Rapport d'étape

PROJET IDENTIFICATION DES MILIEUX HUMIDES : UNE APPROCHE RÉGIONALE ADAPTÉE À L'ABITIBI SUR LA BASE DES DONNÉES DE TÉLÉDÉTECTION ET LiDAR

NUMÉRO DE L'ENTENTE : DCE21-006

Livrable : Spatialiser les classes de milieux humides identifiés

Présenté à :
M. Jean-Pierre Laniel, directeur général
Ministère de l'Environnement et de la Lutte contre les changements climatiques

Par : Osvaldo Valeria IRF-UQAT, Nicole Fenton IRF-UQAT, Philippe Marchand IRF-UQAT et Louis Imbeau VRECC – UQAT.

Version, 25 février 2022

Table des matières

Objectif général

Le principal objectif de ce projet est de proposer une méthode non supervisée de classification des milieux humides permettant de caractériser et de localiser les différents types de milieux humides à l'aide des données LiDAR et satellites (Landsat 8, Sentinel 1 et Sentinel 2) disponibles et vise particulièrement à l'élaboration (production) des cartes d'identification des MH pour l'ensemble de l'Abitibi.

Territoire d'étude

Voir description plus détaillé du territoire d'étude (Figure 1) dans (plan de travail détaillé, projet identification des milieux humides : une approche régionale adaptée à l'Abitibi sur la base des données de télédétection et lidar, numéro de l'entente : dce21-006, version 1er décembre 2021).

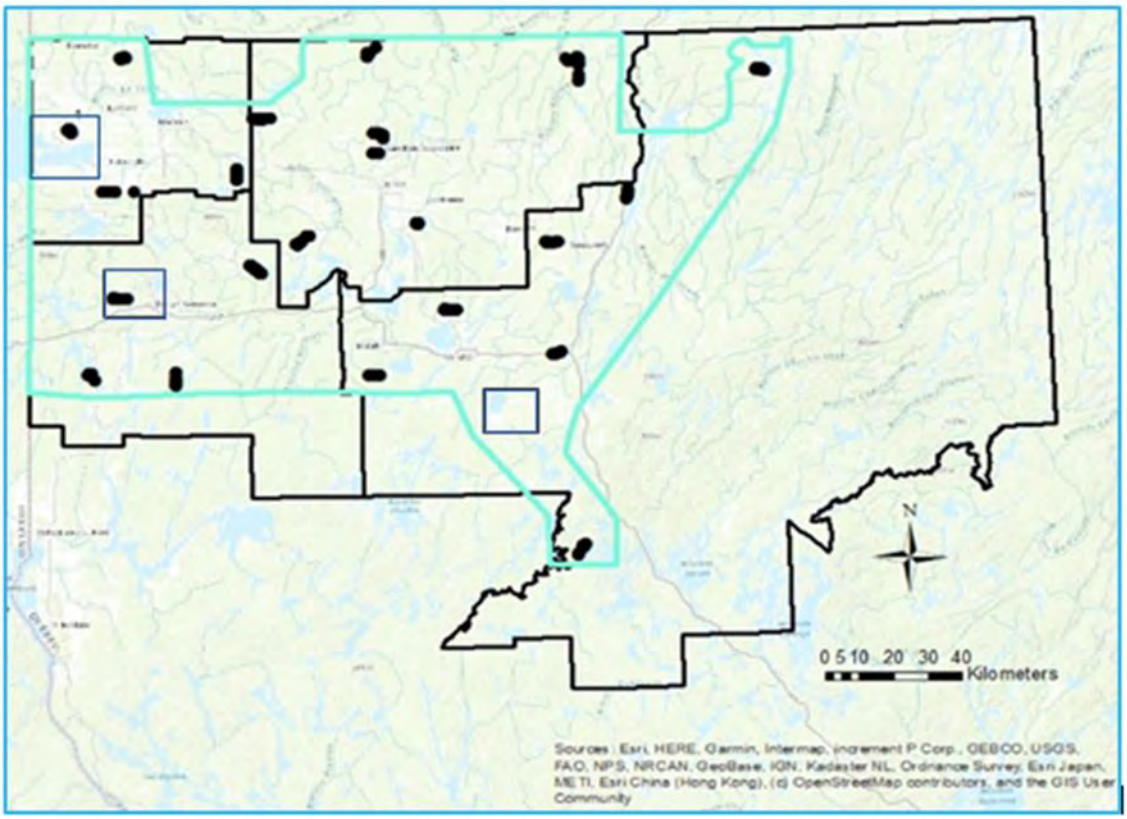


Figure 1 : Carte représentant la localisation des 3 sous-sites d'étude (rectangle noir), une grande portion du territoire (turquoise) et de la région de l'Abitibi délimitée par les lignes noires. Les points noirs correspondent aux données terrain obtenues en 2018 et 2021.

