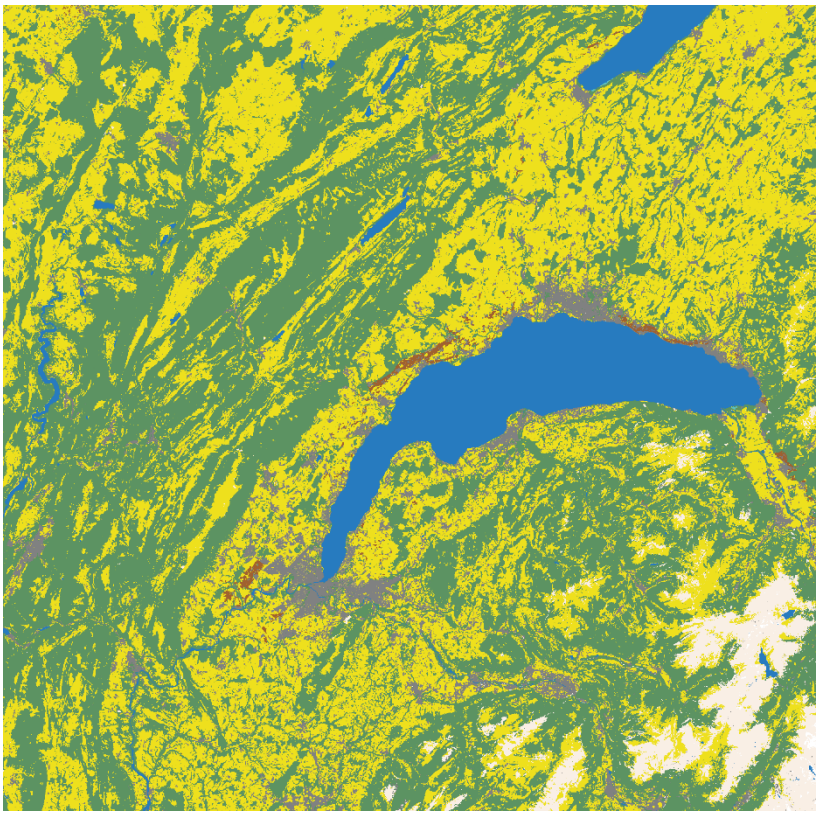
Cartographie de la couverture du sol du bassin genevois et de ses environs en 2018 et en 2022

Classification d'images satellites par apprentissage automatique



Travail de mémoire réalisé dans le cadre du Certificat complémentaire en Géomatique

Présenté par

Daniel Risse

Sous la direction de

Dr. Gregory Giuliani et Dr. Pascal Martin

Janvier 2023

UNIVERSITÉ DE GENÈVE

Table des matières

Remerciements

Plusieurs personnes m'ont aidé à aller au bout du travail présenté ici. Je remercie tout particulièrement Pascal Martin et Gregory Giuliani, qui ont supervisé mon stage et la rédaction de ce mémoire. Tous deux se sont toujours montrés disponibles et réactifs, notamment quand je rencontrais des problèmes parfois difficiles à résoudre. Gregory Giuliani m'a lui-même fourni les données de base de ce travail, sans lesquelles ce dernier aurait été bien plus difficile. Je remercie également Felipe Souza, qui a participé au développement de l'outil informatique utilisé dans le cadre de ce travail. Son aide fut très précieuse et je pense pouvoir affirmer que sans elle, je ne serais pas parvenu à réaliser certains des traitements d'images exposés ici. Pour cela, je lui en suis très reconnaissant. Enfin, je remercie Thierry Froidevaux pour sa réactivité et de m'avoir aidé à accéder au serveur de calcul de l'Université de Genève.

1. Introduction

1.1 Sujet du mémoire et but du stage

Les changements dans la couverture du sol ont un impact conséquent sur le réchauffement climatique et sur l'érosion de la biodiversité. Les surveiller est donc d'une grande importance pour prendre des décisions politiques informées. Or, la tâche n'est pas aisée. En particulier, la récolte d'informations de terrain est fastidieuse et il peut être difficile d'obtenir une information précise et actualisée, raisons pour lesquelles de nouvelles méthodes sont régulièrement explorées.

Le présent mémoire détaille une méthode pour cartographier la couverture du sol qui recourt à la classification d'images satellites par apprentissage automatique. Cette méthode repose sur l'utilisation d'une extension R développée en 2021 par une équipe de chercheurs au Brésil et qui permet de surveiller l'évolution de la couverture du sol avec plus de précision.

Le travail réalisé est lié à un stage effectué entre septembre et décembre 2022 au Système d'Information du Patrimoine Vert (SIPV), dans le cadre du Certificat Complémentaire en Géomatique (CCG) proposé par l'Université de Genève. Coordonné par les Conservatoire et Jardin Botanique de Genève (CJBG), le SIPV est un système d'information qui centralise toutes les données sur la biodiversité végétale de la région genevoise.

La région cartographiée englobe une partie du bassin lémanique, de la Suisse occidentale et de régions françaises limitrophes à la Suisse. Dans le cadre d'un autre stage de géomatique au SIVP, une carte de la couverture du sol avait déjà été réalisée en 2018 pour le Grand Genève, mais son autrice avait utilisé une autre méthode que celle employée ici. À terme, la procédure exposée dans ce travail pourra éventuellement servir à actualiser la carte des milieux naturels, qui elle-même contribuera à l'établissement de l'infrastructure écologique à l'échelle du Grand Genève.

Hors introduction et conclusion, ce mémoire est divisé en quatre parties principales. Dans un premier temps, une revue de la littérature contextualise l'analyse effectuée en exposant les différents concepts mobilisés. Cette section explique dans quel champ de la recherche le présent mémoire s'inscrit, à savoir la classification d'images satellites par apprentissage automatique. Dans un deuxième temps, les objectifs de recherche sont exposés dans une partie problématique, le principal objectif étant d'explorer une méthode susceptible d'actualiser la cartographie de la couverture du sol d'une année à l'autre, afin de mieux surveiller son évolution. Puis, dans un troisième temps, la méthodologie développée pour remplir les objectifs de recherche est présentée. Cette dernière est divisée en trois parties principales : la préparation des données d'entraînement, la classification à proprement parler et l'évaluation statistique de la classification réalisée. Enfin, dans un dernier temps, les résultats présentés sont analysés et discutés. Avant toute chose, il convient néanmoins de présenter l'institution dans laquelle le stage lié à ce mémoire a été effectué.

1.2 Institution hôte

Les CJBG sont une institution publique, financée par la Commune de Genève et dont l'histoire est vieille de plus de 200 ans. De fait, le « Jardin des plantes est fondé en 1817 à l'initiative d'Augustin-Pyramus de Candolle (1778 – 1841) à l'emplacement actuel du parc des Bastions, dans un climat intellectuel propice au développement des sciences et à une époque où plusieurs naturalistes possèdent déjà leurs propres collections de botanique privées (Sigrist et Bungener, 2008). Dans la

conception de son fondateur, il doit alors remplir trois fonctions : l'enseignement, la recherche et l'acclimatation (Sigrist et Bungener, 2008).

Dès 1819, le jardin est ouvert au public et en 1824, un conservatoire botanique est ajouté à l'ensemble, dans le but d'accueillir notamment l'herbier de Candolle et ceux donnés par des botanistes locaux (Sigrist et Bungener, 2008). À l'étroit dans le centre-ville, le Jardin des plantes déménage à son emplacement actuel en 1904, lorsque la Console est construite[1].

Depuis leur déménagement, les CJBG ont connu des agrandissements successifs en 1954 et en 1977, lors de l'acquisition de domaines voisins. Aujourd'hui, cette institution abrite l'un des plus grands herbiers au monde, comptant à peu près six millions de spécimens. Elle se donne comme axes stratégiques de documenter et d'étudier la biodiversité, de conserver, d'enrichir et de mettre à disposition les collections, et de diffuser et de vulgariser les connaissances scientifiques[2].

2. Contexte et concepts théoriques

2.1 La couverture du sol

Le concept central de ce travail est celui de couverture du sol. Ce terme regroupe l'ensemble des éléments naturels ou anthropiques qui recouvrent la surface du globe, comme les constructions, la végétation, le sol nu ou l'eau. Il ne doit pas être confondu avec l'utilisation du sol, qui désigne l'usage que les sociétés humaines font des différents types de couverture du sol, par exemple l'agriculture, l'habitat, la production industrielle ou la détente. Toute carte de la couverture du sol repose sur une nomenclature définissant un certain nombre de classes, comme le Land Cover Classification System (LCCS) de la FAO et le CORINE Land Cover (CLC) de l'AEE.

Cartographier la couverture du sol est un atout essentiel pour gérer efficacement les ressources naturelles (Wulder et al., 2018 ; Santos et al., 2020). En outre, les changements de couverture du sol ont un rôle à jouer dans la perte de biodiversité ou les changements climatiques (Vali et al., 2020 ; Olofsson et al., 2012) et les surveiller doit donc nous permettre de prendre des décisions informées (Giuliani et al., 2022).

En Suisse, les cantons cartographient la couverture du sol dans le cadre de la mensuration officielle, régie par l'Ordonnance fédérale sur la mensuration officielle (OMO). Ils procèdent alors par numérisation, relevés de terrain et géotraitements semi-automatiques. En parallèle, le centre de compétence pour la géoinformation et le traitement numérique des images digitales (GEOSTAT), qui fait partie de l'Office fédéral de la statistique (OFS), réalise environ tous les douze ans une carte de la couverture du sol dans le cadre de sa Statistique de la superficie (AREA). Cette dernière se fonde sur l'interprétation de photographies aériennes, prises par l'Office fédéral de la topographie (swisstopo) et qui couvrent l'intégralité du territoire suisse. Quatre relevés ont été effectués depuis 1979, le plus récent ayant été publié en 2018.

[1] Jardin botanique de Genève : 200 ans d'histoire. Conservatoire et Jardin botaniques de Genève. https://www.cjbg.ch/fr/propos/historique

[2] Nos axes stratégiques. Conservatoire et Jardin botanique de Genève. https://www.cjbg.ch/fr/science/nos-axes-strategiques

Pour effectuer la mesure, l'ensemble du territoire suisse est quadrillé par des lignes verticales et horizontales distantes de 100 mètres et qui constituent une grille d'un peu plus de 4,1 millions de points d'intersection. Ces derniers sont superposés aux photographies aériennes et servent d'échantillons pour leur interprétation, les mêmes points étant gardés d'un relevé à un autre pour améliorer la comparaison historique.

Chaque point se voit alors attribuer une catégorie de couverture (parmi 27 catégories) et d'utilisation du sol (parmi 46 catégories). Puis, à l'aide d'une matrice, chaque combinaison possible de couverture et d'utilisation du sol est transposée dans l'une des 72 catégories de base d'une autre nomenclature, appelée « nomenclature standard » ou NOAS04. Ces catégories de base peuvent être agrégées en différents regroupements de 27, 17 et 4 catégories. AREA fournit une information couvrante, actualisée et historique de la couverture du sol, permettant ainsi de surveiller l'évolution du territoire national.

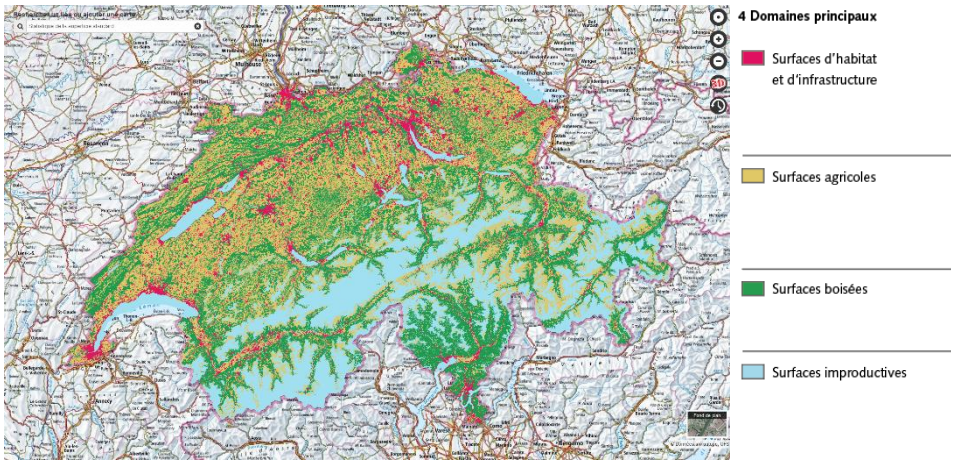


Figure 1 – Swisstopo. (2022). Carte de la statistique de la superficie en Suisse selon la nomenclature NOAS04 [Capture d'écran]. Map.geo.admin.ch. https://map.geo.admin.ch

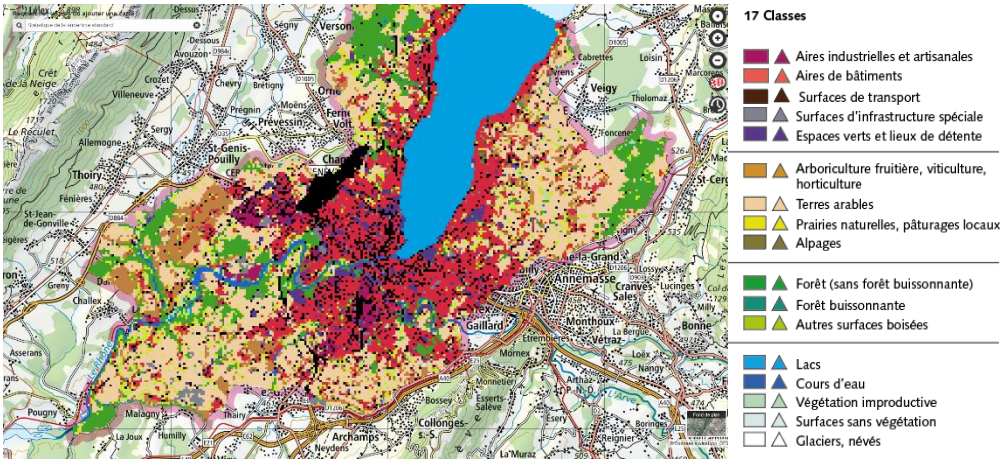


Figure 2 – Swisstopo. (2022). Carte de la statistique de la superficie dans le canton de Genève selon la nomenclature NOAS04 [Capture d'écran]. Map.geo.admin.ch. https://map.geo.admin.ch

La cartographie de la couverture du sol effectuée dans le cadre de la mensuration officielle ne couvre pas la partie française du Grand Genève. De plus, les données ne sont pas gratuites pour tous les cantons. Quant à AREA, la résolution des cartes auxquelles elle aboutit est assez faible (100 * 100 m) et l'intervalle temporel entre deux relevés est important (environ tous les 12 ans). Ainsi, il s'agira d'explorer une méthode susceptible de combler ces lacunes, mais il convient avant cela d'exposer la façon dont sont réalisées les cartes de couverture du sol.

