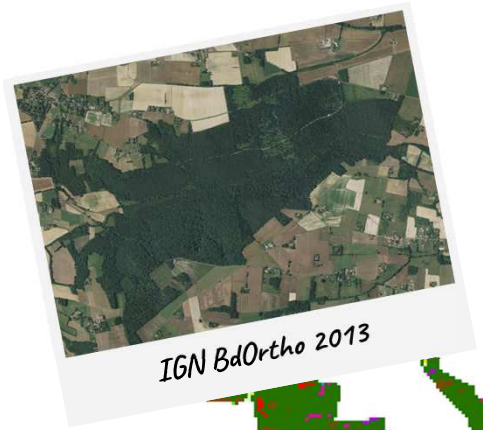
CHÊNE

SAPIN BLANC

PIN NOIR

- Chêne
- Chêne rouge
- Bouleau
- Frêne
- Robinier
- Peuplier
- Saule
- Eucalyptus
- Pin laricio
- Pin maritime
- Pin noir
- Sapin douglas
- Sapin blanc

500m



2006

Chêne
Chêne rouge
Bouleau
Frêne
Robinier
Peuplier
Saule
Eucalyptus

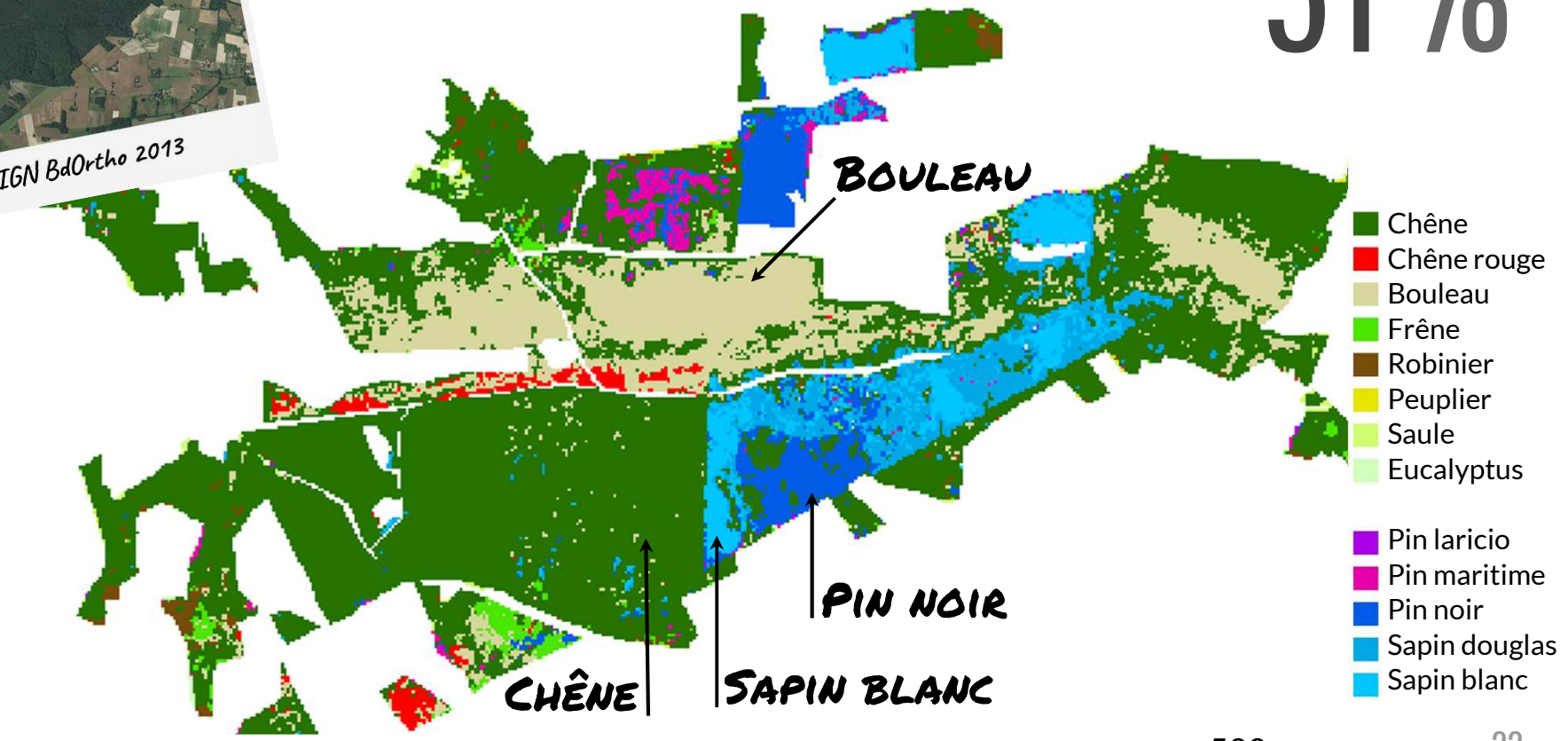
Pin laricio
Pin maritime
Pin noir
Sapin douglas
Sapin blanc

500m

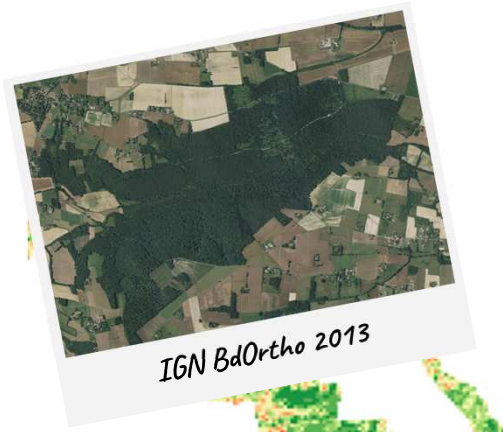
Validation : Spatial Leave-One-Out

kappa

51%



Validation : Spatial Leave-One-Out



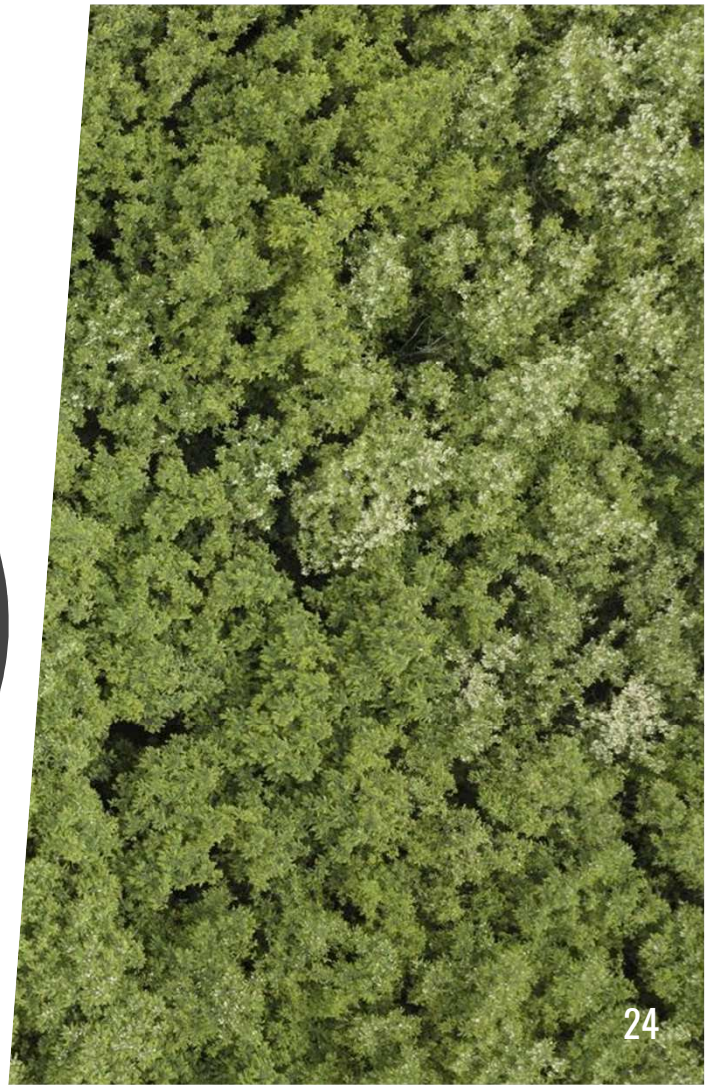
Nombre d'accords (10ans)

10
9
8
7
6
5
4
3
2
1

500m

2.

