Données géographiques d'occupation des sols à grande échelle (OCS GE) sur Arcachon produites et mises à jour par des processus automatisés

Juin 2021

1 Contexte

La production d'un référentiel national, unique et partagé, de données sources d'occupation et usage des sols homogènes, précises et riches (référentiel OCS GE), est indispensable au pilotage des politiques publiques d'aménagement, de protection de la biodiversité et de sauvegarde des terres agricoles.

Le mode de production classique de l'OCS GE ne permet pas un suivi régulier de l'ensemble du territoire dans des délais acceptables, ni à des coûts raisonnables pour la collectivité nationale. C'est pourquoi, en réponse à une commande des ministères de la transition écologique (MTE) ainsi que de l'agriculture et de l'alimentation (MAA), l'IGN est engagé depuis 2019 dans un projet financé par le fond de transformation de l'action publique (FTAP) consistant à mettre en place une chaîne de production automatisée du référentiel OCS GE et du processus de mise à jour associé, dans le cadre de l'observatoire de l'artificialisation des sols. L'IGN pilotera dès 2022 la production de deux millésimes de ce référentiel OCS GE sur le territoire national (des millésimes seront ensuite produits, tous les 3 ans, afin d'assurer la mesure et le suivi du phénomène d'artificialisation des sols dans la durée, au travers d'indicateurs stables).

Ce référentiel d'occupation et usage des sols à grande échelle sur l'ensemble du territoire national sera diffusé en « open data » (« Licence ouverte » en référence à la licence Etalab 2.0) à l'ensemble des parties-prenantes, dont les collectivités territoriales, afin de les accompagner vers une maitrise progressive de la consommation de l'espace et de l'artificialisation des sols dans le cadre de l'objectif de zéro artificialisation nette (ZAN).

2 Les spécifications du produit OCS GE

Pour mémoire, les spécifications de l'OCS GE ont été élaborées par un groupe du travail piloté par le CNIG entre 2012 et 2014. Les principes de l'OCS GE sont les suivants :

- Base de données géographiques décrivant l'occupation et l'usage des sols sur le territoire national.
  - couverture : que voit-on ?
  - usage (ou fonction principale) : à quoi ça sert ?
- Partition du territoire constituée de polygones (chaque point est renseigné, aucun vide, aucun recouvrement)
- Modèle de données (=nomenclature) hiérarchique, emboité : possibilité de détailler plus ou moins le territoire
- Production millésimée sur une référence image
- Compatibilité Inspire

couverture usage

Figure 1 : OCS GE est une partition du territoire. Chaque point est renseigné en deux dimensions (couverture et usage)

- Une ossature socle basée sur le réseau routier et ferré principal de la BD TOPO
- Seuils surfaciques (500m^2 en zone urbaine, 2500m^2 en zone naturelle, agricole et 5000 m^2 en zone forestière; 200m^2 pour les zones bâties) + seuils de largeur
- La mise à jour ne doit montrer que les vrais changements sur le terrain, pas du bruit de mesure ni des différences d'interprétations.

3 Automatisation des processus de production

3.1 Production initiale

La chaîne « classique » de production initiale de l'OCS GE s'appuyait sur un travail conséquent de photo-interprétation sur des images de la BD ORTHO, qui permettait de compléter et de corriger des traitements automatiques réalisés sur des données existantes issues de BD TOPO, RPG et BD Forêt.

L'automatisation de la chaine de production initiale de l'OCS GE est essentiellement basée sur des techniques de télédétection par intelligence artificielle (IA), que l'on mobilise pour cartographier des objets non modélisés dans les bases de données existantes. Cette cartographie automatisée repose sur des outils de segmentation sémantique par *deep learning* c'est-à-dire de classification au pixel, d'images millésimées à très haute résolution dont on intègre les résultats dans une nouvelle chaîne de traitement.