2.2 La classification d'images satellites

Différentes techniques ont été développées pour cartographier la couverture du sol, comme les enquêtes de terrain ou l'interprétation de photos aériennes (Talukdar et al., 2020 ; Karpatne et al., 2022). Ces méthodes sont couteuses et prennent du temps à implémenter (Talukdar et al., 2020). Plus économe en temps et en argent, la classification d'images satellites désigne le regroupement de pixels ayant des valeurs semblables dans des classes thématiques. Dans ce cas, un pixel est vu comme une unité individuelle, à laquelle on associe les valeurs de plusieurs bandes (Jawak et al., 2015). Aujourd'hui, les satellites destinés à l'observation de la Terre sont essentiels pour comprendre l'étendue et les impacts des changements dans la couverture du sol (Camara et al., 2016). Malgré le fait qu'ils soient sensibles à la couverture nuageuse (Xue et al., 2017), ils ont notamment pour avantages d'offrir une large couverture spatiale et de fournir une information actualisée à intervalles réguliers (Chuvieco, 2020).

Traditionnellement, la classification d'une image s'effectue sur la base de la réflectance de chaque pixel, mais il existe également des méthodes que l'on dit orientées objet, où des pixels aux propriétés homogènes sont préalablement regroupés en objets à travers un processus de segmentation, pour ensuite être classés. Dans l'ensemble, la littérature montre que la classification orientée objet présente des avantages importants sur la classification par pixel et ce, surtout avec des images à haute résolution, où la variation propre à chaque classe est plus importante (Jawak et al., 2015). De fait, la classification orientée objet permet d'éviter l'effet « poivre et sel » souvent rencontré dans la classification par pixel, ce qui s'explique par le fait que contrairement à cette dernière, qui utilise uniquement l'information spectrale de l'image, la première utilise également la texture des objets, leur forme et leurs relations avec les régions adjacentes (Jawak et al., 2015). Toutefois, lorsqu'on utilise des images à moyenne résolution, la classification par pixel offre des résultats satisfaisants (Jawak et al., 2015).

Depuis ses débuts, la télédétection a été utilisée pour produire des cartes de la couverture du sol (Pasquarella et al., 2016). Les premiers essais en la matière datent du lancement du premier satellite Landsat de la NASA, en 1972 (Vali et al., 2020), mais les cartes produites étaient limitées à une certaine étendue spatiale (Wulder et al., 2018). Depuis, de nombreuses données satellitaires ont été mises en accès libre, comme c'est le cas des images Landsat ou Sentinel. Cette diffusion des données a offert à la télédétection un regain d'intérêt, puisqu'elle a permis aux chercheurs d'appliquer des méthodes d'apprentissage automatique à la classification (Karpatne et al., 2022 ; Talukdar et al., 2020).

2.3 L'apprentissage automatique appliqué à la classification

La télédétection possède des caractéristiques qui compliquent l'utilisation des algorithmes d'apprentissage automatique, comme la rareté des changements de couverture du sol par rapport à l'étendue de l'image, la rareté des données d'entraînement ou le fait que les données soient hétérogènes dans l'espace et le temps, et souvent dispersées à des échelles différentes ou à travers

plusieurs sources (Karpatne et al, 2022 ; Vali et al., 2021). Malgré cela, l'apprentissage automatique est considéré comme le meilleur outil pour classer des images satellites et pour détecter des changements de couverture du sol (Santos et al., 2021).

On peut séparer les algorithmes en deux principaux types d'apprentissage : le supervisé et le non-supervisé (Talukdar et al., 2020). Le premier repose sur l'entraînement d'un échantillon de données préalablement étiquetées pour classer des données qui n'ont pas de classe attribuée, tandis que le second désigne l'identification de groupes sans entraînement préalable (Jawak et al., 2015).

En général, les algorithmes de classification supervisée fournissent de meilleurs résultats que les algorithmes de classification non-supervisée (Talukdar et al., 2020 ; Pelletier et al., 2016). Toutefois, la précision d'une classification supervisée dépend fortement de la taille et de la qualité de l'échantillon d'entraînement, de sorte qu'un jeu de données précises et nombreuses est préférable, peu importe l'algorithme utilisé (Maxwell et al., 2018 ; Santos et al., 2021 ; Simoes et al., 2021). En particulier, de nombreux travaux ont souligné l'importance d'utiliser des échantillons de bonne qualité et d'effectuer des contrôles avant de réaliser une classification (Santos et al., 2021). De fait, la variabilité inhérente des signatures spectrales au sein d'une même classe peut causer du bruit et amener des échantillons très différents à être classés dans une même catégorie, surtout lorsque les données couvrent une grande aire géographique et une grande période temporelle (Santos et al., 2021). En outre, la performance d'un algorithme donné peut être affectée par le déséquilibre entre les classes et la classification qui en résulte peut sous-estimer les classes les moins abondantes (Maxwell et al., 2018). Il faut également prendre en compte que plus le nombre de classes augmente, plus la classification est difficile à produire (Wulder et al., 2018).

Phase essentielle de l'apprentissage supervisé, la collecte des données d'entraînement est une tâche complexe et fastidieuse (Vali et al., 2020 ; Wulder et al., 2018), qui peut être réalisée de différentes manières : sur le terrain, en interprétant des images de haute résolution, en utilisant des données qui ne sont pas issues de la télédétection ou en utilisant d'autres produits de classification (Wulder et al., 2018). Lorsque l'on bénéficie déjà de données d'entraînement, la classification supervisée comprend généralement trois grandes étapes : l'entraînement du modèle, la classification à proprement parler et l'évaluation de la précision (Jawak et al., 2015). La dernière étape est essentielle si l'on veut donner une appréciation quantitative de la classification. De fait, des estimations de précision qui soient fondées statistiquement et transparentes sont essentielles pour assurer l'intégrité de son travail et effectuer des comparaisons entre différentes méthodes, surtout dans un contexte où la classification d'images satellites est aussi accessible qu'aujourd'hui (Wulder et al., 2018). En particulier, ces estimations doivent être faites à l'aide de données indépendantes de celles utilisées pour entraîner le modèle d'apprentissage (Wulder et al., 2018). À cet égard, il convient de distinguer la validation des données d'entraînement et l'estimation de la précision d'une classification.

2.4 Validation croisée des données d'entraînement et estimation de la précision d'une classification

Un modèle d'apprentissage automatique possède sa propre erreur de prédiction, qui désigne la probabilité pour les données non étiquetées d'être mal classées (Rodriguez et al., 2010). Cette erreur est en général inconnue et doit donc être estimée à partir des données d'entraînement, raison pour laquelle on parle d'erreur de prédiction estimée (Rodriguez et al., 2010). L'estimation peut se faire de plusieurs façons. Une technique connue est la validation croisée à *k*-blocs, qui consiste à partitionner le jeu de données aléatoirement en *k* blocs de tailles semblables que l'on emploie ensuite pour valider le modèle. L'un après l'autre, chaque bloc est utilisé pour tester un modèle généré par les $k - 1$ blocs

restants (Wong et al., 2020). On estime alors l'erreur de prédiction en effectuant la moyenne des erreurs commises pour chaque classificateur généré. Puisque l'erreur estimée peut varier en fonction de l'échantillon utilisé, plusieurs études suggèrent de répéter le processus de validation plusieurs fois, notamment pour éviter le risque de surapprentissage (Wong et al., 2020). Ainsi, on utilise plusieurs ensembles de validation d'un même jeu de données et on calcule le biais et la variance de la performance de validation du modèle.

La validation croisée étant une estimation de l'erreur de prédiction d'un modèle d'apprentissage automatique, elle n'utilise que les données d'entraînement et ne correspond donc pas à une évaluation de la précision d'une classification (Simoes et al., 2021). Pour mesurer la précision, il existe différents indicateurs que l'on peut calculer en s'aidant d'une matrice de confusion : la précision globale, la précision du producteur et la prédiction de l'utilisateur.

Tableau 1 – Matrice de confusion montrant des valeurs fictives en exemple.

		Données de référence (réalité sur le terrain)							
		Surfaces non naturelles	Végétation herbacée	Végétation buissonnante	Végétation d'arbres	Surfaces sans végétation	Plans d'eau et surfaces humides	Total	Précision de l'utilisateur
Données classées	Surfaces non naturelles	11	0	0	0	2	0	13	0,85
	Végétation herbacée	1	18	1	2	1	0	23	0,78
	Végétation buissonnante	2	3	6	3	0	0	14	0,43
	Végétation d'arbres	0	2	1	19	0	0	22	0,86
	Surfaces sans végétation	4	2	1	3	5	2	17	0,29
	Plans d'eau et surfaces humides	0	0	0	0	2	9	11	0,82
	Total	18	25	9	27	10	11	100	
	Précision du producteur	0,61	0,72	0,67	0,70	0,50	0,82	Précision globale	**0,68**

La précision globale indique la proportion de pixels correctement classés sur l'ensemble de la carte. Elle se calcule en divisant la quantité de pixels correctement classés par la quantité totale de pixels. Elle fournit une information de base, mais limitée.

La précision de l'utilisateur désigne, pour une classe donnée, la proportion de pixels qui appartiennent réellement à la classe à laquelle on les a attribués (Olofsson et al., 2012). Pour une ligne donnée du tableau, elle se calcule en divisant le nombre de pixels correctement classés par le nombre total de pixels. Par exemple, la précision de l'utilisateur de la classe « Végétation d'arbres » de notre tableau est de 0.86, ce qui signifie que les points ont 86 % de chances d'appartenir à la classe à laquelle on les a attribués. La précision de l'utilisateur est complémentaire de l'erreur de commission, qui désigne la proportion de points qui, pour une classe donnée d'une classification, n'ont pas été correctement classés. L'erreur de commission se calcule en divisant le nombre de valeurs omises par le nombre total des valeurs dans la ligne.

La précision du producteur désigne la proportion de pixels correspondant à une catégorie de couverture du sol qui ont effectivement été classé dans cette catégorie (Olofsson et al., 2012). Pour une colonne du tableau, elle se calcule en divisant le nombre de pixels correctement classés par le nombre total de pixels. Par exemple, dans notre tableau, la précision du producteur de la classe « Surfaces non naturelles » est de 0.61, ce qui signifie que 61 % des points de la classe ont été correctement classés. La précision du producteur est complémentaire de l'erreur d'omission, qui désigne la proportion des points qui, pour une classe donnée de la réalité de terrain, n'ont pas été correctement classés. L'erreur d'omission se calcule en divisant le nombre de valeurs omises par le nombre total de valeurs dans la colonne. Dans un article publié en 2012, Olofsson et al décrivent une manière différente de calculer la précision du producteur, en utilisant des valeurs estimées et ajustées plutôt qu'un décompte de pixels (Olofsson et al., 2012).

La précision du producteur et la précision de l'utilisateur peuvent être analysées de façon complémentaire. Par exemple, pour la classe « Végétation buissonnante », alors que la précision du producteur est de 0.67, la précision de l'utilisateur est de 0.43. Concrètement, cela signifie que si la 67 % des données de référence ont correctement été identifiés, seuls 43 % des pixels de la classe sont effectivement de la végétation buissonnante.

2.5 Le choix de l'algorithme : les forêts d'arbres décisionnels

La sélection d'un algorithme est compliquée, notamment parce que la littérature en la matière est contradictoire (Maxwell el al, 2018). Malgré cela, certains algorithmes sont plus conseillés que d'autres, parce qu'ils ont été testés par le passé et qu'ils ont montré de bons résultats (Maxwell et al., 2018). Parmi eux, les forêts d'arbres décisionnels, ou *Random forests* en anglais, sont une catégorie de classificateurs dits « ensemblistes », développée par le statisticien états-unien Leo Breiman (Talukdar et al., 2020). En classification ensembliste, plusieurs classificateurs sont entraînés et leurs résultats combinés à travers un vote (Gislason et al., 2006 ; Kulkarni et Lowe, 2016). Dans les cas des forêts d'arbres décisionnels, les classificateurs combinés sont des arbres de décision. Un arbre décisionnel est une division successive des données d'entrée en de multiples embranchements, qui représentent chacun des chemins de décision pour classer les données (Maxwell et al., 2018). L'inconvénient d'un tel algorithme étant le surajustement aux données d'entraînement, combiner plusieurs arbres permet de surmonter les faiblesses d'un classificateur unique (Maxwell el al, 2018).

Dans une forêt d'arbres décisionnels, chaque arbre est construit par un double processus d'échantillonnage aléatoire, à la fois sur les données d'entraînement, qui sont sélectionnées selon un tirage avec remise, et sur les attributs, qui définissent les nœuds intermédiaire de décision. Une fois tous les arbres créés et le modèle entraîné, les données à classer sont passées à travers chaque arbre de décision, qui fournit alors une prédiction de classification. Puis, un simple vote majoritaire est réalisé sur l'ensemble des prédictions pour un pixel donné. Les forêts d'arbres décisionnels sont considérées comme faciles à optimiser puisqu'elles nécessitent de définir principalement deux paramètres : le nombre d'arbres décisionnels et le nombre de variables aléatoires disponibles pour chaque nœud (Maxwell et al., 2018).

Les forêts d'arbres décisionnels sont très précises par rapport à d'autres algorithmes et peuvent être utilisées sur des jeux de données de taille importante (Kulkarni et Lowe, 2016 ; Talukdar et al., 2020). De fait, une étude comparative menée en 2015 sur 30 classifications montre que les forêts d'arbres décisionnels possèdent la précision moyenne de classification la plus élevée (73.19 %), même si elles n'ont été le modèle le plus précis que pour 18 cas sur 30 (Maxwell et al., 2018). La précision de la classification augmente avec le nombre de prédicteurs jusqu'à atteindre un certain point où elle finit

par stagner (Kulkarni et Lowe, 2016). Le nombre d'arbres optimal est donc à déterminer en fonction du modèle, mais il est généralement recommandé d'en utiliser un grand nombre (Maxwell et al., 2018). Bien que certains auteurs préconisent d'en utiliser 500 (Maxwell et al., 2018), d'autres soulignent que la précision de la classification augmente de façon marginale avec le nombre d'arbres et qu'on peut en utiliser 100 sans perdre beaucoup d'information (Pelletier et al., 2016).

2.6 Deux approches pour surveiller l'évolution de la couverture du sol : *space-first, time-later* vs. *time-first, space-later*

L'approche traditionnelle pour rendre compte de l'évolution de la couverture du sol consiste à comparer deux ou plusieurs images classées d'un même endroit prises à des moments différents. Les dates choisies pour la comparaison correspondent en général à des jours de faible couverture nuageuse et lors desquels il est possible de différencier les signatures spectrales des catégories de végétation (Pelletier et al., 2016). D'après Camara et al. (2016), cette approche peut être désignée par le nom de *space-first, time-later*, puisque la dimension temporelle vient dans un second temps. Un inconvénient de cette première approche est que des images à faible couverture nuageuse ne sont pas forcément disponibles aux dates choisies et qu'une image unique ne permet pas toujours de distinguer les différentes catégories de couverture du sol (Pelletier et al., 2016).

Une approche alternative, que Simoes et al désignent sous l'appellation *time-first space-later*, consiste à mettre la dimension temporelle au premier plan, en associant à chaque pixel un ensemble de séries temporelles qui correspondent chacune à l'ensemble des valeurs de réflectance sur une période donnée et pour une bande ou un indice spécifique (Simoes et al., 2021). Dans un second temps, la dimension spatiale est utilisée pour établir des relations de voisinage entre pixels. Avec cette seconde approche, un pixel est relié à ses voisins temporels plutôt qu'à ses voisins spatiaux (Camara et al., 2016). Selon différents auteurs, l'approche *time-first, space-later* est plus efficace pour identifier les changements de couverture du sol (Simoes et al., 2021).

