

## UTILISATION DES IMAGES SATELLITES POUR AMÉLIORER LE REPÉRAGE DES LOGEMENTS À MAYOTTE

Maëlys Bernard (\*), Raya Berova (\*\*), Thomas Faria (\*\*\*)

(\*) Insee, Département de la Démographie

(\*\*) Insee, Direction de la méthodologie et de la coordination statistique et internationale

(\*\*\*) Insee, Direction de la méthodologie et de la coordination statistique et internationale

*maelys.bernard@insee.fr raya.berova@insee.fr thomas.faria@insee.fr*

**Mots-clés :** Enquêtes cartographiques, données satellites, Mayotte, recensement de la population, vision par ordinateur, apprentissage profond

**Domaines :** Sondages, enquêtes

### Résumé

*Malgré leur disponibilité de plus en plus accrue, les images satellites très haute résolution sont jusqu'ici très peu utilisées au sein de la statistique publique. La manipulation de ces données non conventionnelles n'appartient pas encore aux pratiques usuelles des statisticiens publics, bien qu'elles offrent des opportunités significatives pour améliorer ou créer de nouveaux indicateurs statistiques. Cet article vise justement à démontrer qu'il est possible d'intégrer ces données satellitaires dans les processus statistiques afin d'assurer leur qualité. Mayotte constitue un cas emblématique où les difficultés d'accès au terrain et la forte dynamique de développement urbain complexifient l'organisation des enquêtes cartographiques qui permettent de repérer les logements en amont du recensement. Afin d'orienter les travaux de contrôle de la qualité de ces enquêtes, nous avons exploré l'utilisation des images satellites très haute résolution Pléiades (0,5 m de résolution spatiale), en combinaison avec des méthodes d'apprentissage profond, spécifiquement via des modèles de segmentation sémantique. La méthodologie développée consiste en l'entraînement d'un modèle SegFormer, fondé sur une architecture de type Transformer, largement reconnue dans la littérature puisqu'elle est à la base des modèles GPT. Le modèle a été entraîné grâce aux annotations issues du projet CoSIA de l'IGN, qui fournit une référence précise de la couverture du sol. À partir des prédictions du modèle, il est possible de produire des estimations précises des surfaces bâties sur un territoire donné, permettant ainsi d'analyser efficacement leur évolution dans le temps. Les résultats obtenus démontrent une performance élevée dans la détection automatisée des évolutions urbaines dans les DROM. Ces résultats sont disponibles sous la forme d'un tableau de bord interactif permettant aux équipes opérationnelles de cibler efficacement les zones prioritaires pour la réalisation d'enquêtes cartographiques de contrôle sur le terrain.*

*Après avoir décrit la méthode de prédiction de surface du bâti sur les images satellites, cet article présentera un cas pratique d'utilisation de ces prédictions à travers le contrôle qualité des enquêtes cartographiques de la population à Mayotte. Depuis 2021, le recensement à Mayotte fonctionne sur un cycle quinquennal comme en métropole et dans les autres DOM. En amont de la collecte auprès des habitants, une enquête cartographique permet de repérer les logements. Le résultat de l'enquête cartographique détermine la base d'immeubles au sein de laquelle est sélectionné l'échantillon<sup>1</sup> de l'enquête annuelle de recensement. Ainsi, si l'enquête cartographique ne couvre pas bien l'intégralité du bâti du territoire à enquêter, les populations et nombre de logements déduits de l'enquête annuelle de recensement seront sous-estimés du fait de ce défaut de couverture. Les données satellites permettent de contrôler la qualité des enquêtes cartographiques en comparant l'évolution de la couverture de bâti avec celle du nombre de logements observé suite aux enquêtes cartographiques au sein de chaque îlot<sup>2</sup> à enquêter une année donnée. Les îlots sont répartis en cinq groupes de rotation correspondant à l'année à laquelle a lieu l'enquête annuelle de recensement. Afin d'assurer la qualité des enquêtes cartographiques, un deuxième passage sur le terrain a été planifié dans les grandes communes mahoraises pour certains îlots identifiés comme les plus prioritaires. Cet ordre de priorité a été défini par une méthode mobilisant les images satellites pour repérer les plus susceptibles d'être concernés par des manques potentiels dans les enquêtes cartographiques. L'enquête cartographique corrective ciblant ces îlots a permis d'assurer une meilleure couverture de ces enquêtes et d'éviter un risque de sous-estimation des logements.*

## **Abstract**

*Although satellite images are still rarely used in public statistics, their integration into statistical processes offers new opportunities to improve quality and the creation of indicators. The case of Mayotte illustrates this challenge well: difficulties in accessing the terrain and rapid urban development complicate the organization of mapping surveys used to identify dwelling ahead of the census. The use of very high-resolution satellite images, combined with deep learning methods, has made it possible to obtain accurate estimates of built-up areas in a given territory. These results were then used to check the quality of mapping surveys in Mayotte.*

## 1. Introduction

Chaque année, une centaine d'agents recenseurs sont mobilisés dans les Départements et Régions d'Outre-Mer (DROM) afin d'effectuer la collecte du recensement de la population. Avant cette phase de collecte, une enquête cartographique est réalisée dans les DROM afin de mettre à jour le répertoire des immeubles localisés (RIL). Ce répertoire contient une liste exhaustive des logements géolocalisés, parmi lesquels sont sélectionnés les logements à enquêter pour le recensement de l'année en cours. Cette enquête est spécifique aux DROM, car les bases administratives habituellement disponibles en France métropolitaine ne sont pas suffisamment fiables pour alimenter seules ce répertoire.

Un RIL de qualité permet aux enquêteurs de localiser plus facilement les logements à enquêter sur une année donnée. Le calcul de la population dépend du nombre de logements présents dans le RIL. En effet, cette estimation est le produit du nombre moyen de personnes par logement, obtenu via le recensement, et du nombre de logements comptabilisés. L'impact d'un bon RIL sur la qualité des estimations produites par l'Insee est donc considérable.

De ce fait, le contrôle de la qualité de ces enquêtes cartographiques est important pour garantir une bonne qualité du RIL, notamment dans le territoire Mahorais, où les difficultés d'accès au terrain et la forte dynamique de développement urbain complexifient l'organisation des enquêtes cartographiques. Dans ce contexte, l'utilisation d'images satellites s'est avérée très utile.

Par ailleurs, l'enquête cartographique se déroule d'avril à août chaque année et mobilise près de 100 enquêteurs dans les DROM, ce qui en fait une opération coûteuse. L'utilisation de l'imagerie satellite, et notamment des méthodes de détection de logements sur ces images, pourrait également permettre d'améliorer l'organisation de l'enquête cartographique en optimisant notamment le temps de travail des enquêteurs du RP.

Ce travail à partir des données satellitaires s'inscrit donc dans une réelle perspective d'amélioration de la précision du RIL mahorais et d'un meilleur calibrage de la préparation des EAR grâce à un socle d'informations actualisé pour les enquêtes cartographiques.

La comparaison de l'évolution de ces indicateurs de densité du bâti obtenus grâce aux images satellites avec le nombre de logements relevés d'après les enquêtes cartographiques permettra d'identifier les zones les plus divergentes pour les expertiser en bureau en amont et d'orienter la collecte en conséquence.

Enfin, ce projet s'inscrit dans la continuité du *Mémorandum de Varsovie* adopté en 2021 par le Comité du Système statistique européen, European Statistical System Committee (ESSC) [1], qui promeut le recours aux données d'observation de la Terre pour renforcer la qualité, la couverture et la réactivité des statistiques publiques au sein du Système statistique européen.

## 2. Méthodologie

Nous allons dans cette partie revenir sur les concepts d'apprentissage profond (*deep learning*) afin d'expliquer notre démarche et l'application de ces méthodes à nos problématiques précises. Nous définirons ensuite les données que nous avons utilisées pour appliquer notre méthodologie.