Le processus de production initiale est composé des grands modules fonctionnels suivants (cf. schéma n°2 en annexe) :

1. Production **d'annotations**. Les modèles *deep learning* retenus sont composés d'un très grand nombre de paramètres qui sont ajustés via une phase d'apprentissage sur une base d'exemples. Ces jeux de données d'apprentissage, également appelés **annotations**, sont conçus :
   a) en créant d'une part des échantillons raster multicouches composés des 4 canaux (rouge, vert, bleu, infra-rouge) de la BD ORTHO et d'une couche « hauteur » (MNS – MNT),
   b) en combinant des données géographiques existantes (BD TOPO, RPG, BD Forêt) et/ou en saisissant manuellement des éléments sur des ortho images.
2. Paramétrage des modèles *deep learning* : Sur chaque département administratif (ou nouvelle zone de paysage homogène), les modèles *deep learning* sont paramétrés via les **annotations** ou jeux de données d'apprentissage. Il a été décidé, en lien avec les spécifications OCS GE, de distinguer les modèles *deep learning* pour les zones urbaines des modèles pour les zones NAF : Naturelles, Agricoles, Forestières.
3. Inférence des modèles : Les modèles *deep learning* ainsi paramétrés sont appliqués à l'ensemble de la zone, fournissant ainsi des éléments de description du territoire, sous la forme de **cartes de chaleur raster**.
4. Vectorisation des résultats du *deep learning* & 5. Extraction des données millésimées des bases de données existantes & 6. agrégation dans les classes d'OCS GE: Les données issues de l'extraction automatique par *deep learning*, ainsi que les données issues de la BD TOPO, de la BD Forêt ou du RPG sont intégrées dans **une base intermédiaire**, puis converties en données OCS GE, c'est-à-dire en une partition du territoire (sans discontinuité ni vide ni superposition) constituée de polygones comportant les informations de couverture et d'usage du sol calés sur l'ossature dérivée principalement du réseau routier de la BD TOPO, respectant au mieux les seuils de surface et largeur pour chaque classe.

   + de détails sur cette phase en annexe n°1

1. Complètement et corrections manuelles : Le résultat des étapes précédentes est proposé à un photo-interprète pour complétement des zones non renseignées, et corrections des défauts éventuels, en vue de fournir des données conformes aux spécifications.
2. Recette

3.2 Mise à jour

La chaîne « classique » de mise à jour de l'OCS GE s'appuyait essentiellement sur de la photo-interprétation manuelle. La saisie du nouveau millésime s'effectuait en parcourant intégralement à l'aide d'un quadrillage kilométrique le jeu de données existant superposé au nouveau millésime de la BD ORTHO.

L'automatisation de la mise à jour de l'OCS GE repose sur la création d'alertes de détection de changement qui permettent de cibler pour le photo-interprète les zones à modifier. Un soin particulier est apporté à la stabilité de la partition du territoire sur les zones n'ayant subi aucun changement.

Le processus de mise à jour est composé des grands modules fonctionnels suivants :

1. Détection des zones de changement par IA:
   a. Production d'annotations
   b. Paramétrage des modèles *deep learning*
   c. Inférence des modèles

Scénario à court terme

2a. Proposition des zones de changement au photo-interprète pour mise à jour des données

Scénario à moyen terme

2b. Mise à disposition d'un « stylo magique » qui automatise l'intégration du changement.

3. complètement et corrections manuelles :
4. Recette

3.3 Usage

Le recours à la télédétection et l'intelligence artificielle (*deep learning*) permet surtout d'automatiser la production de la dimension « couverture » ; ce n'est pas le cas de la composante « usage ». On utilise actuellement à cet effet les fichiers fonciers, la BD TOPO, le RPG (usage agricole), la BD Forêt (usage sylvicole).

Des données « MAJIC » aux Fichiers fonciers enrichis

Depuis 2009, le Cerema retraite, géolocalise et enrichit les Fichiers fonciers de la Direction Générale des Finances Publiques (DGFiP) pour le compte de la direction générale de l'aménagement du logement et de la nature (DGALN), afin de permettre aux acteurs publics de réaliser facilement des analyses fines et comparables sur leur territoire. Les fichiers fonciers constituent un produit millésimé.