Les séries temporelles d'images satellites désignent des collections calibrées d'images satellites d'une même scène prises à différents moments (Simoes et al., 2021). Les variations temporelles de réflectance que montrent ces séries temporelles peuvent être utilisées pour mieux caractériser les différents types de couverture du sol et pour révéler des processus complexes qu'il serait difficile de montrer en utilisant des approches bi-temporelles ou annuelles (Pasquarella et al., 2016 ; Santos et al., 2021). De plus, les séries temporelles permettent une approche continue dans l'étude des changements en matière de couverture du sol (Pasquarella et al., 2016) et sont, par conséquent, de plus en plus utilisées pour classer des images satellites et détecter les changements de couverture du sol (Santos et al., 2021).

Aujourd'hui, l'accessibilité des données Landsat et Sentinel rend possible l'utilisation de séries temporelles, qui peuvent être extraites à partir de cubes de données. De façon générale, un cube de données d'observation de la Terre est une structure multidimensionnelle qui contient une collection d'images satellites (Santos et al., 2021 ; Simoes et al., 2021), les dimensions incluses étant notamment l'espace et le temps. Plus spécifiquement, un cube de données est une représentation où chaque point (x, y, z) possède un vecteur d'attributs correspondant aux valeurs du point dans chaque bande et indice spectral retenus. Une telle structure représente des volumes importants de données, raison pour laquelle certains auteurs parlent de mégadonnées d'observation de la Terre (Camara et al., 2016).

Dans un article publié en 2021, Simoes et al définissent plus précisément un cube de données d'observation de la Terre comme une structure répondant à des critères spécifiques, le plus important

étant la nécessité pour les tuiles d'être parfaitement alignées les unes aux autres (Simoes et al., 2021). Les structures de données qui suivent cette définition se prêtent à l'utilisation d'algorithmes d'apprentissage automatique, ce qui n'est pas le cas des collections d'images disponibles dans les services cloud, qui doivent donc être préalablement traitées (Simoes et al., 2021). De fait, pour comparer des images d'un même endroit prises à différents moments, il faut que les mesures soient calibrées (Camara et al., 2016).

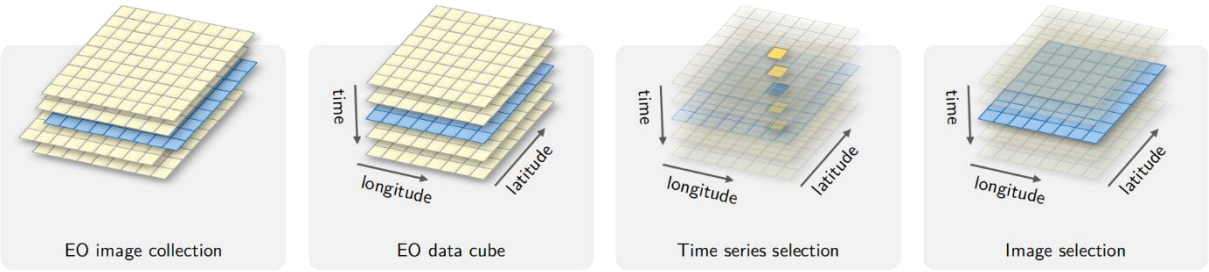


Figure 3 – Simoes et al. (2021). Conceptual view of data cubes [Graphique]. Github. https://e-sensing.github.io/sitsbook/earth-observation-data-cubes.html

En suivant cette définition, Simoes et al. (2021) ont conçu *sits*, un paquet R destiné à l'analyse de séries temporelles d'images satellites, qui recourt à l'apprentissage automatique et adopte l'approche *time-first, space-later*. Les environnements liés à R sont considérés comme d'excellentes plateformes pour implémenter et expérimenter des algorithmes d'apprentissage automatique (Maxwell et al., 2018). *sits* a été conçu par ses auteurs pour englober toutes les étapes clefs de la classification de mégadonnées d'observation de la Terre, allant de la récolte des données sur un serveur à la classification à proprement parler. De fait, chaque étape du processus est incarnée par une fonction spécifique possédant ses propres paramètres, ce qui facilite le travail de classification.

Les principales étapes comprises dans le déroulement des opérations de *sits* sont les suivantes (Simoes et al., 2021) :

1. Sélection d'une collection d'images prêtes à l'analyse sur un fournisseur cloud ;
2. Construction d'un cube de données régulier à partir de la collection d'images ;
3. Création de nouvelles bandes et d'indices spectraux pour compléter la collection d'images ;
4. Extraction de séries temporelles à partir du cube de données, en utilisant des points de mesure dont on connaît la classe ;
5. Contrôle de la qualité des données et filtrage des données aberrantes ;
6. Entraînement d'un modèle d'apprentissage automatique par les séries temporelles ;
7. Utilisation du modèle d'apprentissage automatique pour classer le cube de données et obtenir des probabilités d'appartenance pour chaque classe ;
8. Traitement des données obtenues et filtrage des données aberrantes ;
9. Production d'une carte labellisée ;
10. Evaluation de la précision de la classification réalisée et validation croisée du modèle d'apprentissage.

Pour aider le modèle à apprendre des données et à établir des règles des décisions, on peut utiliser différents attributs associés aux données, tels que leur information spectrale, spatiale ou temporelle (Pelletier et al., 2016). En particulier, les attributs spectraux, comme les indices, aident à séparer les différentes catégories de couverture du sol (Pelletier et al., 2016).

2.7 Les indices spectraux

Les indices spectraux aident à distinguer l'information contenue dans une image et sont à cet égard utiles pour cartographier la couverture du sol. Un indice est une transformation appliquée à plusieurs bandes, sélectionnées pour souligner certaines caractéristiques de la surface terrestre comme la végétation, l'eau ou l'urbain (Kaur et Pandey, 2021). La formule fondatrice de nombreux indices est la différence entre deux bandes, normalisée par leur somme. La somme au dénominateur étant forcément supérieure à la différence au numérateur, le résultat obtenu est compris entre -1 et 1. Cette formule est désignée ci-après comme le quotient normalisé.

$$I = \frac{B_x - B_y}{B_x + B_y}$$

Formule du quotient normalisé, où I désigne l'indice calculé, et B_x et B_y deux bandes spectrales différentes

La formulation d'un indice est fondée sur les signatures spectrales spécifiques aux différents types de couverture du sol, qui absorbent ou reflètent les rayons dans différentes longueurs d'onde (Kaur et Pandey, 2021). Il existe de nombreux indices différents, leur performance dépendant des conditions climatiques et topographiques de l'endroit auquel on les applique, mais aussi de la date d'acquisition des images (Kaur et Pandey, 2021). Ainsi, un indice développé pour une certaine région peut ne pas convenir à d'autres et il est préférable de vérifier son application avant de l'utiliser (Kaur et Pandey, 2021).

Parmi les indices les plus couramment utilisés figurent ceux qui soulignent la végétation, les étendues d'eau et le bâti. En ce qui concerne le végétal, les principales bandes utilisées sont les rayons UV, le visible, le proche infrarouge et l'infrarouge moyen (Xue et al., 2017). Comme l'explique Timothy J. Arkebauer, les feuilles absorbent en grande partie les rayonnements ultraviolet, visible et infrarouge thermique, tandis qu'elles reflètent ou transmettent le rayonnement dans le proche infrarouge. La haute absorption dans le spectre visible est due à la présence de pigments tels que la chlorophylle, qui absorbe plus fortement les rayons bleus et rouges, et qui donne donc aux feuilles leur couleur caractéristique (Arkebauer, 2005). À l'arrivée de l'automne, la concentration en chlorophylles diminue et ce sont d'autres pigments qui donnent aux feuilles leurs couleurs chaudes.

Plus spécifiquement, étant donnée la différence de réflectance entre la bande rouge et la bande du proche infrarouge chez les feuilles, les indices de végétation sont souvent une combinaison entre ces deux bandes. L'indice de végétation par différence normalisée, ou NDVI (*Normalized Difference Vegetation Index*), est le quotient normalisé entre le rouge et le proche infrarouge. Un pixel est considéré comme de la végétation si sa valeur est proche de 1, comme du sol nu si elle est proche de 0 et comme de l'eau si elle est proche de -1. Proposé en 1974, le NDVI est l'un des indices les plus utilisés pour caractériser la croissance et la vigueur de la canopée, mais il a comme inconvénients d'être sensible à la luminosité et à la couleur des sols environnants, à l'ombre de la canopée, à l'atmosphère et à la couverture nuageuse (Xue et al., 2017). Développé en 1995, l'indice de végétation amélioré, ou EVI (*Enhanced Vegetation Index*), corrige certains défauts du NDVI. Notamment, il est plus sensible aux variations de canopée (Somvanshi et Kumari, 2020).

L'indice de l'eau par différence normalisée (NDWI) désigne la différence normalisée entre la bande verte et le proche-infrarouge. Ce choix de bande s'explique par la différence de réflectance entre l'eau et les autres éléments de couverture du sol dans le visible et le proche-infrarouge. De fait, l'eau absorbe fortement les rayons du proche-infrarouge, tandis que la végétation terrestre et les sols nus la reflètent (McFeeters, 1996).

La classification des surfaces bâties est quant à elle une tâche difficile. En particulier, la plupart des indices confondent les surfaces bâties et les sols nus, ces deux catégories ayant des réponses spectrales semblables (Kaur et Pandey, 2021). Cette confusion n'échappe pas à l'indice du bâti par différence normalisée, ou NDBI (*Normalized Difference Built-Up Index*), qui reste malgré cela l'indice le plus utilisé pour l'urbain (Kaur et Pandey, 2021).

Si les attributs spectraux comme les indices aident à séparer les classes, il est difficile de juger leur contribution puisqu'ils augmentent également le temps de calcul et que l'information contenue dans les séries temporelles peut suffire à caractériser les différentes catégories de couverture du sol (Pelletier et al., 2016).

3. Problématique

Le but de ce travail est d'explorer une méthode qui facilite la cartographie de la couverture du sol d'une année à l'autre, en recourant à la classification d'images satellites par apprentissage automatique. L'acquisition des données d'entraînement étant une étape fastidieuse, il serait intéressant de proposer une technique qui ne nécessite pas d'effectuer de nouveaux prélèvements à chaque fois que l'on cherche à cartographier la couverture du sol. Deux cubes de données ont été constitués, le premier plus proche temporellement des données d'entraînement que le second, dans le but d'évaluer si la précision diminue plus on s'éloigne de la création des données d'entraînement. Si tel n'est pas le cas, alors on pourra éventuellement utiliser les mêmes données d'entraînement pour plusieurs années.

Plusieurs outils statistiques seront utilisés dans le but de valider les modèles d'apprentissage automatique utilisés et d'évaluer la précision des classifications réalisées. Les matrices de confusion seront employées dans la validation croisée comme dans l'évaluation de la précision. Le coefficient kappa sera utilisé pour évaluer l'accord entre les classes réelles et les classes estimées lors de la validation croisée. En suivant les recommandations d'Olofsson et al. (2012), le coefficient kappa ne sera pas utilisé pour estimer la précision de la classification. Pour évaluer cette dernière, la précision globale, la précision du producteur et la précision de l'utilisateur seront utilisées. Dans un dernier temps, l'évolution de la couverture du sol entre 2018 et 2022 sera étudiée à l'aide de statistiques de surface.

4. Méthode(s) et données

4.1 Provenance et description des données, zone d'intérêt

Les images analysées dans le cadre de ce travail proviennent des capteurs multi-spectraux (MSI) embarqués à bord des deux satellites Sentinel-2, dont les orbites sont en phase à 180° l'un par rapport à l'autre. Lancés en 2015 et en 2017, ces derniers font partie du programme Copernicus, coordonné et géré par la Commission européenne et l'Agence spatiale européenne (ESA). Leurs capteurs couvrent treize bandes spectrales allant de la région visible du spectre électromagnétique à l'infrarouge courte longueur d'onde. Les satellites Sentinel-2 ont plusieurs intérêts : un temps d'orbite court (5 jours lorsque les deux capteurs sont considérés), une haute résolution spatiale et l'existence de plusieurs bandes dans le proche-infrarouge, facteurs particulièrement utiles pour cartographier la couverture du sol (Phiri et al., 2020 ; Wulder et al., 2018). Néanmoins, ils ne captent pas les rayons dans l'infrarouge thermique, ce qui rend notamment impossible le calcul de certains indices initialement développés pour les satellites Landsat.

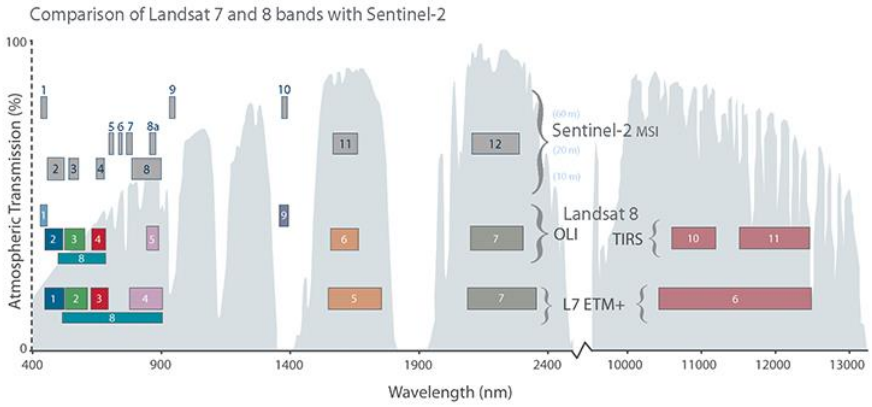


Figure 4 – USGS. (2015). Comparison of Landsat 7 and 8 bands with Sentinel-2. USGS. https://www.usgs.gov

Tableau 2 – Bandes spectrales des satellites Sentinel-2

B01	60	Bleu (côtes et aérosols)
B02	10	Bleu
B03	10	Vert
B04	10	Rouge
B05	20	Proche-infrarouge
B06	20	Proche-infrarouge
B07	20	Proche-infrarouge
B08	10	Proche-infrarouge
B8A	20	Proche-infrarouge
B09	60	Infrarouge courte longueur d'onde
B10	60	Infrarouge courte longueur d'onde
B11	20	Infrarouge courte longueur d'onde
B12	20	Infrarouge courte longueur d'onde

Les données de Sentinel-2 sont accessibles aux utilisateurs à deux niveaux de traitement : la réflectance au sommet de l'atmosphère (L1C) et la réflectance au niveau du sol (L2A), toutes deux orthorectifiées. Le fait que les données aient déjà fait l'objet de prétraitements évite certaines étapes supplémentaires comme la conversion de la radiance en réflectance. Toutefois, par souci d'harmonisation avec de précédents jeux de données, les comptes numériques fournis par l'ESA aux niveaux L1C et L2A sont les valeurs de réflectance multipliées par 10'000[3]. Ainsi, lors de l'affichage dans un logiciel d'information géographique d'une bande Sentinel-2 à une date quelconque, les valeurs obtenues peuvent surprendre. De fait, la réflectance étant une proportion, sa valeur se situe normalement entre 0 et 1. Ainsi, pour afficher les vraies valeurs de réflectance, il est nécessaire de diviser les comptes numériques par 10 000 ou de les multiplier par 0.0001. Heureusement, le paquet *sits* affiche les valeurs ajustées de réflectance et par conséquent, il n'est pas nécessaire de les corriger une fois sur RStudio.