### 2.1. Les modèles d'apprentissage profond pour l'analyse d'images

Dans le domaine de l'apprentissage statistique, les réseaux de neurones profonds en sont un sous-domaine. Si les premiers algorithmes de réseaux de neurones ont été développés dès les années 50, les réseaux de neurones profonds ont, eux, fait leur révolution dans les années 2010 notamment grâce aux développements de nouvelles cartes graphiques (GPU) et d'algorithmes optimisés pour ceux-ci. En effet, leur développement a longtemps été freiné par la quantité de données nécessaires pour obtenir de bons résultats impliquant donc des temps de calculs très longs et des coûts d'annotations prohibitifs. La conception des réseaux de neurones profond est inspirée du fonctionnement du cerveau humain. En effet, l'idée est de représenter les neurones par des fonctions mathématiques très simples qui s'activent si le signal reçu en entrée est suffisamment fort, dans quel cas ils transmettent alors leur « activation ». Un grand nombre de neurones correctement organisés permet alors de réaliser des opérations plus complexes, résultantes des différentes activations.

Si l'apprentissage profond est devenu la solution de facto pour l'analyse d'images par ordinateur, et plus précisément les tâches de segmentation sémantique, c'est en parti grâce aux travaux fondateurs de J. Long, E. Shelhamer, et T. Darrell [2] qui ont développé les réseaux entièrement convolutifs, une extension de l'architecture des réseaux de neurones convolutif proposé par Y. LeCun *et al.* [3].

#### 2.1.1. Réseaux de neurones convolutifs

Les réseaux de neurones convolutifs (CNN) tirent leur inspiration du cortex visuel des animaux et se composent de deux parties (voir Figure 1) :

1. Une première partie composée par des couches de convolutions successives qui permet d'extraire des prédicteurs de l'image en entrée du réseau;
2. Une seconde partie permettant de classifier les pixels de l'image en entrée à partir des prédicteurs obtenus dans la première partie.



Figure 1. – Réseau de neurones convolutif

*Note de lecture : La partie convulsive est représentée par les différents cubes à gauche et la partie classifiante par les rectangles fins et bleus à droite*

Dans la partie convulsive, l'image en entrée du réseau se voit appliquer des filtres appelés convolutions, représentés par des matrices  $A = (a_{ij})$  de petite taille appelées noyaux de convolution. L'image en sortie de l'opération de convolution est obtenue à partir de chaque pixel de l'image en entrée en calculant la somme des pixels avoisinant, pondérée par les coefficients  $a_{ij}$  du noyau de convolution. Cette opération de convolution est illustrée dans la Figure 2, extraite de l'ouvrage P. Kim [4].

$$\begin{array}{|c|c|c|c|} \hline 1 & 1 & 1 & 3 \\ \hline 4 & 6 & 4 & 8 \\ \hline 30 & 0 & 1 & 5 \\ \hline 0 & 2 & 2 & 4 \\ \hline \end{array} \otimes \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{array}{|c|c|c|} \hline 7 & 5 & 9 \\ \hline 4 & 7 & 9 \\ \hline 32 & 2 & 5 \\ \hline \end{array}$$

Figure 2. – Opération de convolution

*Note de lecture : L'image résultante est une image plus petite dont les pixels sont égaux à la somme des pixels de l'image en entrée pondérée par les coefficients du noyau de convolution.*

Une convolution vise ainsi à résumer l'information présente dans une région de l'image. Il est d'ailleurs intéressant de remarquer que lorsque le voisinage d'un pixel ressemble à la structure du noyau de convolution, la valeur de sortie est élevée. Ainsi, si le filtre est, par exemple, un détecteur de lignes verticales, il générera une grande valeur sur les pixels appartenant à des lignes verticales. Un filtre de convolution permet donc d'obtenir une sorte de carte de caractéristiques de l'image. Une couche convulsive correspond finalement à l'application d'un nombre  $n_f$  de filtres différents en parallèle. Ainsi, en sortie d'une couche, il y a autant d'images (carte de caractéristiques) que de filtres appliqués. Ces  $n_f$  cartes deviennent les canaux d'entrée de la couche suivante.

Les CNN se composent de couches successives, chacune ayant un rôle spécifique. Les premières couches détectent des motifs simples dans l'image (bords, textures, lignes horizontales ou verticales...) tandis que les couches intermédiaires et profondes, plus spécialisées,

combinent ces motifs simples pour repérer des formes de plus en plus complexes (motifs, objets, visages...). La Figure 3 schématise un réseau de neurones convolutif.



Figure 3. – Représentation d'une convolution

*Note de lecture : Les premières couches convolutives reconnaissent les formes simples tandis que les couches les plus profondes à droite reconnaissent des formes plus concrètes.*

Après une couche de convolution, la taille des cartes de caractéristiques reste identique à celle de l'image, ce qui peut rapidement entraîner un volume de données important et rendre le modèle coûteux à entraîner. Pour pallier cela, une méthode couramment utilisée est le *pooling*. Cette opération, représentée par la Figure 4, consiste à diviser chaque carte de caractéristiques en petites régions, puis à ne conserver, pour chacune, que la valeur maximale. Lorsqu'il s'agit de prendre la valeur maximale, on parle de *max pooling* mais d'autres opérations peuvent être utilisées comme le minimum, la moyenne ou la médiane. Le *pooling* permet ainsi de réduire la dimension spatiale des données tout en conservant l'information la plus pertinente. Cette réduction favorise également une certaine invariance locale aux petites translations ou déformations de l'image, et contribue à limiter le nombre de paramètres et la complexité de calcul du réseau, tout en mettant en valeur les motifs les plus discriminants. D'autres opérations telles que le padding détaillé dans la Figure 5 permettent malgré tout de contrôler la dimension de l'image en sortie.

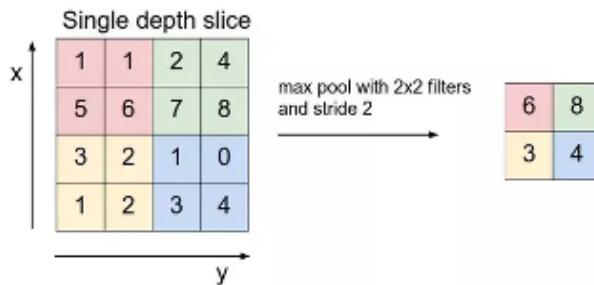


Figure 4. – Représentation du max pooling

*Note de lecture : L'image résultante de l'opération de max-pooling est égale au maximum des pixels dans chacune des zones dessinées sur l'image de gauche.*



Figure 5. – Représentation du padding

*Note de lecture : l'ajout d'une bande de 0 autour de l'image avant d'appliquer la convolution permet de limiter la réduction de dimension de l'image résultante.*

Une fonction d'activation non linéaire, comme la ReLU (Rectified Linear Unit), est appliquée systématiquement après chaque couche de convolution (cf. Figure 6). Inspirée du fonctionnement des neurones biologiques, cette opération introduit une non-linéarité dans le réseau. Sans cette étape, l'empilement de couches convolutives ne ferait qu'appliquer une combinaison linéaire de filtres, ce qui limiterait la capacité du modèle à apprendre des relations complexes. L'ajout de fonctions d'activation non linéaires permet ainsi au réseau d'extraire et de modéliser des motifs variés et des structures non linéaires présentes dans les données, ce qui est essentiel pour traiter efficacement des images. L'utilisation de fonctions d'activation non linéaires ne se limite pas aux réseaux convolutifs, c'est justement une composante fondamentale de l'ensemble des architectures d'apprentissage profond.

$$f(u) = \max(0, u)$$

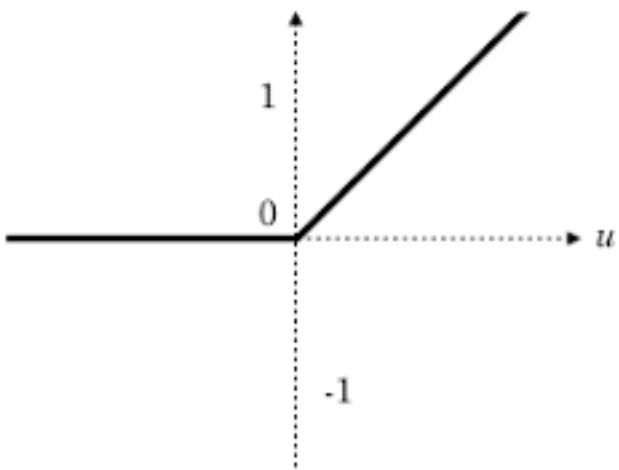


Figure 6. – Fonction d'activation ReLU

*Note de lecture : la fonction ReLU « s'active » quand l'argument est positif.*

Contrairement aux approches classiques du machine learning, la particularité des réseaux convolutifs réside dans l'automatisation de l'extraction des caractéristiques (*feature extraction*) de l'image. Dans un cadre traditionnel, cette étape est généralement réalisée de manière manuelle ou repose sur l'expertise métier : on choisit alors à l'avance quels prédicteurs utiliser (par exemple, des statistiques sur certaines bandes spectrales ou des filtres définis à la main). À l'inverse, dans un réseau de neurones convolutif, ce sont les poids des noyaux de convolution eux-mêmes qui sont appris automatiquement au cours de l'entraînement. Ces

coefficients font partie du vecteur de paramètres du modèle, noté  $\theta$ . Ainsi, une fois le réseau entraîné, l'ensemble des filtres optimaux  $\theta^*$  représente la meilleure façon de transformer et d'extraire les informations pertinentes de l'image. Les cartes de caractéristiques produites par l'enchaînement de ces filtres alimentent ensuite la partie classifiante du réseau. Cette approche permet au modèle de découvrir de manière autonome les motifs les plus discriminants pour la tâche visée, sans intervention manuelle dans la conception des prédicteurs.