Livrable : Spatialiser les classes de milieux humides identifiés. - cartes avec classes de MH et résolutions spatiales.

Méthodologie

Données d'entrée

Une stagiaire BRPC (Elainie Voyer-Leblanc) a procédé à la correction corrections radiométriques, élimination des nuages via l'acquisition de nouvelles scènes, éliminé l'incohérence de certaines valeurs dans les couches, vérifié la qualité de chacun des descripteurs voir annexe 1 et complété le manque d'information de certaines sections du territoire.

Les données ont été générées pour la superficie presque complète des MRC Abitibi-Ouest, Abitibi, Abitibi-Est et Rouyn-Noranda, mais la classification est actuellement limitée à la zone d'étude déjà délimitée dans les rapports précédents (qui équivaut à un peu plus de la moitié de la superficie des nouvelles données).

Nous utilisons 33 couches matricielles dans la version actuelle de la classification.

Landsat 8 (résolution 30 m, 15 couches)

- Bandes 2, 3, 4 pour le printemps, été et automne (9 couches)
- NDVI pour le printemps, été et automne (3 couches)
- Indice de brillance pour le printemps, été et automne (3 couches)

Radar (9 couches)

- Sentinel 1 (résolution de 10 m) : Bandes VH et VV pour le printemps, été et automne (6 couches)

- Palsar (résolution de 30 m) : Polarisations HH, HV, HVHH pour le printemps (3 couches)

LiDAR (résolution de 5 m, 9 couches)

- Altitude
- Courbure : standard, planiforme et longitudinale (3 couches)
- Exposition : Est-Ouest et Nord-Sud (2 couches) *
- Modèle de hauteur de canopée
- Pente
- TPI

* L'angle d'exposition de la pente, une variable circulaire (0 à 360 degrés), a été transformé en deux variables linéaires (nord-sud et est-ouest) avec une transformation cosinus et sinus, respectivement. Les deux variables sont 0 lorsque la pente est nulle.

Couches générées mais non utilisées actuellement:

- **Sentinel 2** (mêmes couches que Landsat) : Dans la nouvelle version des données, entre 20% (été) et 50% (automne) de la superficie étudiée est manquante en raison de la présence de nuages dans les données Sentinel 2. Lorsque plus d'années de données seront disponibles pour ce satellite, réduisant la superficie où les données sont manquantes, il serait utile de les inclure, car les données ne sont pas redondantes avec Landsat (Sentinel 2 a une meilleure résolution spectrale, notamment).

- **TWI** : Cette couche provenant des données du MFFP a plus de 50% de données manquantes. Elle pourrait être utilisée lorsque plus de données seront disponibles pour la région d'étude.
- **Bande 5 de Landsat (NIR)** : Cette bande est fortement corrélée avec l'indice de brillance. Vu que nous avons déjà le NDVI qui est basée sur la différence entre les bandes NIR et rouge et l'indice de brillance qui est basé sur leur moyenne quadratique, la bande NIR elle-même est redondante.
- **Direction du flux hydrologique** : Ce n'est pas une couche utile pour identifier les milieux humides.
- **Modèle d'épaisseur potentielle de la couche organique** : Cette couche est catégorielle plutôt que numérique ce qui rend plus difficile son utilisation dans une classification hiérarchique avec la distance euclidienne. Aussi, il s'agit d'un modèle dérivé des autres couches de télédétection, elle est fortement corrélée notamment avec la pente.

Approche

Préparation des données matricielles pour la classification

Les 33 couches ont été ramenées à une grille commune correspondant aux données Landsat (résolution de 30 m) en projection UTM (zone 17N). Pour les couches avec une résolution plus fine (5 ou 10 m), nous avons d'abord divisé la grille Landsat pour obtenir une grille de 5 ou 10 m, puis nous avons utilisé la fonction « resample » du package raster dans R pour rapporter les données à cette grille (avec interpolation basée sur la cellule voisine la plus proche), avant d'agréger les données à 30 m en prenant la moyenne.

- Nous avons utilisé la couche «Utilisation du territoire 2018» du MELCC, aussi interpolée à la même grille avec «resample», pour masquer de nos données toutes les cellules correspondant à la catégorie «Aquatique».
- Toutes les couches sont normalisées à une moyenne de 0 et un écart-type de 1 pour placer chaque variable sur la même échelle pour le calcul de la distance euclidienne utilisée dans la classification hiérarchique. Nous utilisons cette mesure de distance et le critère de Ward pour toutes les applications de la classification hiérarchique.