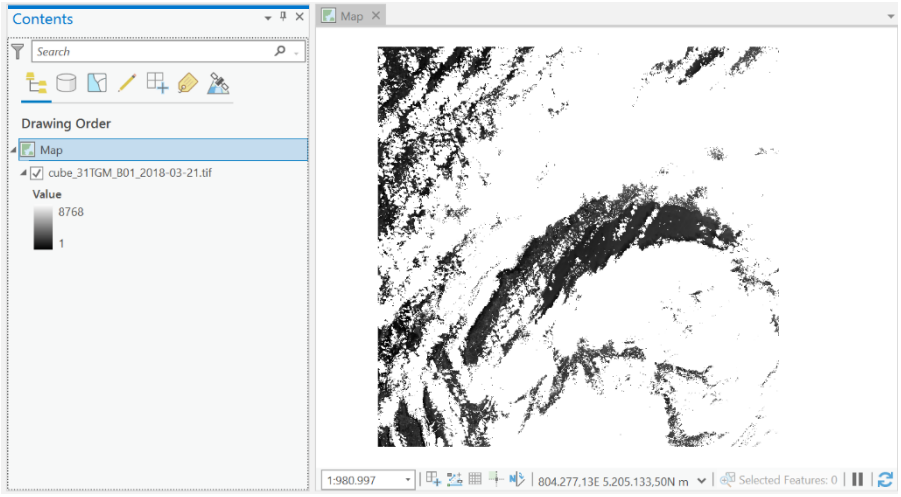


Figure 5 – Exemple d'une bande Sentinel-2 avec les valeurs brutes de réflectance au niveau L2A

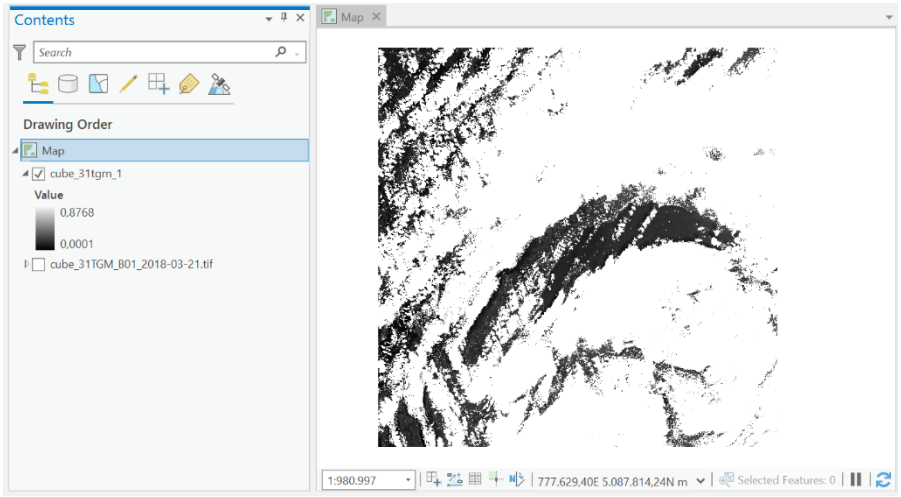


Figure 6 – Exemple d'une bande Sentinel-2 après conversion des valeurs de réflectance

[3] Page : https://docs.sentinel-hub.com/api/latest/data/sentinel-2-l2a/

Les données L2A, utilisées dans le cadre de ce travail, sont accessibles au format GeoTIFF auprès des services et produits de cloud Amazon (AWS). Si l'on télécharge directement les images sur la plateforme en accès libre de Copernicus, certaines d'entre elles présentent les valeurs de réflectance au niveau de l'atmosphère (niveau L1C), ce qui est le cas pour les images de l'année 2018. Pour obtenir la réflectance au niveau du sol, il faut alors transformer les données à l'aide d'un algorithme proposé par l'ESA, écrit en Python et exécutable sur l'interface en ligne de commande de Windows. À nouveau, ce problème n'existe pas avec le paquet *sits*, grâce auquel on peut accéder directement aux donnée L2A.

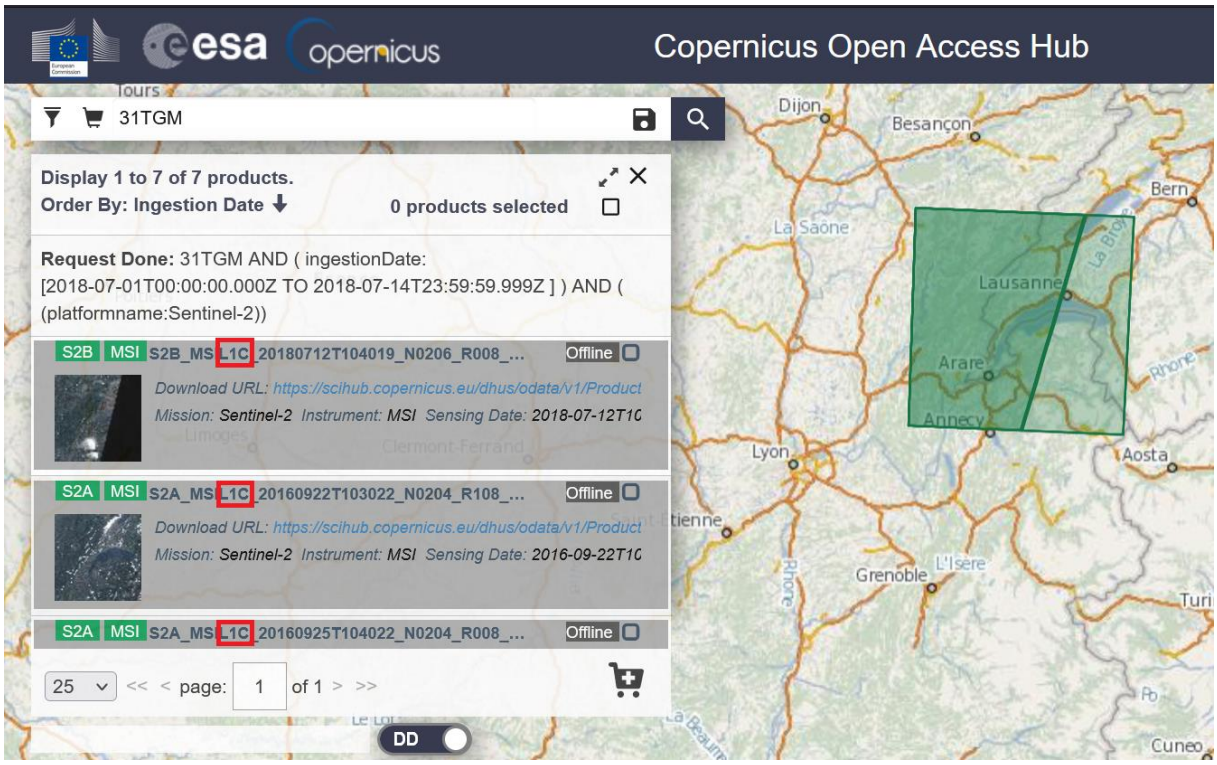


Figure 7 – Copernicus. (2022). Exemple des résultats d'une requête effectuée sur la plateforme Copernicus, affichant uniquement des images au niveau de traitement L1C [Capture d'écran]. Copernicus Open Access Hub. https://scihub.copernicus.eu/dhus/#/home

Le but de ce travail étant de comparer deux classifications réalisées sur deux cubes différents, deux principaux jeux de données ont été constitués : l'un pour l'année 2018, avec 58 dates, et l'autre pour l'année 2022, avec 61 dates. Le deuxième satellite de Sentinel-2 (Sentinel-2B) ayant été lancé en mars 2017, l'année 2018 a été choisie comme premier jeu de données parce qu'elle présente plus de dates de prise de vue que l'année 2017. Malgré cela, la série d'images collectées pour 2018 ne commence qu'à partir du mois de mars et par conséquent, elle contient moins de dates que la collection d'images pour 2022, qui se termine quant à elle à la fin octobre. Cette différence en termes d'intervalles temporels et de dates est importante à relever puisqu'elle rend impossible l'entraînement de deux cubes de données à partir d'un même modèle d'apprentissage automatique. De fait, pour que la classification aboutisse, le modèle d'apprentissage et le cube de données doivent avoir les mêmes intervalles temporels. La zone d'intérêt choisie correspond à la tuile 31 TGM du système de quadrillage Sentinel-2. Cette dernière couvre notamment l'étendue du Grand Genève et s'étale sur plusieurs cantons suisses et régions françaises.

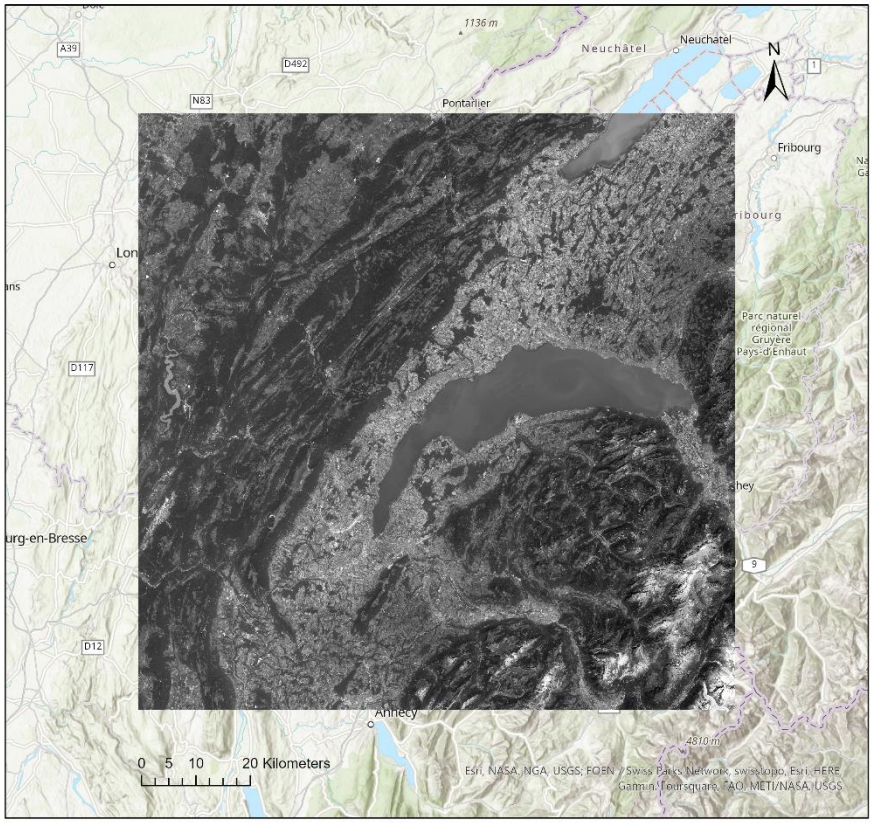


Figure 8 – Carte montrant l'étendue de la tuile 31 TGM du système de quadrillage de Sentinel-2, utilisée pour la classification

4.2 Méthode utilisée

La méthode employée comporte trois grandes étapes. Dans un premier temps, les données d'entraînement sont extraites sur FME, logiciel d'extraction et de transformation de données spatiales. Ce dernier propose un grand nombre de transformateurs faciles d'utilisation et possède une interface visuelle aisée à comprendre, qui offre notamment la possibilité de scinder l'espace de travail en marque-pages. Dans un second temps, les données d'entraînement sont utilisées pour entraîner un modèle de forêt d'arbres décisionnels, qui est ensuite employé pour classer une collection d'images satellites. Cette deuxième étape est réalisée en utilisant les fonctions dédiées du paquet *sits*, sur RStudio. Enfin, dans un troisième temps et toujours en utilisant les fonctions dédiées du paquet *sits*, une validation croisée à *k*-blocs est effectuée pour chaque modèle d'apprentissage et la précision de chaque classification est évaluée grâce à des matrices de confusion. L'ensemble des traitements ont été réalisés sur un serveur Dell Precision 7920 Rack de l'Université de Genève, avec Windows comme système d'exploitation. Ce dernier dispose de deux supports, accueillant chacun un multiprocesseur à 18 cœurs de 3.10 GHz et deux processeurs logiques chacun, totalisant 36 cœurs et 72 processeurs logiques, ainsi que de 256 GB de RAM. Etant donné le volume important de données à traiter, le parallélisme et la mémoire vive sont des facteurs essentiels dans la rapidité de traitement.

4.2.1 Extraction des données d'entraînement et de validation sur FME

Les données d'entraînement utilisées dans le cadre de ce travail sont les points d'échantillonnage de la Statistique suisse de la superficie (AREA), décrite dans la partie introductive de ce mémoire. La nomenclature retenue est celle relative à la couverture du sol, dénommée NOLC04 et qui propose différents niveaux d'agrégations, à 6 et à 27 catégories. Des essais peu concluants de classification ont été effectués avec les 27 catégories. De fait, certaines classes n'avaient pas été prédites dans les classifications résultantes, peut-être parce qu'elles présentaient des signatures spectrales trop différentes. Ainsi, une nomenclature intermédiaire à 10 catégories a été choisie à la place. Cette dernière est accessible sur le portail cartographique de la Confédération.

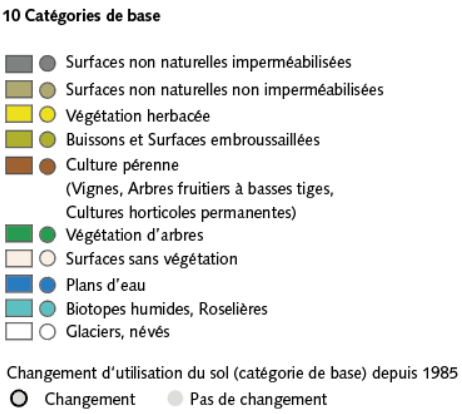


Figure 9 – Swisstopo. (2022). Les dix catégories de base de la nomenclature de couverture du sol NOLC04. [Image]. Map.geo.admin.ch. https://map.geo.admin.ch

Pour les cantons de Genève et de Vaud, les dernières photographies aériennes ont été prises entre 2012 et 2014. Plusieurs années se sont donc écoulées entre la création des échantillons et la classification des images satellites. Les points couverts par la zone d'intérêt ont été sélectionnés et exportés à l'aide du modèle FME ci-dessous.

