

CoSIA
Couverture du Sol par Intelligence Artificielle

Mise à jour : 20 avril 2023

Documentation technique

SOMMAIRE

- Qu'est-ce que CoSIA ? p.1
- Quelles sont les couvertures décrites par CoSIA ? p.2
- À partir de quelles données est produite CoSIA ? p.5
- Comment est produite CoSIA ? p.6
- Pourquoi certaines données sont-elles erronées ? p.8
- À quel métriques puis-je me référer ? p.9
- Lexique p.17

Descriptif

Les cartes CoSIA décrivent la couverture du sol, soit la nature du sol, selon 16 classes (bâtiment, surface d'eau, conifère, culture, broussaille...). Cette description du sol est produite pour tout le territoire français (métropole et DROM) et avec une haute résolution spatiale de 20 cm par pixel.

Les cartes CoSIA sont un produit de l'IGN qui interviennent actuellement dans la conception de l'OCSGE. Leur résolution spatiale et leur finesse sémantique peuvent également aider dans la production d'autres cartographies et au calcul d'autres indicateurs comme la végétation en ville, les haies & bocages, les trames vertes & bleues ou encore intervenir dans la réalisation de MOS locaux ou d'un OCSGE plus fin.

Pour produire ces cartes, on utilise des processus d'intelligence artificielle et plus particulièrement l'apprentissage profond (deep learning). Ces cartes sont alors dites de "prédiction" car elles sont obtenues à partir d'un modèle numérique d'IA qui estime statistiquement pour chaque pixel son appartenance à une classe, et peuvent ne pas refléter de manière exhaustive la réalité du terrain. Il existe des marges d'erreurs qui sont référencées pour chaque classe.



Quelles sont les couvertures décrites par CoSIA ?

p.2

Nomenclature

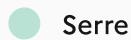
Les cartes CoSIA sont composées de 16 classes : bâtiment, zone imperméable, zone perméable, piscine, serre, sol nu, surface d'eau, neige, conifère, feuillus, coupe, broussaille, pelouse, culture, terre labourée, vigne et autre. Elles sont définies et illustrées dans la table ci-dessous.

CLASSE	DESCRIPTION	ILLUSTRATION
● Bâtiment	Bâtiment ou autre type de construction. Exemples : tour, château d'eau, silo, auvent... <u>Attention</u> : un simple mur n'est pas considéré comme un bâtiment.	
● Zone imperméable	Zone non construite, munie d'un revêtement la rendant imperméable (asphalte, béton...). <u>Exemples</u> : route, terrain de sport revêtu, parking...	
● Zone perméable	Terrain stabilisé et compacté, partiellement ou totalement perméable, et recouvert de matériaux minéraux (pierres, terre, graviers). <u>Exemples</u> : voies ferrées, chemins, carrières, salines, chantiers, enrochements rocheux, cours de ferme, cimetières (sauf zones goudronnées ou végétalisées), terrains de sports non revêtus...	



Piscine

Bassin de piscine non couverte.



Serre

Construction, pérenne ou non, en verre ou en plastique, translucide, close ou couverte à vocation le plus souvent agricole.



Sol nu

Zone naturelle non végétalisée.
Exemple : sable, galets, rochers, lapiaz...

On peut aussi classer en sol nu d'anciennes zones artificialisées retournant à la nature (carrière).



Surface eau

Surface naturelle couvertes d'eau au moment de la prise de vues aérienne.



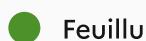
Neige

Surface couverte de neige ou de glace au moment des prises de vues aériennes.



Conifère

Peuplement de conifères ou conifères isolés.



Feuillu

Peuplement de feuillus ou feuillus isolés.



Coupe

Zone à vocation forestière récemment exploitée (coupe à blanc, coupe claire) ou récemment plantée (semis, fourrés).



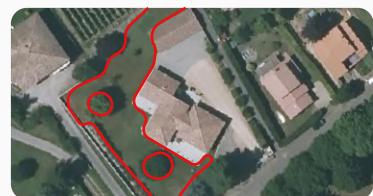
Broussaille

Terrain couvert d'arbustes et de sous-arbrisseaux.
Exemples : landes montagnardes, garrigues, maquis non boisés, terrains en friches, landes marécageuses.