Amélioration de l'usage en zones bâties grâce aux fichiers fonciers

En milieu urbain, les fichiers fonciers permettent de raffiner l'usage mixte 235 sur les zones bâties, l'usage 235 ne faisant aucune distinction entre les différents bâtiments (industriels, tertiaires ou résidentiels).

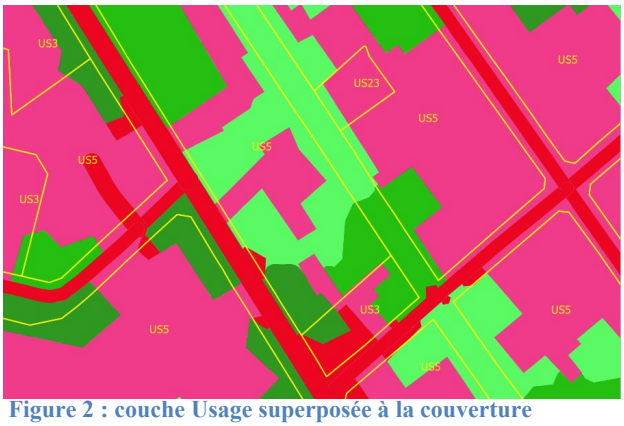

Figure 2 : couche Usage superposée à la couverture

Une couche des usages est créée à partir de la table TUP créée par le Cerema (Table unifiée des parcelles). La mise en cohérence des deux couches (afin que chaque polygone possède les deux informations de couverture et d'usage conformément aux spécifications) entraîne un découpage des entités issues de la couche couverture par la couche des usages. Il résulte que certaines superficies ne respectent plus les Unité minimum de collecte qui sont de 200 m² pour les zones bâties (CS1.1.1.1) et 500m² pour les autres classes en zone construite.

- Si l'un des trois usages est présent à plus de 90% dans la surface, on conserve cet usage simple.
- Dans le cas contraire, on passe à l'usage mixte US235.

Concernant le traitement des surfaces inférieures aux seuils lors de la mise en cohérence US/CS, la règle est de donner la priorité à la dimension « couverture » et au respect des géométries lors du découpage entre US2, US3 et US5.

Au-delà de l'emploi des fichiers fonciers, le renseignement automatisé de la couche Usage est encore à l'étude.

3.4 Perspectives

L'IA ne constitue aujourd'hui qu'une partie des processus mis en œuvre néanmoins l'ambition est, qu'à terme, la qualité des modèles IA permette de minimiser les dépendances vis à vis des bases de données IGN (et ainsi permettre de générer une OCS aussitôt que possible après la mise à disposition d'un nouveau millésime d'orthoimage et de MNS).

4 Jeu de données « Arcachon 2020 »

Le jeu de données « Arcachon 2020 » est issu du processus prototype dans son état de novembre 2020. Les données n'ont pas encore la qualité cible décrite dans les spécifications, et la nomenclature est focalisée sur les classes impactant le suivi de l'artificialisation des sols. Il a été supposé que la classe CS11 (surfaces anthropisées) pourrait être proche de la définition future de l'artificialisation des sols.

