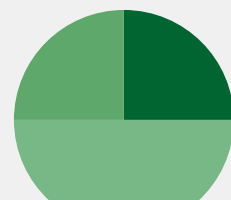




Actualités OFS



02 Espace et environnement

Neuchâtel, septembre 2023

Recours à l'intelligence artificielle pour la production de la statistique de la superficie de l'OFS

Fonctionnement, capacités et limites de performance de l'outil d'apprentissage profond ADELE

Depuis 2022, une intelligence artificielle contribue pour la première fois à la production d'une statistique publique en Suisse. En effet, l'outil d'apprentissage profond ADELE permet de partiellement automatiser l'interprétation des images aériennes de la statistique de la superficie, une tâche particulièrement chronophage. ADELE montre de manière exemplaire les nouvelles possibilités que l'intelligence artificielle offre à la statistique publique et comment son utilisation peut se concrétiser dans la pratique. Cependant, cet outil met également en évidence la difficulté d'associer le fonctionnement de l'intelligence artificielle et l'exigence de produire des statistiques précises livrant des séries temporelles fiables à long terme.

1 Statistique de la superficie de l'OFS: contenu et but

La statistique de la superficie établie par l'OFS fournit des informations sur l'état et l'évolution de l'utilisation et de la couverture du sol en Suisse. Elle constitue ainsi un instrument indispensable pour observer le territoire sur le long terme. Les résultats de cette statistique permettent en particulier d'évaluer si et dans quelle mesure l'évolution de l'utilisation du sol répond aux objectifs de la politique de développement territorial en Suisse et à la volonté d'une utilisation économe du sol en tant que ressource.

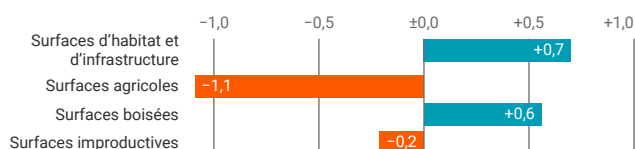
Actuellement, les résultats de quatre relevés recourant à une méthodologie homogène sont disponibles: 1985, 1997, 2009 et 2018 (voir encadré à droite), ces années correspondant chacune à la clôture d'une période de relevé pluriannuelle. Une nouvelle campagne a été lancée en 2021. De premiers résultats partiels (pour les cantons de Genève et de Vaud) ont été publiés durant l'été 2023. Les résultats pour l'ensemble de la Suisse sont attendus pour fin 2027.

Résultats et documents méthodologiques détaillés relatifs à la statistique de la superficie

Le graphique ci-dessous montre un exemple d'analyse basée sur les données de la statistique de la superficie: entre 1985 et 2018, une surface équivalente à 0,7 m² a été construite chaque seconde en Suisse. Autrement dit, cette surface a été transformée en surface d'habitat et d'infrastructure. Vous trouverez d'autres graphiques, tableaux détaillés, textes explicatifs et documents méthodologiques relatifs à la statistique de la superficie sur le site Web de l'OFS www.landuse-stat.admin.ch

Évolution de l'utilisation du sol entre 1985 et 2018, en m² par seconde

Gains et pertes par domaine principal



Source: OFS – Statistique de la superficie (AREA)

© OFS 2023

2 Sources de données et processus de relevé traditionnel

Point de départ: les images aériennes

La statistique de la superficie se base sur des images aériennes numériques à haute résolution couvrant la totalité du territoire suisse qui sont mises à la disposition de l'OFS par l'Office fédéral de topographie. Ces images sont prises depuis des avions spécialement équipés qui survolent la Suisse à intervalles réguliers. Leur but premier est d'établir et de mettre à jour les cartes nationales suisses; leur utilisation pour la statistique de la superficie représente l'une des nombreuses utilisations secondaires.

Plus de quatre millions de points d'échantillonnage

Pour les besoins de la statistique de la superficie, un réseau de points d'échantillonnage équidistants de 100 mètres est superposé aux images aériennes. La statistique s'appuie ainsi sur 4,1 millions de points, qui sont géoréférencés et permanents: leur position reste exactement la même d'un relevé à l'autre. Pour chacun de ces points d'échantillonnage, l'utilisation et la couverture du sol sont déterminés lors de chaque relevé sur la base des images les plus récentes. Finalement, les résultats sont extrapolés à l'ensemble du territoire.