Pelouse

Surface non agricole couverte de végétation herbacée.
Exemples : pelouse naturelle, d'agrément, terrain de sport, clairière...



Culture

Zone herbacée à vocation agricole incluant les cultures et les prairies.



Terre labourée

Zone de culture sans végétation au moment de la prise de vues aérienne.



Vigne

Plantations de vignes.



Autre

Zone non classée (classe inconnue, ombre dense...).



À partir de quelles données est produite CoSIA ?

p.5

Données entrantes

Les cartes CoSIA sont produites à partir de trois bases de données de l'IGN : la BD Ortho, le MNT RGE-Alti et le MNS ainsi qu'à partir d'annotations manuelles.



Annotations

L'IGN a fait produire 2500 km² d'annotations par photo-interprétation réparties sur 63 départements. À partir des prises de vue aériennes, chaque pixel (de 20cm²) est annoté et appartient à l'une des 16 classes de couverture du sol.

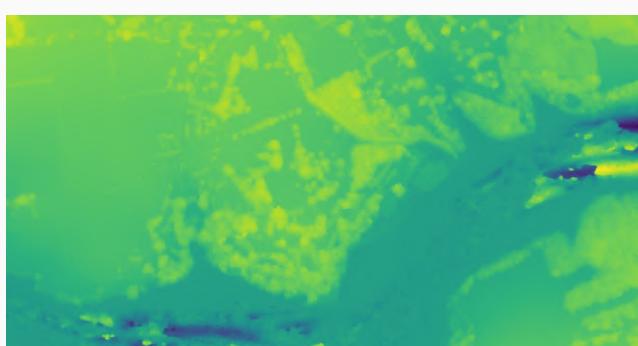
<https://geoservices.ign.fr/ressources-ia-de-couverture-du-sol>



Données Ortho

La BD ORTHO® 20 cm et l'Ortho IRC sont utilisées. Elles sont produites tous les trois ans (1/3 des départements français par an) à partir des prises de vues aériennes opérées par l'IGN. Les canaux rouge, vert et bleus sont utilisés ainsi que l'infrarouge pour détecter la végétation.

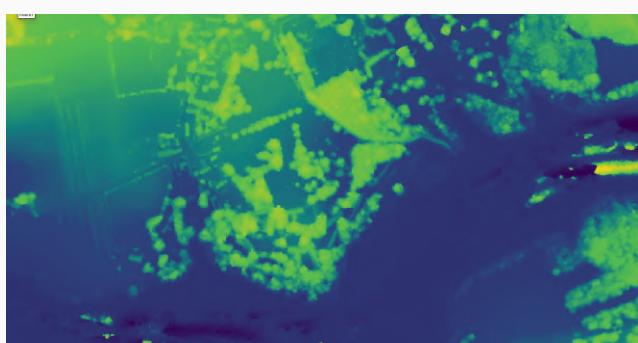
<https://geoservices.ign.fr/bdortho>



MNS

Le modèle numérique de surface (MNS) maillé décrit le relief du territoire : la forme du sol et du sursol. Il fournit l'altitude du sol en terrain dégagé et au sommet des toits et des arbres dans les zones construites et arborées. Il est également produit à parti des prises de vues aériennes opérées par l'IGN.

<https://geoservices.ign.fr/rgealtri>



MNT RGE ALTI

Le modèle numérique de terrain (MNT) maillé décrit le relief du territoire : la forme et l'altitude normale de la surface du sol.

<https://geoservices.ign.fr/rgealtri>

Comment sont produites les cartes CoSIA ?

p.6

La production de COSIA en 8 étapes.