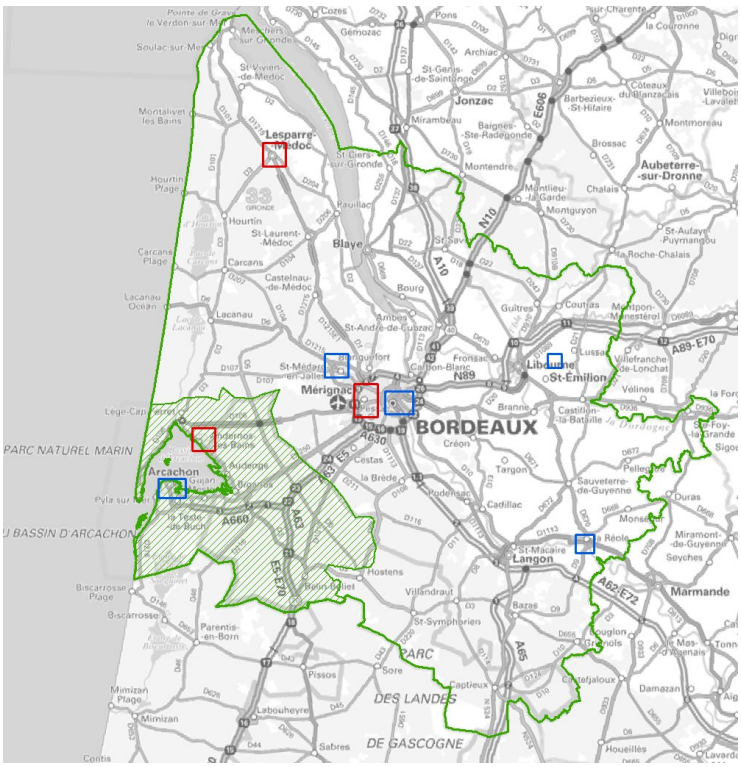

Figure 3 : zone test

Surface : 1500 km²

Paysage particulier :

- Littoral (complexe)
- Plantations de conifères
- Peu d'agricole
- Peu de Vignes

Millésimes des ortho-images : 2015/2018

Le prototype (novembre 2020) apporte des résultats convaincants pour renseigner la dimension « Couverture » de l'OCS GE

- En zone urbaine : la majorité des classes sont bien détectées par l'IA (mention spéciale aux bâtiments >>80%) même si des confusions (non spécifiques à l'IA car déjà présentes en photo-interprétation traditionnelle) persistent entre certaines classes (bitume/matériaux minéraux par exemple, végétation à améliorer);
- En zone NAF : des axes de progrès sont constatés, qui diminueront via les leviers suivants :
  - L'augmentation du vivier des annotations qui rendra les modèles *deep learning* plus robustes.
  - Le recours aux séries temporelles Sentinel qui permettra d'améliorer la discrimination entre des classes ambigües.

Le prototype (novembre 2020) a mis en exergue les difficultés pour renseigner la dimension « Usage » de l'OCSGE

- l'IA n'est pas –encore- adaptée pour renseigner la dimension « Usage » ;
- Les fichiers fonciers permettent d'obtenir l'usage et notamment de distinguer les usages 2, 3, 5. Mais les nombreuses erreurs contenues (actualité des données) génèrent un bruit et une surcharge de travail importante dans le cadre de la production d'une OCSGE à 2 dimensions ✉ à ce stade, la photo-interprétation reste largement indispensable pour obtenir un haut niveau de qualité sur les usages.

A noter : Les résultats du prototype de novembre 2020 ont fait l'objet d'un effort particulier en photo-interprétation, pour compenser autant que possible les lacunes des processus dans leur état d'alors.

4.1 Bilan par classe de couverture

✓ CS1.1.1.1 - Zones bâties	Constitution IA et BDTopo : bons résultats
✓ CS1.1.1.2 - Zones non bâties	Constitution IA et BDTopo : bons résultats

~ CS1.1.2.1 - Matériaux minéraux	Confusion avec zones non-bâties et sols nus > correction photo-interprétation
? CS1.1.2.2 - Matériaux composites	Trop peu de données d'entrée pour la détection automatique > photo-interprétation
✓ CS1.2.1 - Sols nus	Constitution IA et BDTopo : bons résultats
~ CS1.2.2 - Surfaces d'eau	Certaines surfaces d'eau non détectées (asséchées) > photo-interprétation
? CS1.2.3 - Névés et Glaciers	Pas de cas sur le prototype > pas de détection pour le moment
~ CS2.1.1.1 - Feuillus	Pas encore de distinction feuillus/conifères/mixte dans la détection automatique
~ CS2.1.1.2 - Conifères	Pas encore de distinction feuillus/conifères/mixte dans la détection automatique
~ CS2.1.1.3 - Formations Arborées Mixtes	Pas encore de distinction feuillus/conifères/mixte dans la détection automatique
x CS2.1.2 - Formations arbustives	Qualité de la détection à évaluer (processus NAF[1])
~ CS2.1.3 - Autres formations ligneuses	Qualité de la détection des vignes en cours (processus NAF)
~ CS2.2.1 - Formations herbacées	Des confusions avec les coupes et jeunes plantations (NAF)
? CS2.2.2 - Autres formations non ligneuses	Pas de cas sur le prototype