Interprétation visuelle des images aériennes

Traditionnellement, c'est une personne qui détermine l'utilisation et la couverture du sol aux points d'échantillonnage: des interprètes spécialement formés examinent les images aériennes sur un écran stéréoscopique offrant une restitution 3D et attribuent à chaque point l'une des catégories d'utilisation et de couverture prédéfinies. Pour ce faire, les experts examinent sur les images les environs immédiats du point à interpréter: se trouve-t-il sur

un bâtiment? Si c'est le cas: de quel type de bâtiment s'agit-il? Un bâtiment d'habitation, un bâtiment industriel ou un bâtiment agricole? La statistique de la superficie distingue 46 types d'utilisation du sol et 27 types de couverture du sol (p. ex. «bâtiments» ou «buissons»). Chacune des catégories d'utilisation du sol peut être attribuée à l'une des quatre catégories d'utilisation principale, à savoir: surfaces d'habitat et d'infrastructure, surfaces agricoles, surfaces boisées et surfaces improductives.

Informations auxiliaires aidant à l'interprétation

Pour faciliter la classification des points d'échantillonnage, les interprètes peuvent consulter des informations auxiliaires sur deux écrans supplémentaires. Ainsi, par exemple, le registre fédéral des bâtiments et des logements permet de distinguer les maisons individuelles des immeubles résidentiels, le modèle topographique du paysage de la Suisse renseigne sur la largeur des routes et l'inventaire des surfaces agricoles utiles aide à identifier des cultures spécifiques comme les vergers ou les vignobles. En dépit de ces informations et d'autres données disponibles, l'interprète n'est pas toujours en mesure de déterminer avec certitude à l'écran l'utilisation ou la couverture du sol exacte de certains points d'échantillonnage. Dans ce cas, une visite sur le terrain s'impose.



Interprétation des images aériennes au moyen d'un écran stéréoscopique 3D. L'interprète peut consulter des informations auxiliaires sur les écrans latéraux.



Image aérienne montrant des points d'échantillonnage et les codes d'utilisation du sol constatés. Le code 106/11 signifie «surfaces compactées sur aires de bâtiments publics».

3 Recours à l'intelligence artificielle pour la statistique de la superficie: potentiels et défis

En dépit des données disponibles et de l'utilisation d'outils d'assistance spécifiques (p. ex. contrôle automatique de la plausibilité durant la saisie), l'interprétation visuelle des images aériennes représente une tâche laborieuse au vu des 4,1 millions de points d'échantillonnage à traiter. Dans ce contexte, il était logique que l'on recherche une solution pour automatiser ces travaux à l'aide de logiciels de reconnaissance d'images. Le champ des possibilités dans le domaine de la reconnaissance automatique d'images s'est considérablement étendu ces dernières années grâce au recours à l'intelligence artificielle sous forme d'apprentissage automatique (machine learning, voir explication plus bas).

Le recours à l'intelligence artificielle (IA) pour la statistique de la superficie semble d'autant plus séduisant que d'énormes quantités de données sont d'ores et déjà disponibles pour entraîner les modèles utilisés par ces logiciels: dans le cadre des précédents relevés de la statistique de la superficie, des millions d'images aériennes ont été interprétées par des experts et l'ensemble de ces images, avec leur interprétation, ont été conservées sous forme numérique. Autrement dit, le traitement manuel des données d'apprentissage (qui suppose une myriade de clics de souris) nécessaire à l'entraînement de la reconnaissance automatique des images, a déjà été effectué.

Se basant sur le potentiel présumé, l'OFS a réalisé dès 2017 une étude de faisabilité visant à clarifier si et dans quelle mesure l'interprétation des images aériennes de la statistique de la superficie pouvait être automatisée, du moins partiellement, grâce au recours à l'intelligence artificielle. Ceci également dans le but de réduire à l'avenir les intervalles entre les relevés. Si les potentiels pour la statistique de la superficie se sont vite révélés considérables, il en a été de même pour les défis:

- Les exigences de précision quant à la reconnaissance automatique d'images sont ici incomparablement plus élevées que pour de nombreuses autres utilisations, avant tout commerciales.
- L'immense quantité de points d'échantillonnage à déterminer exige d'importantes capacités de calcul et beaucoup de temps. En estimant que la détermination d'un point dure 1 seconde, le calcul pour 4,1 millions de points requiert pas moins de 47 jours!
- Sur les images, certaines catégories d'utilisation et de couverture du sol se ressemblent beaucoup. Il s'avère alors indispensable de recourir à des informations auxiliaires pour déterminer la bonne catégorie. Si l'on veut mettre à profit l'intelligence artificielle, celle-ci doit donc être capable de traiter d'autres informations, en plus des images, ce qui nécessite une architecture de programme plus complexe, augmentant d'autant la charge de calcul déjà très importante.
- La fréquence de répartition varie fortement entre les différentes catégories d'utilisation ou de couverture du sol. Certaines catégories sont si rares qu'elles n'offrent pas suffisamment d'informations de référence pour permettre à l'intelligence artificielle d'effectuer un «apprentissage» optimal. En outre, le risque existe que les catégories rares soient même

complètement ignorées par les algorithmes d'apprentissage automatique. Il apparaît dès lors nécessaire d'équilibrer la fréquence des catégories en appliquant des méthodes statistiques.