1

Prises de vues aériennes

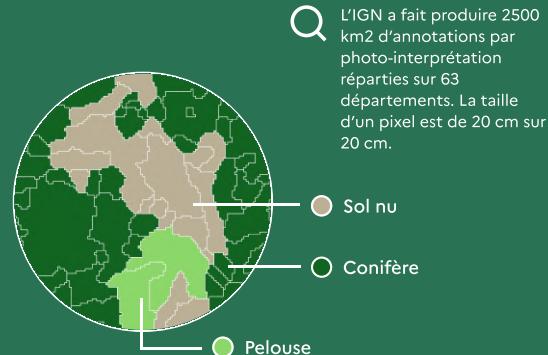
La production de CoSIA se fait par département et à partir des données ortho issues des prises de vues aériennes (un tiers des départements sont couverts chaque année).



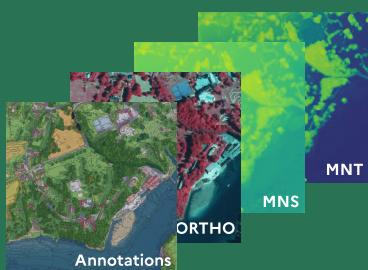
2

Annotation du territoire

À partir des données ortho, on crée des groupes de pixels dont les couleurs sont proches (on parle de segmentation d'images) et l'on assigne à chaque groupe de pixel l'une des 16 classes de couverture du sol.



3



Génération d'un jeu de données

Les annotations au format raster sont regroupées avec la BD Ortho RVB IR, le MNT et le MNS pour obtenir un jeu de données. On a alors pour chaque pixel un certain nombre d'informations.

4

Conception d'un modèle d'intelligence artificielle

On crée un modèle numérique (un ensemble de règles) et on utilise pour cela l'intelligence artificielle et des méthodes d'apprentissage profond (deep learning). Ces technologies servent à détecter des objets sur un jeu de données (ici bâtiment, arbre, plan d'eau...).

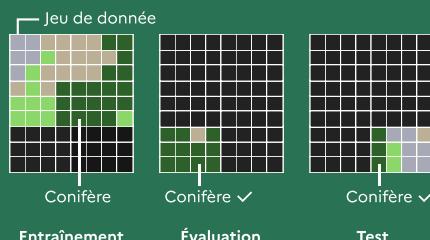
5

Entraînement du modèle

Le modèle d'intelligence artificielle prédit pour chaque pixel du jeu de données la probabilité qu'il appartienne à chacune des classes. Le modèle s'entraîne : il teste et affine un très grand nombre de paramètres jusqu'à ce qu'il produise un résultat le plus proche des annotations.



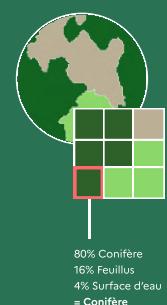
Pour l'entraînement du modèle, le jeu de données est divisé en trois parties. Le modèle s'entraîne sur la partie 1. Il est ensuite appliqué sur la partie 2 pour évaluer les paramètres choisis et est enfin testé sur la partie 3 pour s'assurer de sa performance.



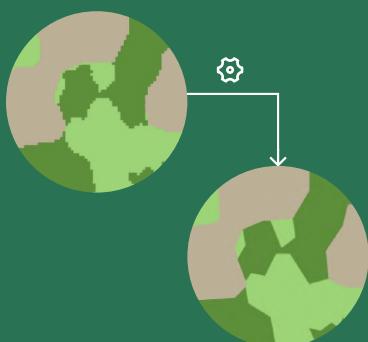
6

Inférence France entière

Le modèle, correctement paramétré, est appliqué (ou inféré) sur l'ensemble de la France par département. La probabilité la plus élevée détermine dans chaque cas la classe du pixel.



7



Vectorisation

En data-science, les données obtenues grâce au modèle sont appelées cartes de prédition (car le modèle produit une estimation). Ces données, au format raster, sont ensuite simplifiées et vectorisées (suppression des pixels isolés et lissage des contours des polygones). Cette simplification réduit fortement le poids de la donnée et offre un meilleure usage.

8

Mise à disposition

CoSIA est disponible depuis les sites de l'IGN. La donnée millésimée est mise à jour tous les trois ans pour un même département.