Étude de l'impact de l'autocorrélation spatiale



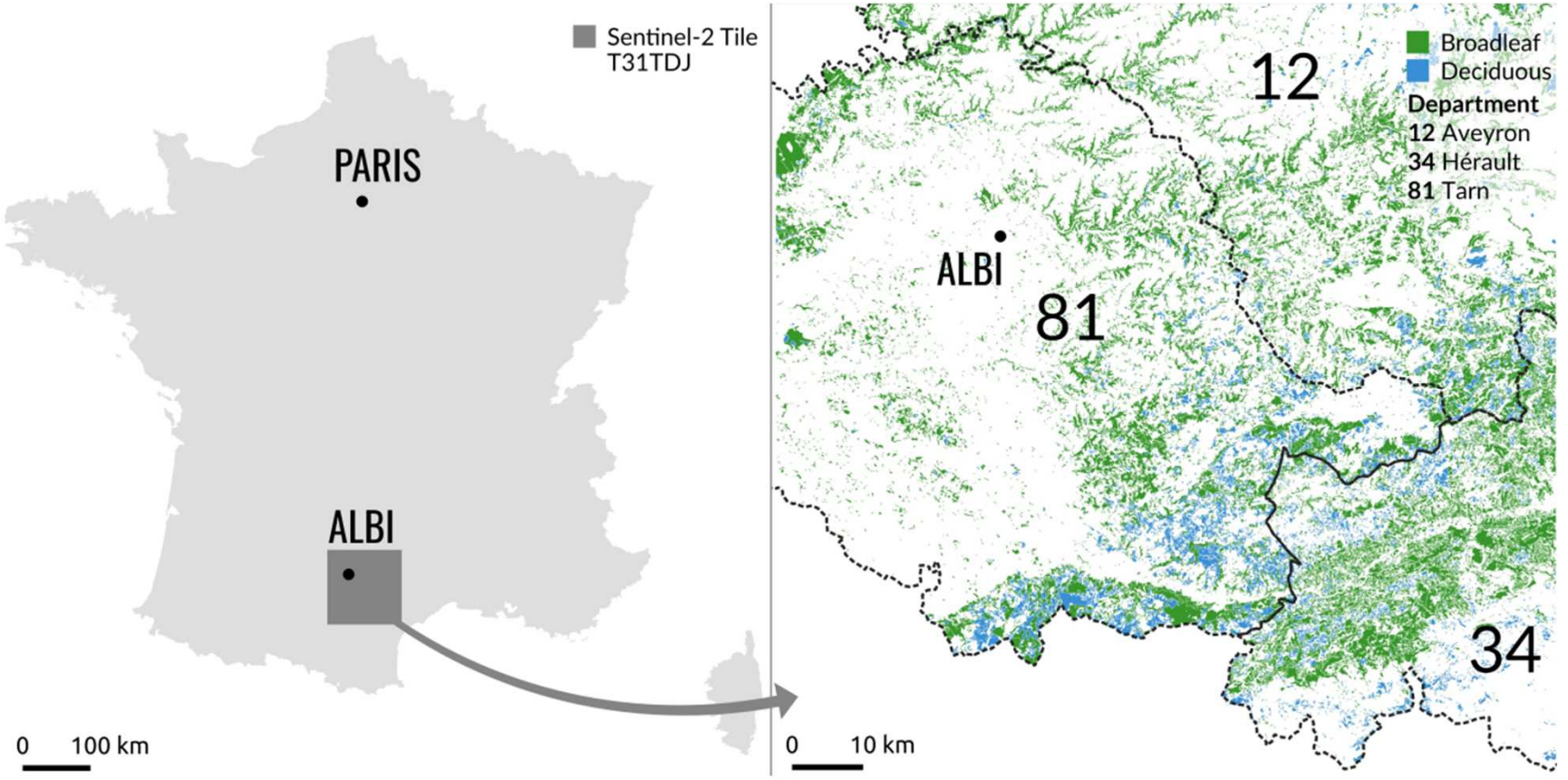
Introduction

- Comparer et estimer les différentes méthodes de validation/entraînement
- Valider le modèle avec des données indépendantes spatialement
- Proposer une prise en compte systématique et donc facilement accessible

Hypothèse :

L'augmentation du nombre de peuplements (et donc d'échantillons) diminue la surestimation de l'indice de qualité issu des méthodes d'échantillonnages.

Site d'étude

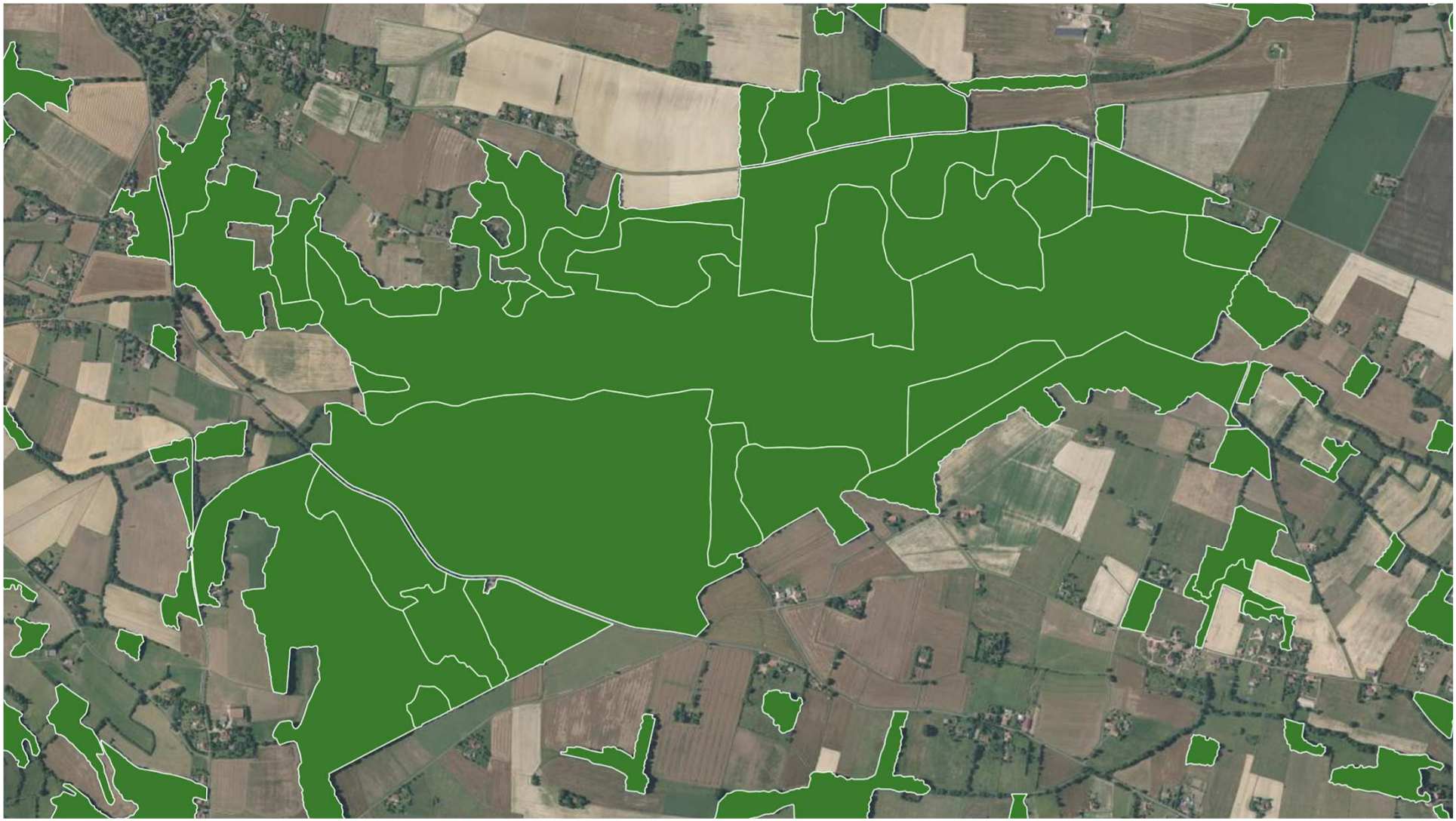

Sentinel-2 Tile
T31TDJ
PARIS
ALBI
0 100 km
12
Broadleaf
Deciduous
Department
12 Aveyron
34 Hérault
81 Tarn
ALBI
81
34
0 10 km