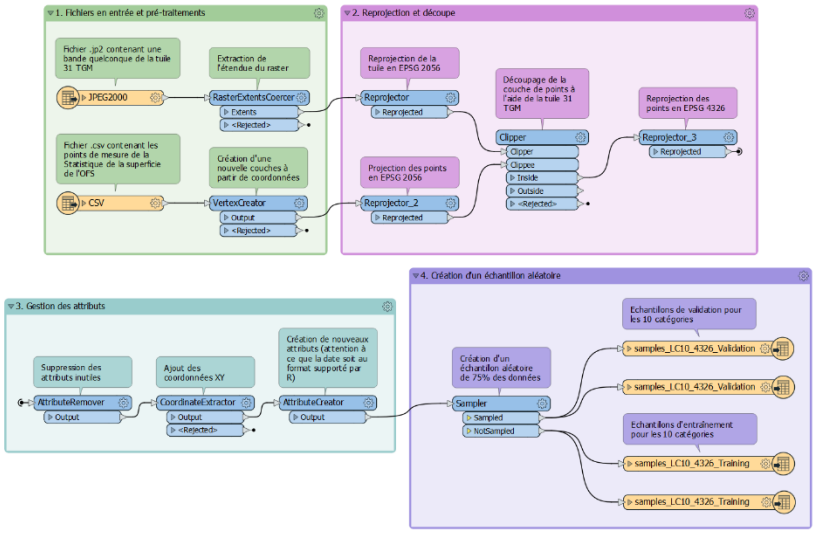


Figure 10 – Modèle FME utilisé pour extraire les données d'entraînement

Le modèle a été scindé en quatre parties, ou marque-pages. Le premier regroupe la lecture des deux fichiers de base et les prétraitements qui leurs sont appliqués. Les points AREA ont été téléchargés sur le site de l'OFS, sous la forme d'un fichier .csv, et constituent le premier fichier de base. Ce dernier contient les points d'échantillonnage sur l'ensemble de la Suisse (plus de 4,1 millions d'entrées), chacun d'entre eux pouvant être identifié grâce au couple de coordonnées (*x*, *y*) qui lui est associé dans le système de coordonnées suisse (MN95). Entre autres, les attributs pour chaque point sont la classe qui lui est attribuée pour chaque nomenclature de couverture du sol, d'utilisation du sol et de statistique de la superficie. À partir de ce fichier .csv, une couche de point est créée grâce au transformateur VertexCreator. Le second fichier de base est une bande de la tuile 31 TGM. Puisqu'il s'agit uniquement de récupérer l'étendue de la tuile, l'information spectrale n'est ici pas importante et on peut donc choisir n'importe quelle bande et n'importe quelle date. L'étendue du raster est extraite grâce au transformateur RasterExtentsCreator.

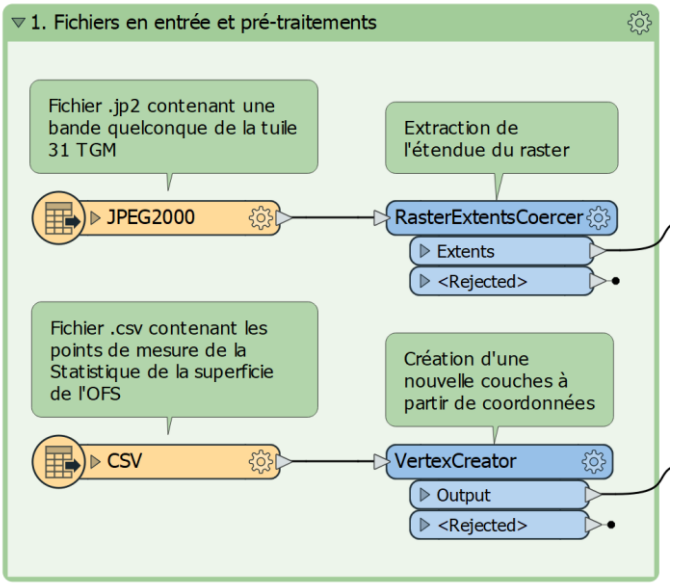


Figure 11 – Premier marque-page du modèle FME : lecture des fichiers de base et prétraitements

Le deuxième marque-page regroupe les transformateurs qui servent à projeter les données dans le système de coordonnées adéquat et à les découper selon l'étendue souhaitée. Puisque les coordonnées de longitude et de latitude des points AREA sont dans le système de coordonnées CH1903+ / MN95 (EPSG : 2056), ils doivent d'abord être projetés dans ce dernier à l'aide du transformateur Reprojector. La même projection est appliquée à la tuile 31 TGM, initialement référencée selon le système de coordonnées WGS84 (EPSG : 4326). Puis, la couche de points est découpée selon l'étendue de la tuile 31 TGM, pour ne retenir que les points contenus dans la zone d'intérêt. Ces derniers sont ensuite reprojetés en WGS84, qui correspond au système de coordonnées dans lequel les tuiles Sentinel-2 sont référencées. Ce dernier point est important puisque les coordonnées sont précisées en degrés décimaux, unique format reconnu par *sits* pour la lecture des points d'entraînement et de validation.

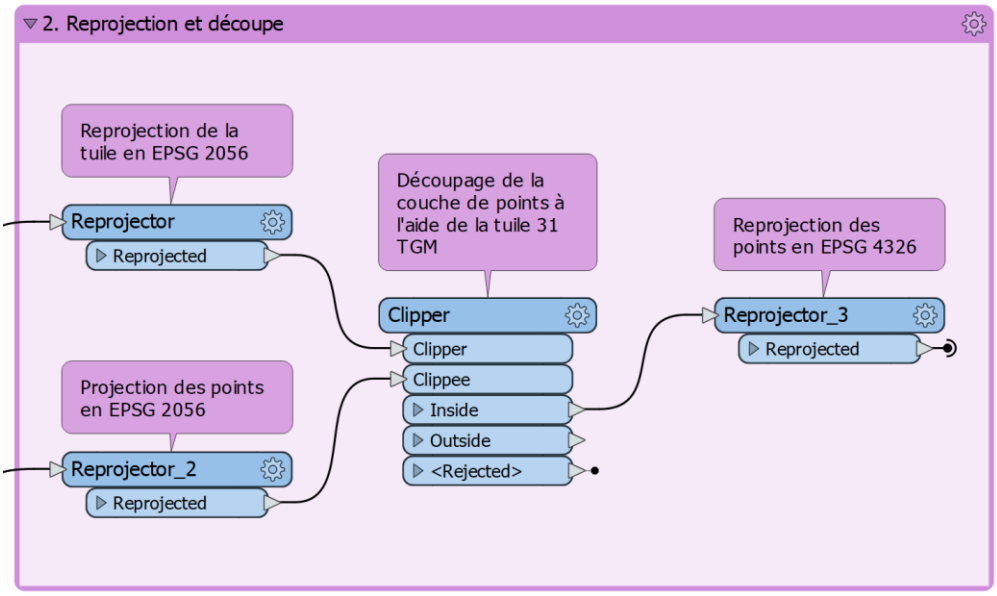


Figure 12 – Deuxième marque-page du modèle FME : reprojection des données et découpage en utilisant la tuile sentinel-2

Le troisième marque-page regroupe les transformateurs destinés à la gestion des attributs de la couche de points. Cette partie est nécessaire puisque lors de l'extraction des séries temporelles avec *sits*, les données fournies en entrée doivent ne posséder que cinq attributs : la longitude, la latitude, l'étiquette du point pour une nomenclature donnée et les dates de début et de fin pour lesquelles l'étiquette est considérée comme valide. Dans un premier temps, les attributs inutiles sont supprimés et seul celui qui contient le code dans la nomenclature souhaitée est gardé. Dans un second temps, les coordonnées de longitude et de latitude en EPSG 4326 sont rajoutées au tableau. Enfin, trois autres attributs sont ajoutés. Les deux premiers sont les dates susmentionnées, tandis que le dernier est un attribut de type textuel qui associe à chaque code NOLC04 une étiquette en français.

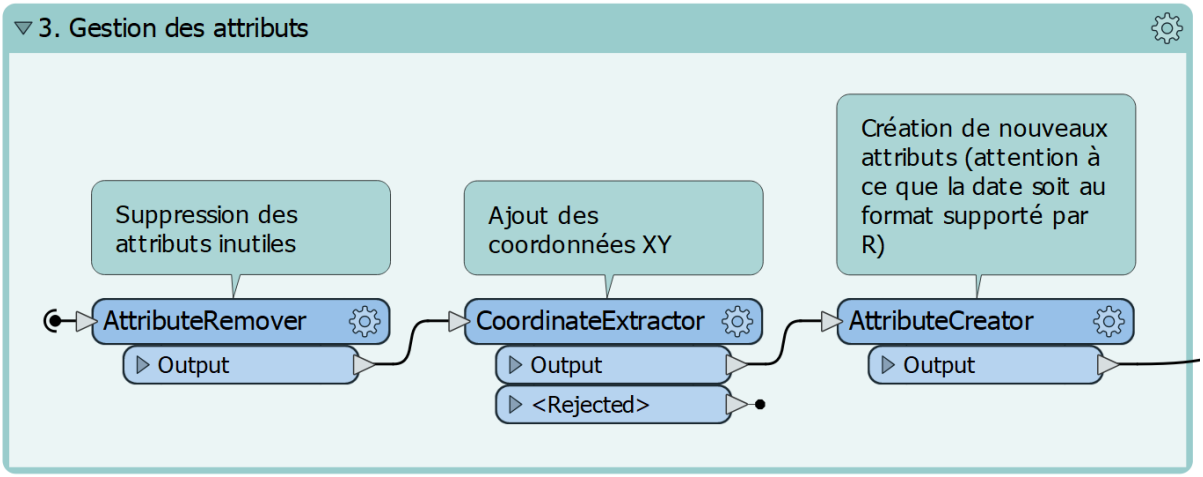


Figure 13 – Troisième marque-page du modèle FME : gestion des attributs

Le quatrième et dernier marque-page est destiné à créer deux échantillons aléatoires à partir de la couche de points. Le premier, qui contient 25% des points, sera utilisé pour estimer la précision de la classification une fois réalisée, tandis que le second, qui contient 75% des points, sera destiné à entraîner le modèle d'apprentissage automatique.

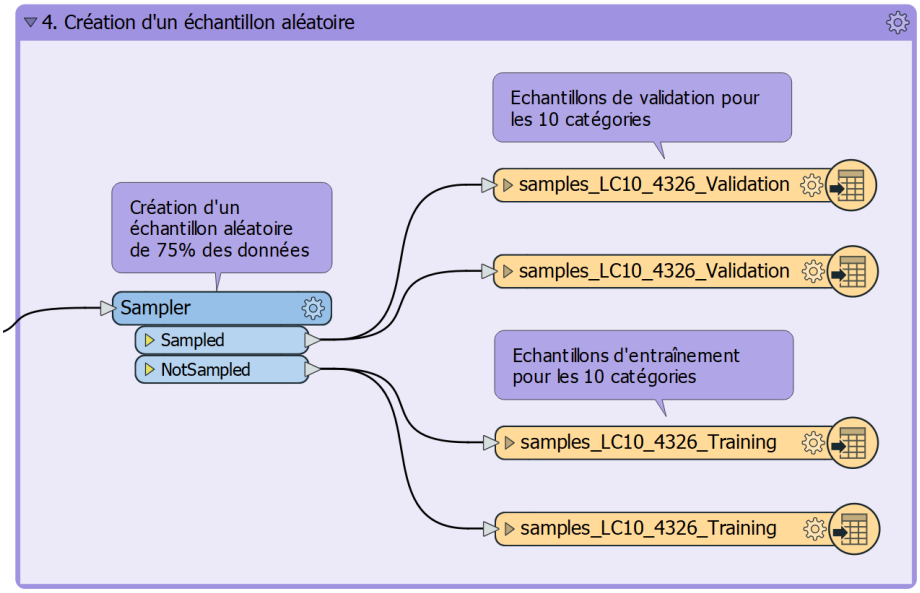


Figure 14 – Quatrième marque-page du modèle FME : création d'un échantillon aléatoire d'entraînement contenant 75% des données de base

En tout, 410'111 points d'échantillonnage ont été gardés, dont 307'584 points d'entraînement et 102'527 points de validation. Les points d'entraînement sont répartis de manière inégale dans les 10 catégories de couverture du sol. Puisque la Statistique de la superficie est suisse, aucun point d'entraînement n'a été récolté en France, ce qui permet notamment d'évaluer la capacité du modèle à prédire des valeurs en dehors du territoire national.

4.2.2 Classification des images satellites sur RStudio

Structure du projet

Afin que le code source contenant les instructions utilisées dans le cadre de ce travail pour générer la classification soit compréhensible et utilisable par d'autres personnes, une attention toute particulière a été apportée à la mise en page du code et à la structure du dossier qui contient tous les fichiers nécessaires au travail.

La classification des images satellites, la validation croisée des données d'entraînement et l'évaluation de la précision de la classification ont toutes les trois étés regroupées dans un même projet sur RStudio nommé sits_GVA. Enregistrer son travail dans un projet permet de référencer les fichiers avec des chemins relatifs, ce qui permet un meilleur partage du code source. Le dossier contient le projet sits_GVA.Rproj en lui-même, un fichier texte de type « lisez-moi », un fichier .RData dans lequel est enregistré l'environnement de travail et quatre sous-dossiers. Le dossier « Data » contient les données en entrée, à savoir les fichiers contenant les points d'entraînement et de validation, tandis

que les dossiers « Output_2018 » et « Output_2022 » contiennent les cubes de données et les produits générés sur *sits* tels que les classification d'images. Le dossier « Script » contient le code source utilisé pour l'analyse.

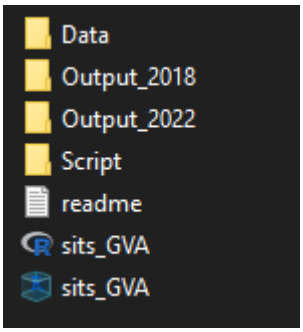


Figure 15 – Structure du projet sits_GVA

Le déroulement des opérations résumé dans les sections suivantes suit dans les grandes lignes celui décrit par Simoes et al pour le paquet *sits*. Pour une meilleure compréhension des paramètres inhérents aux fonctions proposées par sits, on peut se référer à l'aide offerte par la documentation R. Avant d'interpréter les instructions relatives à la classification à proprement parler, il peut être utile de détecter le nombre de cœurs disponibles dans l'ordinateur utilisé. Plusieurs méthodes existent, mais il y a un moyen simple de le faire sur RStudio, en utilisant la fonction detectCores() du paquet *parallel*, qui renvoie le nombre de processeurs logiques. Le nombre de cœurs est ensuite stocké dans une variable qui sera réutilisée par la suite, ce qui évite de devoir redéfinir le nombre de cœurs à chaque fonction lorsque l'on change d'ordinateur.

```
> parallel::detectCores()
[1] 72
```

Figure 16 – Résultat de l'instruction parallel::detectCores() lorsqu'elle est interprétée par l'ordinateur

Le code source à proprement parler a également été subdivisé en plusieurs sections grâce au pliage de code proposé par RStudio. Cette fonctionnalité permet d'avoir une vue d'ensemble du code, détail particulièrement efficace lorsque ce dernier contient beaucoup d'instructions. L'ensemble du code est subdivisé en 7 parties, à la fin desquelles l'environnement de travail est à chaque fois enregistré.