4.2 Bilan par classe d'usage

~ US1.1 – Agriculture	A revoir avec le développement du processus NAF
~ US1.2 – Sylviculture	A revoir avec le développement du processus NAF
US1.3 - Activités d'extraction	Information BD TOPO
x US1.4 - Pêche et aquaculture	Pas d'informations Fichiers fonciers > Photo-interprétation
? US1.5 - Autres productions primaires	Pas de cas sur le prototype
US2 – Production secondaire	Distinction US2, 3 et 5 bonne grâce aux Fichiers fonciers
US3 – Production tertiaire	Distinction US2, 3 et 5 bonne grâce aux Fichiers fonciers
US5 – Usage résidentiel	Distinction US2, 3 et 5 bonne grâce aux Fichiers fonciers
US235 – Secondaire, tertiaire, résidentiel	Pas d'infos fichiers fonciers ou multi usages

5 Propriété des données, licence

Ces données, dont l'IGN détient la propriété, sont diffusées sous « Licence ouverte » en référence à la licence Etalab 2.0.

6 Format des données

Ces données (emprise, millésime 2015 et millésime 2018) sont diffusées au format Shapefile.

1 *Le modèle NAF actuel s'appuie sur la même technologie que le modèle urbain mais avec des classes donc des jeux d'annotation différents. Il est implémenté à ce jour (juin 2021) et permet dorénavant de meilleures performances sur ces zones.*

7 ANNEXES

Annexe 1 : Vectorisation des résultats du *deep learning*, Extraction des données millésimées des bases de données existantes, agrégation dans les classes d'OCS GE

Compilation des données existantes

Les informations millésimées des bases de données existantes sont intégrées dans les classes de couverture ou d'usage de l'OCS GE. Les zones blanches dans la figure 4 ci-dessous montrent les zones contenant des informations non décrites dans les bases de données utilisées. Ces zones blanches seront remplies par des techniques de télédétection par intelligence artificielle.

Figure 4 : état de l'intégration des données existantes (BD Uni, BD TOPO, RPG…) dans la nomenclature OCS GE.

Choix des modèles IA

Les modèles de *deep learning* (et tous les codes et techniques associés) ont été sélectionnés en fonction de leur maturité et de leur adaptation à la télédétection. Ces modèles, de type réseaux « convolutionnels » mono-temporels, s'appuient sur l'architecture Unet, sur un framework PyTorch.

Deux modèles multi-classes ont été développés : un pour les zones urbaines et un pour les zones Naturelles, Agricoles et Forestières (NAF). Cette distinction s'appuie sur deux réalités :

- Dans la nomenclature, les Unités Minimales d'Intérêt (seuils) de surface ne sont pas les mêmes que l'on se situe en zone urbaine (500m²) ou en zone NAF (2500m²).
- Les modèles *deep learning* envisagés entre les deux types de paysages ne sont pas les mêmes et le choix des classes à détecter est différent.

Le choix d'un modèle multi-classes, plutôt que de multiples modèles mono-classe, facilite la mise au point et la maintenance. Il s'accompagne toutefois d'une contrainte assez forte sur les jeux d'annotations (vérité-terrain) qui doivent :

- tous décrire la totalité des classes
- représenter de manière équilibrée les différentes classes

La production des annotations, ou données « Vérités »

La création d'annotations[2] (données vérités ou données d'apprentissage, indispensables en intelligence artificielle) est une activité chronophage. On espérerait en début de projet limiter le temps dédié à cette activité en utilisant les bases de données géographiques existantes, qui se sont malheureusement révélées inadéquates pour la problématique de l'OCS GE. Les spécifications des bases existantes entrainent en effet des généralisations ou des représentations qui ne sont pas assez fidèles à l'image. Par exemple un tronçon de route est un linéaire dans la BD TOPO et non pas une surface au plus près du bord de la route. L'algorithme de *deep learning* a besoin d'apprendre que la largeur de la chaussée peut être variable (carrefours). Les premières expérimentations ont montré que la qualité (représentativité, justesse) des annotations permettait de s'affranchir raisonnablement de la nécessité d'avoir des millions d'échantillons.