En sortie d'un réseau de neurones convolutif classique, la partie dite classifiante permet d'associer une catégorie à l'image analysée. Par exemple, on peut vouloir déterminer si une image représente un chien ou un chat. Pour cela, la sortie de la partie convulsive (c'est-à-dire les cartes de caractéristiques extraites par les filtres) est d'abord transformée en un seul vecteur grâce à une opération de mise à plat (*flattening*). Ce vecteur est ensuite traité par un réseau de neurones dense, appelé aussi *fully connected*, qui produit en sortie un vecteur de scores  $x = (x_0, \dots, x_n)$  lorsqu'on cherche à classer l'image parmi n classes possibles.

Pour interpréter ces scores comme des probabilités, on applique la fonction *softmax* :

$$\text{Softmax } (x_i) = \frac{\exp x_i}{\sum_{j=0}^n \exp x_j}.$$

Cette opération convertit les scores en une distribution de probabilité, dont la somme vaut 1, et permet de sélectionner la classe associée à la probabilité la plus élevée comme prédiction finale. La fonction *softmax* est particulièrement utile car elle est infiniment dérivable, ce qui garantit la dérivation de tout le modèle  $f_\theta$  par rapport à ses paramètres  $\theta$ . Cela rend possible l'utilisation de la descente de gradient pour ajuster les poids du réseau au cours de l'apprentissage.

### 2.1.2. Segmentation sémantique

Comme nous l'avons vu, les réseaux de neurones convolutifs (CNN) sont particulièrement bien adaptés aux tâches de classification d'images. Historiquement, ils ont démontré leur efficacité sur des problèmes comme la reconnaissance de chiffres manuscrits (ex. : MNIST<sup>1</sup>), où l'objectif est d'assigner une étiquette à une image dans son ensemble.

Dans notre projet, la problématique est différente : il s'agit d'identifier automatiquement les zones bâties sur des images satellites. Autrement dit, on ne cherche pas à classifier une image globalement, mais à déterminer pour chaque pixel s'il appartient à une catégorie donnée (bâti ou non bâti, par exemple). Cette tâche s'inscrit dans le cadre de la segmentation sémantique, dont l'objectif est de produire une prédiction dense, c'est-à-dire une carte de labels de même dimension que l'image d'entrée.

La segmentation sémantique a connu un tournant majeur avec l'introduction des *Fully Convolutional Networks* (FCN), notamment dans l'article fondateur de J. Long, E. Shelhamer, et T. Darrell [2]. Ce travail a proposé une adaptation des CNN à la segmentation en éliminant la nécessité d'aplatiser l'image avant la classification, une étape qui, dans les CNN traditionnels, conduit à la perte de la structure spatiale de l'image (c'est-à-dire les relations entre pixels voisins), pourtant cruciale pour localiser précisément les objets.

Dans les CNN classiques, on empile des couches de convolution qui extraient des caractéristiques de plus en plus abstraites, et on réduit progressivement la résolution spatiale à l'aide

---

<sup>1</sup>[https://fr.wikipedia.org/wiki/Base\\_de\\_données\\_MNIST](https://fr.wikipedia.org/wiki/Base_de_données_MNIST)

du *pooling*. La profondeur de l'image (le nombre de canaux) augmente, mais sa taille spatiale (hauteur × largeur) diminue. Finalement, les données sont aplatis et envoyées dans une ou plusieurs couches entièrement connectées pour produire une prédiction globale. Mais cette architecture n'est pas compatible avec une prédiction localisée pixel à pixel.

L'idée clé des FCN est de supprimer ces couches entièrement connectées et l'étape d'aplatissement, et de les remplacer par des convolutions  $1 \times 1$ . Cette opération permet de transformer chaque vecteur de caractéristiques (par pixel) en un score de classe, tout en conservant la structure spatiale. On obtient alors, en sortie, une carte par classe, où chaque pixel contient une prédiction de probabilité.

Un problème subsiste toutefois : à cause des opérations de *pooling* successives, la sortie du réseau est plus petite que l'image d'entrée. Or, pour produire une prédiction par pixel à la résolution d'origine, il est nécessaire de restaurer la taille initiale. Pour cela, J. Long, E. Shelhamer, et T. Darrell [2] proposent l'utilisation de la convolution transposée (souvent appelée déconvolution, bien que ce terme soit techniquement incorrect), qui agit comme une opération inverse de la convolution en augmentant la résolution spatiale. La Figure 7 illustre schématiquement ce mécanisme.



Figure 7. – Représentation de la déconvolution

Pour résumer, la plupart des modèles de segmentation adoptent une architecture en U, composée de deux parties complémentaires :

- La branche descendante (encodeur), qui extrait progressivement des représentations de plus en plus abstraites de l'image. Elle applique des couches convolutives et des opérations de *pooling* successives afin de réduire la résolution tout en enrichissant la représentation sémantique. Le résultat est un embedding (vecteur de caractéristiques) condensé, moins volumineux que l'image initiale. Cette partie est similaire aux CNN.
- La branche montante (décodeur), qui reconstruit une carte de sortie à la même résolution que l'image d'entrée, en utilisant des opérations d'*upsampling* comme la convolution transposée. Cette partie vise à propager l'information sémantique vers les pixels, tout en rétablissant progressivement la structure spatiale.

Il est important de noter que les deux branches sont apprises : les poids des filtres convolutifs et transposés sont ajustés durant l'entraînement, via rétropropagation, comme dans n'importe quel réseau de neurones profond.

Parmi les extensions les plus remarquables aux FCN, on peut citer le modèle U-Net proposé par O. Ronneberger, P. Fischer, et T. Brox [5]. Dans les architectures FCN classiques, les couches de convolution couplée au *pooling* induisent une perte progressive de détails spatiaux fins, ce qui tend à produire des contours flous dans les cartes de segmentation. L'idée novatrice des

auteurs réside dans sa structure symétrique en U, où chaque niveau de la partie montante est connecté à son équivalent dans la partie descendante via des connexions appelées *skip connections*. Ces connexions permettent au décodeur de récupérer directement les cartes de caractéristiques locales extraites aux étapes de l'encodeur, riches en détails de bas niveau (textures, contours), tout en conservant le contexte global appris dans les couches profondes. La Figure 8 illustre cette mécanique de fusion multi-niveaux.

Une autre avancée majeure est le modèle DeepLabv3, introduit par L.-C. Chen et al. [6]. Ce dernier cherche également à atténuer la perte d'information due à l'opération de *pooling*, mais en adoptant une approche différente : l'utilisation de convolutions dilatées (*atrous convolutions*). Le principe est d'étendre la taille du noyau de convolution en insérant des espaces (zéros) entre les éléments du filtre. Cela permet d'augmenter le champ réceptif (vision globale du réseau) sans augmenter le nombre de paramètres à estimer.



Figure 8. – Représentation schématique du U-Net

Malgré les avancées considérables apportées par les architectures convolutionnelles, ces dernières présentent plusieurs limitations structurelles. Par nature, une convolution opère sur un voisinage local (défini par la taille du noyau), ce qui restreint la capacité du réseau à capturer des dépendances à longue distance dans l'image. Pour élargir le champ de vision du modèle, il faut empiler plusieurs couches convolutives, ce qui augmente la profondeur du réseau... et donc son coût computationnel.