Classification hiérarchique avec connectivité

À une résolution de 30 m, notre région d'étude comporte près de 20 millions de cellules après avoir masqué les données manquantes et les cellules aquatiques. Une classification hiérarchique globale devient difficile à réaliser en pratique avec >100 000 points, car la classification hiérarchique globale requiert de consulter les distances entre chaque paire de points pour déterminer quels points ou quels groupes doivent être combinés à chaque étape.

Comme première étape, nous avons donc réalisé une classification hiérarchique en spécifiant une matrice de connectivité. Dans ce cas, à chaque étape de la classification, chaque cellule ou groupe ne peut être combiné qu'à une cellule ou un groupe de cellules voisin (donc initialement, chaque cellule n'a que quatre voisins). Cette première étape crée des groupes de cellules voisines ou «patches» qui à chaque étape de la classification ont un niveau comparable d'homogénéité. Nous utilisons la courbe montrant la variance intra-patch en fonction du nombre de patches pour identifier

différents nombres de patchs qui sont «optimaux» dans un certain sens (ex. les endroits dans la courbe où la variance diminue rapidement avant d'arriver à ce nombre et plus lentement ensuite).

Après avoir choisi un nombre de patches, nous faisons la moyenne des indices à l'intérieur de chaque patch. Nous pouvons ainsi réduire nos 20 millions de points à un nombre plus raisonnable de polygones pour réaliser une classification hiérarchique globale. Cela simplifie aussi la carte résultante tout en respectant le fait que certaines portions du territoire sont plus homogènes que d'autres. Autrement dit, la résolution spatiale « effective » de la classification sera plus fine aux endroits où les indices varient à une échelle spatiale plus fine, car les patchs choisies y seront plus petites.

Classification hiérarchique globale

La classification hiérarchique globale est réalisée à partir de ces données moyennes au niveau des patches. Si le nombre de patches est trop grand pour l'algorithme de classification hiérarchique de base, nous utilisons l'algorithme BIRCH qui est moins coûteux en mémoire et en temps de calcul. Celui-ci produit un premier ensemble de groupes sur la base d'un seuil de distance et calcule la moyenne des indices dans les groupes, avant de classifier ces groupes avec l'algorithme de base.

Pour la classification globale, nous comparons la variance intra-classe pour un nombre de classes k entre 2 et 50. La variance intra-classe diminue toujours avec le nombre de classes, donc nous cherchons les valeurs de k pour qui sont des « coudes » dans la courbe de la variance intra-classe vs. k,

donc lorsque la variance diminue plus rapidement quand on ajoute des classes jusqu'à k que lorsqu'on en ajoute au-delà de k.

Pour la classification, nous utilisons le langage Python plutôt que R, en particulier le package scikit-learn et ses fonctions «AgglomerativeClustering» pour la classification hiérarchique (avec connectivité spatiale ou non), «Birch» pour l'algorithme Birch, ainsi que «grid_to_graph» pour produire la matrice de connectivité à partir de la couche matricielle montrant la zone d'étude (avec un masque pour indiquer les cellules incluses / exclues). Nous utilisons aussi les serveurs de Calcul Canada pour le stockage des données complètes et la réalisation des étapes les plus coûteuses en temps de calcul.

Résultats et discussion

Classification hiérarchique avec connectivité (segmentation)

Pour cette première étape, nous avons réalisé la classification et calculé la variance intra-patch pour une séquence logarithmique du nombre de patches variant entre 1000 et 1 000 000. Cette variance V est égale à la somme des distances au carré entre les indices de chaque cellule et la valeur moyenne des indices pour la patch contenant cette cellule, divisée par (n – k), où n est le nombre de cellules et k le nombre de patches. Vu cette variation de k sur plusieurs ordres de grandeur, nous considérons le taux de réduction de la variance sur une échelle logarithmique, i.e. la pente de log V vs. log k, ce qui est équivalent au % de changement de V pour une augmentation de 1 % de k. Comme nous pouvons voir sur la Fig. 2, ce taux devient plus négatif avec k (donc la réduction relative accélère), mais on peut aussi identifier des optimums locaux (plus difficiles à voir sur le graphique

pour des k élevés) qui offrent des choix potentiels pour le nombre de patches.