Q Zoom sur la simplification

Pourquoi simplifier la donnée ?

Pour répondre aux usages métiers, la donnée est vectorisée. Mais la résolution initiale des pixels de 20 cm par 20 cm alourdit considérablement la donnée. On procède alors à une simplification pour que son poids reste raisonnable et qu'elle

puisse facilement être exportée et manipulée au sein de logiciels SIG. Par exemple, la métropole de Tours sans simplification est d'environ 1.6 Gb. Avec le procédé de simplification, on réduit son poids à 186 Mb, soit par 8.

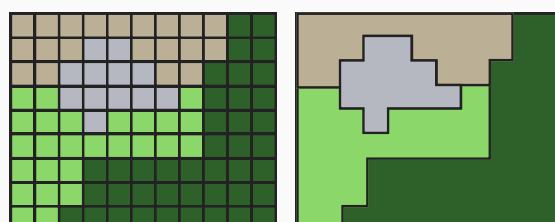
1 Suppression de pixels

On supprime les pixels isolés et les petits groupes de pixels par méthode Sieve. Seules les petites surfaces, correspondant à des objets inférieurs à 1m, sont supprimées.



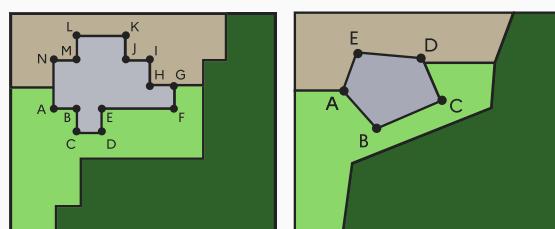
2 Vectorisation de la donnée

Les données sont vectorisées. Les polygones suivent les formes des pixels et ont une forme "d'escalier". Cette opération participe à alourdir la donnée.



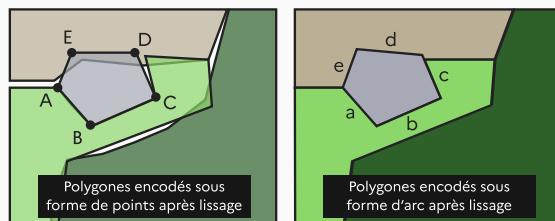
3 Simplification des polygones

Les polygones sont lissés en suivant l'algorithme avec un coefficient de lissage de 0,5 m. La valeur indique approximativement la distance maximale qu'un contour peut changer après lissage.



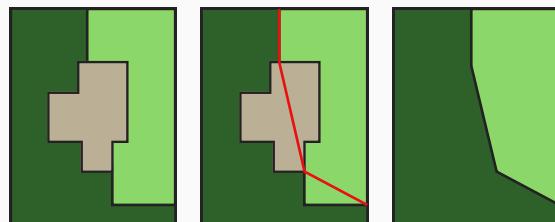
+ Cohérence Topologique

Pour garder une cohérence topologique entre les polygones adjacents lors du lissage et éviter des trous ou des recouvrements, les polygones sont encodés comme une suite d'arcs au lieu de points. Puis, après simplification des arcs, ils sont de nouveau encodés comme une suite de points standards pour faciliter leur manipulation.



+ Effets de bord

On note parfois dans la méthode utilisée, des effets de bords. Des polygones avec une géométrie particulière et une largeur étroite (inférieur à deux fois le coefficient de lissage, soit environ 1m) sont supprimés ou il existe des recouvrements. Mais ces cas restent rares.



5 paramètres à prendre en compte

pour mieux comprendre les technologies d'intelligence artificielles et de deep learning et leurs limites.

1 Les données d'entrée

Les modèles d'intelligence artificielle sont des modèles qui appliquent des traitements à une donnée. Si l'on a une donnée entrante qui possède des biais ou des inexactitudes, ceux-ci peuvent se retrouver au sein des données sortantes. S'il y a des erreurs dans le jeu de données, elles pourront induire des résultats erronés dans CoSIA. Par exemple, utiliser les données MNS et MNT en entrée a entraîné une meilleure détection de manière générale mais cela peut produire des erreurs là où les données sont de moins bonne qualité (bruit sur les zones d'eau).