Par conséquent, si les CNN sont très efficaces pour modéliser les détails locaux, ils peinent à capturer des relations spatiales globales. Par exemple, deux parties d'un même objet spatialement éloignées (comme les ailes d'un avion ou les extrémités d'une route) sont rarement mises en relation par un CNN standard. De plus, les réseaux convolutionnels ne sont pas

naturellement invariants au positionnement global dans l'image : une voiture située en haut à gauche ou en bas à droite peut être traitée différemment, bien qu'il s'agisse du même objet.

La révolution amorcée par les Transformers dans le traitement du langage naturel (NLP), à travers l'article fondateur de A. Vaswani *et al.* [7], a rapidement essaimé vers d'autres domaines, dont celui de la vision par ordinateur. En 2020, les Vision Transformers (ViT), introduits par A. Dosovitskiy *et al.* [8], chercheurs chez Google, ont bouleversé les approches traditionnelles de la vision artificielle. Leur particularité réside dans l'adoption du même mécanisme fondamental que celui utilisé en NLP : le *self-attention*. Ce mécanisme permet au modèle de se concentrer sur différentes parties d'une image (comme il le ferait avec des mots dans une phrase) en les pondérant dynamiquement par leurs importances, afin de mieux en comprendre la structure et le contenu.

Contrairement aux CNN qui traitent directement l'image comme une grille de pixels 2D, les Vision Transformers commencent par diviser l'image en petits blocs réguliers, appelés *patches*. Chaque *patch* est ensuite aplati et encodé en vecteur (*embedding*), exactement comme un mot l'est dans un modèle de langage. Il faut donc s'imaginer qu'une image devient une séquence de vecteurs de la même manière qu'une phrase est une séquence de mots. Cependant, contrairement aux mots dans une phrase, les *patches* d'image n'ont pas d'ordre explicite ou structure syntaxique. Pour préserver la structure spatiale, on ajoute à chaque embedding un encodage positionnel (*positional encoding*), qui fournit au modèle une information sur la position d'origine du *patch* dans l'image.

Finalement, une fois tous les *patches* encodés, ils sont passés dans le mécanisme de self-attention, qui permet à chaque *patch* de s'informer de tous les autres *patches*, en attribuant à chacun une pondération calculée dynamiquement selon leur pertinence. Ce mécanisme permet de capturer à la fois des dépendances locales (entre *patches* voisins) et globales (entre régions éloignées), ce qui est particulièrement utile pour des objets de grande taille ou des structures étendues.

Ce traitement est généralement répété à travers plusieurs couches de Transformer (notées «  $L \times$  » dans la Figure 9), permettant un enrichissement progressif des représentations à chaque niveau. À la sortie, chaque *patch* est représenté par un vecteur qui incorpore à la fois son propre contenu et son contexte global.

On obtient alors de nouveaux vecteurs qui sont des représentations des *patches* individuels, enrichis par le contexte global. Il est important de noter qu'on empile généralement plusieurs couches Transformer (par exemple, «  $L \times$  » signifie  $L$  couches dans la Figure 9). Dans la configuration originale du ViT pour la classification d'image, un token spécial est ajouté à la séquence dès l'entrée. Après passage dans les couches Transformer, le vecteur final associé à ce token est utilisé comme représentation globale de l'image, que l'on transmet à une couche entièrement connectée pour effectuer la classification.



Figure 9. – Représentation schématique du Vision Transformer (ViT)

Le succès des Vision Transformers (ViT), initialement conçus pour des tâches de classification d’images, a rapidement suscité un intérêt croissant pour leur adaptation à des tâches plus complexes, notamment la segmentation sémantique. Cependant, l’utilisation directe des ViT pour la segmentation se heurte à plusieurs limitations :

1. Les ViT standards ne conservent pas la structure spatiale de manière aussi précise que les CNN, et l’encodage positionnel appris se révèle souvent insuffisant pour localiser finement les objets au niveau du pixel.
2. Contrairement aux CNN qui construisent des représentations à différentes échelles (petits détails et vue globale), le ViT standard traite tous les *patches* à la même résolution.
3. Le mécanisme d’attention a un coût quadratique avec la taille des *patches*. Plus on veut de détails (donc des *patches* plus petits), plus ça devient lourd à calculer.

Pour répondre à ces défis, le modèle SegFormer, proposé par E. Xie, W. Wang, A. L. Yuille, A. Anandkumar, et J. M. Alvarez [9] chez NVIDIA, introduit une approche hybride qui combine intelligemment les forces des CNN (efficacité locale et structure hiérarchique) avec celles des Transformers (apprentissage du contexte global via self-attention). Le modèle SegFormer, est défini par deux composantes principales (cf. Figure 10) un encodeur (hiérarchique) basé sur des couches Transformer empilées inspiré des CNN et un décodeur très simple pour produire la carte de segmentation.

Contrairement au ViT standard qui applique le même traitement à tous les patches, SegFormer adopte une structure à plusieurs niveaux de résolution, analogue à celle des CNN. L’image est traitée par une succession de blocs Transformer, où chaque niveau réduit progressivement la résolution spatiale, construisant ainsi une hiérarchie d’abstractions.

De plus, le partitionnement de l’image se fait en *patches* chevauchants (*Overlapping Patch Merging*). Contrairement aux ViT où les *patches* sont non recouvrants (et donc spatialement disjoints), ici chaque *patch* inclut une portion de ses voisins, ce qui permet de préserver la continuité spatiale. Grâce à cette conception, et à la structure hiérarchique à plusieurs résolu-

tions, il n'est plus nécessaire d'utiliser un encodage positionnel explicite : la position est captée implicitement par le recouvrement et la profondeur du traitement.

Finalement, l'une des particularités du Segformer est également son décodeur très simple qui contraste avec les décodeurs complexes de l'U-Net ou du DeepLab. En effet, il prend les sorties des différents niveaux - 4 dans le papier - de l'encodeur qui représentent des résolutions différentes. Il les projette dans un espace commun, les interpole à la même taille, puis les concatène pour produire la carte de segmentation via quelques couches simples. Cette conception permet au SegFormer de rester relativement léger en nombre de paramètres, tout en offrant des performances compétitives. En conséquence, le modèle est rapide à entraîner, efficace à l'inférence, et bien adapté aux contraintes de déploiement en production.



Figure 10. – Représentation schématique du Segformer

Outre le SegFormer, d'autres modèles de segmentation récents peuvent être considérés comme l'état de l'art. Par exemple, on peut citer SETR, proposé par S. Zheng *et al.* [10], qui fut l'un des premiers à adapter une architecture Transformer pure à la segmentation sémantique. Le Swin Transformer, introduit par Z. Liu *et al.* [11], repose quant à lui sur une approche hiérarchique utilisant des fenêtres glissantes à déplacement progressif (*shifted windows*), permettant de mieux capter les structures à différentes échelles tout en conservant une efficacité computationnelle élevée. Enfin, le Segment Anything Model v2 (SAM-2), récemment développé par Meta (N. Ravi *et al.* [12]), représente une avancée majeure dans la segmentation universelle. Il permet de segmenter automatiquement n'importe quel objet dans une image, à partir d'un simple point d'indication ou d'un prompt, sans entraînement spécifique à un domaine donné.

Dans le cadre de notre projet, après avoir évalué plusieurs architectures, notamment DeepLabv3, notre choix s'est porté sur le modèle SegFormer-B5<sup>2</sup>, avec lequel nous avons obtenu les meilleurs résultats en termes de précision de segmentation, tout en maintenant des temps de calcul raisonnables. Nous avons utilisé la version pré-entraînée du modèle proposée par NVIDIA, que nous avons ensuite spécialisée via un apprentissage supervisé sur nos images satellites des DROM. Ce transfert d'apprentissage nous a permis de bénéficier à la fois

<sup>2</sup><https://github.com/NVlabs/SegFormer>

des connaissances générales acquises sur de grands jeux de données, et d'une adaptation fine aux caractéristiques particulières de la géographie des DROM.