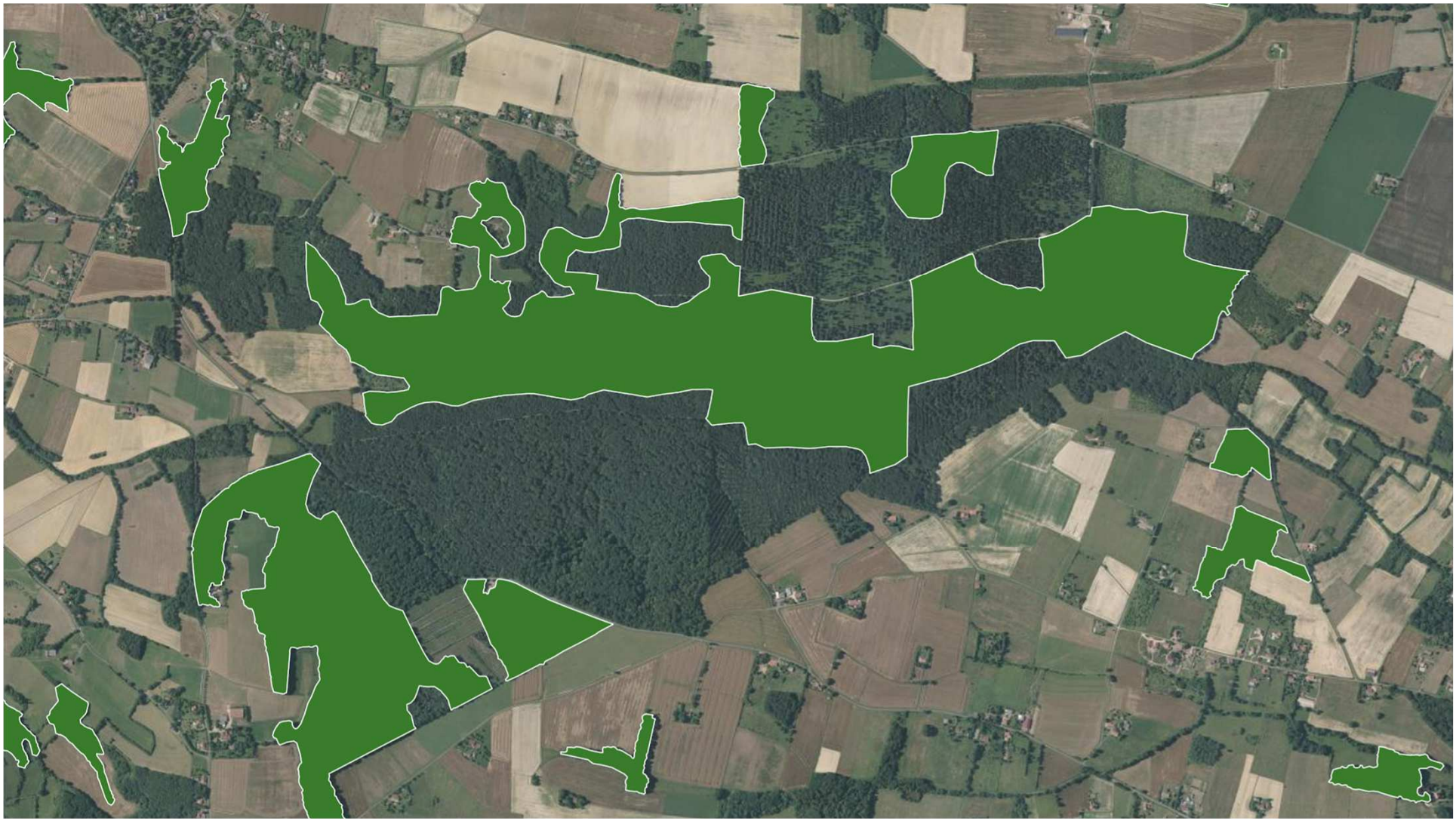
Données images et références

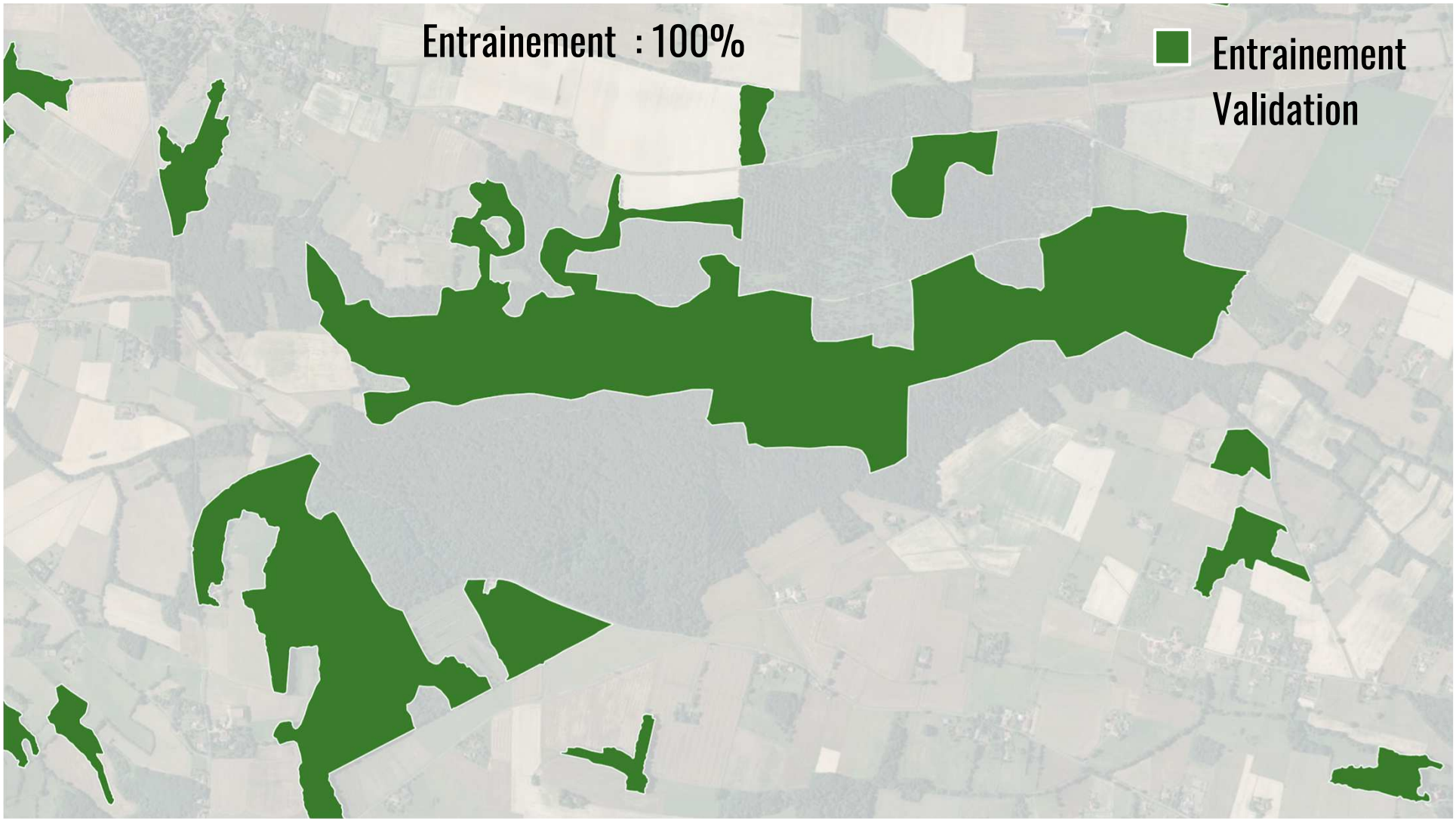
- Sentinel-2 2016 : 4 dates <5% nuages (17 juillet, 26 août, 5 septembre et 15 octobre)
- Interpolation linéaire si nuage
- Si NDVI < 0.4 en Juillet : Suppression du pixel dans le jeu de données

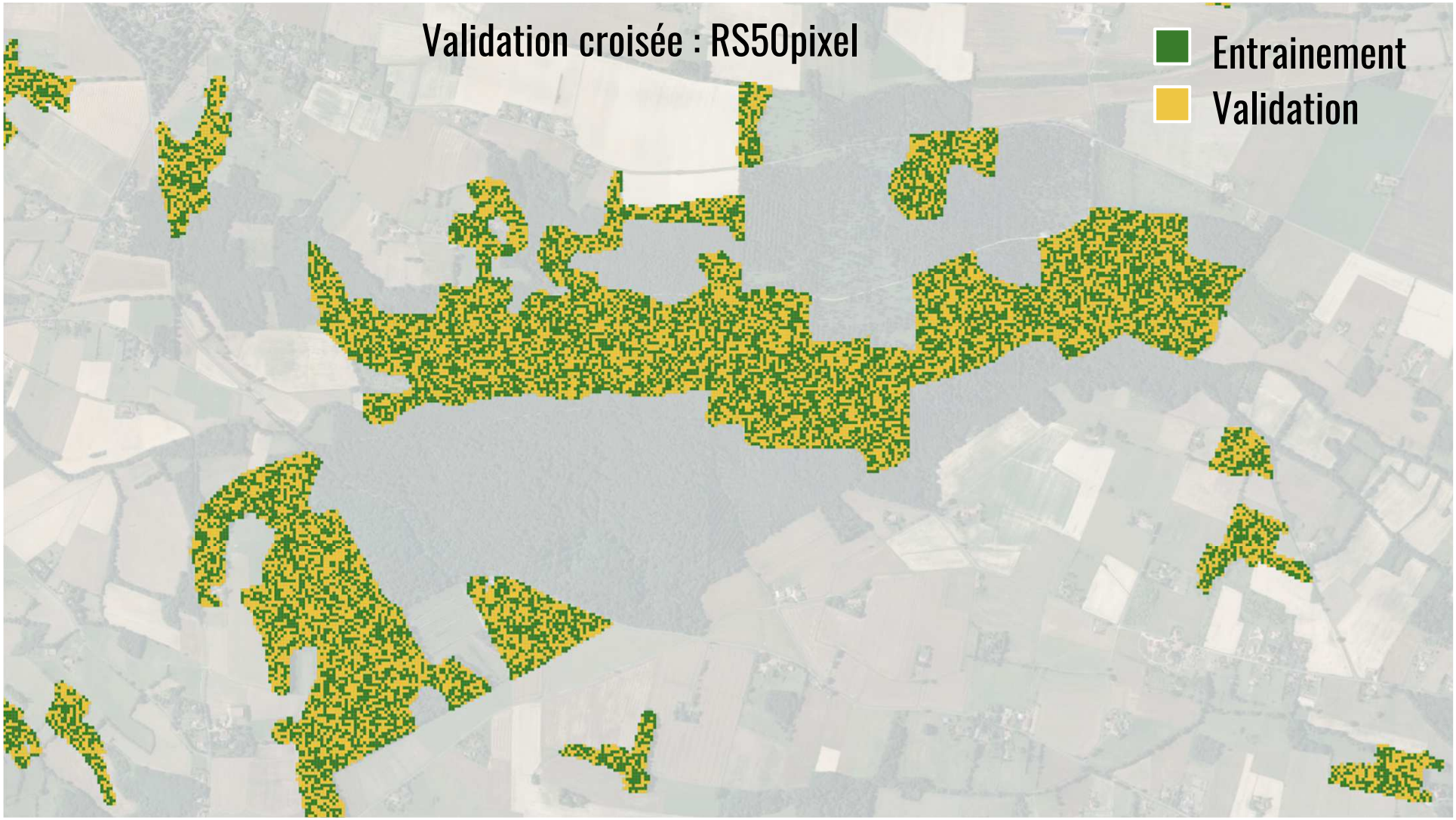
Deux classes : **feuillus et résineux**

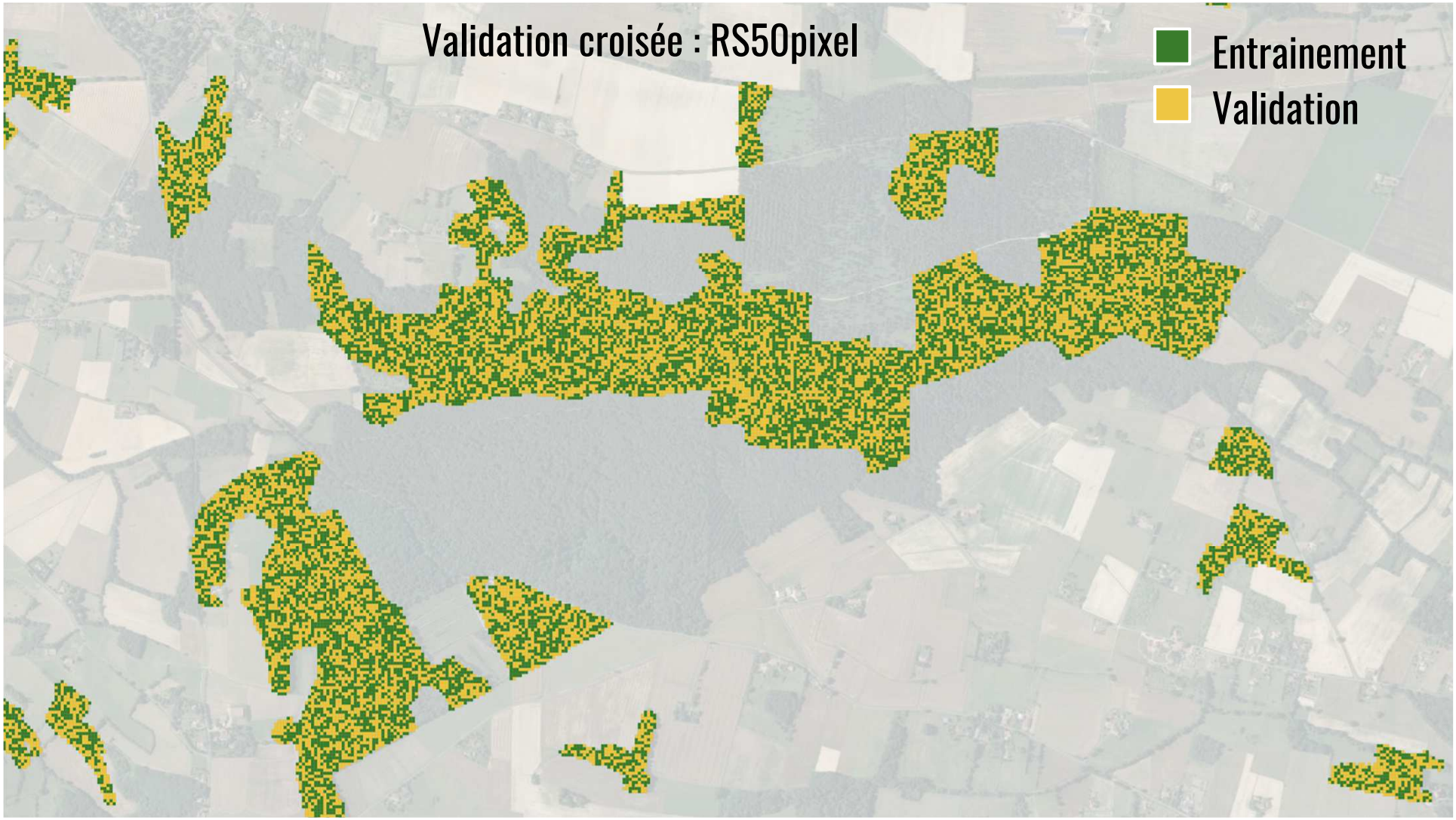
			Broadleaf	Conifers
2015	**Training** (Department 34)	pixels	4,257,112	1,119,079
		stands	2,841	1,972
2006	**Validation** (Department 81)	pixels	7,046,056	2,216,920
		stands	4,210	2,002
2008	**Validation** (Department 12)	pixels	5,485,172	836,075
		stands	3,092	966