4 L'outil d'IA ADELE: structure et fonctionnement théorique

Précisons d'emblée qu'une automatisation intégrale de l'interprétation des images aériennes à l'aide de l'intelligence artificielle ne sera guère possible à court comme à moyen terme, notamment à cause des difficultés décrites plus haut. L'OFS a néanmoins décidé de développer un outil d'IA en collaboration avec des partenaires compétents. En effet, même une automatisation partielle des travaux d'interprétation assurera un gain de temps considérable. Étant donné que le module principal de cet outil utilise la méthode de l'apprentissage profond («deep learning»), il a reçu l'acronyme «ADELE» pour «**A**realstatistik **DE**ep **L**Earning».

Les paragraphes suivants décrivent le principe de la détermination automatique de l'occupation et de l'utilisation du sol par ADELE, indépendamment des limites pratiques de l'outil. Il est ensuite expliqué comment ADELE est associé dans la pratique avec l'interprétation des images par des experts, de sorte à accélérer la production de la statistique de la superficie, sans péjorer ni la qualité ni la fiabilité des résultats.

4.1 Élément central de l'outil ADELE: un module d'apprentissage profond permettant l'interprétation «visuelle» des images aériennes

Ainsi que décrit plus haut, l'utilisation et l'occupation du sol aux points d'échantillonnage sont déterminées par les interprètes humains principalement sur la base des informations contenues dans les images aériennes: l'expert procède à une interprétation visuelle. Il en va de même avec ADELE: la compétence centrale de l'outil consiste à examiner les mêmes images aériennes que celles qui étaient interprétées jusqu'à présent par l'humain et à repérer, autour des points d'échantillonnage, des motifs permettant d'attribuer chacun des points à l'une des 46 catégories d'utilisation du sol et à l'une des 27 catégories d'occupation du sol de la statistique de la superficie. Concrètement, un extrait d'image correspondant à une surface de 50 x 50 mètres est pris en compte pour chaque point d'échantillonnage, ce qui représente une image composée de 200 x 200 pixels, soit 40 000 pixels de trois couleurs. Les valeurs correspondantes de chaque couleur constituent les données de base pour la classification de l'image par ADELE.

Mais comment cette interprétation fonctionne-t-elle? La réponse à cette question est simple et difficile à la fois: pour déterminer l'utilisation et l'occupation du sol, ADELE applique un modèle statistique constitué d'un très grand nombre de paramètres que l'outil ajuste lui-même. En effet, ADELE n'a pas été pourvu par des programmeurs humains de règles formelles telles que «couleur blanche = glacier». Car il serait tout bonnement impossible de programmer manuellement l'ensemble des règles

Apprentissage profond (deep learning)

L'apprentissage profond (de l'anglais «deep learning») est le nom donné à un certain type d'apprentissage automatique et donc à un sous-ensemble du vaste domaine de l'intelligence artificielle. Les principes de base de l'apprentissage profond sont connus depuis des décennies. Pendant longtemps toutefois, l'application effective de cette méthode a été entravée par des capacités de calcul limitées et un manque de données d'entrée disponibles pour le processus d'apprentissage. Depuis qu'il a été possible de surmonter ces obstacles, l'apprentissage profond connaît un essor considérable: il permet d'obtenir, notamment dans la reconnaissance automatique d'images et de la parole, des résultats encore inconcevables il y a quelques années seulement.

Le principe fondamental de l'apprentissage profond s'inspire du fonctionnement du cerveau humain, raison pour laquelle les algorithmes correspondants sont appelés «réseaux neuronaux artificiels». Mais le matériel utilisé est celui d'un processeur d'ordinateur normal.

Mode de fonctionnement

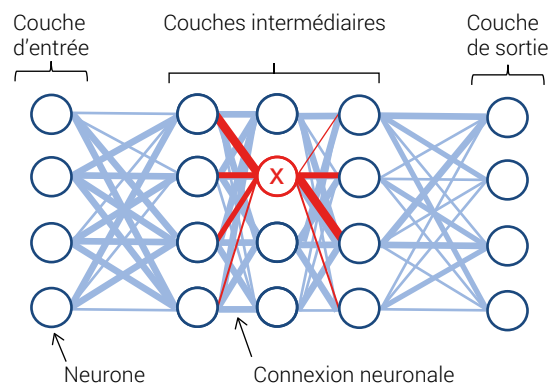
Les réseaux neuronaux artificiels de l'apprentissage profond se composent d'une grande quantité de neurones simulés (cellules nerveuses artificielles) qui sont organisés en couches et qui échangent entre eux des signaux de charge (G2). La première couche est appelée couche d'entrée. C'est là que sont stockées les données d'entrée qui doivent être examinées par le réseau de neurones artificiels. Dans le cas de l'interprétation des images aériennes de la statistique de superficie, la couche d'entrée correspond aux 120 000 valeurs de couleur pour les 40 000 pixels de la partie d'image examinée (voir chapitre 4.1). Chaque valeur de couleur mesurée est donc utilisée comme signal de sortie d'un neurone de la couche d'entrée.