2 La caractérisation des données

Pour le modèle, certaines classes sont plus faciles que d'autres à interpréter par leur couleur, altitude, géométrie ou voisinage. Les bâtiments ou les piscines sont par exemple facilement reconnaissables par leur couleur, hauteur et géométrie. En revanche, les cultures, broussailles et pelouses peuvent être plus difficiles à distinguer par leur géométries plus organiques et leurs couleurs similaires.

3 La quantité et diversité des données

Certaines classes sont plus représentées sur l'ensemble du territoire comme les bâtiments ou les cultures. Le modèle aura pu identifier une diversité de représentations de ces classes. Au contraire, le modèle aura été moins exposé aux neiges ou aux serres et aura plus de mal à classifier les pixels s'y apparentant.

4 La détection par vue aérienne

Une autre problématique liée au classement est la nature de l'image. Si un parking est planté et couvert par les feuilles d'un arbre, sa surface sera réduite ou morcelée. Le modèle n'est entraîné que pour détecter ce qui est visible sur les images. Les cartes de couverture du sol sont donc à utiliser avec d'autres bases de données pour obtenir des cartographies cohérentes.

5 Les zones sensibles

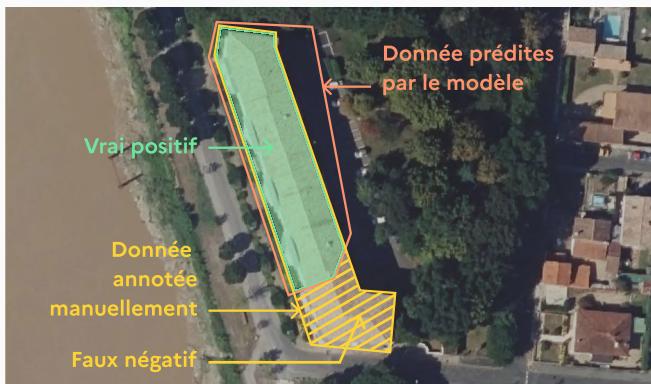
Certaines zones considérées comme sensibles tels que les aéroports ou les bases militaires sont floutées dans les données entrantes. Le modèle ne peut prédire pour ces zones des données correctes.

À quelles métriques puis-je me référer ?

p.9

Métriques

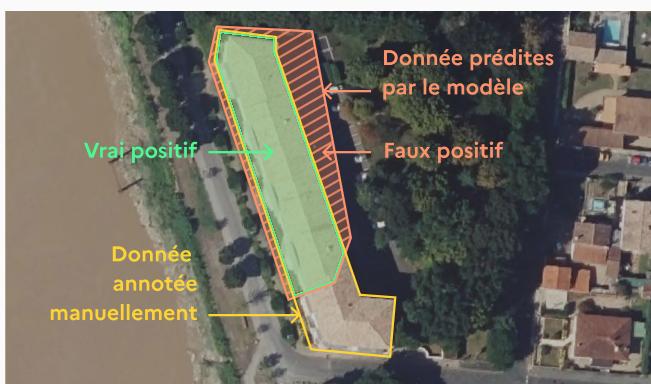
Les métriques disponibles mesurent la performance du modèle sur chaque classe. Elles sont calculées à partir des annotations réparties à travers la France et ne prévalent donc pas pour un département en particulier.



Rappel (ou sous-détection)

Cette métrique indique le nombre de pixels bien classés sur l'ensemble des pixels qui ont été annotés.