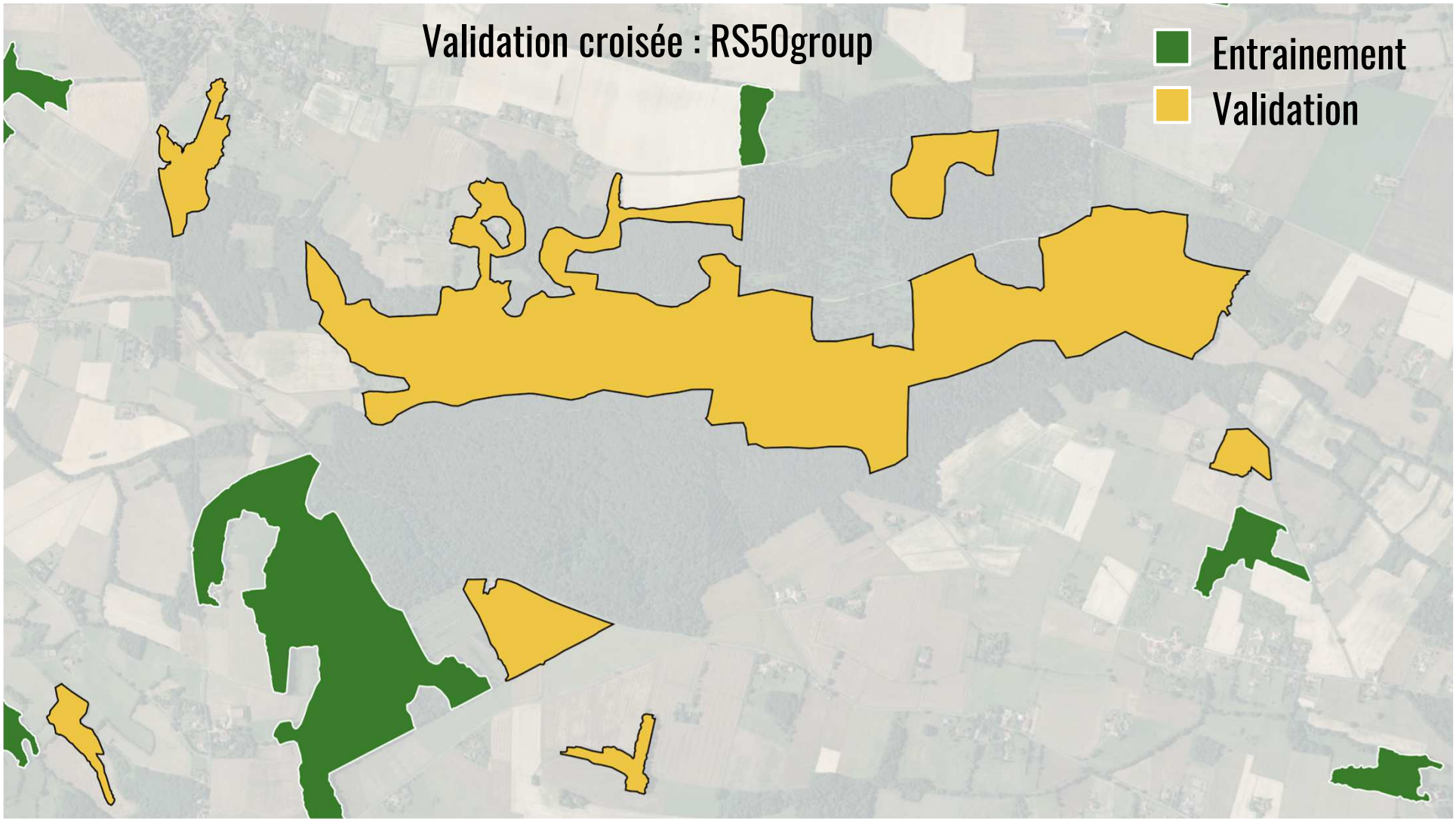


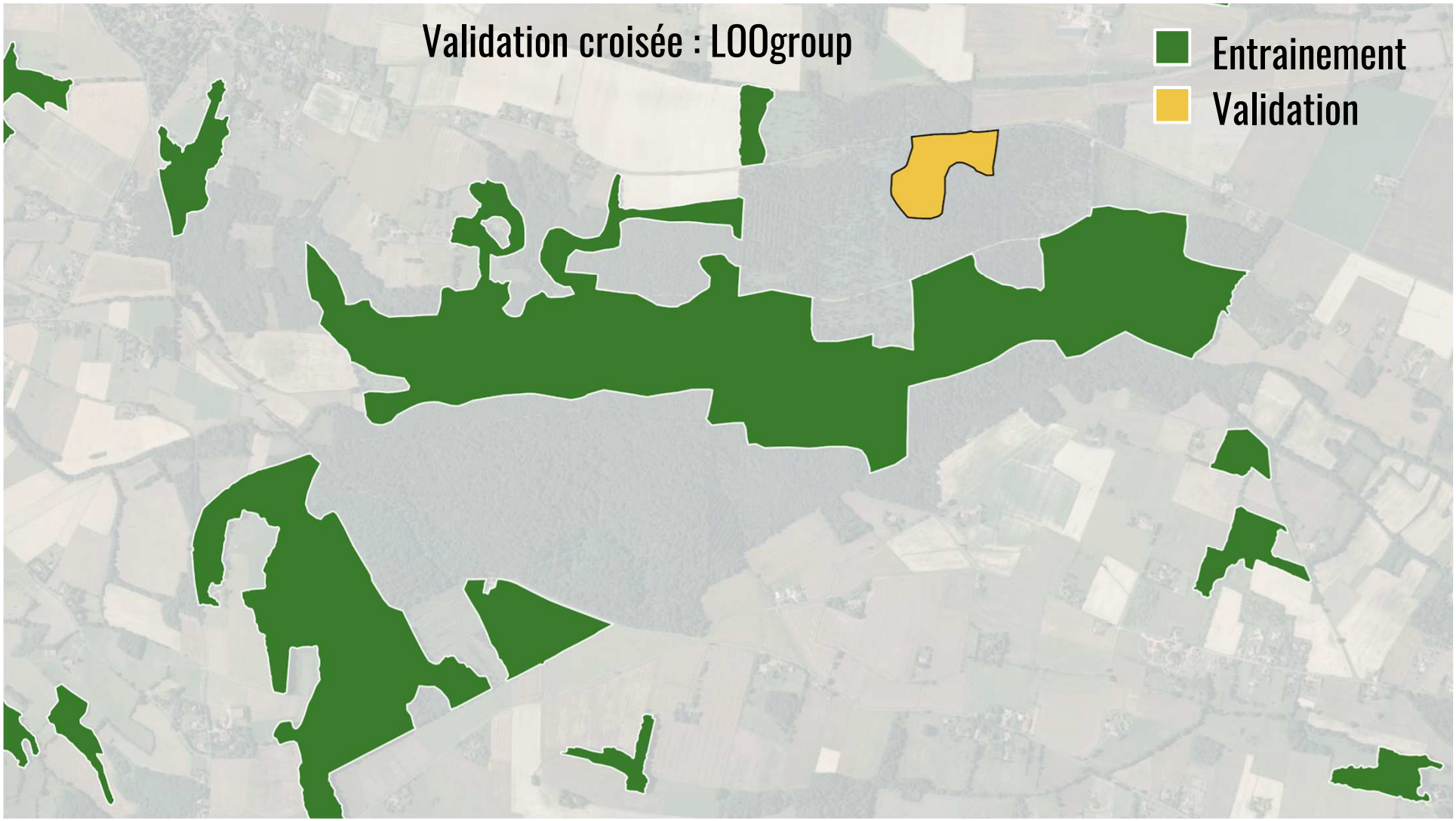


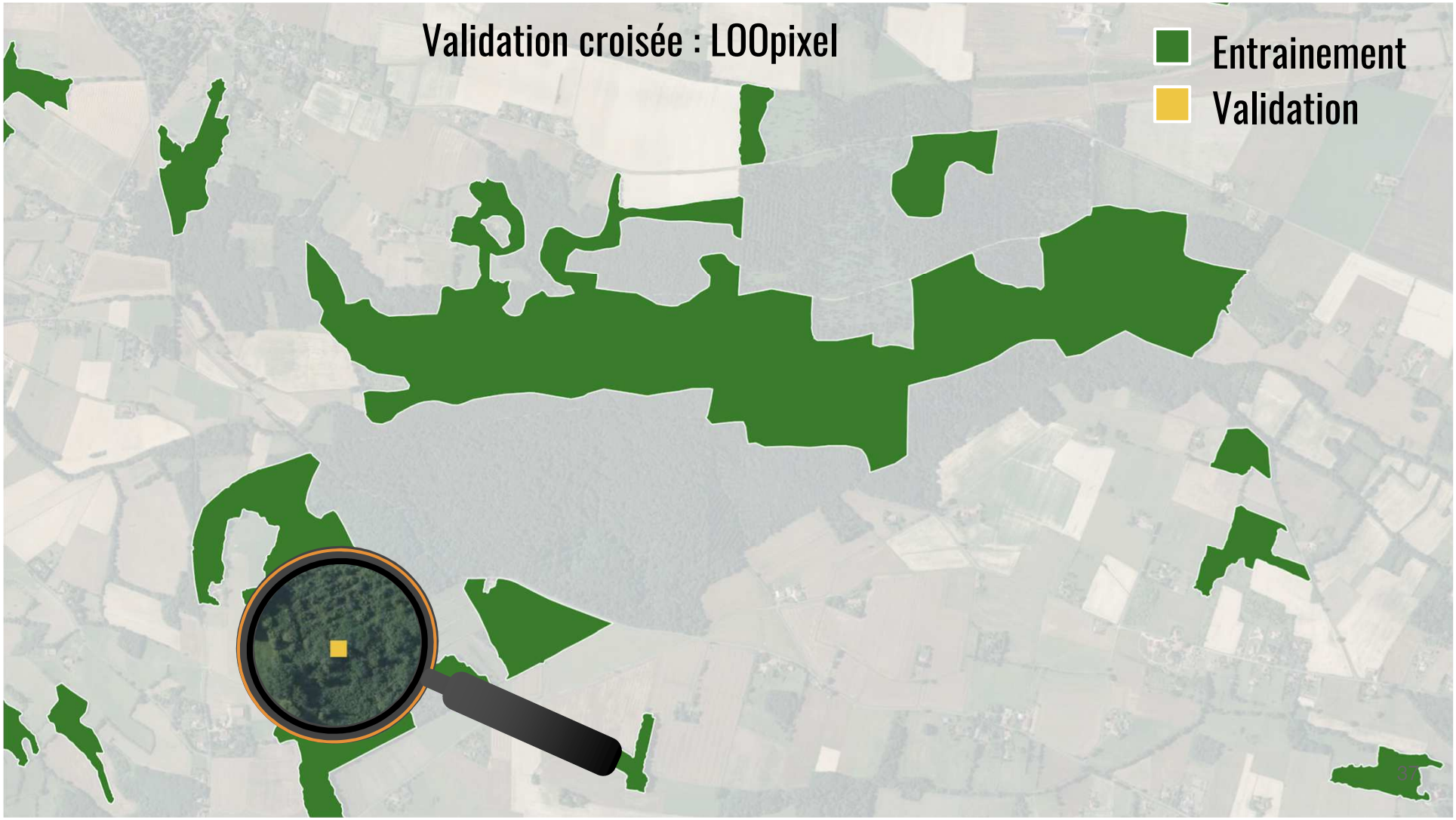

Entrainement : 100%
Entrainement
Validation

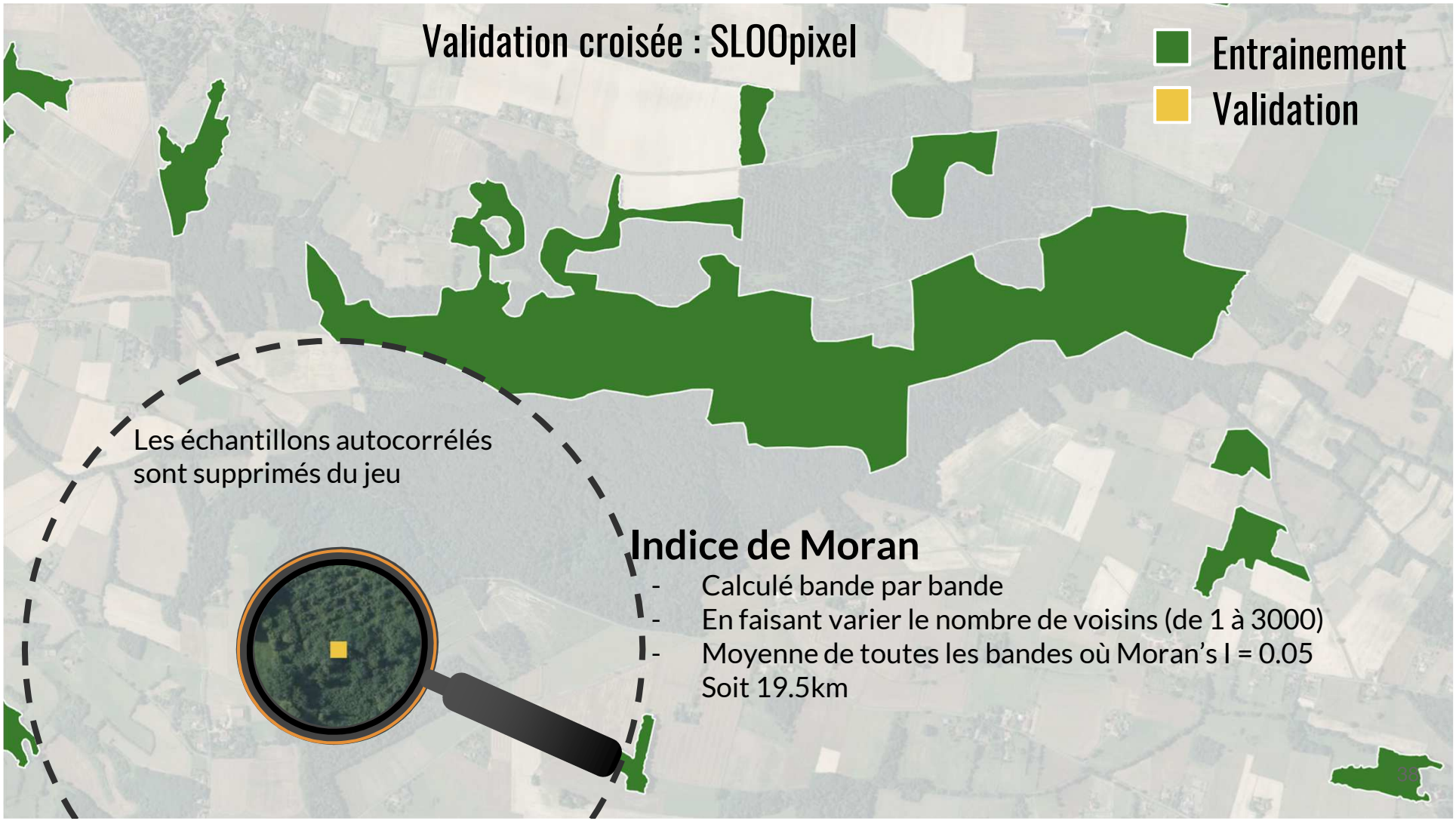

Validation croisée : RS50pixel
Entrainement
Validation


Validation croisée : RS50pixel
Entrainement
Validation


Validation croisée : RS50group
Entrainement
Validation


Validation croisée : LOOgroup
Entrainement
Validation


Validation croisée : LOOpixel
Entrainement
Validation

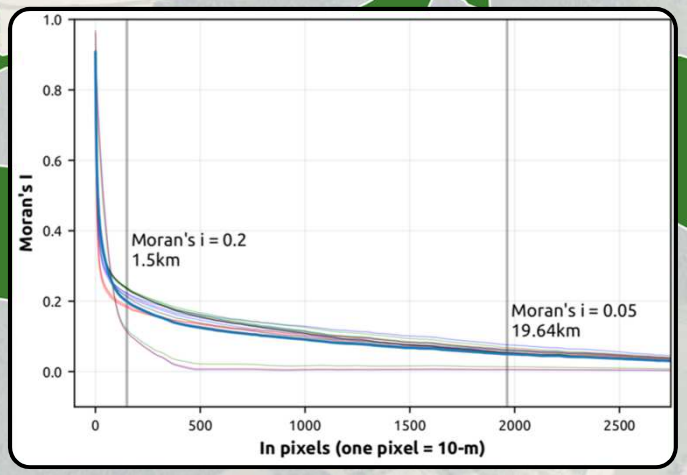

Validation croisée : SLOOpixel