Le traitement des données à proprement parler a lieu dans les couches *intermédiaires* du réseau. Le nombre de ces couches peut être très élevé et le réseau peut donc être «profond», d'où l'appellation «apprentissage profond». Les réseaux modernes d'apprentissage peuvent en outre présenter des structures de convolution compliquées, qui augmentent encore leur performance. C'est le cas également pour le type de «réseau neuronal convolutif (convolutional neural network)» utilisé pour le relevé de la statistique de la superficie.

Un neurone «X» donné se trouvant dans l'une des couches intermédiaires reçoit systématiquement des signaux des neurones de la couche située en amont. En fonction de l'intensité de «l'émission» de ces derniers et de leur connexion avec le neurone X, la somme des signaux arrivant à celui-ci est plus ou moins forte. Le neurone X amplifie ou atténue le signal de charge reçu avant de le transmettre aux neurones de la couche suivante. Le signal est alors fractionné: selon la force de la connexion, un neurone donné de la couche suivante reçoit une part plus ou moins grande du signal de charge émis par le neurone X. Le point crucial est que la force de chaque

Réseau de neurones artificiels avec trois couches intermédiaires

G2



Source: OFS – Statistique de la superficie (AREA)

© OFS 2023

connexion au sein du réseau neuronal artificiel peut varier séparément, tout comme la «puissance d'amplification ou d'atténuation du signal» de chaque neurone. L'outil comprend donc d'innombrables petits «leviers» permettant de modifier le flux des signaux à travers les différentes couches du réseau et de déterminer quelles parties du signal émis par la couche d'entrée parviennent à un neurone donné de la dernière couche. Cette dernière couche est appelée la couche de sortie. Dans le cas de la statistique de la superficie, ses neurones correspondent aux différentes catégories d'utilisation ou d'occupation du sol. Elle se compose donc de 46 neurones dans le réseau servant à déterminer l'utilisation du sol, et de 27 neurones dans celui servant à déterminer la couverture du sol. Le but est alors de paramétrer la puissance d'amplification ou d'atténuation des neurones des couches intermédiaires ainsi que la force des connexions entre ces neurones de telle sorte que, pour chaque image aérienne en entrée, le neurone de la couche de sortie qui reçoit le signal le plus fort soit celui qui correspond à la bonne catégorie d'utilisation ou de couverture du sol.

Si, par exemple, le réseau neuronal reçoit en entrée un extrait d'une image aérienne avec un point d'échantillonnage se trouvant sur un lac, seul le neurone dans la couche de sortie représentant la catégorie d'utilisation «Lacs» devrait, autant que possible, recevoir un signal et donc être «activé».

Processus d'apprentissage

Les paramètres définis dans les couches intermédiaires contiennent donc les «règles» selon lesquelles le réseau interprète les données reçues. Toutefois le réseau neuronal ne se voit pas imposer ces règles par un programmeur humain, il les élabore lui-même. C'est précisément cette particularité qu'on appelle «intelligence artificielle».

La définition de ces règles suppose un «apprentissage» très long, exigeant une immense puissance de calcul et nécessitant d'énormes quantités de données interprétées au préalable par des experts. Avant de lancer l'apprentissage, une certaine force est attribuée de manière aléatoire à chaque connexion neuronale du réseau. De même, la puissance d'amplification ou d'atténuation du signal de chaque neurone est aussi paramétrée de manière aléatoire. Ensuite, on laisse le réseau traiter les données d'entraînement selon ces règles aléatoires. Dans le cas de la statistique de la superficie, on a fourni au réseau des millions d'extraits d'images aériennes provenant de relevés antérieurs. Les catégorisations générées par le réseau avec ces paramètres purement aléatoires sont bien sûr inutilisables pour l'interprétation d'images. Mais à ce stade, cela n'a aucune importance puisque l'objectif de ce premier passage est tout autre: il permet d'établir, pour chaque image aérienne et pour chaque neurone de la couche de sortie, dans quelle mesure la valeur calculée s'écarte de la valeur réelle, à savoir de celle qui a été déterminée par un expert. Sur la

base de ces «erreurs», le réseau est en mesure de lancer une nouvelle phase de calcul, encore plus complexe, pour mieux se paramétrer. Autrement dit: il apprend. Pour ce faire, le réseau suit le chemin inverse: de la couche de sortie, il remonte vers la couche d'entrée («rétropropagation» des erreurs constatées). L'ensemble de règles que le réseau crée alors dans ses couches intermédiaires est tellement vaste et complexe que l'humain ne peut que difficilement l'appréhender.