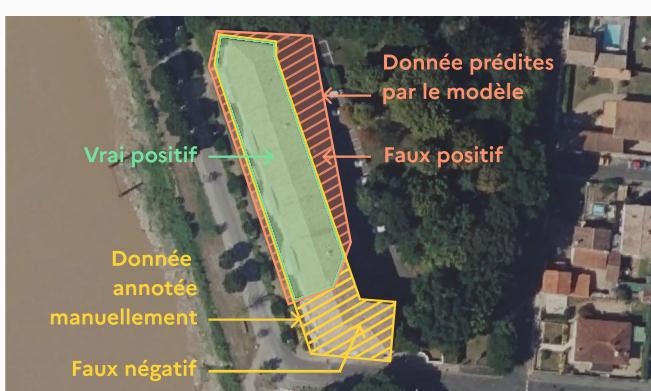
$$\frac{\text{Vrai positif}}{\text{Vrai positif} + \text{Faux négatif}}$$



Précision (ou sur-détection)

Cette métrique indique le nombre de pixels bien classés sur l'ensemble des pixels prédits par le modèle.

$$\frac{\text{Vrai positif}}{\text{Vrai positif} + \text{Faux positif}}$$



IoU Intersection over Union

Cette métrique indique le nombre de pixels bien classés sur l'ensemble des pixels qui ont été annotés et des pixels prédits par le modèle.

$$\frac{\text{Vrai positif}}{\text{Vrai positif} + \text{Faux négatif} + \text{Faux positif}}$$

Rappel

- Données en (%) • France entière • Millésimes: 2019, 2020, 2021

Données de référence	bâtiment	91,12	5,31	0,91	0,01	0,07	0,07	0,30	0,00	0,02	0,35	0,00	0,06	1,76	0,02	0,00	0,00	100
	Zone imperméable	3,53	83,38	7,36	0,03	0,01	0,18	0,48	0,00	0,09	1,05	0,00	0,21	3,54	0,08	0,07	0,01	100
Zone perméable	1,45	11,76	67,12	0,00	0,02	4,01	0,87	0,01	0,27	1,15	0,04	0,57	10,51	1,28	0,82	0,11	100	
Piscine	6,09	16,88	0,68	66,26	0,24	0,06	3,09	0,00	0,05	0,48	0,00	0,11	6,03	0,03	0,00	0,00	100	
Serre	30,47	3,47	5,93	0,03	49,02	0,13	0,11	0,00	0,01	0,66	0,01	0,12	4,35	4,34	0,19	0,57	100	
Sol nu	0,16	2,30	25,14	0,00	0,00	47,82	7,48	0,01	0,57	0,66	0,19	2,13	9,90	1,93	1,67	0,05	100	
Surface eau	0,42	1,37	2,14	0,10	0,02	0,93	90,15	0,00	0,04	0,82	0,01	0,31	2,31	0,93	0,47	0,00	100	
Neige	16,82	4,65	36,98	0,01	0,01	26,26	5,39	7,09	0,00	0,00	0,00	0,00	0,02	0,00	2,77	0,00	100	
Conifère	0,03	0,16	0,16	0,06	0,00	0,21	0,20	0,00	64,93	28,09	0,14	3,12	2,87	0,06	0,00	0,02	100	
Feuillu	0,99	0,36	0,19	0,00	0,03	0,06	0,18	0,00	3,44	86,97	0,08	3,97	3,96	0,50	0,01	0,16	100	
Coupe	0,01	0,01	2,67	0,00	0,00	1,48	0,20	0,00	5,30	12,42	6,74	31,05	31,66	7,65	0,08	0,74	100	
Brousaille	0,13	0,65	0,99	0,00	0,00	1,34	0,54	0,01	2,58	17,48	0,35	43,38	30,77	1,76	0,01	0,1°	100	
Pelouse	0,51	1,45	2,45	0,00	0,01	1,17	0,53	0,00	1,19	5,15	0,16	4,28	72,61	10,08	0,21	0,20	100	
Culture	0,02	0,10	1,04	0,00	0,02	0,10	0,16	0,00	0,06	0,95	0,05	0,66	24,98	66,75	4,40	0,70	100	
Terre labourée	0,02	0,96	5,08	0,00	0,01	0,43	0,42	0,00	0,03	0,44	0,03	0,06	2,33	44,60	44,89	0,70	100	
Vigne	0,01	0,06	0,79	0,00	0,02	0,10	0,03	0,00	0,32	1,51	0,51	1,48	4,46	19,33	0,96	70,44	100	
bâtiment		Zone imperméable	Zone perméable	Piscine	Serre	Sol nu	Surface eau	Neige	Conifère	Feuillu	Coupe	Brousaille	Pelouse	Culture	Terre labourée	Vigne	Total	
Données prédites par le modèle																		

Légende

Données de référence : partie des annotations manuelles qui ont servi à évaluer le modèle (voir étape 5 de la production de CoSIA, page 8).