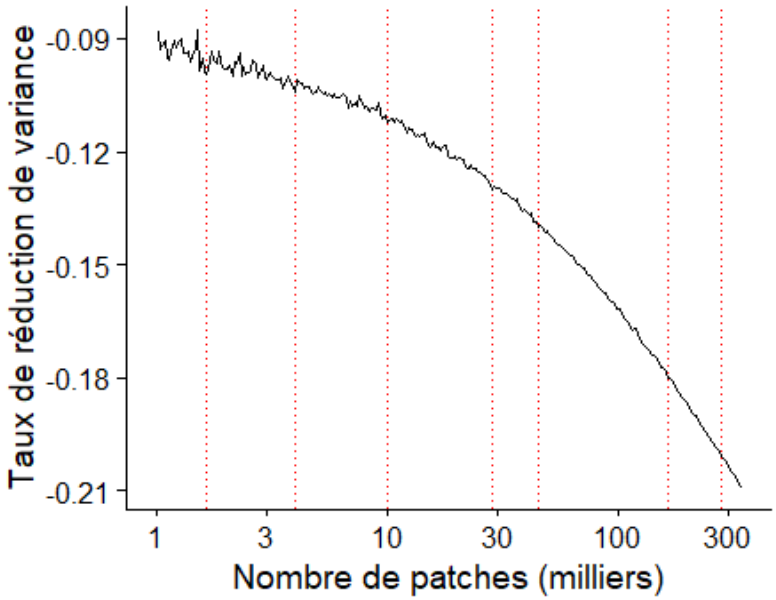


Figure 2: Taux de réduction de la variance intra-patch (% de réduction de la variance pour une augmentation de 1% du nombre de patches) en fonction du nombre de patches. Les lignes pointillées rouges représentent des optimums locaux.

Pour la suite, nous retenons les segmentations en 28 500 et en 278 500 patches (valeurs arrondies à la centaine près des optimums ci-dessus). Celles-ci correspondent à une taille moyenne d'environ 701 et 72 cellules / patch, ou de 63 et 6 hectares / patch, respectivement. La segmentation en 28 500 patches correspond à la résolution minimale que nous considérons ici, tandis que celle à 278 500 patches vise à tester la différence obtenue en augmentant la résolution par un facteur d'environ 10.

Avant de poursuivre, mentionnons certaines limites pratiques de cette méthode. Tout d'abord, elle requiert une connectivité complète des cellules non-masquées. Il faut donc éliminer les îles, mais nous avons aussi perdu certaines portions en marge de la région étudiée qui étaient séparées du reste par le réseau hydrographique (voir Fig. 4). Ce problème peut être résolu en ajoutant des «ponts» de cellules non-masquées sur les rivières au besoin.

Aussi, avec 20 millions de cellules à l'entrée, la classification requiert environ 50 heures de calcul et 42 Go de mémoire vive. La mémoire vive nécessaire augmente linéairement avec le nombre de cellules, mais le temps de calcul augmente de façon quadratique, si bien que d'étendre la méthode aux 4 MRC au complet (38 millions de cellules) pourrait demander environ 180 heures de calcul. Ce problème peut être résolu en divisant la zone d'étude en sections avant la segmentation, en prenant soin de choisir un nombre de patches amenant à une variance comparable entre chaque section et en acceptant des bordures artificielles aux limites des sections.

Bien sûr, il serait aussi possible de choisir un autre algorithme pour la segmentation de patches homogènes, tout en gardant la classification hiérarchique à la prochaine étape.

Classification hiérarchique globale

Nous utilisons d'abord la segmentation en 28 500 patches pour effectuer la classification hiérarchique, puis nous comparons le changement de la variance intra-classe V en variant le nombre de classes k de 2 à 50 (Fig. 3).

Dans ce cas, la réduction de variance (i.e. le gain en homogénéité des classes) lié à l'ajout d'une classe diminue lorsque k augmente. Il est utile donc de choisir des valeurs de k précédant une chute rapide de ce gain entre k et k + 1. Selon les résultats en Fig. 3, nous choisissons donc des k possibles de 8, 18, 33 et 39 classes.

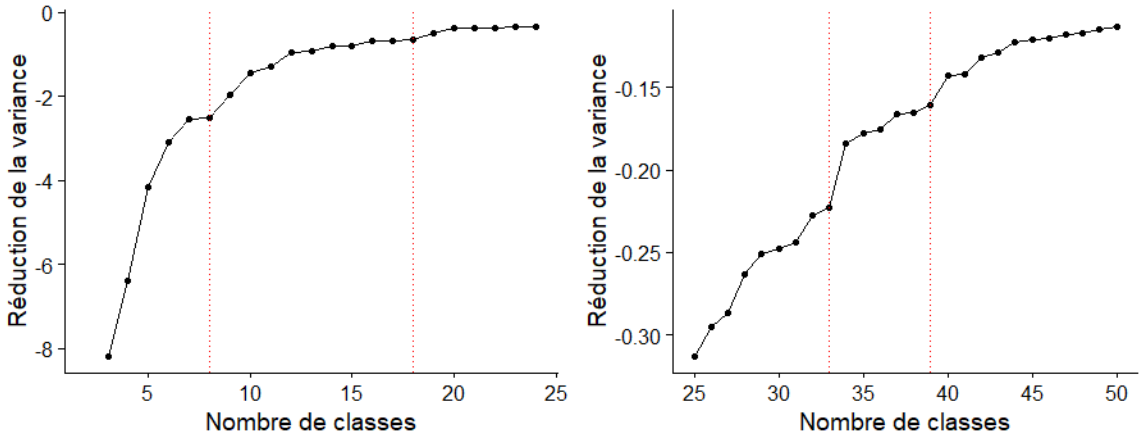


Figure 3: Réduction de la variance intra-classe pour l'ajout d'une classe entre 2 et 50 classes, à partir de la segmentation en 28 500 patches. Les lignes pointillées rouges représentent les différents choix de k considérés ici (8, 18, 33 et 39 classes).