Afin de raccourcir le processus d'apprentissage, tout en améliorant la qualité d'interprétation des images, la statistique de la superficie utilise un réseau neuronal qui a déjà été soumis à un entraînement de base avec d'autres types d'images. Dès le moment où il commence à s'entraîner avec les données de la statistique de la superficie, il est donc déjà capable de reconnaître certaines caractéristiques de l'image, telles que les arêtes ou les angles, et peut transférer ce «savoir» aux images aériennes utilisées pour la statistique de la superficie (ce que l'on appelle «l'apprentissage par transfert»).

nécessaires à l'interprétation correcte des images aériennes. La programmation d'ADELE recourt à un procédé plus évolué: l'homme a donné à l'outil la capacité d'«auto-apprendre» à déterminer correctement l'utilisation et l'occupation du sol en se basant sur les images aériennes plus anciennes de la statistique de la superficie et sur leurs interprétations faites par l'humain.

Pour ce faire, le composant central d'ADELE recourt à un réseau neuronal artificiel qui fonctionne selon le principe de l'apprentissage profond («deep learning»), évoqué plus haut (l'encadré en page 4 fournit de plus amples informations sur cette forme particulière de l'intelligence artificielle). Le type de réseau utilisé dans le cas d'ADELE est un réseau neuronal convolutif (CNN).

Lorsque l'on soumet au module entraîné un point d'échantillonnage pour interprétation, le résultat du calcul n'est pas une réponse définitive du style: «le point X correspond à la catégorie d'utilisation A». Le programme fournit en revanche une valeur de probabilité (probabilité de prédiction) distincte pour chacune des 46 catégories d'utilisation du sol et des 27 catégories d'occupation du sol. Ce résultat peut ensuite être utilisé pour nourrir le système en vue d'étapes de calcul subséquentes.

4.2 Intégration de données auxiliaires à l'aide de «forêts d'arbres décisionnels»

Comme expliqué au début de l'article, lorsque l'interprétation est faite par un expert, celui-ci peut faire appel à toute une panoplie d'informations auxiliaires pour arriver à une classification correcte des images aériennes. Il en va de même pour ADELE: les résultats de l'interprétation de l'image par le module d'apprentissage profond sont associés avec des informations auxiliaires

lors d'une itération de calcul. Cela permet de préciser la catégorisation. Actuellement, ADELE utilise les données auxiliaires suivantes:

- données altimétriques: altitude, pente, hauteur de végétation, etc.;
- données spécifiques provenant du modèle topographique du paysage de swisstopo;
- données spectrales (entre autres infrarouges) mesurées par satellite au point d'échantillonnage à différentes saisons;
- autres géodonnées: mensuration officielle, zones à bâtir, surfaces agricoles, etc.

Pour associer l'interprétation des images par apprentissage profond avec les différentes données auxiliaires, une autre méthode d'apprentissage automatique appelée «forêt aléatoire» (voir explications dans l'encadré ci-dessous) est également mise en œuvre. Trois modèles de forêt aléatoire sont entraînés séparément:

- le premier détermine l'utilisation du sol au point d'échantillonnage,
- le second définit la couverture du sol,
- le troisième se contente de décider si l'utilisation et l'occupation du sol au point d'échantillonnage ont changé par rapport au dernier relevé (oui/non). La tâche de ce dernier modèle est donc plus simple que celle des deux premiers. Il en résulte des prédictions qui sont d'autant plus pertinentes (plus grande précision).

Forêt d'arbres décisionnels

On désigne par «forêt d'arbres décisionnels (random forest)» un procédé d'apprentissage automatique souvent utilisé lorsque des objets doivent être «classés» selon certaines caractéristiques, autrement dit affectés à une classe d'objets définie. Ce procédé s'avère particulièrement performant pour traiter des caractéristiques d'objets très différentes et sans rapport direct les unes avec les autres. C'est précisément le cas des données auxiliaires très hétérogènes (voir chapitre 4.2) utilisées pour attribuer les points d'échantillonnage de la statistique de la superficie aux différentes catégories d'utilisation et d'occupation du sol.