Données prédites par le modèle : la prédiction de la classe pour chaque pixel des données de référence.

Lecture du tableau

Ce tableau se lit en ligne selon l'énoncé suivant : "Parmi tous les pixels annotés comme [classe] dans les annotations de référence, XX % ont été prédits par le modèle comme [classe]."

Exemple : Parmi tous les pixels annotés comme bâtiment dans les annotations de référence; 91,12 % ont été prédits par le modèle comme bâtiment.

Précision

- Données du modèle (%) • France entière • Millésimes: 2019, 2020, 2021

		Données prédites par le modèle															
		bâtiment	Zone imperméable	Zone perméable	Piscine	Serre	Sol nu	Surface eau	Neige	Conifère	Feuillu	Coupe	Broussaille	Pelouse	Culture	Terre labourée	Vigne
Données de référence	bâtiment	88,99	3,22	0,85	1,93	4,06	0,20	0,33	0,14	0,02	0,09	0,01	0,07	0,48	0,01	0,00	0,00
	Zone imperméable	5,72	83,90	11,36	9,93	0,54	0,82	0,89	0,09	0,16	0,44	0,03	0,42	1,62	0,06	0,33	0,08
	Zone perméable	1,36	6,82	59,75	0,57	1,42	10,89	0,92	5,14	0,26	0,28	1,51	0,66	2,77	0,56	2,41	0,42
	Piscine	0,03	0,05	0,00	67,80	0,07	0,00	0,02	0,00	0,00	0,00	0,00	0,00	0,01	0,00	0,00	0,00
	Serre	0,80	0,06	0,15	0,17	80,05	0,01	0,00	0,02	0,00	0,00	0,01	0,00	0,03	0,05	0,02	0,06
	Sol nu	0,07	0,68	11,38	0,00	0,11	66,03	4,06	2,03	0,28	0,08	3,73	1,26	1,33	0,43	2,49	0,09
	Surface eau	0,37	0,74	1,77	16,38	1,01	2,34	89,49	0,57	0,03	0,18	0,22	0,33	0,57	0,38	1,28	0,01
	Neige	0,56	0,10	1,16	0,08	0,01	2,52	0,20	91,92	0,00	0,00	0,00	0,00	0,00	0,00	0,29	0,00
	Conifère	0,03	0,12	0,18	0,06	0,00	0,73	0,28	0,00	73,63	8,72	6,85	4,63	0,97	0,04	0,21	0,09
	Feuillu	0,33	0,79	0,66	0,37	6,59	0,62	0,71	0,01	12,60	80,65	12,00	17,60	3,98	0,84	0,10	2,34
	Coupe	0,00	0,00	0,26	0,00	0,00	0,44	0,02	0,00	0,56	0,33	29,04	4,00	0,92	0,37	0,03	0,31
	Broussaille	0,13	0,39	0,91	0,15	0,21	3,76	0,59	0,01	2,48	4,40	13,97	52,32	8,40	0,80	0,03	0,41
	Pelouse	1,55	2,73	7,10	2,47	1,99	10,32	1,85	0,01	3,74	4,08	20,94	16,23	62,36	14,38	1,96	2,40
	Culture	0,05	0,14	2,22	0,08	3,36	0,68	0,41	0,02	0,13	0,56	4,68	1,87	15,90	70,59	31,24	6,32
	Terre labourée	0,01	0,25	2,01	0,00	0,14	0,52	0,20	0,04	0,01	0,05	0,57	0,03	0,27	8,72	58,88	1,17
	Vigne	0,00	0,01	0,23	0,00	0,43	0,09	0,01	0,00	0,10	0,01	6,46	0,56	0,38	2,77	0,92	86,29
	Total	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100

Légende

Données de référence : partie des annotations manuelles qui ont servi à évaluer le modèle (voir étape 5 de la production de CoSIA, page 8).