Nous avons ensuite créé des couches vectorielles (polygones basés sur les patches) et matricielles pour la classification avec ces 4 choix de k. Afin de décrire la composition des différentes classes, nous avons fait une tabulation croisée des couches matricielles avec la couche d'utilisation du territoire du MELCC, autant au niveau des grandes catégories (ex. : forêt, agriculture, milieux humides) que des classes détaillées.

Le Tableau 1 montre la fraction de l'aire d'étude occupée et la composition avec k = 8 classes et la distribution de ces classes dans la région est illustrée à la Figure 4. La grande majorité de la région d'étude se trouve dans la classe 2 (82.8%), composée principalement de forêts et de milieux humides et la classe 3 (15.8%), composée principalement de milieux agricoles et humides. La classe 1 (0.7%) est composée de forêts et de sols nus, tandis que les classes 4 à 8 (0.7%) sont des milieux anthropiques parfois mélangés à des forêts (classe 6) ou milieux humides (classe 8). Dans la division en 33 classes que nous utiliserons plus loin, 16 des 33 classes proviennent des classes 4 à 8 identifiées ici. La subdivision importante de ces classes très minoritaires pourrait être due au fait que les milieux anthropiques présentent une grande variation au niveau des indices considérés et sont donc divisés plus tôt dans la classification hiérarchique.

Tableau 1. Composition des classes pour la classification en 8 classes.

		Composition de la classe (% par catégorie)						
Classe	**% des cellules**	Agricole	Anthro.	Coupe/ régén.	Forêt	Humide	Sol nu/ lande	**Classes d'utilisation détaillées (3 principales)**
1	0.7%	0%	1%	1%	80%	1%	17%	Forêt mixte (46%), forêt de conifères (19%), sol nu (17%)
2	82.8%	1%	1%	7%	48%	43%	1%	Forêt mixte (19%), forêt de feuillus (14%), tourbière (8%)
3	15.8%	36%	6%	15%	10%	33%	0%	Tourbière (28%), Culture pérenne (19%), Agri. Indif. (13%)
4	0.6%	4%	86%	1%	6%	2%	1%	Mine et déchet (33%), zone dével. (28%), carrière (11%)
5	0.1%	4%	92%	1%	2%	1%	0%	Mine et déchet (34%), carrière (32%), zone développée (16%)
6	0.0%	0%	40%	1%	49%	2%	8%	Mine et déchet (37%), forêt mixte (32%), forêt de conifères (10%)
7	0.0%	3%	89%	0%	5%	3%	0%	Zone dével. (40%), mine et déchet (25%), zone industrielle (14%)
8	0.1%	0%	41%	0%	8%	49%	0%	Marais (39%), mine et déchet (31%)