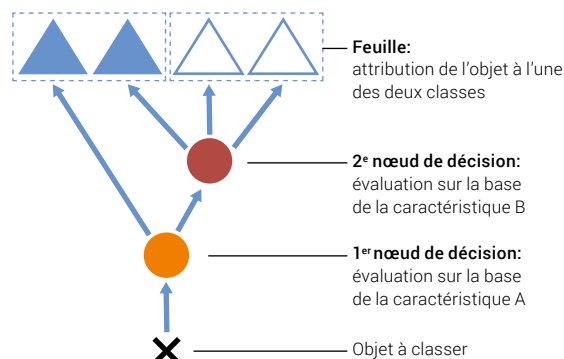
Mode de fonctionnement

Le terme «forêt d'arbres décisionnels» est une traduction usuelle du concept anglais «random forest». Le terme «forêt» est utilisé ici parce que la méthode classifie un objet donné (un point d'échantillonnage dans le cas de la statistique de la superficie) en consultant les décisions d'un grand nombre d'arbres de décision qui forment ensemble une «forêt». La classe finalement retenue est celle à laquelle la majorité des arbres de décision de la forêt attribuent l'objet en question.

Les arbres décisionnels sont des enchaînements de décisions du type «si X, alors Y», dans lesquels on peut intégrer l'ensemble des caractéristiques connues d'un objet: un point d'échantillonnage donné de la statistique de la superficie est-il situé sur un bâtiment, selon les données auxiliaires de la mensuration officielle? Si c'est le cas, il devrait s'agir du type d'occupation du sol «bâtiment». Mais si ce n'est pas le cas, on recourt à une autre caractéristique du point d'échantillonnage, puis à une autre encore, etc. En représentant ces chaînes de décision graphiquement (G3), on observe que les nombreuses ramifications dessinent une structure arborescente – d'où le nom «d'arbre de décision». Dans le cas de l'outil d'IA ADELE, 280 caractéristiques au total sont prises en compte par les arbres de décision. Il s'agit d'une part des 73 valeurs de probabilité générées par l'interprétation des images aériennes par apprentissage profond (voir chapitre 4.1) pour les 46 catégories d'utilisation du sol et les 27 catégories d'occupation du sol, et d'autre part d'un total de 207 caractéristiques supplémentaires tirées de diverses données auxiliaires.

Exemple d'arbre de décision simple

G3



Source: OFS – Statistique de la superficie (AREA)

© OFS 2023

Processus d'apprentissage

Les arbres de décision doivent être entraînés à l'aide de grandes quantités de données issues des relevés antérieurs: sur la base, d'une part, des caractéristiques disponibles pour chaque point d'échantillonnage et, d'autre part, des attributions de classes par les experts, le système optimise la structure des branches des arbres de décision jusqu'à ce que les «décisions» des arbres se rapprochent le plus possible de celles qui avaient été prises par l'homme.

Lors de l'entraînement des nombreux arbres de décision, on veille à ce que la différence entre leurs structures de branches soit relativement élevée: en effet, la «forêt aléatoire» ne doit en aucun cas ressembler à une monoculture, mais compter des arbres de décision d'essences très variées. Cette approche permet d'étayer plus largement les classifications finales pour lesquelles sont justement prises en compte les décisions de l'ensemble des arbres de la forêt. Pour arriver à une forêt «diversifiée», on fait varier de manière aléatoire l'ordre – et donc l'importance – des différentes caractéristiques pour les différents arbres de décision, d'où l'appellation «forêt aléatoire».

5 Utilisation d'ADELE dans la pratique

5.1 Limites de performance et mesures à prendre

Jusqu'ici le fonctionnement théorique de l'outil d'intelligence artificielle ADELE a été expliqué. Dans la pratique, il n'est toutefois pas possible de confier entièrement à ADELE la détermination de l'utilisation et de l'occupation du sol, et ce pour deux raisons surtout.

Tout d'abord, pour une part importante des catégories, l'outil d'IA n'atteint pas le niveau de précision nécessaire pour assurer une mise à jour cohérente des séries temporelles de la statistique de la superficie. Une situation qui ne devrait guère évoluer, ni à moyen, ni à long terme, même si ADELE est continuellement optimisé: certaines catégories sont tout simplement trop similaires pour que l'outil puisse les distinguer, tandis que d'autres (les parcs p. ex.) présentent une diversité de formes et de couleurs que l'IA ne maîtrise pas. Par conséquent, des tests sont réalisés pour déterminer quelles catégories d'utilisation et de couverture du sol l'outil ADELE est en mesure de détecter avec

la précision requise. Au final, seule cette partie de l'output fourni par ADELE pourra être intégrée dans la publication des résultats de la statistique de la superficie. En outre, il n'est possible de prendre en compte que les points d'échantillonnage pour lesquels la probabilité de classification correcte, calculée par l'outil ADELE, dépasse un seuil défini. Tous les autres points d'échantillonnage continuent à être interprétés par des experts.