## 2.2. Données

### 2.2.1. Images satellites

Une image satellite est une matrice à trois dimensions. Chaque élément de cette image est un **pixel**, qui correspond à une surface au sol (par exemple 10m\*10m). Le pixel contient plusieurs valeurs numériques. Ces valeurs expriment l'intensité du rayonnement solaire reflété dans chaque **bande spectrale** pour ce pixel. Une bande spectrale correspond à une portion du spectre électromagnétique, qui peut être par exemple le bleu, le rouge, le vert, le proche infrarouge. Donc, pour une image satellite de  $n*m$  pixels avec les trois bandes du visible (rouge, vert, bleu), chaque pixel aura trois valeurs différentes représentant l'intensité dans chacune des bandes du visible. Ainsi, un pixel qui représente 10m\*10m au sol donnera une couleur.



Figure 11. – Exemple d'un pixel d'une image satellite

Il existe de nombreux produits satellitaires disponibles. Nous nous sommes surtout concentrés sur deux d'entre eux : **Pléiades** et **Sentinel2**.



Figure 12. – Pléiades (gauche) vs Sentinel2 (droite) à Mamoudzou, Mayotte (2024)

Les images satellites **Pléiades** constituent une ressource précieuse dans notre cas d'usage. Ce sont des images à très haute résolution, spécifiquement conçues pour l'observation fine des

territoires. Elles offrent trois bandes spectrales dans le visible (**rouge, vert, bleu**) auxquelles s'ajoute une quatrième bande dans le proche infrarouge (NIR), bien que cette dernière ne soit pas disponible dans nos données actuelles. Leur résolution spatiale remarquable de **0,5m** par pixel permet une détection très précise des objets au sol, ce qui est essentiel pour nos traitements d'analyse fine.

Ces images peuvent être obtenues soit via les **archives gratuites** (sous conditions d'accord), soit par acquisition à la demande, un service payant encadré par une licence Airbus©, avec des délais d'environ 6 à 8 mois par département.

Dans le contexte du cyclone Chido, qui a frappé Mayotte en décembre 2024, un plan d'urgence a permis l'accès gratuit à une mosaïque d'images Pléiades post-cyclone couvrant l'île, seulement deux mois après la catastrophe. Une **mosaïque** est une image composite constituée d'assemblages de prises de vues réalisées à différentes dates. Elle permet d'obtenir une couverture complète du territoire, avec un minimum de zones nuageuses et une meilleure homogénéité visuelle. Ce travail de reconstitution, mené par **l'Institut National de l'information géographique et forestière** (IGN), est crucial pour garantir des données exploitables, notamment dans le cadre de l'entraînement de modèles d'analyse automatique. Pour mener à bien le projet données satellites, un accord a été signé avec l'IGN afin de fournir à l'Insee des mosaïques archivées des DROM. En pratique, une mosaïque par DROM est réalisée (et livrée) chaque année, mais en cas spécial comme pour le cyclone Chido, plusieurs mosaïques peuvent être constituées en un an.

Mais Pléiades n'est pas notre seule source d'imagerie satellite. Les satellites **Sentinel-2**, développés par l'Agence spatiale européenne (ESA) dans le cadre du programme Copernicus, offrent une alternative open source, particulièrement intéressante pour les analyses à large échelle ou à forte fréquence temporelle. Contrairement à Pléiades, Sentinel-2 capte **treize bandes spectrales**, réparties entre le visible, le proche infrarouge (NIR) et l'infrarouge à ondes courtes (SWIR), ce qui ouvre la voie à une multitude d'indices et d'analyses thématiques (comme la détection de la végétation, de l'humidité, ou des matériaux).

En revanche, la résolution spatiale de Sentinel-2 est moindre : selon les bandes, elle varie entre **10 m**, 20 m et 60 m, ce qui rend l'observation de petits objets ou de détails fins plus difficile. Néanmoins, ces images ont l'avantage d'être acquises automatiquement **tous les cinq jours et gratuites**, garantissant une fréquence de revisite élevée et une mise à disposition régulière des données, y compris en période de crise.

Ainsi, les images Pléiades semblent être les plus adaptées pour de la détection de bâtiment dans les DROM. Les mosaïques disponibles par territoire et millésimes sont décrites Figure 13.

	2017	2018	2019	2020	2021	2022	2023	2024	2025
Guadeloupe	x	x	x		x				
Martinique					x				
Guyane					x	x	x		
La Réunion		x			x	x			
Mayotte	x	x	x	x	x	x	x	x	
Bonus : Saint Martin							x		

Figure 13

### 2.2.2. Annotations

Afin d'entraîner un modèle de segmentation sémantique, les images doivent être annotées pour que le modèle apprenne afin de reproduire, sur de nouvelles images, la même méthode de segmentation. Le coût d'une annotation à la main des bâtiments sur l'ensemble des images satellites utilisées pour l'entraînement étant prohibitif, il est donc nécessaire d'explorer des sources permettant d'automatiser cette tâche. Il faut toutefois noter, que sans labellisation manuelle des bâtiments au sein des images Pléiades utilisées pour l'entraînement, un décalage entre les images et les annotations existe. Ceci altère la qualité de l'entraînement.

La **BDTOPO** est une base de type géométrique produite par l'IGN pour l'ensemble du territoire français. Celle-ci liste, entre autres, l'ensemble des constructions sous forme de polygones sur un territoire donné.

Une autre source a été explorée : le **Répertoire des Immeubles Localisés (RIL)**, produite par l'Insee. Il est donc possible de produire des zones tampons autour des points localisant les bâtiments et de s'en servir comme une approximation pour labelliser les logements. Cette base de données n'englobe donc pas tous les bâtiments, mais uniquement ceux qui sont habités. Dans les DROM, les données sont en quantité limitée, puisque chaque année seulement un cinquième de la base est mis à jour, c'est-à-dire que les données peuvent dater d'il y a plus de quatre ans. Cette source ne permet pas de faire des labellisations adaptées pour l'entraînement et a été écartée.



Figure 14. – Zone en Guyane Pleiades © CNES\_2022, points RIL 2022 (gauche) VS polygones BDTOPO 2022 (droite)

Les deux sources précédentes offrent des annotations binaires : bâtiment oui/non. Il est également possible d'utiliser des annotations multiclasses. La base **COSIA**, générée par IA à partir des photographies aériennes de l'IGN à une résolution de 20 cm, offre des polygones multiclassés précis, disponibles également pour les DROM. Parmi ces classes, la classe bâtiment est incluse, ce qui constitue un avantage pour les performances du modèle. C'est cette source qui a été retenue pour l'entraînement.



Figure 15. – Zone à Mayotte Pleiades © CNES\_2023, couverture du sol COSIA 2023 et légende COSIA

La Figure 16 informe de la couverture du sol COSIA disponible par millésime et par territoire.

	2017	2018	2019	2020	2021	2022	2023	2024	2025
Guadeloupe	x				x			x	
Martinique	x				x				
Guyane		x			x		x		
La Réunion	x				x			x	
Mayotte						x			

Figure 16

### 3. Résultats

Les mosaïques d'images Pléiades et les annotations COSIA de Mayotte (2023), Martinique (2022) et Guadeloupe (2022) ont été utilisées pour l'entraînement du modèle. Des zones de test ont été sélectionnées manuellement au sein de ces couples images annotations, pour représenter une diversité de contextes (quartiers résidentiels, bidonvilles, zones naturelles, centres urbains).

Une image Pléiade telle qu'elle est livrée par l'IGN contient 2000\*2000 pixels. Elle est trop volumineuse pour être prise en entrée d'un réseau de neurones avec les capacités usuelles disponibles de mémoire vive du processeur graphique **GPU** (Graphics Processing Unit). Pour prévenir cela, l'ensemble des images des mosaïques sont découpées en plusieurs sous-images de 250\*250 pixels traitées indépendamment, appelées des tuiles.

Comme un pixel sur une image Pléiades équivaut à  $0.5 \times 0.5 \text{m}^2$ , alors **un tuile a une superficie de 15 625m}^2**. La Figure 17 montre le nombre de tuiles par mosaïque envoyés pour l'entraînement mais également le nombre de tuiles gardés pour l'évaluation. Pour mieux comprendre ce que cela représente, la superficie associée est exprimée en km<sup>2</sup>.

	train ( <b>superficie km</b> <sup>2</sup> )	test ( <b>superficie km</b> <sup>2</sup> )
Guadeloupe (2022)	>100 000 ( <b>1 600</b> )	1280 ( <b>20</b> )
Martinique (2022)	>70 000 ( <b>1 100</b> )	1230 ( <b>19</b> )
Mayotte (2023)	23 000 ( <b>357</b> )	840 ( <b>13</b> )

Figure 17

La fonction de perte (loss) utilisée est la **CrossEntropyWeighted**, avec un poids de 2 pour la classe bâtiment. Cette loss est une version pondérée de la CrossEntropy, qui permet de donner plus d'importance aux classes rares, comme les bâtiments, afin de compenser leur sous-représentation dans les données. En effet, sur un territoire, la surface de bâtiment est relativement faible comparé à la surface de verdure ou de sol nu.