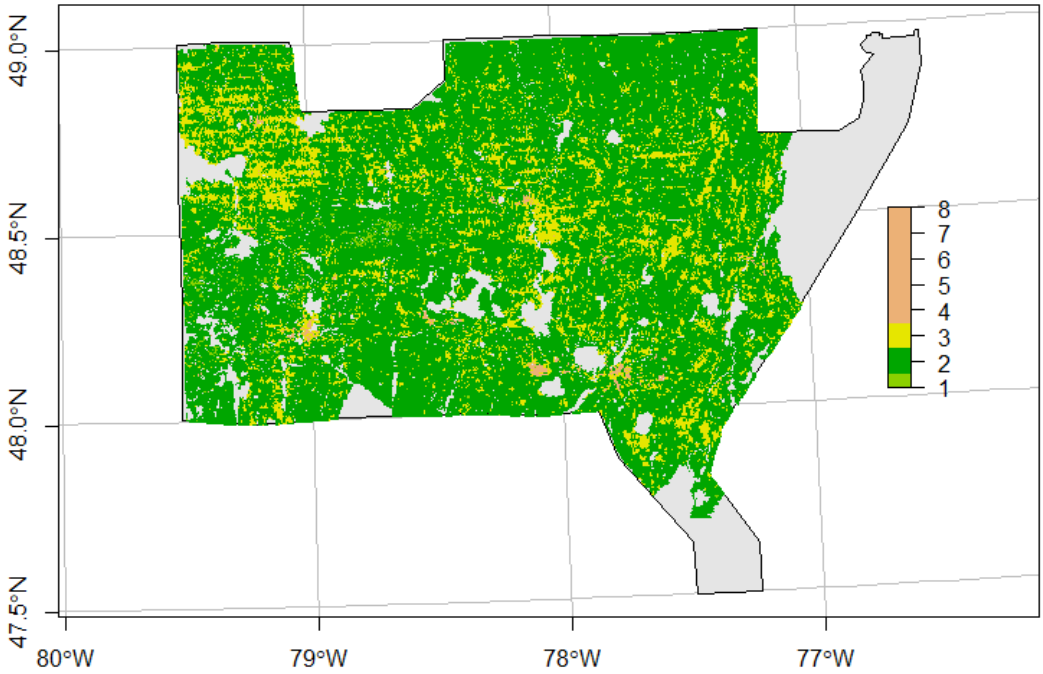


Figure 4. Carte des classes dans la région d'étude pour la classification à 8 classes. Une seule couleur a été attribuée aux classes 4 à 8 (majoritairement anthropiques). Les zones grises représentent les cellules masquées (données manquantes, réseau hydrologique) ou les parties de l'aire d'étude non-connectées en raison de l'eau (nord-est et sud-est).

Nous regardons donc maintenant la sous-classification en 33 classes, mais seulement pour les sous-classes provenant des classes 2 et 3 de la classification en 8 classes (ces deux classes comptent pour 99.9% de la superficie des milieux humides selon la carte d'utilisation du territoire, le reste étant presque tout situé dans la classe 8).

Tableau 2. Composition des classes pour la classification en 33 classes, pour les sous-classes des classes 2 et 3 de la classification à 8 classes. Le % des cellules réfère à l'aire d'étude totale.

Classe 2 sur 8

		Composition de la classe (% par catégorie)						
Classe	**% des cellules**	Agricole	Anthro.	Coupe/ régén.	Forêt	Humide	Sol nu/ lande	**Classes d'utilisation détaillées (3 principales)**
13	19.8%	0%	1%	3%	69%	26%	1%	Forêt mixte (36%), forêt de feuillus (19%), forêt de conifères (9%)
27	25.7%	0%	0%	2%	38%	59%	0%	Forêt de conifères / marécage (26%), forêt de conifères (20%), forêt de conifères / tourbière (14%)
8	17.1%	0%	1%	17%	19%	63%	0%	Tourbière (25%), Forêt de conifères / tourbière ombr. (12%), coupe (9%)
26	0.5%	2%	4%	1%	11%	83%	0%	Tourbière (41%), Marais (26%), Arbuste / Tourbière (9%)
24	18.1%	3%	1%	9%	63%	23%	1%	Forêt de feuillus (33%), forêt mixte (18%), coupe (7%)
4	1.7%	0%	0%	2%	84%	12%	2%	Forêt mixte (49%), forêt de feuillus (19%), forêt de conifères (14%)

Classe 3 sur 8

		Composition de la classe (% par catégorie)						
Classe	**% des cellules**	Agricole	Anthro.	Coupe/ régén.	Forêt	Humide	Sol nu/ lande	**Classes d'utilisation détaillées (3 principales)**
9	3.6%	2%	9%	50%	19%	18%	1%	Coupe (27%), plantation (16%), tourbière (7%)
19	4.4%	1%	0%	2%	0%	97%	0%	Tourbière (94%)
20	2.1%	40%	4%	13%	30%	14%	0%	Culture pérenne (25%), forêt de feuillus (14%), Agri. Indiff. (11%)
22	4.4%	92%	2%	0%	4%	2%	0%	Culture pérenne (49%), Agri. Indif. (38%)
10	0.7%	47%	27%	13%	10%	3%	1%	Avoine (14%), Agri. Indif (12%), Culture pérenne (12%), zone dével. (12%)
32	0.6%	83%	12%	0%	4%	1%	0%	Avoine (27%), agri. Indif. (21%), culture pérenne (20%)
21	0.0%	5%	45%	26%	17%	5%	2%	Zone dével. (23%), plantation (18%), route (11%)

Les sous-classes de la classe 2 (forêt / milieux humides) incluent des classes principalement forestières (4, 13 et 24 : de 63% à 84% de forêts), des classes dominées par la tourbière (8 et 26 : de 63% à 83% de milieux humides) et une classe (27) composée de forêts de conifères mélangées à des marécages et milieux humides (Tableau 2, haut).

Les sous-classes de la classe 3 (agricole / milieux humides) incluent des classes fortement agricoles (22 et 32 : 83% à 92% agricole), une classe de tourbières (19 : 97% humide), une classe dominée par des coupes, régénérations et plantations (9 : 50% coupe et régénération) et des classes mixtes (20 : agricole et forêt, 10 : agricole et développement, 21 : développement et plantations) (Tableau 2, bas).