Il a été décidé, en lien avec les spécifications OCS GE, de distinguer les labellisations (production d'annotation) en zones urbaines et en zones Naturelles, Agricoles, Forestières (NAF). Le modèle NAF actuel s'appuie sur la même technologie que le modèle urbain (UNet) mais avec des classes donc des jeux vérité différents.

Sorties du deep learning

Les sorties du *deep learning* sont des cartes de chaleur par objet à détecter (bâti, route, etc.). Plus le modèle de *deep learning* est certain qu'un pixel appartient à une classe, plus la valeur de ce pixel sera élevée.

2 Annotations = données Vérités = données d'apprentissage

Figure 5 : cartes de chaleur vectorisées représentant le résultat d'une segmentation sémantique par *deep learning* (modèle urbain)

Post-traitements

La dérivation des données sources vers les données finales de l'OCS GE, conformes aux spécifications, consiste, à partir des différentes données source (*deep learning*, BD TOPO, RPG) à obtenir une partition du territoire (sans discontinuité ni vide ni superposition) constituée de polygones comportant les informations de couverture et d'usage du sol calés sur l'ossature dérivée principalement du réseau routier de la BD TOPO, respectant au mieux les seuils de surface et largeur pour chaque classe.

Ceci implique :

1. d'harmoniser les données entre elles,
2. de combler les surfaces non renseignées,
3. de croiser les données de couverture avec la couche des usages calculée indépendamment,
4. de transformer les données pour les rendre conformes aux spécifications (respect des seuils minimaux).

Les cartes de chaleurs issues du *deep learning* sont vectorisées (outil GRASS). Les traitements intègrent la suppression des objets inférieurs à un seuil de surface, ainsi qu'une opération de "fermeture" morphologique (cf. Figure 5 : cartes de chaleur vectorisées représentant le résultat d'une segmentation sémantique par deep learning (modèle urbain)).

La vectorisation est gérée par classe, ce qui demande d'arbitrer les éventuelles superpositions pendant les traitements d'intégration aux autres données (BD Uni, BD Forêt) vers des classes d'OCS GE.

Un travail de comblement des zones non renseignées est opéré (cf. Figure 6). Une grille au pas de 5 m permet de « généraliser » les résultats de la détection. Pour chaque maille de la grille, la couverture majoritaire est retenue parmi les classes : matériaux minéraux, herbacée, ligneux.