La **CrossEntropyWeighted** pour un pixel  $i$  s'écrit :

$$\text{CEWeighted}_i = -w_c \cdot \log(p_{i,c^*})$$

avec :

- $c^*$  : classe vraie du pixel  $i$ ,
- $p_{i,c^*}$  : probabilité prédite pour cette classe,
- $w_c$  : poids associé à la classe

La loss finale sur tous les pixels est la moyenne :

$$\text{Loss} = \frac{1}{N} \sum_{i=1}^N \text{CEWeighted}_i$$

Il est toutefois difficile de juger de l'efficacité de l'entraînement uniquement avec cette métrique. Pour avoir une idée plus concrète des performances du modèle, une autre métrique, largement utilisée en segmentation d'images, a été implémentée: l'**Intersection Over Union** (IOU), représentée en Figure 18.



Figure 18. – Présentation du calcul de l'IOU

L'IOU se définit comme le rapport entre l'intersection des annotations et des prédictions et leur union. Plus l'IOU est proche de 1, plus les prédictions correspondent à la réalité. Dans la pratique, une IOU considérée comme très satisfaisante se situe autour de 0.6, car un certain décalage entre les annotations et les prédictions est inévitable. Néanmoins, cela n'empêche pas d'obtenir des prédictions fiables et qualitatives.

Pour notre modèle, l'IOU moyenne lors de l'entraînement se trouve autour de **0.81** pour l'ensemble des classes, et plus précisément de **0.84** pour la classe bâtiment. Dans les zones de test, on retrouve des IOU respectivement de **0.78** et **0.76**.

La Figure 19 illustre une portion de zone de test issue de la mosaïque Mayotte 2023, avec en Figure 20 la réalité terrain COSIA 2023 (gauche) et les prédictions produites par le modèle entraîné (droite).



Figure 19. – Zone de test Mayotte Pleiades © CNES\_2023 et légende COSIA



Figure 20. – Couverture de sol COSIA 2023 VS prédictions

### 3.1. Inférence

Dans le cadre de notre projet de détection automatique des zones bâties à partir d'images satellites, il est essentiel de distinguer les phases d'entraînement du modèle et d'inférence. L'entraînement correspond à la phase exploratoire du projet, mobilisant des ressources matérielles conséquentes (notamment des GPU) et nécessitant de nombreux choix méthodologiques (prétraitement des images, ajustement des hyperparamètres, choix du modèle, etc.). Cette étape, une fois réalisée, n'est plus réellement centrale et peut être améliorée à la marge de manière indépendante.

À l'inverse, la phase d'inférence s'apparente à une mise en production opérationnelle du modèle entraîné. Une fois les poids du modèle estimés, la génération de prédictions devient relativement peu coûteuse, notamment grâce à la possibilité d'exécuter les inférences sur CPU avec un temps de latence acceptable. L'enjeu devient alors de concevoir un dispositif fiable, reproduitible et facilement mobilisable pour mettre à disposition les résultats du modèle, en particulier lorsque de nouvelles images satellites sont acquises.

Actuellement, les inférences sont réalisées, comme l'entraînement, sur des tuiles disjointes de taille  $250 \times 250$  pixels. Ce découpage présente plusieurs inconvénients. Elle introduit des artefacts de bordure, limite la vision contextuelle du modèle, et empêche une prise en compte fluide des objets situés à la jonction de deux tuiles.

Afin d'améliorer la cohérence spatiale des prédictions, nous mettons en œuvre une stratégie d'inférence par fenêtre glissante (*sliding window*). Celle-ci consiste à réaliser plusieurs prédictions pour un même pixel, à partir de fenêtres décalées les unes par rapport aux autres. Les probabilités ainsi obtenues sont ensuite moyennées, ce qui permet de lisser les effets de bord, de renforcer la stabilité des contours et de réduire le bruit.

Cette inférence multiple est complétée par un pipeline de post-traitement multiclasse. Celui-ci repose d'abord sur une stratégie de repli contextuel : pour les pixels dont la probabilité maximale est inférieure à un certain seuil, la classe est réattribuée en fonction du voisinage local (via vote majoritaire ou moyenne pondérée des probabilités). Ensuite, des opérations de morphologie mathématique sont appliquées classe par classe afin de supprimer les artefacts et lisser les masques de segmentation. Enfin, un filtrage par taille minimale est mis en œuvre, avec des seuils spécifiques à chaque classe (par exemple, un bâtiment peut être plus petit

qu'une zone d'eau). Ce traitement permet d'éliminer les objets aberrants ou trop petits pour être statistiquement significatifs.

Afin d'industrialiser l'ensemble de la phase d'inférence, une API a été développée. Elle encapsule l'ensemble des étapes précédentes (inférence, post-traitement) et permet une exploitation simple et flexible du modèle par les différentes équipes concernées. Trois points d'entrée sont disponibles :

1. fourniture directe d'une image satellite à prédire,
2. sélection d'une zone géographique via coordonnées GPS (*bounding box*) et d'une année d'observation,
3. saisie d'un identifiant d'îlot (entité géographique infra-communale) et d'une année d'observation pour obtenir les prédictions correspondantes.

L'API intègre un mécanisme de cache qui évite de recalculer une prédition déjà effectuée pour une même zone géographique, un même modèle et une même année. L'ensemble est déployé sur CPU et permet une chaîne de traitement entièrement automatisée, assurant une inférence rapide dès la réception de nouvelles images satellites.

Conçue comme un outil évolutif au service des équipes métiers de l'Insee, l'API est amenée à s'adapter aux besoins exprimés. Par exemple, à la suite d'un besoin identifié concernant le calcul de statistiques par îlot, une nouvelle fonctionnalité a été ajoutée pour retourner non seulement la carte de prédictions, mais aussi des indicateurs agrégés, tels que les surfaces bâties par îlot.

Néanmoins, la diffusion de résultats via une API n'est pas toujours la modalité la plus adaptée aux usages internes. C'est pourquoi deux modes de restitution complémentaires ont été développés : d'une part, la production de fichiers Parquet contenant les résultats structurés pour exploitation statistique ; d'autre part, une application interactive cartographique, destinée aux agents de terrain. Cette application s'appuie sur un GeoServer pour diffuser les images satellites, les cartes de segmentation, ainsi que les évolutions détectées (créations, destructions de bâti), facilitant ainsi le croisement avec les informations d'enquête et les validations sur le terrain.



Figure 21. – Schéma de la pipeline

## 3.2. Mise à disposition pour les statisticiens

### 3.2.1. Vers de nouveaux indicateurs statistiques ?

Contrairement à d'autres applications de l'apprentissage automatique déjà déployées à l'Insee, comme les modèles de classification de textes pour la codification automatique dans des nomenclatures, les sorties brutes d'un modèle de segmentation d'images ne constituent pas, en soi, des données statistiques directement exploitables ou diffusables. Une carte de segmentation s'apparente davantage à un support intermédiaire, qui doit faire l'objet d'une interprétation, d'un traitement spatial et d'une contextualisation pour devenir une information statistique pertinente. Comme le souligne le Mémorandum de Varsovie, ces nouvelles formes de données issues de sources non traditionnelles peuvent cependant jouer un rôle essentiel : elles peuvent soit conduire à la construction de nouveaux indicateurs, soit servir de proxies statistiques utiles pour améliorer la qualité ou la finesse d'indicateurs existants. En ce sens, le développement d'un modèle de segmentation performant n'a de valeur, dans le contexte de la statistique publique, que s'il permet effectivement de produire de l'information utile, fiable, et intégrable dans les systèmes d'observation existants. À défaut, l'usage d'un tel outil par les statisticiens resterait discutable, car sans finalité opérationnelle claire.