Deuxièmement, une automatisation intégrale de la détermination de l'utilisation et de la couverture exempte d'«accompagnement» humain n'est pas un but en soi. À l'avenir également, ADELE devra être régulièrement alimenté avec des données d'entraînement récentes – à savoir des catégorisations interprétées par des experts. Cela a pour but :

- d'enseigner au système à utiliser des données auxiliaires entièrement nouvelles, ou du moins restructurées ;
- d'éviter que le système n'accumule certaines erreurs d'interprétation au fil des ans ;
- de régulièrement adapter les règles de décision d'ADELE aux «changements physiques» des catégories elles-mêmes. Ainsi par exemple, par rapport aux images actuelles, l'aspect visuel des «aires industrielles et artisanales» était sensiblement différent sur les images aériennes d'il y a quelques décennies, et il continuera vraisemblablement d'évoluer. Tôt ou tard les règles de décision développées par ADELE sur la base des données d'apprentissage actuelles ne seront donc plus appropriées pour identifier cette catégorie d'utilisation.

Pour faire face à la deuxième série de problèmes, et donc garantir la qualité du relevé à long terme, il s'agit de sélectionner un certain nombre de points d'échantillonnage pour lesquels l'utilisation et l'occupation du sol continueront à l'avenir à être déterminées par des experts. Concrètement, près d'un million de tels points ont été sélectionnés pour la statistique de la superficie.

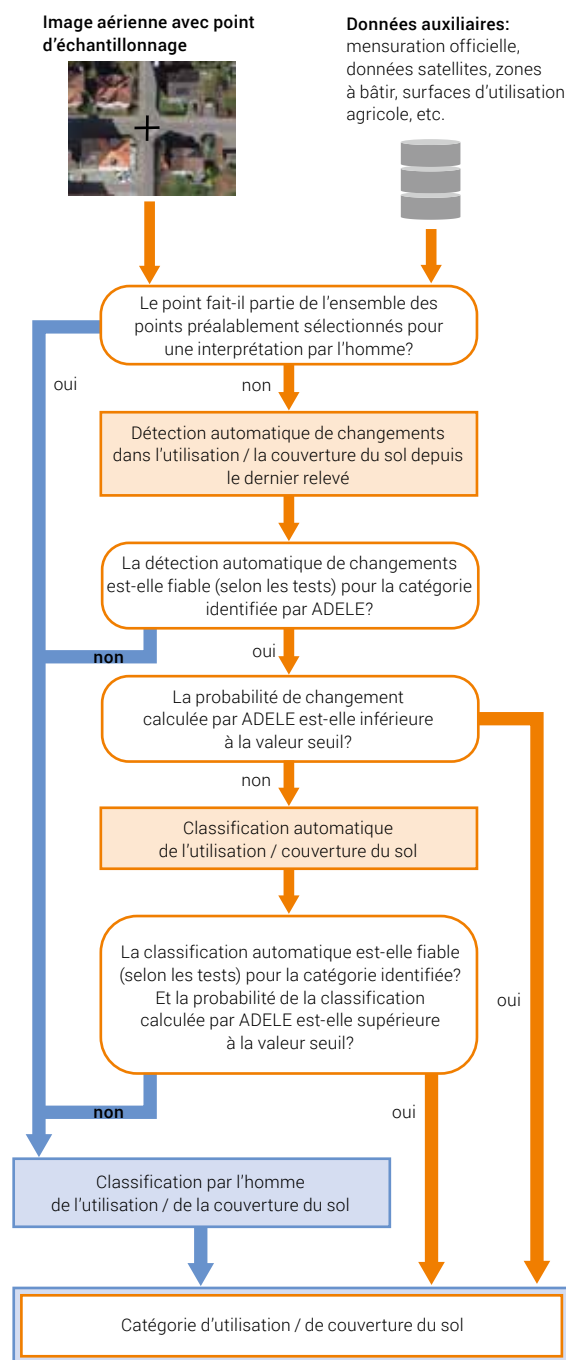
5.2 Déroulement de la détermination semi-automatisée de l'utilisation

Comme décrit ci-dessus, il importe de délimiter de façon réfléchie l'utilisation d'ADELE et de l'associer avec le travail d'interprétation fait par l'humain afin de garantir la constance de la qualité des résultats statistiques. Un processus précis a été défini à cet effet (G4). Celui-ci prescrit de vérifier tout d'abord si le point d'échantillonnage à déterminer fait partie du million de points qui, pour des raisons d'assurance qualité à long terme, nécessitent une interprétation par un expert (voir chapitre 5.1). Si ce n'est pas le cas, ADELE vérifie automatiquement si l'utilisation et l'occupation du sol à ce point ont changé par rapport au dernier relevé, la réponse ne pouvant dans un premier temps être que oui ou non. Un algorithme de forêt aléatoire (voir chapitre 4.2) dont l'unique tâche est de détecter un éventuel changement a été entraîné spécifiquement à cet effet. Cet algorithme s'appuie sur les résultats du module d'apprentissage profond ainsi que sur différentes données auxiliaires (ici, essentiellement des images satellites). S'il est possible d'exclure de manière quasi certaine tout changement, les catégories déterminées lors du dernier relevé sont reprises de manière définitive et la classification automatique