Données prédites par le modèle : la prédiction de la classe pour chaque pixel des données de référence.

Lecture du tableau

Ce tableau se lit en colonne selon l'énoncé suivant : "Parmi tous les pixels prédis comme [classe] par le modèle, XX % sont annotés comme [classe] dans les annotations de référence."

Exemple : Parmi tous les pixels prédis comme bâtiment par le modèle, 88,99 % sont annotés comme bâtiment dans les annotations de référence.

IoU

France entière Millésimes: 2019, 2020, 2021

Données du modèle

IoU par classe en %

- Batiment 81,89
- Zone imperméable 71,88
- Zone perméable 46,22
- Piscine 50,40
- Serre 44,16
- Sol nu 38,38
- Surface eau 81,52
- Neige 7,05
- Conifère 55,69
- Feuillu 71,96
- Coupe 5,78
- Broussaille 31,09
- Pelouse 50,49
- Culture 52,23
- Terre labourée 34,18
- Vigne 63,35

Conclusions

- Le modèle prédit avec précision les classes bâtiment, zone imperméable, surface d'eau et feuillu.
- Le modèle prédit avec difficulté les classes sol nu, neige, coupe, broussaille et terre labourée.

Ces définitions proviennent du glossaire de la CNIL et sont disponibles à cette URL : <https://www.cnil.fr/fr/glossaire>

Annotation

L'annotation est le procédé par lequel les données sont décrites manuellement afin d'être caractérisées, par exemple en attribuant à une image de chien l'étiquette correspondante. On parle aussi de vérité terrain ou groundtruth.

Apprentissage automatique (ou Machine learning)

L'apprentissage automatique (machine learning en anglais) est un champ d'étude de l'intelligence artificielle qui vise à donner aux machines la capacité d'« apprendre » à partir de données, via des modèles mathématiques. Plus précisément, il s'agit du procédé par lequel les informations pertinentes sont tirées d'un ensemble de données d'entraînement. Le but de cette phase est l'obtention des paramètres d'un modèle qui atteindront les meilleures performances, notamment lors de la réalisation de la tâche attribuée au modèle. Une fois l'apprentissage réalisé, le modèle pourra ensuite être déployé en production.

Apprentissage profond (ou Deep learning)

L'apprentissage profond est un procédé d'apprentissage automatique (machine learning) utilisant des réseaux de neurones possédants plusieurs couches de neurones cachées. Ces algorithmes possédant de très nombreux paramètres, ils demandent un nombre très important de données afin d'être entraînés.

Donnée d'entrée

Dans le domaine de l'intelligence artificielle, une donnée d'entrée est une donnée utilisée pour l'apprentissage automatique ou la prise de décision du système d'IA (en phase de production).

Donnée de sortie

Dans le domaine de l'intelligence artificielle, une donnée de sortie est une valeur représentant tout ou partie de l'opération effectuée par le système d'IA à partir des données d'entrée.

Échantillon

Dans le domaine de l'intelligence artificielle, l'échantillon est une fraction représentative d'une population ou d'un univers statistique.

Entraînement

L'entraînement est le processus de l'apprentissage automatique pendant lequel le système d'intelligence artificielle construit un modèle à partir de données.

Intelligence artificielle

L'intelligence artificielle est un procédé logique et automatisé reposant généralement sur un algorithme et en mesure de réaliser des tâches bien définies. Pour le Parlement européen, constitue une intelligence artificielle tout outil utilisé par une machine afin de « reproduire des comportements liés aux humains, tels que le raisonnement, la planification et la créativité ».

Modèle

Le modèle d'IA est la construction mathématique générant une déduction ou une prédiction à partir de données d'entrée. Le modèle est estimé à partir de données annotées lors de la phase d'apprentissage (ou d'entraînement) du système d'IA.