Entrainement
Validation
Les échantillons autocorrélés sont supprimés du jeu
Indice de Moran
- Calculé bande par bande
- En faisant varier le nombre de voisins (de 1 à 3000)
- Moyenne de toutes les bandes où Moran's I = 0.05
Soit 19.5km

Validation croisée : SLOOpixel

Entrainement

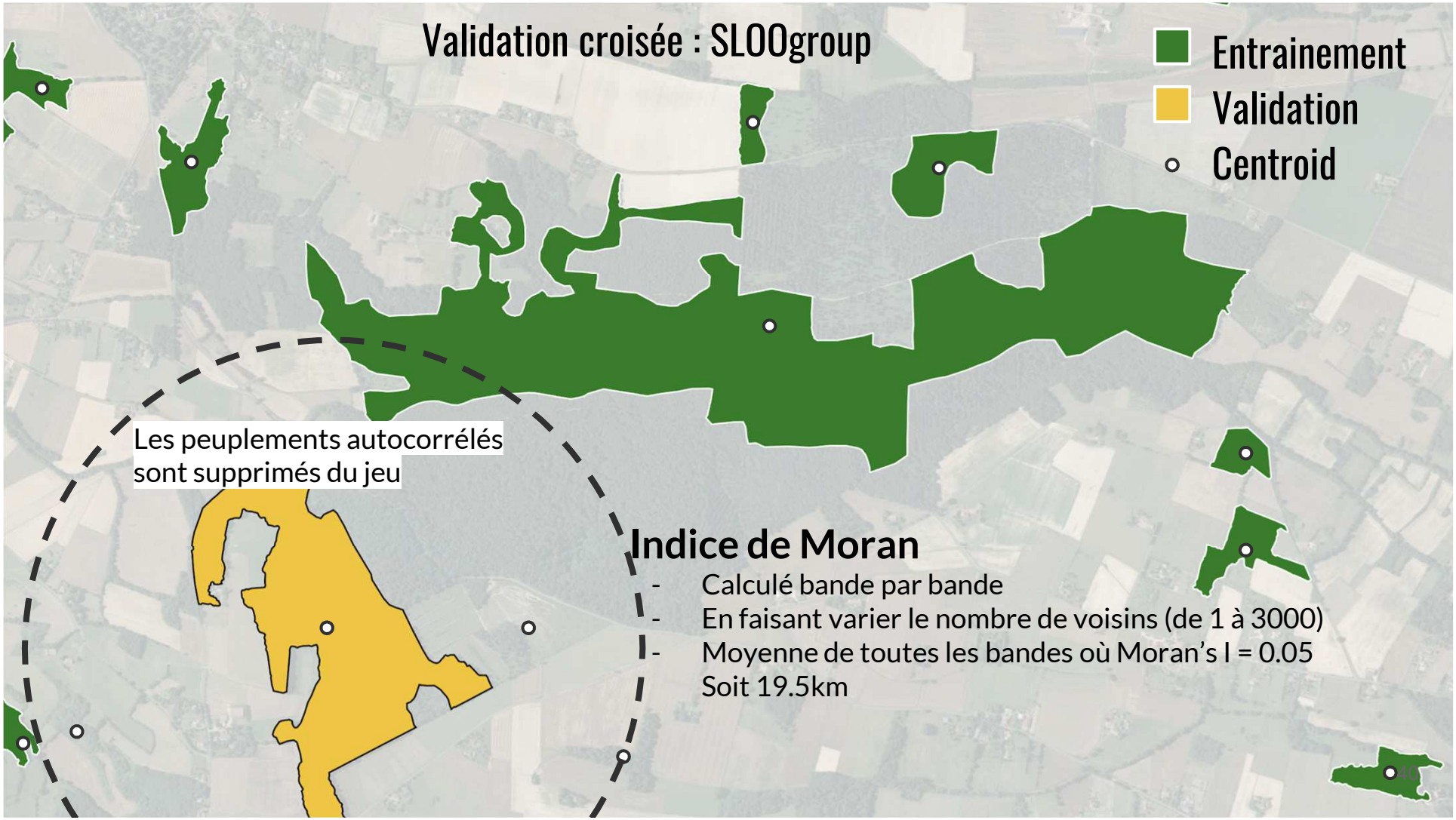
Validation

Les échantillons autocorrélés sont supprimés du jeu



Indice de Moran

- Calculé bande par bande
- En faisant varier le nombre de voisins (de 1 à 3000)
- Moyenne de toutes les bandes où Moran's I = 0.05
  Soit 19.5km