Classification semi-automatique de l'utilisation et de la couverture du sol dans la statistique de la superficie: déroulement du processus

G4

Les étapes en orange sont prises en charge par l'outil d'IA ADELE



Source: OFS – Statistique de la superficie (AREA)

© OFS 2023

se termine. Si, en revanche, il existe une certaine probabilité de changement, l'outil d'IA lance une détermination automatique de l'utilisation. Le résultat de celle-ci n'est toutefois validé que si la catégorie identifiée a été approuvée pour la détermination automatique, c'est-à-dire si les tests préalables ont montré que le fonctionnement d'ADELE est fiable dans ce cas précis. Pour terminer, la probabilité de prédiction calculée par ADELE pour le point d'échantillonnage en question doit dépasser un seuil donné. Si ces conditions ne sont pas toutes remplies, l'utilisation et la couverture du sol continuent d'être déterminées par des experts.

6 Résultats et perspectives

Actuellement (état en 2023), le recours à l'outil d'apprentissage profond ADELE pour la statistique de la superficie permet de catégoriser de façon automatique environ 27% des points d'échantillonnage. Cela représente environ 1,1 million de points qui n'exigent plus d'interprétation par un expert; le gain de temps est donc considérable. Notons toutefois que jusqu'à présent, ADELE n'est utilisé que pour la détection automatique du changement: pour l'instant, l'outil ne fournit pas encore de catégorisation directe.

À l'avenir, les mesures suivantes permettront d'accroître la part des points déterminés automatiquement:

- utilisation de données additionnelles: données issues du registre des bâtiments, images aériennes infrarouges, séries chronologiques plus longues pour les images satellites, données sur l'utilisation et la couverture du sol de points d'échantillonnage voisins, etc.;
- utilisation d'un réseau de neurones mieux pré-entraîné pour le module d'apprentissage profond;
- optimisation des «hyperparamètres» dans les algorithmes d'IA offrant par exemple un meilleur paramétrage du nombre de (couches de) neurones pour le module d'apprentissage profond et du nombre d'arbres de décision pour les modèles de forêt aléatoire.

Au fil des relevés, le volume des données d'entraînement augmentera, ce qui contribuera également à accroître progressivement le degré d'automatisation. Notons toutefois que, même à long terme, il ne sera guère possible de catégoriser plus de la moitié des points selon un procédé automatisé. Cette «valeur limite» découle aussi d'une réflexion sur les ressources: plus le degré d'automatisation est élevé, plus il devient difficile et donc coûteux de l'augmenter encore. À partir d'un certain point, il sera plus rationnel d'allouer directement les ressources qu'exigerait une automatisation accrue à l'interprétation des points d'échantillonnage par des experts. D'autant que le traitement des données de base par l'humain (la consultation répétée des images aériennes à l'écran, les discussions au sein de l'équipe, les visites sur le terrain décrites plus haut) ne sert pas uniquement à la catégorisation en tant que telle. Il permet également d'approfondir la compréhension des données et des catégories de la statistique de la superficie et, par conséquent, de développer une expertise qui reste indispensable pour l'interprétation correcte des résultats statistiques.

Éditeur:	Office fédéral de la statistique (OFS)
Renseignements:	section Géoinformation, OFS, arealstatistik@bfs.admin.ch
Contenu:	Gillian Milani, OFS; Romain Douard, OFS; Ernest Hennig, OFS; Lena Willi-Tobler, OFS; Michèle Gillard, OFS; Anton Beyeler, OFS
Rédaction:	Ferenc Biedermann, OFS
Série:	Statistique de la Suisse
Domaine:	02 Espace et environnement
Langue du texte original:	allemand
Traduction:	Services linguistiques de l'OFS
Mise en page:	Publishing et diffusion PUB, OFS
Graphiques:	Publishing et diffusion PUB, OFS
Images:	Images aériennes: ©swisstopo Photos: © Arealstatistik
En ligne:	www.statistique.ch
Imprimés:	www.statistique.ch Office fédéral de la statistique, CH-2010 Neuchâtel, order@bfs.admin.ch, tél. +41 58 463 60 60 Impression réalisée en Suisse
Copyright:	OFS, Neuchâtel 2023 La reproduction est autorisée, sauf à des fins commerciales, si la source est mentionnée.
Numéro OFS:	2262-2300

Les informations publiées ici contribuent à mesurer la réalisation des objectifs de développement durable (ODD).



Système d'indicateurs MONET 2030

www.statistique.ch → Trouver des statistiques → Développement durable → Système d'indicateurs MONET 2030