À ce stade du projet, l'enjeu est donc de déterminer les indicateurs statistiques pouvant être dérivés des sorties du modèle. Cette démarche doit reposer sur une co-construction entre les équipes métiers de l'Insee, qui expriment les besoins statistiques concrets, et la division Méthodes pour la géographie et la confidentialité (DMGC), appuyée par le SSP Lab, qui assure l'évaluation de la faisabilité technique, la pertinence méthodologique, et le développement des outils associés.

Plusieurs cas d'usage ont d'ores et déjà été identifiés. L'un d'eux, détaillé en Section 4, concerne la construction d'indicateurs d'évolution du bâti. Étant donné que le modèle attribue une classe à chaque pixel, et que chaque pixel représente une surface de  $0,25m^2$ , il est

possible de calculer la surface bâtie dans une zone géographique donnée à une date donnée. En comparant ces surfaces sur plusieurs dates, on peut alors quantifier les dynamiques de construction ou de disparition de bâtiments. Croisées avec d'autres sources de données, comme le Répertoire d'immeubles Localisés (RIL), ces surfaces bâties peuvent fournir des proxies de l'évolution du parc de logements et, par extension, de la population. De telles approches ouvrent des perspectives intéressantes pour l'enrichissement des dispositifs de suivi, en particulier dans les zones où les données administratives sont lacunaires ou peu à jour comme c'est le cas dans certains DROM.

Concrètement, nous avons produit des fichiers millésimés pour chacun des DROM, retraçant l'évolution de la surface bâtie par îlot, pour chaque année disponible. Cette base de données constitue une ressource structurée et directement exploitable pour des analyses statistiques, déjà mobilisée par la division Méthodes et Traitements des Recensements (DMTR) pour évaluer la qualité de l'enquête cartographique. Ces premiers résultats illustrent le potentiel d'une telle approche pour renforcer la couverture, la réactivité et la granularité des indicateurs mobilisés dans les travaux de la statistique publique territorialisée.

### **3.2.2. Une application interactive pour les agents de terrain**

Si les résultats agrégés issus du modèle de segmentation, notamment les bases d'évolution des surfaces bâties, trouvent des applications dans les analyses post-enquête cartographique, l'appui à la production des chiffres du recensement ou encore l'évaluation de politiques publiques comme l'objectif Zéro Artificialisation Nette (ZAN), un autre besoin a rapidement émergé du terrain. Les agents en charge de l'organisation des enquêtes cartographiques ont exprimé le besoin de pouvoir anticiper la charge de travail que représente une zone, notamment dans les territoires où l'habitat précaire est très dynamique. Dans ces zones, les évolutions rapides du bâti peuvent rendre les opérations d'enquête particulièrement complexes à planifier et à gérer. Plus précisément, ces agents souhaitent pouvoir visualiser les changements intervenus entre deux campagnes d'enquête, afin d'identifier les secteurs à forte évolution et de prioriser efficacement les ressources. Ils ont ainsi besoin d'un outil leur permettant une lecture intuitive et rapide des transformations spatiales, en amont du travail de terrain.

Pour répondre à ces attentes, une application web interactive a été développée. Son objectif principal est de rendre accessibles, de manière simple et ergonomique, les résultats issus du modèle de segmentation aux agents de terrain. L'application permet de visualiser l'ensemble des images satellite disponibles pour les différents départements d'outre-mer (Guadeloupe, Martinique, Guyane, La Réunion, Mayotte), selon différents millésimes. Les utilisateurs peuvent ainsi comparer les images entre deux années données afin d'identifier visuellement les évolutions notables. Au-delà de la simple visualisation des images, l'application intègre également les cartes de segmentation générées par le modèle. Ces cartes fournissent une lecture globale et à grande échelle de la répartition des classes (bâti, végétation, etc.). Afin d'optimiser la détection des évolutions, des masques de changement ont été calculés entre deux années à partir des cartes de segmentation. Ces masques mettent en évidence les zones de création ou de transformation du bâti, facilitant ainsi l'identification des secteurs à fort changement.



Figure 22. – Application web, statistiques sur le bâti (Mayotte 2017/2018)

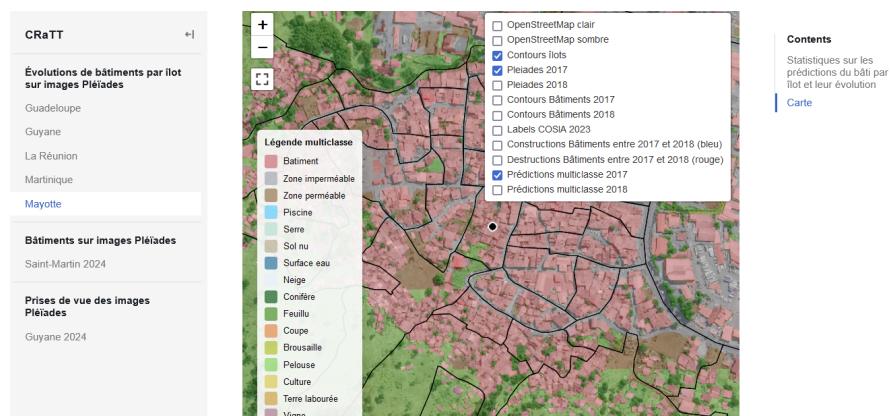


Figure 23. – Application web, visualisation des prédictions (Mayotte 2017)

L’application intègre également des indicateurs statistiques directement liés aux îlots. L’application contient les délimitations des îlots, et l’utilisateur peut accéder aux statistiques associées en cliquant sur un îlot spécifique. Une table de synthèse recense l’ensemble des statistiques disponibles pour tous les îlots, avec la possibilité d’effectuer des recherches ciblées. Il est ainsi possible de rechercher un îlot par son identifiant, d’accéder à ses données et de recentrer automatiquement la carte. Les statistiques de surface, présentées en Section 3.2.1, sont ainsi pleinement exploitables dans l’outil. L’application est accessible en ligne à l’adresse suivante : <https://inseefrlab.github.io/satellite-images-webapp/>.

L’application a été conçue comme un outil modulaire, destiné à évoluer en fonction des besoins exprimés par les utilisateurs. Toute remontée de terrain est prise en compte pour améliorer son ergonomie et ses fonctionnalités. L’objectif est de construire un outil aligné sur les usages opérationnels des enquêtes cartographiques dans les DROM.

L’un des enjeux majeurs de cette approche reste la qualité des données d’entrée. La rapidité d’acquisition des images ainsi que leur qualité (notamment l’absence de nuages) sont des conditions indispensables à l’utilité et la fiabilité des résultats. De ce point de vue, un partenariat renforcé avec l’Institut national de l’information géographique et forestière (IGN) apparaît comme indispensable. L’IGN dispose à la fois des expertises nécessaires en prétraitement d’images satellite et de ressources pour garantir un flux de données pertinent pour les usages statistiques.

## **4. Application : contrôle de la qualité des enquêtes cartographiques du recensement de la population à Mayotte**

### **4.1. Le recensement de la population à Mayotte**

Depuis 2004, le recensement de la population est réalisé, en France, par sondage lors d'enquêtes annuelles de recensement (EAR). La méthodologie employée diffère selon la taille de la commune : dans les communes de moins de 10 000 habitants, un recensement exhaustif de la population est effectué tous les 5 ans alors que, dans les communes de 10 000 habitants ou plus, une interrogation d'environ 8 % des logements de la commune est réalisée chaque année.

À Mayotte, le recensement exhaustif, réalisé tous les 5 ans sur l'ensemble du territoire, est resté en vigueur jusqu'en 2017. Lorsque que Mayotte est devenu un département français, la volonté des acteurs locaux a été d'avoir un recensement similaire à celui des autres départements. La première EAR a donc eu lieu à Mayotte en 2021, avec un fonctionnement sur un cycle quinquennal, comme en métropole et dans les autres DROM. En 2025, du fait du cyclone Chido, la cinquième et dernière enquête du premier cycle n'a pas pu être réalisée. Il a alors été décidé de réaliser de nouveau un recensement exhaustif de la population, au plus tôt, afin de pouvoir authentifier de nouvelles populations de référence. Ce recensement exhaustif aura lieu du 27 novembre 2025 au 10 janvier 2026. Les travaux mentionnés ci-dessous sont antérieurs à cette opération, ils font référence aux enquêtes ayant eu lieu de 2021 à 2024.