Validation croisée : SLOOgroup
Entrainement
Validation
Centroid
Les peuplements autocorrélés sont supprimés du jeu
Indice de Moran
- Calculé bande par bande
- En faisant varier le nombre de voisins (de 1 à 3000)
- Moyenne de toutes les bandes où Moran's I = 0.05
Soit 19.5km

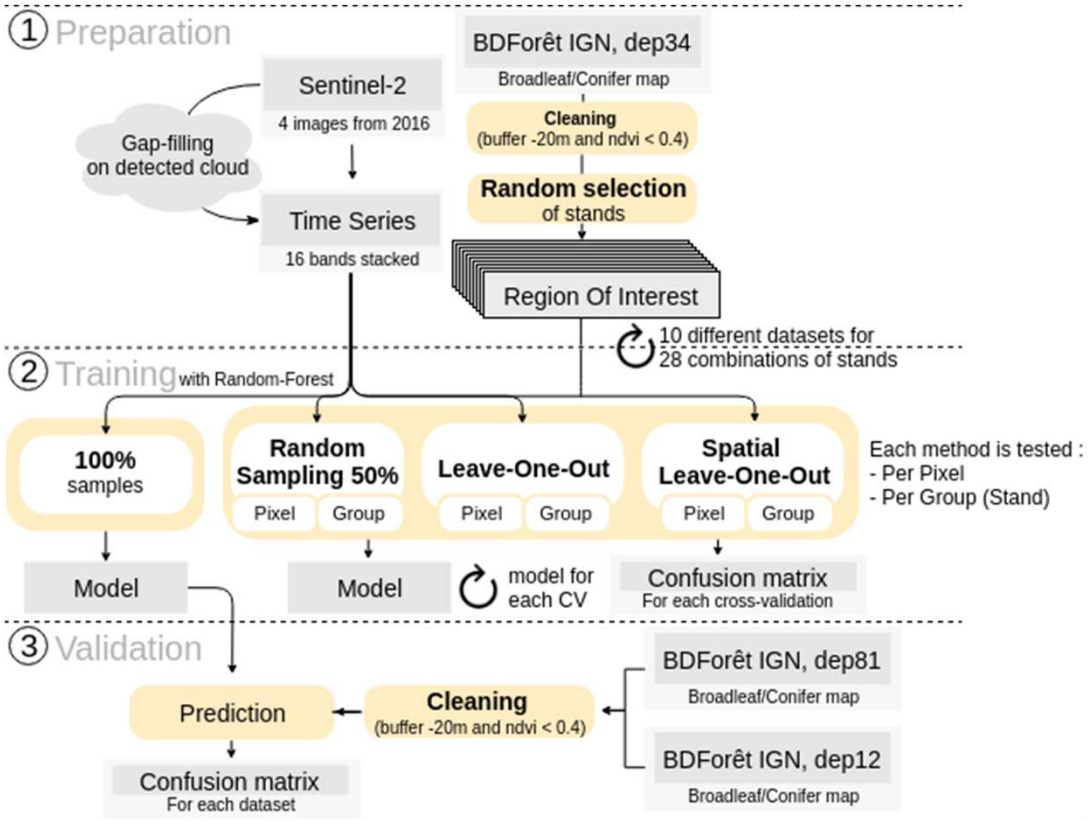
Méthode / Chaîne de traitement

Traitement des images
et sélection aléatoire
des peuplements forestiers

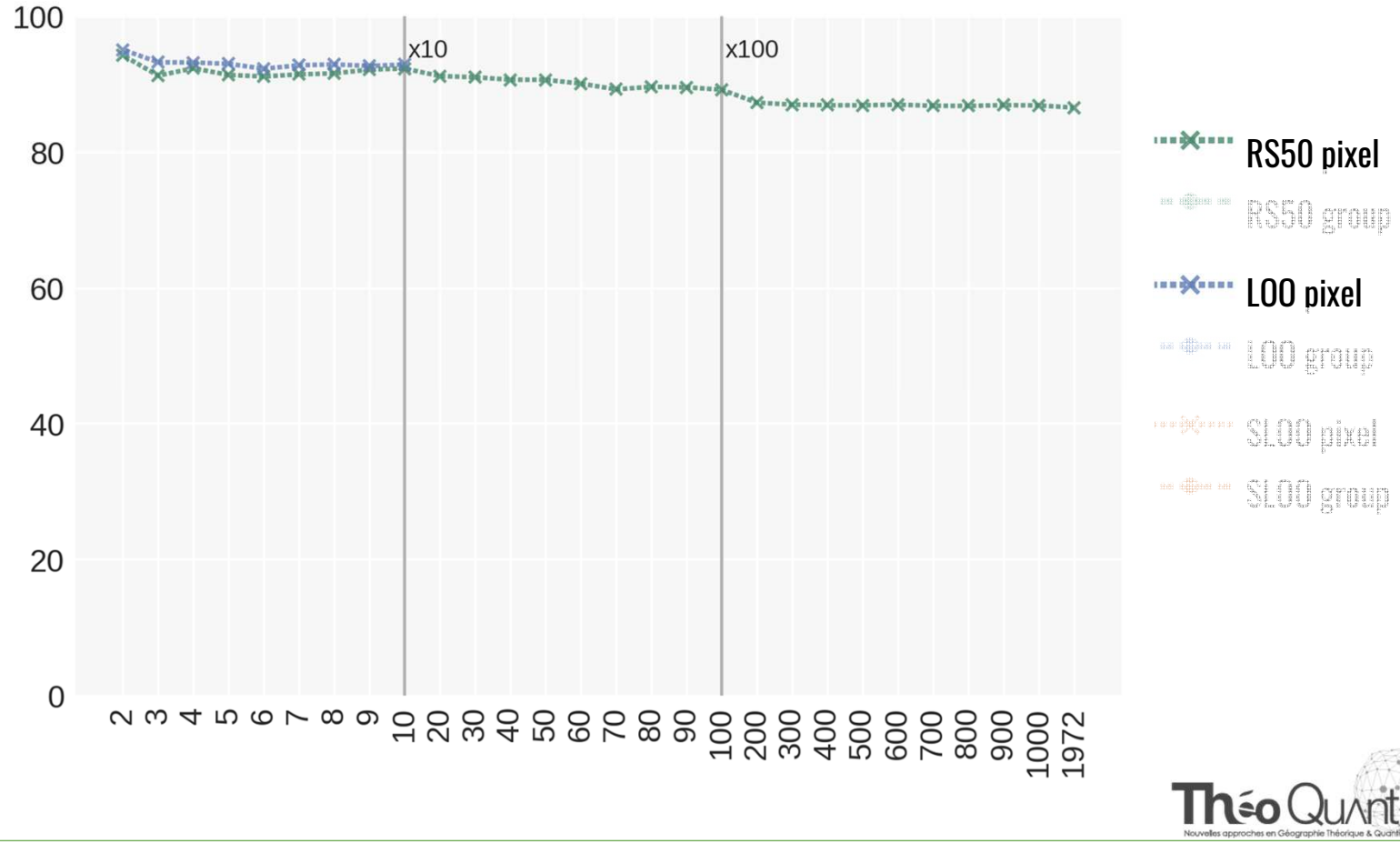
De **2 à 2000 peuplements**
10 jeux à chaque combinaison.

Apprentissage automatique avec l'algorithme **Random-Forest** et les différentes méthodes de **validations croisées**.

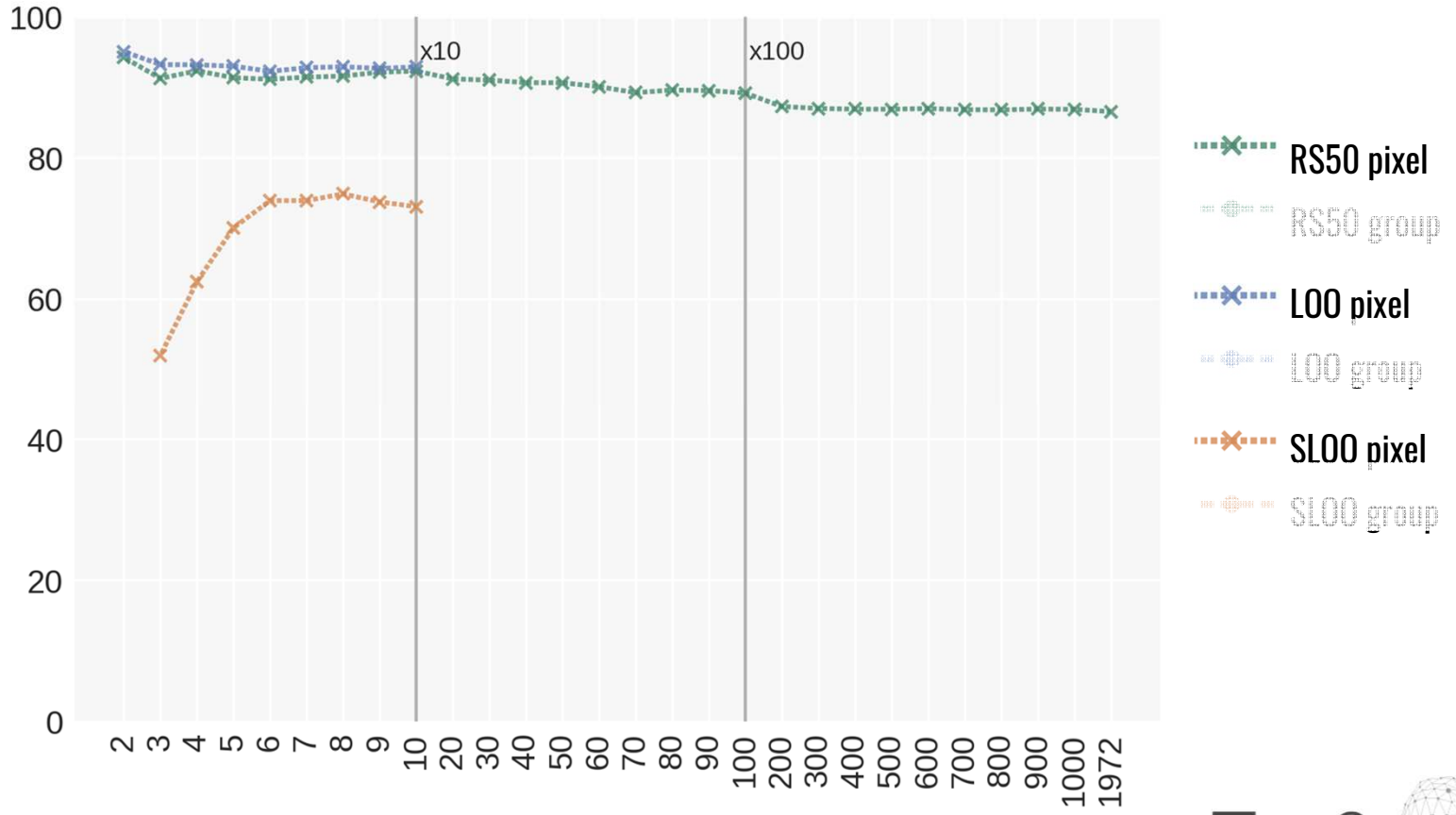
Prédiction du **modèle de référence** obtenu avec **100% des échantillons**.