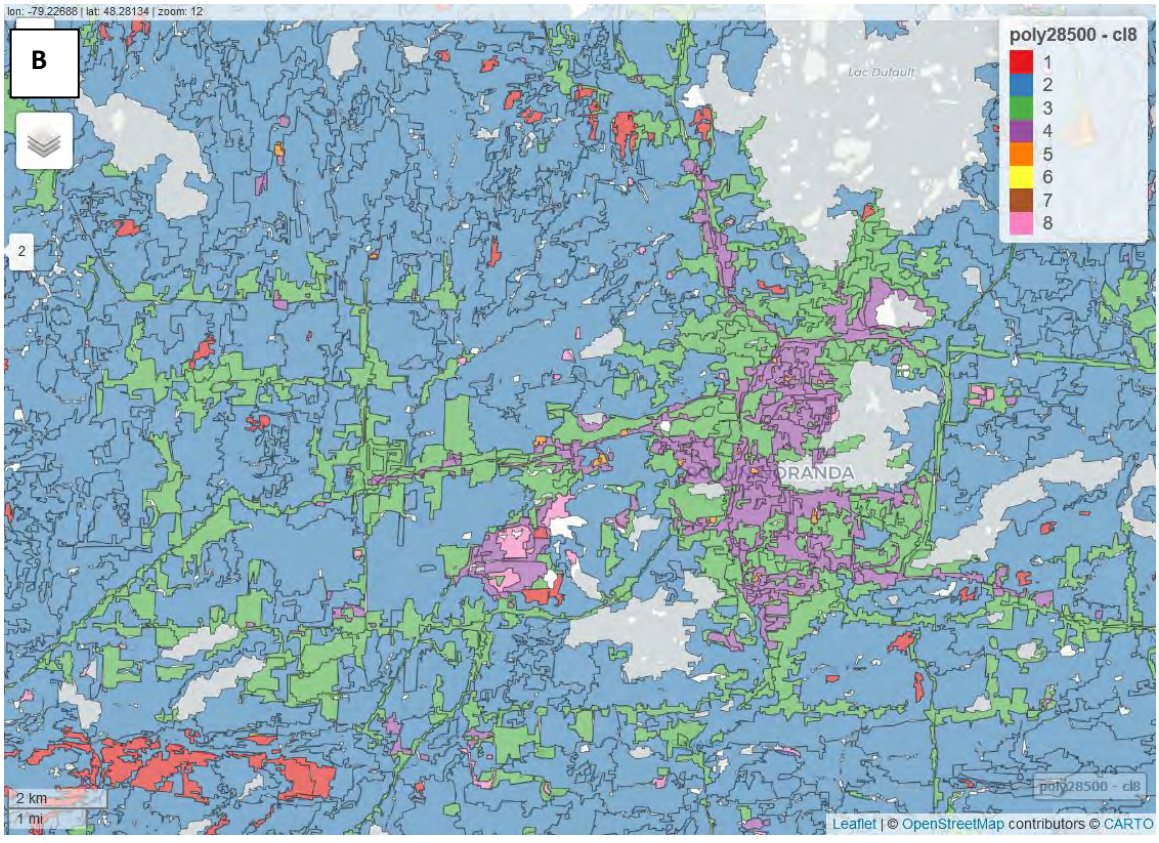
Nous ne présentons pas les sous-classes des autres classes du Tableau 1 pour la classification en 33 classes, car elles sont moins liées à ce projet. Par exemple, les sous-classes des classes anthropiques pourront différencier les mines des zones résidentielles et industrielles. Notons que la classe 8 (anthropique et marais) ne se trouve pas davantage subdivisée en passant à 33 ou 39 classes. Les classes présentées au Tableau 2 sont aussi essentiellement les mêmes en passant à 39 classes (seule la classe forestière 4 est divisée en deux).

Afin d'illustrer de façon plus concrète les classes, la Figure 5 plus loin montre un gros plan des différentes classifications (8 classes, ainsi que les sous-classes des classes 2 et 3 avec 33 classes) pour le secteur des environs de Rouyn-Noranda.

Segmentation en 278 500 patches

Pour la classification basée sur la segmentation en 278 500 patches, il n'était malheureusement pas possible d'utiliser l'algorithme de classification hiérarchique de base (cela aurait requis notamment 600 Go de mémoire vive). Nous avons donc utilisé l'algorithme Birch mentionné dans les méthodes avec un seuil de 2 pour le rayon des groupes, ce qui a mené au regroupement des 278 500 patches en près de 56 000 groupes, qui ont ensuite été classifiés avec l'algorithme de base. Malheureusement, cette approche produit des classes beaucoup moins équilibrées : avec 8 classes, 93% du territoire se trouve dans une seule classe; avec 32 classes, 62% du territoire se trouve dans 1 classe et 83% dans 2 classes). Dans ce cas, nous jugeons que la classification produite par la segmentation moins fine est préférable. Il reste à voir si des ajustements à cet algorithme ou un autre algorithme permettraient d'obtenir des meilleurs résultats avec une segmentation plus fine.