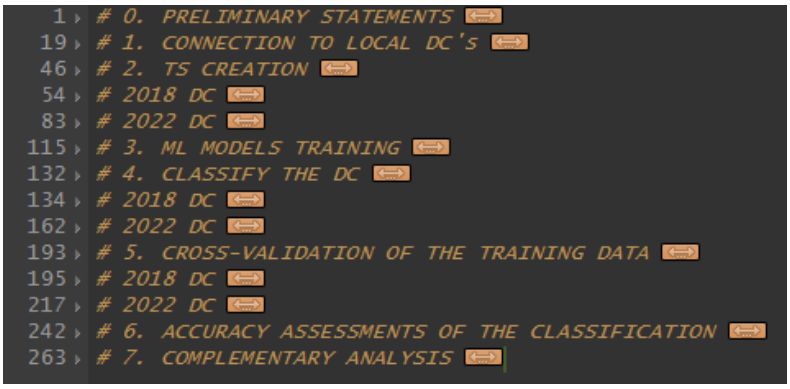


Figure 17 – Structure générale du code utilisé pour l'analyse

Création des cubes de données

Comme expliqué précédemment, la première et la deuxième étape consistent à sélectionner une collection d'images prêtes à l'analyse sur un fournisseur cloud, puis à construire un cube de données réguliers à partir de la collection d'images sélectionnées. Ces deux premières étapes doivent être répétées pour chacune des deux années retenues pour la comparaison.

Malheureusement, si la première étape a abouti avec succès, la deuxième n'a produit aucune image régularisée, même après plusieurs heures de calcul sur un serveur à 72 processeurs logiques. Plusieurs tentatives ont été effectuées, sans succès. Selon Felipe Souza, qui a participé au développement du paquet *sits*, ce problème pourrait être causé par le fait que les données AWS (*Amazon Web Services*) sont situées sur un serveur aux Etats-Unis et que leur téléchargement depuis la Suisse prend du temps. Une solution consisterait à télécharger les images Sentinel sur la plateforme en accès libre de Copernicus, puis à régulariser l'ensemble des images sur *sits*. Néanmoins, Gregory Giuliani est parvenu à régulariser sur un ordinateur Mac deux cubes pour 2018 et 2022 qui ont pu être utilisés dans le cadre de ce travail, épargnant ainsi le téléchargement des images sur internet.

Il convient de préciser qu'au terme de la deuxième étape, qui consiste à régulariser une collection d'images prêtes à l'analyse pour créer un cube de données, ce dernier ne contient pas les mêmes intervalles temporels que la collection d'images. De fait, la fonction sits_regularize() contient un paramètre temporel nommé « period » que l'utilisateur doit configurer pour indiquer à l'ordinateur quels intervalles temporels garder dans la collection d'images prêtes à l'analyse.

```r
# Get an ARD collection from a cloud service
ard_2021 <- sits_cube(
  source = "AWS",
  collection = "SENTINEL-S2-L2A",
  tiles = "31TGM",
  bands = c("B02", "B03", "B04", "B08", "B8A", "CLOUD"),
  start_date = "2021-07-01",
  end_date = "2021-08-31"
)

# Regularize a cube
cube_2021 <- sits_regularize(
  cube = ard_2021,
  period = "P5D",
  res = 10,
  output_dir = "./Output",
  multicores = 72
)
```

Figure 18 – Les deux premières étapes de sits : sélection d'une collection d'images prêtes à l'analyse (sits_cube()) et régularisation pour créer un cube de données (sits_regularize()). La deuxième étape contient l'argument « period », qui indique à l'ordinateur les intervalles temporels à garder, ici une image tous les cinq jours.

Lors de la régularisation des cubes de données, toutes les bandes de Sentinel-2 ont été retenues, à l'exception de la bande 10. Par la suite, seules certaines de ces bandes seront utilisées pour l'entraînement des modèles d'apprentissage automatique et la classification des cubes de données.

Le produit de la fonction sits_regularize(), qui sert à régulariser une collection d'images prêtes à l'analyse, est une série d'images stockées dans un dossier de sortie. Dans notre cas, les rasters pour 2018 sont stockés dans le dossier Output_2018 et les rasters pour 2022 dans le dossier Output_2022. Sans les indices, le cube de 2018 comptabilise 696 images (58 dates avec 12 bandes chacune) et celui de 2022, 732 images (61 dates avec 12 bandes chacune). Une fois que les cubes de données ont été régularisés et déposés dans un dossier, on peut aisément y accéder grâce à la fonction sits_cube(). Chacun des cubes est alors assigné à une variable différente dans R.

```
# 1. CONNECTION TO LOCAL DC's ================================================

# Cube for 2018
cube_2018 <- sits_cube(
  source = "AWS",
  collection = "SENTINEL-S2-L2A-COGS",
  data_dir = "./Output_2018",
  parse_info = c("X1", "tile", "band", "date")
)

# Show timeline of the 2018 cube
sits_timeline(cube_2018)

# Cube for 2022
cube_2022 <- sits_cube(
  source = "AWS",
  collection = "SENTINEL-S2-L2A-COGS",
  data_dir = "./Output_2022",
  parse_info = c("X1", "tile", "band", "date")
)

# Show timeline of the 2022 cube
sits_timeline(cube_2022)

# Save the environment into a .RData file
save.image("C:/Users/risse/Documents/R/sits_GVA/sits_GVA.RData")
```

Figure 19 – Capture d'écran montrant le code source pour la première étape, qui consiste à importer dans l'environnement de travail des cubes de données préalablement régularisés

La fonction sits_cube() permet d'importer une série d'images enregistrées dans un dossier local, en utilisant certains caractères génériques avec le paramètre « parse_info ». La fonction sits_timeline() permet quant à elle de visualiser les dates contenues dans les cubes de données.

Création des indices

Une fois les cubes importés sur RStudio, on peut procéder à la création de nouveaux indices spectraux, susceptibles de nous aider à discriminer l'information dans le cadre de la classification des images. Quatre indices ont été calculés pour chacun des deux cubes générés : le NDVI, l'EVI, le NDWI et le NDBI. Le NDVI, le NDWI et le NDBI ayant déjà été calculés pour l'année 2018 par Gregory Giuliani, il ne restait qu'à créer les mêmes indices pour l'année 2022, ainsi que l'EVI pour les deux années. Initialement, l'EVI a été développé pour les satellites Landsat, mais on peut aisément trouver une formule applicable à Sentinel-2 sur internet[4].

[4] Page : https://www.indexdatabase.de/db/si-single.php?sensor_id=96&rsindex_id=16

Extraction des séries temporelles

Une fois les indices générés, on peut passer à l'extraction des séries temporelles. À cette étape, les coordonnées de chaque point de mesure de la Statistique de la superficie, dont on connaît la classe, sont utilisées pour extraire du cube de données des séries temporelles. Le nombre de séries temporelles pour chaque point équivaut au nombre de bandes sélectionnées dans le cube, indices compris. Dans notre cas et en suivant les conseils de Felipe Souza, seules les bandes 02, 03, 04, 08, 8A, 12 et les quatre indices créés précédemment ont été sélectionnés. Chaque point est associé à dix séries temporelles, soit une pour chaque bande.

Les séries temporelles en elles-mêmes sont définies en fonction du nombre de dates contenues dans le cube de données. Par exemple, pour le cube de 2018, chaque série temporelle s'étend du 21 mars au 31 décembre. Les séries temporelles sont stockées dans un tableau à trois dimensions où chaque entrée, ou point de mesure, possède une valeur pour chaque date et pour chaque bande.

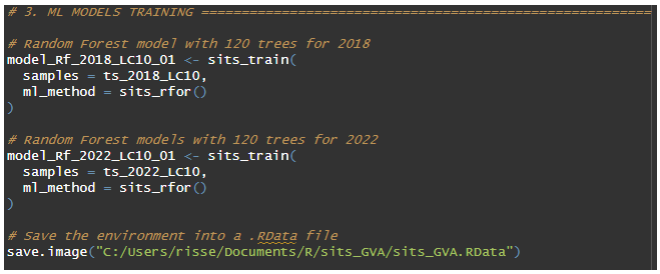
Lors de la création des séries temporelles, certains points d'entraînement n'ont pas été retenus. De fait, le nombre d'occurrences dans les séries temporelles est plus petit que le nombre de points d'entraînement. Une première hypothèse est que l'erreur soit due à un problème de projection, or des essais ont été effectués avec deux systèmes de projection différents (EPSG 4326 et EPSG 32631), sans que cela n'y change quelque chose. Une autre hypothèse est que le transformateur qui sert à découper la couche de points sur FME inclut des points à la frontière du raster, qui ne sont alors pas inclut dans les séries temporelles sur *sits*. En d'autres termes, il pourrait s'agir d'une incompatibilité entre logiciels. Toutefois, cette erreur n'a pas d'incidence sur la suite du travail.

```
# 2. TS CREATION ===============================================================

# Training and validation points were generated using FME
# .shp and .csv outputs of the FME model were pasted in the Data folder

# Read the .shp file containing the training points for LC17
samples_Training_LC10 <- sf::st_read("./Data/samples_LC10_Training.shp")

# 2018 DC ------------------------------------------------------------------

# Obtain TS for 2018 with LC17
ts_2018_LC10 <- sits_get_data(
  cube = cube_2018,
  samples = samples_Training_LC10,
  bands = c("B02", "B03", "B04", "B08", "B8A", "B12", "NDVI", "EVI", "NDWI", "NDBI"),
  multicores = system_cores,
  output_dir = "./Output_2018",
  progress = TRUE
)

# Show label summary for 2018 time series
summary_2018 <- sits_labels_summary(ts_2018_LC10)

# Write label summary into an .xlsx file
write.xlsx(summary_2018, file = "./ts_Summary_2018.xlsx")
```

Figure 20 – Capture d'écran montrant le code source pour la deuxième étape, qui consiste à utiliser les points d'entraînement pour extraire du cube de données des séries temporelles. Seuls les indices et les bandes 02, 03, 04, 08, 8A et 12 ont été retenus. La même procédure doit être effectuée pour 2022.

Entraînement du modèle d'apprentissage automatique

Une fois les séries temporelles extraites du cube de données, on peut les utiliser pour entraîner les modèles d'apprentissage automatique. Dans notre cas, les modèles sont basés sur des forêts d'arbres décisionnels, décrites dans la revue de la littérature. Les forêts d'arbres décisionnels sont les algorithmes les plus utilisés pour classer des images Sentinel-2 (Phiri et al., 2020). Le paramètre par défaut de 120 arbres de la fonction *sits* liée n'est pas changé, compte tenu des conclusions présentées dans l'article de C. Pelletier, abordé précédemment (Pelletier et al., 2016).



```
# 3. ML MODELS TRAINING ================================================

# Random Forest model with 120 trees for 2018
model_Rf_2018_LC10_01 <- sits_train(
  samples = ts_2018_LC10,
  ml_method = sits_rfor()
)

# Random Forest models with 120 trees for 2022
model_Rf_2022_LC10_01 <- sits_train(
  samples = ts_2022_LC10,
  ml_method = sits_rfor()
)

# Save the environment into a .RData file
save.image("C:/Users/risse/Documents/R/sits_GVA/sits_GVA.RData")
```

Figure 21 – Capture d'écran montrant le code source pour la troisième étape, qui consiste à entraîner deux modèles d'apprentissage automatique basés sur des forêts d'arbres décisionnels.

Classification des images, suppression des données aberrantes et production de cartes étiquetées

Une fois les forêts d'arbres décisionnelles entraînées, on peut classer les données des cubes de 2018 et 2022 grâce à la fonction sits_classify(). Le résultat de la fonction est un cube dont les différentes couches sont des cartes montrant la probabilité d'appartenance de chaque pixel à chaque classe. Chaque pixel est alors assigné à la classe pour laquelle il possède la probabilité d'appartenance la plus élevée.

Comme énoncé dans la revue de la littérature, la classification par pixel peut engendrer un effet visuel dérangeant. En outre, en raison de la variabilité spectrale intrinsèque à chaque classe, il se peut que des pixels soient mal classés. Pour ces raisons, *sits* inclut une fonction de lissage des données aberrantes qui utilise le contexte spatial des cubes de probabilités générés précédemment. Le lissage des données supprime les données aberrantes en partant du postulat que les pixels proches les uns des autres tendent à avoir la même étiquette.

Dans *sits*, la dernière étape, consécutive au lissage des données aberrantes, consiste à produire une carte étiquetée à l'aide de la fonction sits_label_classification(). Le produit de cette dernière est une carte labellisée, enregistrée dans le dossier de sortie sélectionné en tant que fichier .tif. On peut alors importer la carte produite dans un logiciel d'information géographique. Les deux cartes présentées ci-après ont par exemple été importée et mises en page à l'aide du logiciel ArcGIS Pro.

```
# 4. CLASSIFY THE DC ===================================================

# 2018 DC ------------------------------------------------------------------

# Classify the DC and store the results in a  probability cube
probs_2018_LC10 <- sits_classify(
  data = cube_2018,
  ml_model = model_Rf_2018_LC10_01(),
  multicores = system_Cores,
  memsize = system_Memory,
  output_dir = "./Output_2018",
  version = "v1"
)

# Apply bayesian smoothing to the probability cube
bayes_2018_LC10 <- sits_smooth(
  cube = probs_2018_LC10,
  multicores = system_Cores,
  memsize = system_Memory,
  output_dir = "./Output_2018",
  version = "v1"
)

# Create the classified map
map_2018_LC10 <- sits_label_classification(
  cube = bayes_2018_LC10,
  output_dir = "./Output_2018",
  version = "v1"
)
```

Figure 22 – Capture d'écran montrant le code source pour la quatrième, lors de laquelle la classification est réalisée. La même procédure doit être répétée pour 2022.