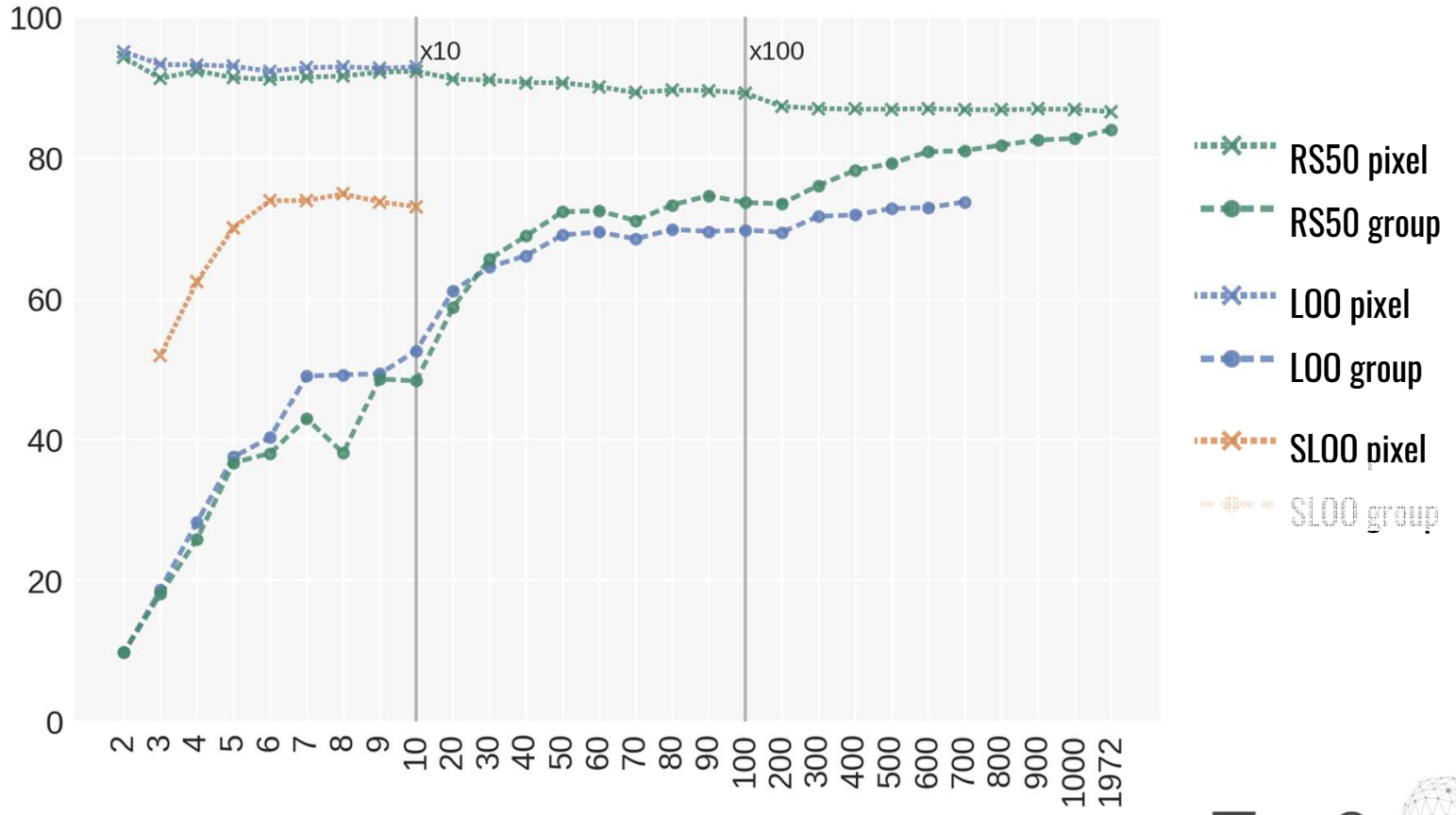
Résultats par méthode



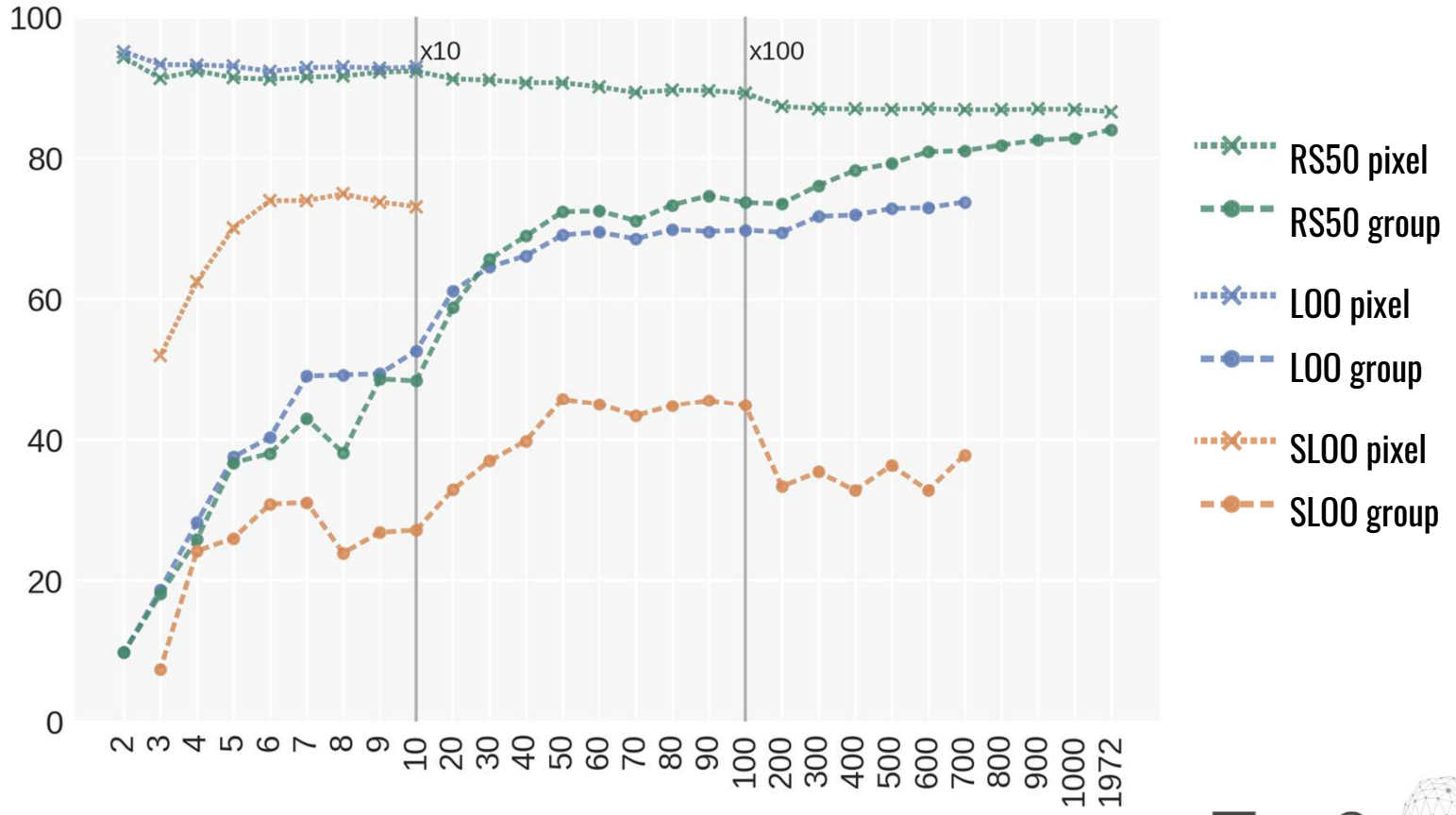
Résultats par méthode



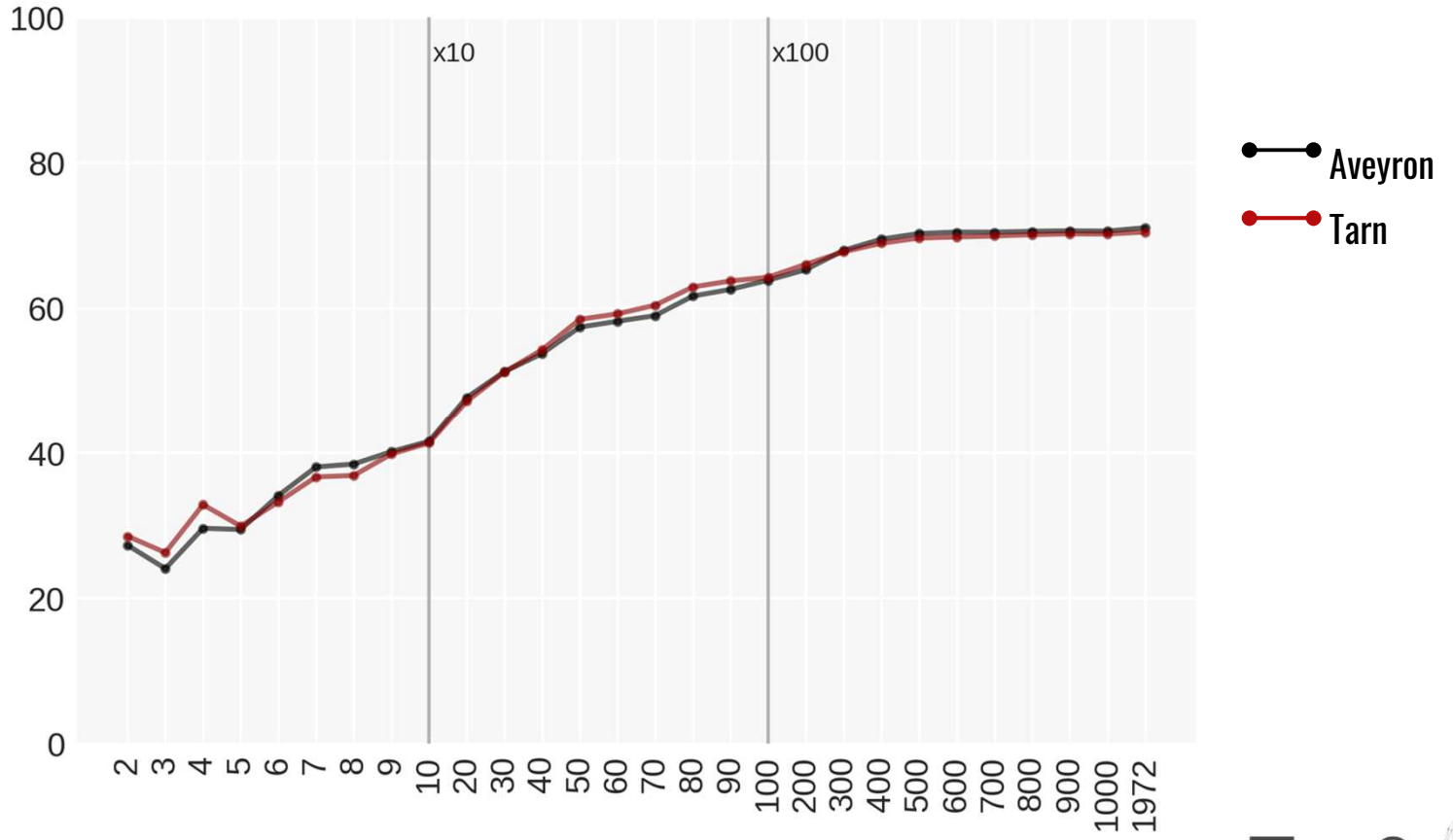
Résultats par méthode



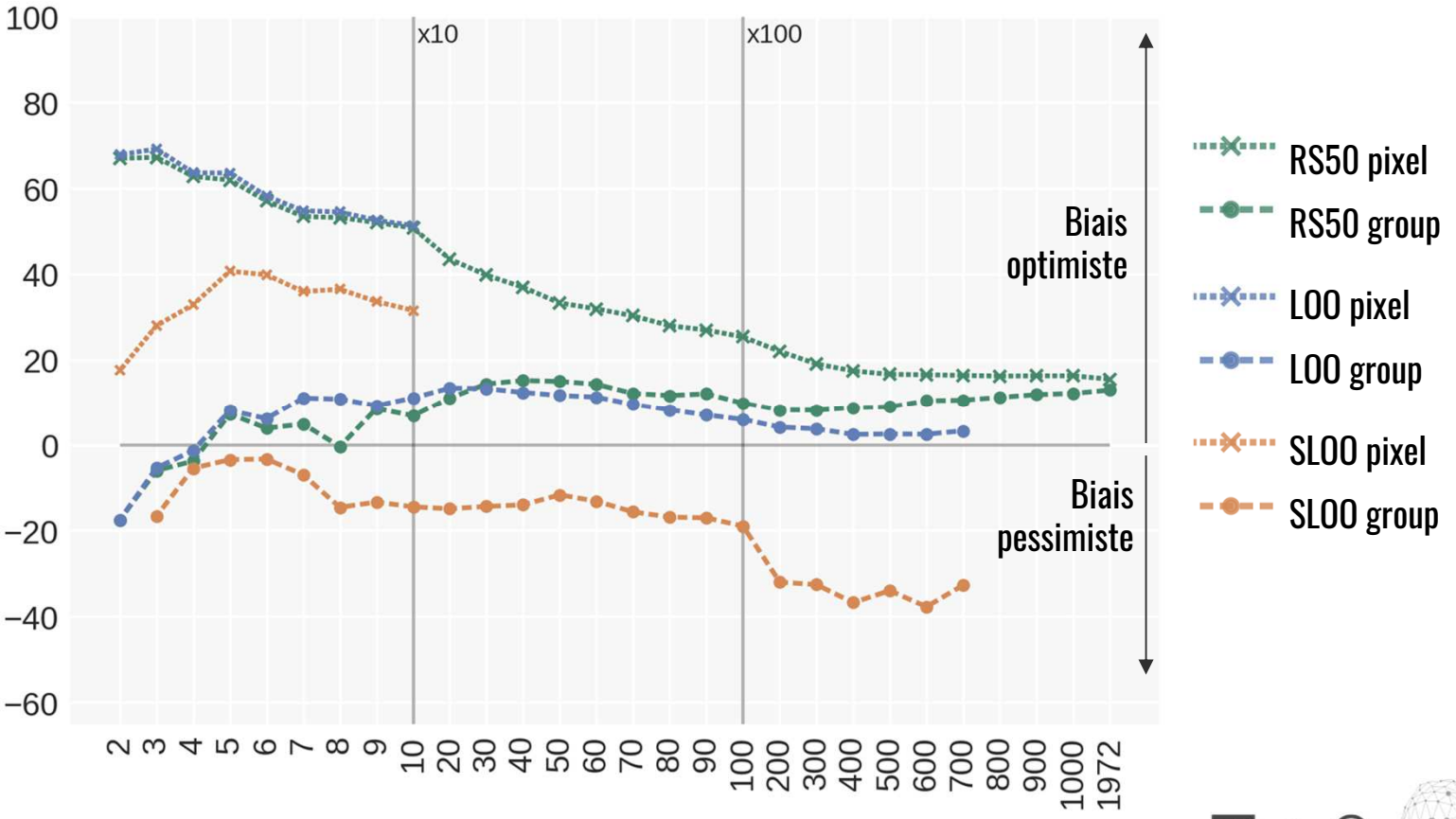
Résultats par méthode



Résultats des validations du modèle RS100 (100% échantillons)



Résultats comparés à la prédiction de l'Aveyron



Conclusions

- **Évaluation des méthodes**
  - **À éviter :**
    - RS50 pixel : **Biais optimiste très important** (diminue + il y a d'échantillons)
    - SLOO group : **Biais pessimiste important** (très instable)
  - **À privilégier**
    - Si très peu de peuplements (2 ou moins) : **Spatial Leave-One-Out par pixel**
    - **Séparation par peuplement recommandée** (LOO ou RS50)

De manière générale :