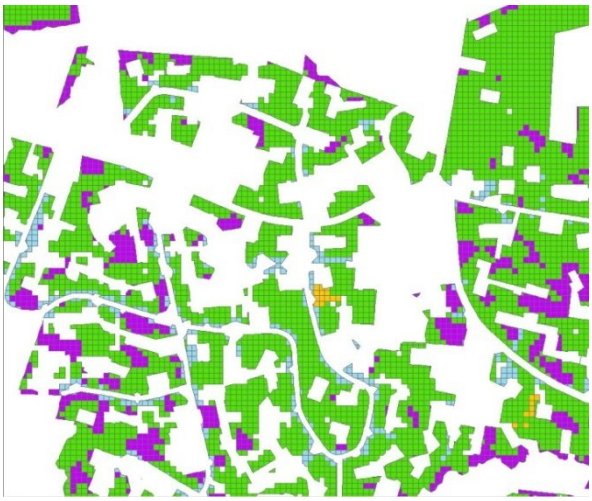

Figure 6 : comblement des zones non renseignées

Annexe 2

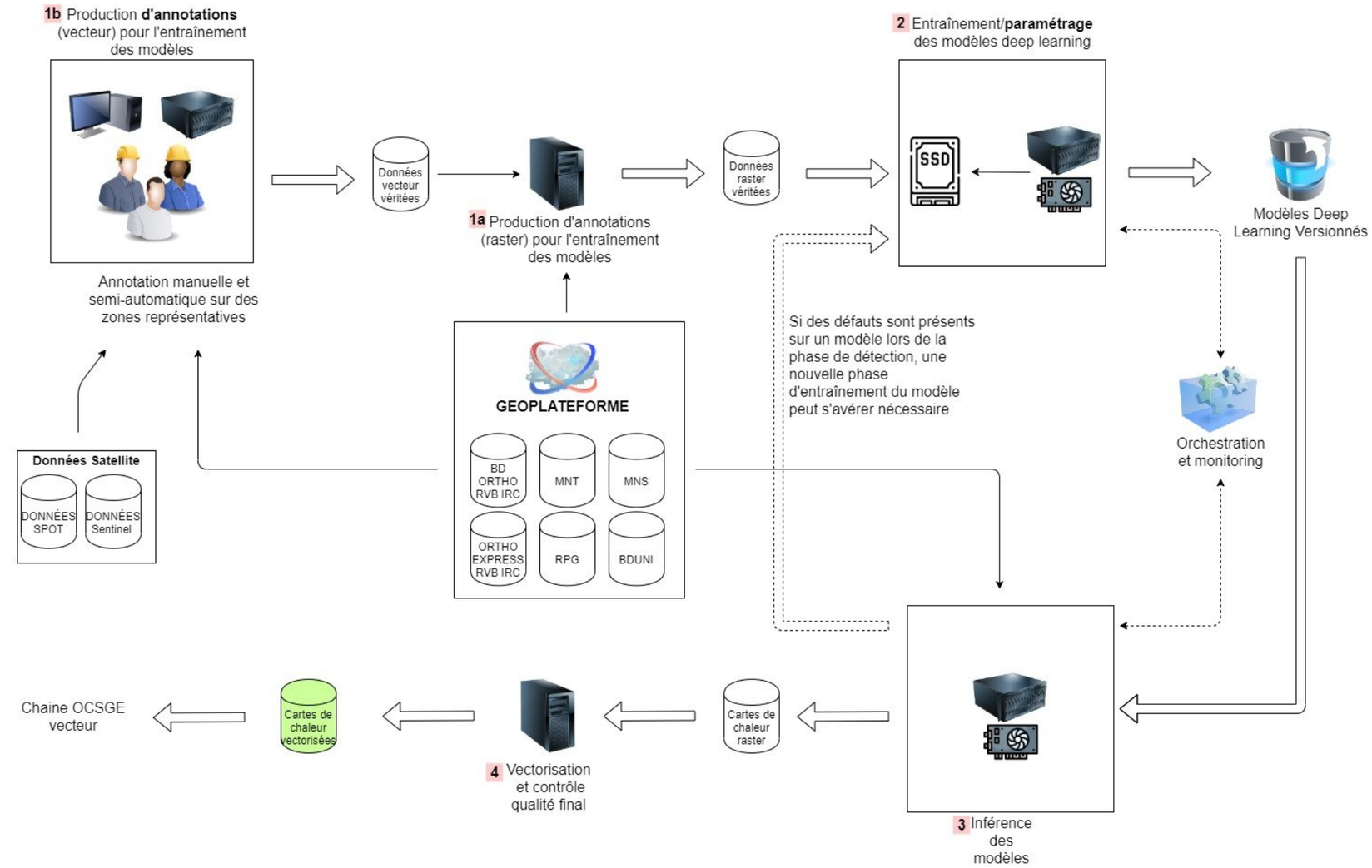

Schéma n°1 : partie IA de la chaine automatisée de production initiale

Les données produites automatiquement sont ensuite soumises à photo-interprétation manuelle, pour complétement et corrections, en vue de respecter les spécifications.