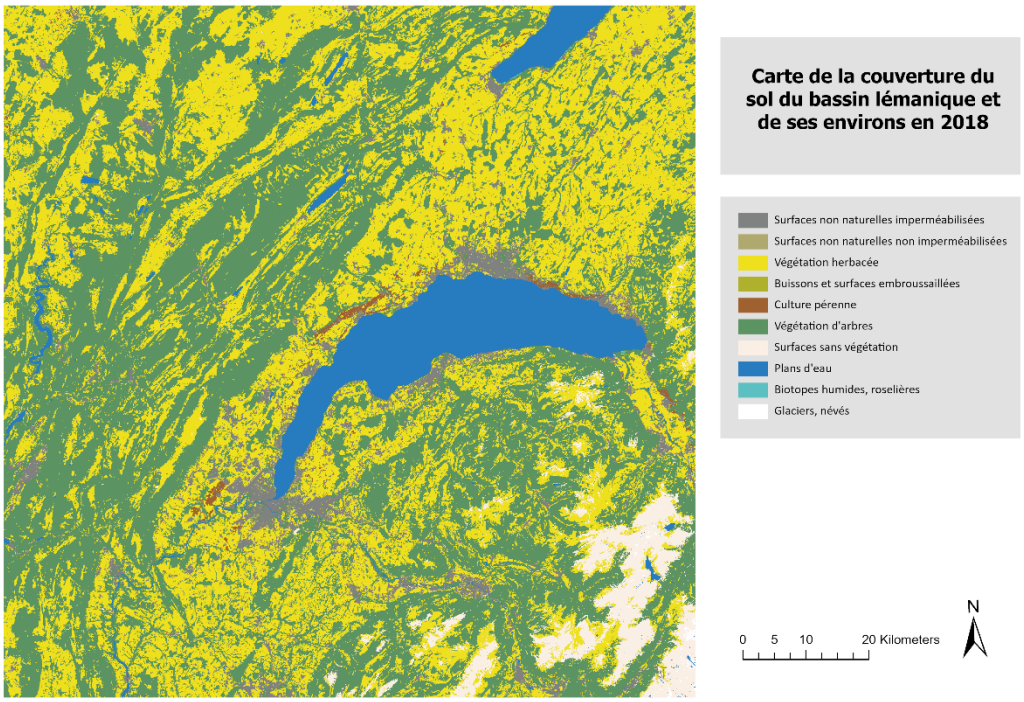


Figure 23 – Résultat de la classification pour 2018

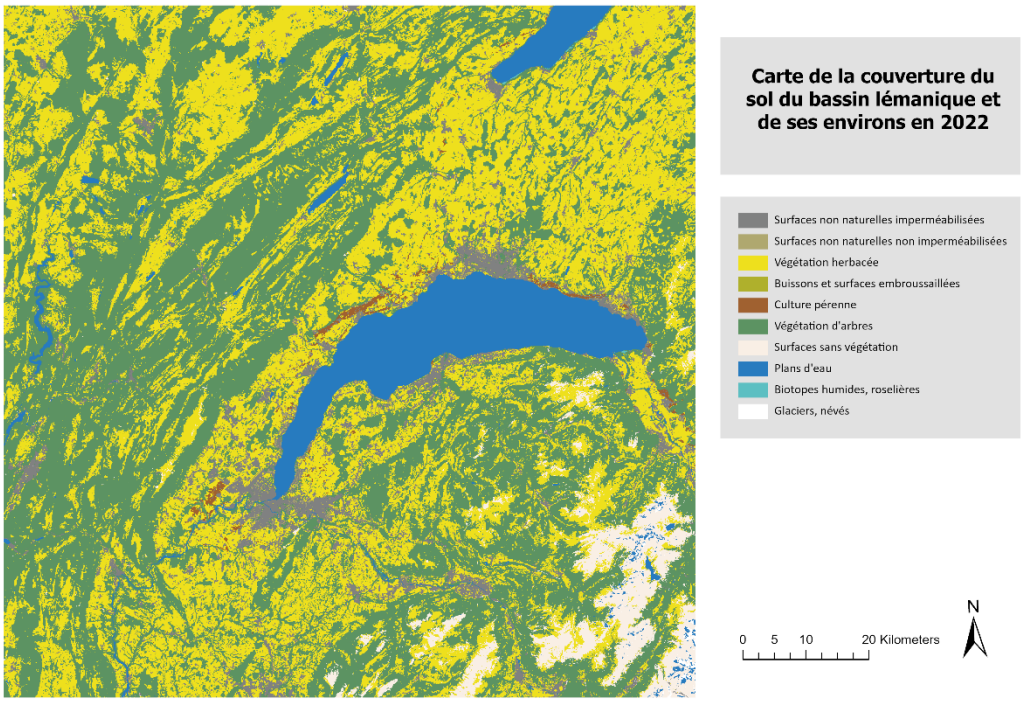


Figure 24 – Résultat de la classification pour 2022

4.2.3 Validation croisée et évaluation de la précision de la classification

Sur sits, la validation croisée des données d'entraînement peut être réalisée grâce à la fonction sits_kfold_validate(). Les paramètres à définir sont notamment le nombre de blocs, la méthode d'apprentissage automatique à employer et le nombre de cœurs disponibles pour le parallélisme. Le résultat obtenu est une liste d'éléments, dont une matrice de confusion et le coefficient kappa. Ces résultats peuvent être exportés dans un tableau Excel grâce à la fonction sits_to_xlsx().

```
# 5. CROSS-VALIDATION OF THE TRAINING DATA ====================================

# 2018 DC ====================================================================

# Validation for 2018
val_Rf_2018_LC10_01 <- sits_kfold_validate(
  ts_2018_LC10,
  folds = 5,
  ml_method = sits_rfor(),
  multicores = system_Cores
)
```

Figure 25 – Capture d'écran montrant la cinquième étape, lors de laquelle est réalisée la validation croisée à k-blocs des données d'entraînement. La même procédure doit être répétée pour 2022.

L'évaluation de la précision se fait quant à elle grâce à la fonction sits_accuracy(). Ses paramètres sont la carte dont la précision doit être évaluée et le fichier .csv contenant les points de validation. À cette étape, différentes erreurs ont empêché l'obtention des statistiques de précision. La première erreur concernait le format du fichier .csv fourni dans les paramètres de la fonction. De fait, il manquait un nom à la première colonne du document. La deuxième erreur concernait la projection même des données de validation. Le paquet sits exige que ces dernières soient précisées en degrés décimaux, alors que le fichier produit par le modèle FME affiche des coordonnées en mètres. Grâce à une aide externe, les deux erreurs ont pu être résolues et les statistiques de précision obtenues.

```
# 6. ACCURACY ASSESSMENTS OF THE CLASSIFICATION ===============================

accuracy_2018 <- sits_accuracy(
  data = map_2018_LC10_01,
  validation_csv = "./Data/samples_LC10_Validation_rep.csv"
)

errorMatrix_2018 <- write.csv(accuracy_2018$error_matrix,
                              file = "./errorMatrix_2018.csv")

accuracy_2022 <- sits_accuracy(
  data = map_2022_LC10,
  validation_csv = "./Data/samples_LC10_Validation_rep.csv"
)

errorMatrix_2022 <- write.csv(accuracy_2022$error_matrix,
                              file = "./errorMatrix_2022.csv")

# Save the environment into a .RData file
save.image("C:/Users/risse/Documents/R/sits_GVA/sits_GVA.RData")
```
s

Figure 26 – Capture d'écran montrant la cinquième étape, lors de laquelle est évaluée la précision des classifications réalisées.

5. Résultats et discussion

La validation croisée aboutit à des coefficients kappa très proches de 0.826 pour 2018 et de 0.832 pour 2022, montrant dans les deux cas un niveau de concordance élevé entre les valeurs prédites et la réalité du terrain. Les forêts d'arbres décisionnelles générées ont donc une forte valeur prédictive.

Une fois la classification réalisée, les cartes produites sont assez semblables à large échelle. On remarque par exemple la ressemblance globale dans la classification des zones urbaines, des plans d'eau, des forêts ou des sites de culture pérenne tels que les vignes. Néanmoins, on peut visuellement distinguer certaines différences notables, tel le fait que le modèle pour 2022 a prédit plus de forêts dans le Jura français (ouest de la carte) ou le fait que ce même modèle a produit des plans d'eau dans les Alpes (sud-est de la carte), là où le modèle pour 2018 a prédit des surfaces sans végétation ou des glaciers et des névés. D'autres écarts mineurs peuvent être dus aux différences dans les modèles d'apprentissage automatique.

Si l'on excepte les erreurs de prédiction dans les Alpes pour la carte de 2022, il est toutefois difficile de juger visuellement si l'une des deux cartes est plus précise que l'autre et il est donc nécessaire d'utiliser des indicateurs statistiques pour s'en faire une meilleure idée. Avec *sits*, la précision de l'utilisateur est calculée à partir du nombre de pixels par classe, tandis que la précision du producteur et la précision globale sont calculées à partir de valeurs estimées et corrigées.

Tableau 3 – Matrice de confusion entre réalité de terrain (en colonnes) et données classifiées (en lignes) pour 2018

2018	Biotopes humides, roselières	Buissons et surfaces embroussaillées	Culture pérenne	Glaciers, névés	Plans d'eau	Surfaces non naturelles imperméabilisé	Surfaces non naturelles non imperméabilisé	Surfaces sans végétation	Végétation d'arbres	Végétation herbacée	Total	Utilisateur
Biotopes humides, roselières	26	0	0	0	1	0	0	0	0	0	27	0,963
Buissons et surfaces embroussaillées	0	11	0	0	0	0	0	3	1	1	16	0,688
Culture pérenne	0	2	849	0	0	60	14	2	5	24	956	0,888
Glaciers, névés	0	0	0	42	0	0	0	12	0	0	54	0,778
Plans d'eau	13	4	0	0	11403	24	5	10	10	3	11472	0,994
Surfaces non naturelles imperméabilisé	3	48	108	0	48	4053	1715	143	196	549	6863	0,591
Surfaces non naturelles non imperméabilisé	0	1	3	0	1	129	285	1	11	5	436	0,654
Surfaces sans végétation	0	110	0	144	10	20	1	2116	40	184	2625	0,806
Végétation d'arbres	138	1005	51	0	111	522	273	250	28449	2057	32856	0,866
Végétation herbacée	171	448	537	0	49	2037	1119	496	2168	40178	47199	0,851
Producteur	0,020	0,006	0,346	0,248	0,962	0,540	0,049	0,748	0,950	0,907	Précision globale	0,852

Tableau 4 - Matrice de confusion entre réalité de terrain (en colonnes) et données classifiées (en lignes) pour 2022

2022	Biotopes humides, roselières	Buissons et surfaces embroussaillées	Culture pérenne	Glaciers, névés	Plans d'eau	Surfaces non naturelles imperméabilisé	Surfaces non naturelles non imperméabilisé	Surfaces sans végétation	Végétation d'arbres	Végétation herbacée	Total	Utilisateur
Biotopes humides, roselières	45	0	0	0	3	0	0	0	0	0	48	0,938
Buissons et surfaces embroussaillées	0	23	0	0	0	0	0	1	1	3	28	0,821
Culture pérenne	0	2	953	0	0	60	18	2	7	35	1077	0,885
Glaciers, névés	0	0	0	53	0	0	0	3	0	0	56	0,946
Plans d'eau	20	6	0	55	11418	29	7	90	13	9	11647	0,980
Surfaces non naturelles imperméabilisé	2	55	73	0	39	4122	1695	151	205	602	6944	0,594
Surfaces non naturelles non imperméabilisé	0	1	2	0	0	190	403	3	21	17	637	0,633
Surfaces sans végétation	0	114	0	78	14	16	0	2090	60	194	2566	0,814
Végétation d'arbres	132	1016	44	0	105	517	296	252	28346	1899	32607	0,869
Végétation herbacée	152	412	476	0	44	1911	993	441	2223	40242	46894	0,858
Producteur	0,028	0,015	0,899	0,357	0,964	0,557	0,074	0,715	0,950	0,908	Précision globale	0,855

La précision globale est assez proche pour les deux classifications réalisées, avec 85,2 % des valeurs correctement classées en 2018, contre 85,6 % en 2022. La précision de l'utilisateur est en moyenne relativement élevée par rapport à la précision du producteur, ce qui signifie que dans la plupart des cas, les pixels appartiennent réellement à la classe à laquelle on les a attribués. En 2018 comme en 2022, les surfaces non naturelles imperméabilisées ont la précision de l'utilisateur la plus basse (0.591 en 2018, 0.594 en 2022), tandis que les plans d'eau ont la précision de l'utilisateur la plus élevée (0.994 en 2018, 0.980 en 2022). Ce constat n'est pas étonnant, compte tenu de l'hétérogénéité des signatures spectrales des zones artificialisées et de l'homogénéité de ces signatures pour les plans d'eau.

La précision du producteur varie dans de plus grandes proportions, avec certaines valeurs très faibles pour les classes « Biotopes humides, roselières » (0.028 en 2018, 0.081 en 2022) et « Buissons et surfaces embroussaillées » (0.006 en 2018, 0.015 en 2022). Dans ce genre de cas, la réalité du terrain est alors sous-estimée puisque la majorité des pixels qui auraient dû être classés dans l'une de ces deux classes ne l'ont pas été.

La faible précision de l'utilisateur pour certaines classes peut s'expliquer par le faible nombre de points d'apprentissage qui ont été retenus pour ces dernières. Les diagrammes ci-dessous montrent une corrélation relativement élevée entre le nombre de points par classe et la précision du producteur. Ce lien est conforme au constat relevé dans la revue de la littérature, selon lequel un trop grand déséquilibre entre le nombre de points par classe risque d'aboutir à un modèle qui sous-estime les classes les moins représentées. Ainsi, dans notre cas, plus le nombre de points d'une classe est bas, plus on peut s'attendre à ce qu'elle soit sous-estimée lors de la classification. Cette observation doit néanmoins être considérée avec précaution, compte tenu du faible nombre de classes retenu.

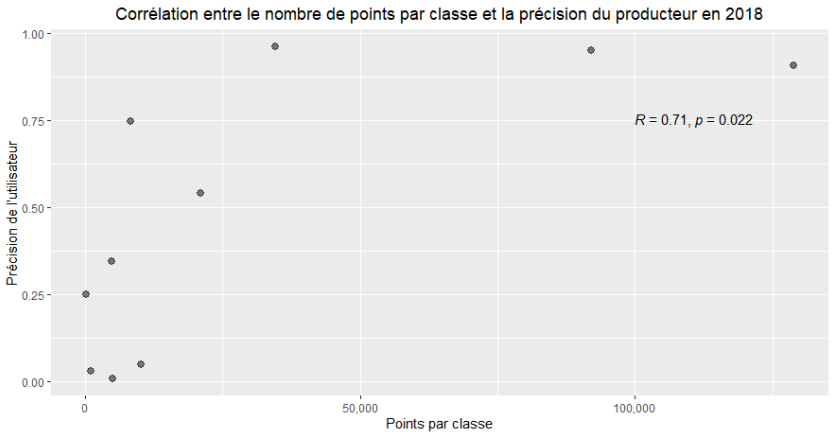


Figure 27 – Diagramme de corrélation entre le nombre de points et la précision du producteur en 2018

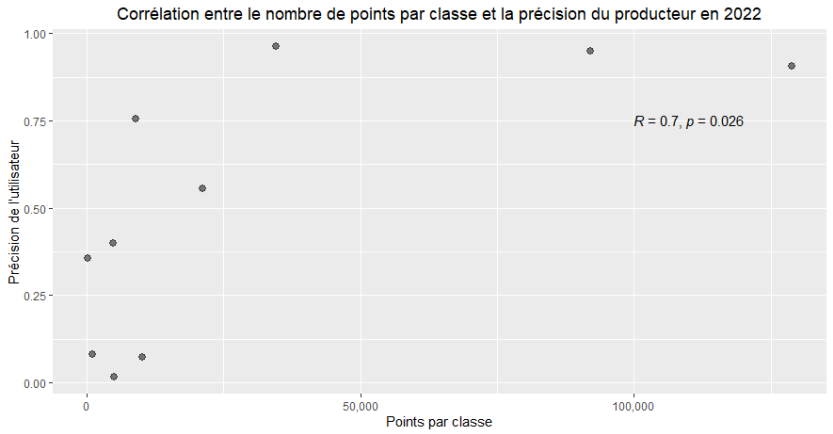


Figure 28 – Diagramme de corrélation entre le nombre de points et la précision du producteur en 2022

L'un des principaux intérêts de s*its* est de mesurer l'évolution de la couverture du sol. De fait, l'un des résultats produits par la fonction sits_accuracy() est une estimation ajustée de l'aire occupée par chaque classe de couverture du sol, que l'on peut alors utiliser pour comparer les deux classifications produites. Dans le tableau ci-dessous, on remarque une diminution de la surface dédiée aux biotopes humides, aux cultures pérennes, aux glaciers et à la végétation herbacée. On constate également une diminution des surfaces imperméables, ce qui est surprenant au premier regard. Il est possible qu'une partie des surfaces imperméables en 2018 aient été classées dans les surfaces sans végétation en 2022. Les surfaces embroussaillées et les forêts ont quant à elles augmenté. L'augmentation en surface des plans d'eau peut quant à elle être due à une simple erreur, constatée précédemment lors de la visualisation des cartes produites.