- **Privilégier une validation croisée séparant spatialement les peuplements**
- **mais… un LOO group peut être très coûteux (calcul), préférer un RS50 Peuplement.**

Comment inciter au changement ?

MUSEO TOOLBOX

Bibliothèque python pour l'apprentissage automatique à partir de vecteur ou de raster (basée sur Scikit-learn).

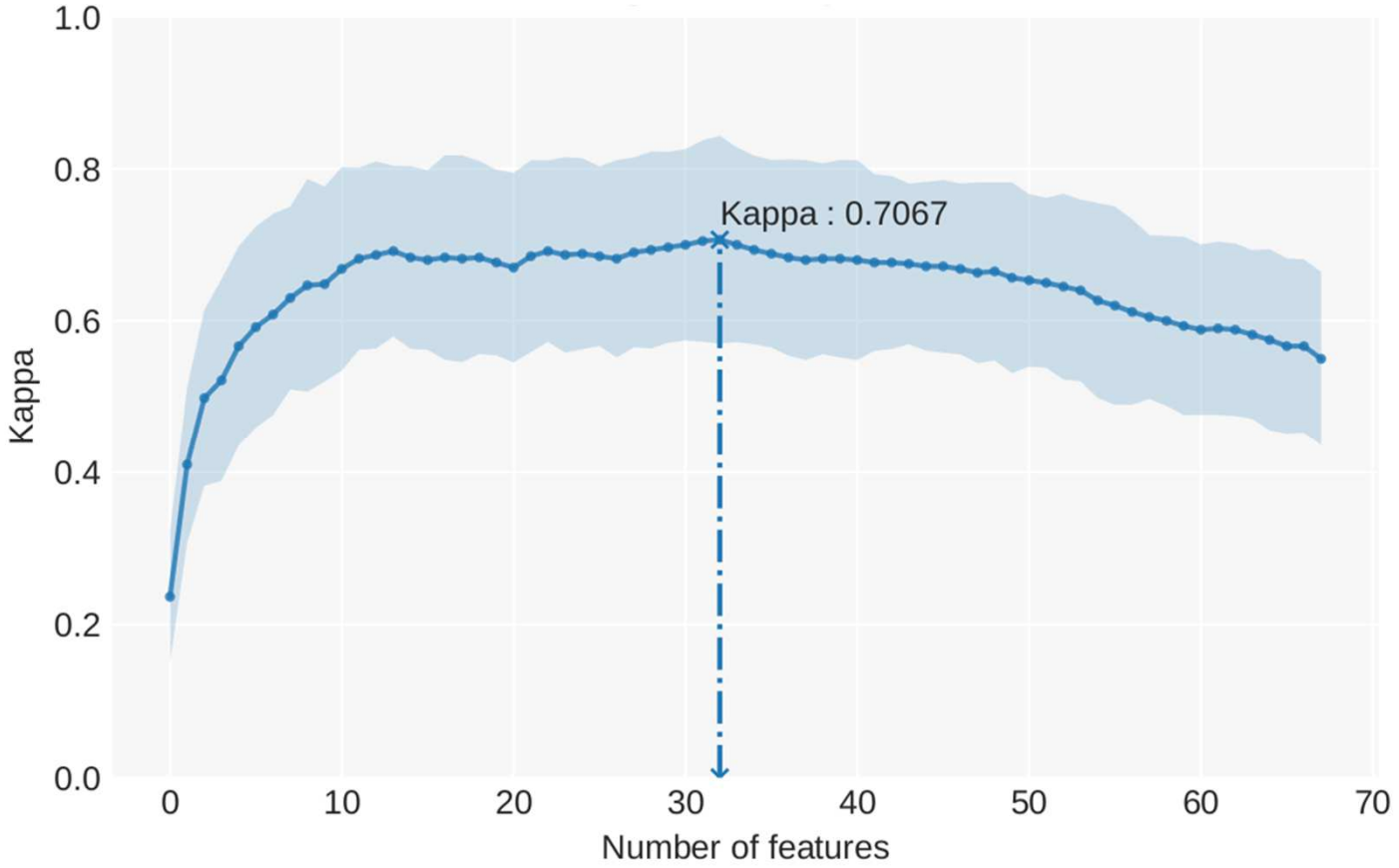
L'objectif ? Rendre **+ accessible** les classifications d'images, et fournir des **validations croisées spatialisées**.

Documentée (avec notebooks python), elle sera publiée en même temps que l'article (licence GNU v3.0).

Permet de reproduire l'ensemble des méthodes de ma thèse (et bien +).

Probable : Portage dans Qgis

Quelle performance de classification ?



Par Nicolas Karasiak

Merci de votre attention

www.karasiak.net

@nkarasiak

nicolas.karasiak@inra.fr