A
2 km
1 mi
poly28500 - cl8
Leaflet | Tiles © Esri — Source: Esri, i-cubed, USDA, USGS, AEX, GeoEye, Getmapping, Aerogrid, IGN, IGP, UPR-EGP, and the GIS User Community


lon: -79.22688 | lat: 48.28134 | zoom: 12
B
poly28500 - cl8
1
2
3
4
5
6
7
8
Lac Dufault
2 km
1 mi
Leaflet | © OpenStreetMap contributors © CARTO


C
poly1 - cl33
13
24
26
27
4
8
Lac Dufault
ROUYN-NORANDA
2 km
1 mi
poly1 - cl33
Leaflet | © OpenStreetMap contributors © CARTO

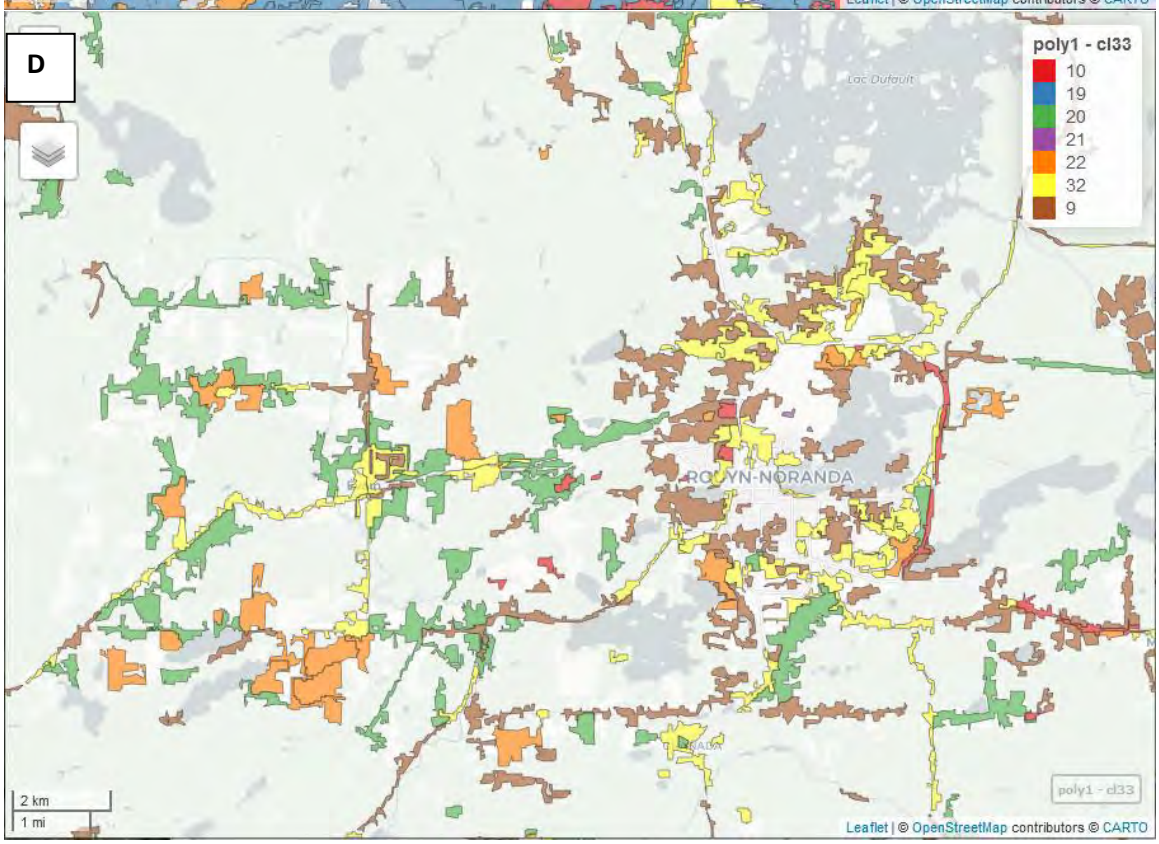

D
poly1 - cl33
10
19
20
21
22
32
9
Lac Dufault
ROUYN-NORANDA
2 km
1 mi
poly1 - cl33
Leaflet | © OpenStreetMap contributors © CARTO

Figure 5: (A) Image satellite des environs de Rouyn-Noranda. (B) Classification en 8 classes pour la même image. (C) Classification en 33 classes, sous-classes de la classe 2 sur 8. (D) Classification en 33 classes, sous-classes de la classe 3 sur 8.