	2018	2022	Différence absolue	Différence relative (%)
Biotopes humides, roselières	4247,83	4226,907	-20,93	-0,49
Buissons et surfaces embroussaillées	23442,87	23918,112	475,25	2,03
Culture pérenne	11120,37	10245,517	-874,86	-7,87
Glaciers, névés	2891,14	2471,308	-419,83	-14,52
Plans d'eau	76679,69	77623,6	943,91	1,23
Surfaces non naturelles imperméabilisées	66944,88	66725,152	-219,73	-0,33
Surfaces non naturelles non imperméabilisées	32895,49	32692,872	-202,62	-0,62
Surfaces sans végétation	42711,26	43595,583	884,33	2,07
Végétation d'arbres	494726,85	502323,042	7596,19	1,54
Végétation herbacée	449943,61	441781,908	-8161,71	-1,81

6. Conclusion

Le but de ce travail était d'explorer une procédure de cartographie de la couverture du sol dans le bassin lémanique, en utilisant la classification d'images satellites par apprentissage automatique. Le paquet *sits*, récemment développé par une équipe de chercheurs au Brésil, propose justement des fonctions qu'il est facile de paramétrer à cette fin.

Plus précisément, l'objectif était d'implémenter une méthode qui facilite la cartographie de la couverture du sol et de son évolution d'une année en l'autre, en évitant de produire de nouvelles données d'entraînement. Si la précision de la classification ne diminue pas avec les années, alors on peut utiliser des données d'entraînement qui ont été récoltées plusieurs années avant la production d'une classification. Pour atteindre l'objectif fixé, deux cubes de données issues des satellites Sentinel-2 ont été générés, l'un pour 2018 et l'autre pour 2022.

Au bout d'une procédure de plusieurs étapes, deux cartes ont été produites. Visuellement, ces dernières sont très proches l'une de l'autre, même si l'on peut constater quelques différences ici et là. Cette proximité visuelle se confirme d'un point de vue statistique, avec une précision globale de 0,852 pour 2018 et de 0,856 pour 2022. Ces résultats montrent que la précision globale de la classification ne diminue pas de façon significative entre 2018 et 2022. Au contraire même, elle augmente. De telles conclusions sont prometteuses quant à l'utilisation de données d'entraînement pour prédire la classe de données éloignées dans le temps, mais elles ne suffisent pas. De fait, les deux cubes ne sont éloignés que de quatre années et peu de changements de la couverture du sol interviennent dans une période aussi courte.

Il conviendrait de répéter le travail effectué avec des cubes de données temporellement plus éloignés l'un de l'autre et en utilisant des données d'entraînement plus anciennes, pour constater si la procédure utilisée permet de détecter l'évolution de la couverture du sol de manière adéquate. Heureusement, les données produites par la Statistique de la superficie en Suisse remontant jusqu'à 1979, on peut les utiliser pour générer des données d'entraînement plus anciennes. Etant donné que les satellites Sentinel-2 n'ont été mis en service qu'à partir de 2015, on peut envisager l'utilisation de données produites par les satellites Landsat de la NASA. Si ces derniers ont une résolution spatiale plus faible (15 m avec la bande panchromatique à partir de Landsat 7), ils ont été mis en service bien avant les satellites Sentinel-2.

D'autres pistes peuvent être proposées pour améliorer cette fois la précision des cartes de couverture du sol pour une année donnée. Premièrement, il conviendrait d'utiliser un échantillon d'entraînement plus équilibré, avec un nombre de points qui varie peu d'une classe à l'autre. Cela éviterait notamment que certaines classes soient sous estimées dans la classification, comme cela a été le cas pour les classifications réalisées dans le cadre de ce travail. Deuxièmement, il faudrait effectuer un contrôle qualité des données aberrantes au préalable, ce qui éviterait que des signatures spectrales très différentes soient inclues dans une même classe. Ces deux premières propositions, qui concernent les données d'entraînement en elles-mêmes, pourraient déjà apporter quelques améliorer la classification de manière significative.

On peut également envisager des pistes d'amélioration en proposant d'autres bandes et d'autres indices sur lesquels entraîner les modèles d'apprentissage automatique. Dans le cadre de ce travail, les mêmes bandes et les mêmes indices ont été utilisés pour réaliser les deux cartes présentées, mais on pourrait tester différents modèles en y retirant ou en y ajoutant certaines bandes. Eventuellement, on peut également ajouter à la classification d'autres indices plus performants, notamment pour le bâti et pour les plans d'eau. Cette proposition permettrait de vérifier dans quelle mesure l'ajout de bandes supplémentaires améliore réellement la précision d'une classification sans trop augmenter le temps de calcul nécessaire. Si l'enlèvement de bandes diminue peu la précision de la classification, alors on peut raisonnablement les abandonner pour diminuer le temps de calcul et faciliter le travail.

En outre, on peut améliorer les cartes réalisées en testant d'autres modèles d'apprentissage automatique. Si les forêts d'arbres décisionnels sont performantes, d'autres algorithmes ont également fait leurs preuves, telles les machines à vecteur de support. En particulier, les algorithmes d'apprentissage profond ont montré des performances prometteuses voire supérieures en matière de classification. Envisager d'autres algorithmes permettrait de déterminer lesquels sont les plus adéquats dans le contexte du bassin lémanique.

Enfin, d'autres améliorations concernent le paquet *sits* lui-même. Bien que ce dernier soit relativement aisé à comprendre et à prendre en main, il ne propose pas de classification orientée objet, alors que la littérature a montré qu'elle était plus performante que la classification par pixel. En outre, *sits* ne propose pas d'intégrer des données auxiliaires de pente, d'exposition ou d'altitude, susceptibles d'améliorer la prédiction de certaines classes de couverture du sol et ce, notamment en matière de végétation. Heureusement, les développeurs de *sits* sont conscients de ces lacunes et travaillent, selon les dires de Felipe Souza, à les combler.

Une fois les différentes pistes d'améliorations explorées, on peut envisager différents élargissements de la procédure proposée, telle qu'une cartographie à l'échelle de tout le territoire suisse ou l'utilisation d'autres nomenclatures. Ces propositions sont susceptibles d'augmenter la portée de la méthode proposée.

Malheureusement, peu des solutions envisagées ont pu être explorées dans le cadre de mon stage et ce, pour différentes raisons. D'abord, le stage n'a duré que trois mois et demi, ce qui m'a peu laissé le temps d'explorer et ce, surtout si l'on compte le temps de formation dédié à la compréhension même de *sits*. Ensuite, le matériel informatique à disposition a constitué une réelle limite tout au long de mon stage. De fait, le serveur de l'Université de Genève ayant souffert d'une panne l'automne dernier, je n'ai pu en bénéficier qu'à la fin de mon travail, devant me dépêcher d'effectuer les traitements qu'il m'était incombé de faire. Malgré cela, les quelques mois passés aux CJBG m'ont permis d'approfondir certains domaines de compétences de manière considérable, que cela soit en programmation, en télédétection ou en informatique de manière générale.

7. Bibliographie

Arkebauer, T. J. (2015). Leaf Radiative Properties and the Leaf Energy Budget (J. L. Hatfield & J. M. Baker, Éds.; p. 93-103). American Society of Agronomy, Crop Science Society of America, and Soil Science Society of America. https://doi.org/10.2134/agronmonogr47.c5

As-syakur, Abd. R., Adnyana, I. W. S., Arthana, I. W., & Nuarsa, I. W. (2012). Enhanced Built-Up and Bareness Index (EBBI) for Mapping Built-Up and Bare Land in an Urban Area. Remote Sensing, 4(10), 2957-2970. https://doi.org/10.3390/rs4102957

Camara, G., Assis, L. F., Ribeiro, G., Ferreira, K. R., Llapa, E., & Vinhas, L. (2016). Big earth observation data analytics : Matching requirements to system architectures. Proceedings of the 5th ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data - BigSpatial '16, 1-6. https://doi.org/10.1145/3006386.3006393

Chuvieco, E. (2020). : An Environmental Approach, Third Edition (3e éd.). CRC Press. https://doi.org/10.1201/9780429506482

Gislason, P. O., Benediktsson, J. A., & Sveinsson, J. R. (2006). Random Forests for land cover classification. Pattern Recognition Letters, 27(4), 294-300. https://doi.org/10.1016/j.patrec.2005.08.011

Giuliani, G., Rodila, D., Külling, N., Maggini, R., & Lehmann, A. (2022). Downscaling Switzerland Land Use/Land Cover Data Using Nearest Neighbors and an Expert System. Land, 11(5), 615. https://doi.org/10.3390/land11050615

Jawak, S. D., Devliyal, P., & Luis, A. J. (2015). A Comprehensive Review on Pixel Oriented and Object Oriented Methods for Information Extraction from Remotely Sensed Satellite Images with a Special Emphasis on Cryospheric Applications. Advances in Remote Sensing, 4(3), Art. 3. https://doi.org/10.4236/ars.2015.43015

Karpatne, A., Jiang, Z., Vatsavai, R. R., Shekhar, S., & Kumar, V. (2016). Monitoring Land-Cover Changes : A Machine-Learning Perspective. IEEE Geoscience and Remote Sensing Magazine, 4(2), 8-21. https://doi.org/10.1109/MGRS.2016.2528038

Kaur, R., & Pandey, P. (2022). A review on spectral indices for built-up area extraction using remote sensing technology. Arabian Journal of Geosciences, 15(5), 391. https://doi.org/10.1007/s12517-022-09688-x

Kulkarni, A., & Lowe, B. (2016). Random Forest Algorithm for Land Cover Classification. https://www.semanticscholar.org/paper/Random-Forest-Algorithm-for-Land-Cover-Kulkarni-Lowe/2731ce300dcadd70302742113cfc65554e2cbe1b

Maxwell, A. E., Warner, T. A., & Fang, F. (2018). Implementation of machine-learning classification in remote sensing : An applied review. International Journal of Remote Sensing, 39(9), 2784-2817. https://doi.org/10.1080/01431161.2018.1433343

McFEETERS, S. K. (1996). The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features. International Journal of Remote Sensing, 17(7), 1425-1432. https://doi.org/10.1080/01431169608948714

Olofsson, P., Foody, G. M., Stehman, S. V., & Woodcock, C. E. (2013). Making better use of accuracy data in land change studies : Estimating accuracy and area and quantifying uncertainty using

stratified estimation. Remote Sensing of Environment, 129, 122-131. https://doi.org/10.1016/j.rse.2012.10.031

Pasquarella, V. J., Holden, C. E., Kaufman, L., & Woodcock, C. E. (2016). From imagery to ecology : Leveraging time series of all available Landsat observations to map and monitor ecosystem state and dynamics. Remote Sensing in Ecology and Conservation, 2(3), 152-170. https://doi.org/10.1002/rse2.24

Pelletier, C., Valero, S., Inglada, J., Champion, N., & Dedieu, G. (2016). Assessing the robustness of Random Forests to map land cover with high resolution satellite image time series over large areas. Remote Sensing of Environment, 187, 156-168. https://doi.org/10.1016/j.rse.2016.10.010

Phiri, D., Simwanda, M., Salekin, S., Nyirenda, V., Murayama, Y., & Ranagalage, M. (2020). Sentinel-2 Data for Land Cover/Use Mapping : A Review. Remote Sensing, 12(14), 2291. https://doi.org/10.3390/rs12142291

Rodríguez, J. D., Pérez, A., & Lozano, J. A. (2010). Sensitivity analysis of kappa-fold cross validation in prediction error estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(3), 569-575. https://doi.org/10.1109/TPAMI.2009.187

Santos, L. A., Ferreira, K. R., Camara, G., Picoli, M. C. A., & Simoes, R. E. (2021). Quality control and class noise reduction of satellite image time series. ISPRS Journal of Photogrammetry and Remote Sensing, 177, 75-88. https://doi.org/10.1016/j.isprsjprs.2021.04.014

Santos, L., Ferreira, K. R., Picoli, M., & Camara, G. (2020). Self-Organizing Maps in Earth Observation Data Cubes Analysis (A. Vellido, K. Gibert, C. Angulo, & J. D. Martín Guerrero, Éds.; Vol. 976, p. 70-79). Springer International Publishing. https://doi.org/10.1007/978-3-030-19642-4_7

Sigrist, R., & Bungener, P. (2008). The first botanical gardens in Geneva (c. 1750–1830) : Private initiative leading science. Studies in the History of Gardens & Designed Landscapes, 28(3-4), 333-350. https://doi.org/10.1080/14601176.2008.10404723

Simoes, R., Camara, G., Queiroz, G., Souza, F., Andrade, P. R., Santos, L., Carvalho, A., & Ferreira, K. (2021). Satellite Image Time Series Analysis for Big Earth Observation Data. Remote Sensing, 13(13), Art. 13. https://doi.org/10.3390/rs13132428

Somvanshi, S. S., & Kumari, M. (2020). Comparative analysis of different vegetation indices with respect to atmospheric particulate pollution using sentinel data. Applied Computing and Geosciences, 7, 100032. https://doi.org/10.1016/j.acags.2020.100032

Talukdar, S., Singha, P., Mahato, S., Shahfahad, Pal, S., Liou, Y.-A., & Rahman, A. (2020). Land-Use Land-Cover Classification by Machine Learning Classifiers for Satellite Observations—A Review. Remote Sensing, 12(7), Art. 7. https://doi.org/10.3390/rs12071135

Vali, A., Comai, S., & Matteucci, M. (2020). Deep Learning for Land Use and Land Cover Classification Based on Hyperspectral and Multispectral Earth Observation Data : A Review. Remote Sensing, 12(15), 2495. https://doi.org/10.3390/rs12152495

Wong, T.-T., & Yeh, P.-Y. (2020). Reliable Accuracy Estimates from k-Fold Cross Validation. IEEE Transactions on Knowledge and Data Engineering, 32(8), 1586-1594. https://doi.org/10.1109/TKDE.2019.2912815

Wulder, M. A., Coops, N. C., Roy, D. P., White, J. C., & Hermosilla, T. (2018). Land cover 2.0. International Journal of Remote Sensing, 39(12), 4254-4284. https://doi.org/10.1080/01431161.2018.1452075

Xue, J., & Su, B. (2017). Significant Remote Sensing Vegetation Indices : A Review of Developments and Applications. Journal of Sensors, 2017, 1-17. https://doi.org/10.1155/2017/1353691

Zioti, F., Ferreira, K. R., Queiroz, G. R., Neves, A. K., Carlos, F. M., Souza, F. C., Santos, L. A., & Simoes, R. E. O. (2022). A platform for land use and land cover data integration and trajectory analysis. International Journal of Applied Earth Observation and Geoinformation, 106, 102655. https://doi.org/10.1016/j.jag.2021.102655

8. Figures et tables