À Mayotte ainsi que dans l'ensemble des autres DROM, en l'absence de sources administratives fiables pour alimenter le répertoire d'immeubles localisés (RIL), des enquêtes cartographiques conduites par des enquêteurs de l'Insee sont réalisées chaque année entre cinq et dix mois avant l'EAR pour repérer l'ensemble des logements en amont de la collecte. Pour étaler la charge d'enquête dans le temps, les enquêtes ne sont menées chaque année que sur un cinquième du territoire de chacune des grandes communes. Pour ce faire, la commune a été divisée en îlots, qui est un ensemble d'habitations contigües délimitées par des frontières facilement repérables sur le terrain (voie, cours d'eau...). Ces îlots ont été répartis en cinq groupes de rotation (GR) correspondant à l'année à laquelle a lieu l'enquête annuelle de recensement. Les îlots d'un groupe qui feront l'objet de l'enquête de recensement l'année N feront l'objet d'une enquête cartographique préalable durant l'année N-1.

Le résultat de cette enquête cartographique détermine la base de sondage d'adresses (BSA) dans laquelle est tiré l'échantillon de l'EAR pour les communes de plus de 10 000 habitants. Cet échantillon représente 40 % des logements des îlots du groupe à enquêter. Toutefois, les spécificités du terrain mahorais ont fait que le plan de sondage habituellement employé dans les DROM a dû être adapté<sup>3</sup>. En effet, 39 % des logements recensés au RP 2017 sont des habitations de fortune (zones de bangas). Ce type d'habitat évolue très rapidement et pose des problèmes de repérage sur le terrain. Cela peut affecter les estimateurs produits par la méthode classiquement appliquée dans les grandes communes des DROM, principalement à cause des défauts de couverture (à la hausse ou à la baisse) engendrés. En effet, de nombreuses habitations de fortune ont rendu inadaptées les techniques d'échantillonnage qui supposent que la situation soit cohérente entre l'enquête cartographique réalisée au printemps et à l'été N-1 et l'enquête de recensement réalisée au début de l'année N. De ce fait, les habitations

<sup>3</sup>Pour plus d'informations sur la définition du plan de sondage à Mayotte, se référer au document de travail n°2022-10.

en dur<sup>4</sup> et les habitations de fortune ont été traitées différemment. Les habitations de fortune ont été recensées exhaustivement pour que l'enquête correspondent bien à la réalité observée au moment de la collecte. L'EAR porte par ailleurs sur 40 % des logements en dur. Au total, près de deux tiers des logements du groupe de rotation ont été enquêtés à Mayotte une année donnée. Pour les petites communes mahoraises, le recensement était exhaustif, comme dans en métropole et dans les autres DROM.

## 4.2. Contrôler la couverture de l'enquête cartographique avec les données du bâti

### 4.2.1. Objectif : Identifier les zones sous-couvertes éventuelles dans l'enquête cartographique

L'enquête cartographique, qui permet de définir la base de sondage, joue donc un rôle dans la détermination des poids de sondage. Si l'enquête cartographique omet des zones à enquêter une année donnée, les populations et nombre de logements déduits de l'EAR - et du recensement de la population (RP) auquel elle contribue - seront sous-estimés. Cela concerne aussi bien les logements en dur, qui sont recensés par sondage mais également les logements non durs pour lesquels le recensement est exhaustif, même en grande commune. En effet, il peut être difficile pour les agents recenseurs de ratisser l'ensemble du territoire au moment de la collecte pour recenser des logements non durs qui n'auraient pas été repérés à l'enquête cartographique, notamment dans des zones isolées.

Comme la base de sondage de l'EAR est définie à l'aide de l'enquête cartographique, il a été décidé en 2024 d'effectuer un contrôle de la qualité de cette enquête. Pour ce faire, les données satellites décrites en partie I ont été mobilisées afin de comparer l'évolution de la couverture de bâti obtenue à l'aide du modèle de prédiction avec celle du nombre de logements observé entre la collecte de 2017 et de les enquêtes cartographiques des EAR 2021 et 2022. Cette comparaison a été réalisée en agrégeant les données de surface au sein de chaque îlot, pour une année d'enquête donnée.

Il est possible que la surface bâtie évolue différemment de l'évolution du nombre de logements. C'est le cas si un bâtiment commercial s'implante quelque part : cette implantation laisse inchangée l'évolution des logements des enquêtes cartographiques alors qu'elle est comptabilisée dans les données du bâti issues des données satellites. Malgré ces limites, l'utilisation des données satellites permet d'identifier efficacement les zones où, du fait de difficultés de terrain, la qualité de l'enquête cartographique peut-être jugée de moins bonne qualité.

### 4.2.2. Méthode : Identifier les îlots problématiques en comparant évolution du bâti et des logements

Nous mobilisons ici les données satellites afin d'identifier les îlots potentiellement sous-couverts dans les enquêtes cartographiques de 2021 et 2022. Une fois ce travail réalisé, une enquête cartographique complémentaire a été mise en place dans ces îlots en 2024.

#### Données utilisées

Deux évolutions ont été comparées, nécessitant l'utilisation de deux sources de données :

<sup>4</sup>Modalités 2, 3 et 4 de l'aspect du bâti (case traditionnelle, maison ou immeuble en bois, maison ou immeuble en dur).

1. La prédiction de surface de bâti en 2017, 2021 et 2022 issue du modèle de segmentation sémantique mobilisé en partie I est utilisée pour mesurer l'évolution du bâti issue des données satellites entre 2017 et 2021 puis entre 2017 et 2022.
2. L'évolution du nombre de logements entre la collecte 2017 et l'enquête cartographique de 2020 (pour l'EAR 2021) et 2021 (l'EAR 2022).

#### Construction d'un score/indicateur de cohérence

Pour définir un indicateur qui priorise les îlots à examiner, le rapport entre l'évolution du bâti et celle des logements en dur d'une part et l'ensemble des logements d'autre part a été calculé.

Dans un premier groupe d'îlots, les deux évolutions sont concordantes (rapport proche de 1). Dans un deuxième groupe d'îlots, plus petit, l'évolution issue des enquêtes cartographiques est plus importante que celle issue des données satellites (rapport inférieur à 1). Dans un troisième et dernier groupe d'îlots, l'évolution du bâti issu des données satellites est plus importante que l'évolution issue des enquêtes cartographiques (rapport supérieur à 1). L'objectif étant d'évaluer une potentielle sous-couverture du bâti lors de l'enquête cartographique, on s'intéresse uniquement aux îlots du troisième groupe.

##### **4.2.2.1. Définition d'un score total *score***

Un score total appelé *score* a été défini de manière à identifier les îlots pour lesquels il semble y avoir le plus de sous-couverture. Plus celui-ci est élevé, plus la sous-couverture est supposée importante.

Ce score total représente la somme de trois scores explicités ci-dessous. Ces scores ont été calculés pour 2 sous-catégories de logements :

- 1. pour l'ensemble des logements
- 2. pour les logements dits « durs » uniquement

Pour chacun de ces sous-ensembles,  $score = score_{évolution} + score_{sens} + score_{distance}$  et vaut entre 0 et 5.

##### **Une évolution du nombre de logements cartographiés moindre que l'évolution du bâti ?**

Pour chaque îlot, le  $score_{évolution}$  vaut :

- 1 si l'évolution du nombre de logements dans le recensement est plus faible que celle de la surface bâtie (données satellites)
- 0 sinon

##### **Une évolution du nombre de logements cartographiés opposée à l'évolution du bâti ?**

Pour chaque îlot, le  $score_{sens}$  vaut :

- 1 si l'évolution du nombre de logements dans le recensement est dans le sens opposé de celle de la surface bâtie (données satellites)
- 0 sinon

##### **Importance du différentiel d'évolution entre logements cartographiés et bâti ?**

On mesure ici l'importance du différentiel de taux d'évolution entre le nombre de logements dans le recensement et la surface bâtie. Pour cela, on calcule la différence en valeur absolue

entre l'évolution de la collecte et du bâti. On calcule ensuite la distribution de ces différences pour l'ensemble des îlots à examiner (premier quartile, médiane et dernier quartile).

Pour chaque îlot,  $score_{distance}$  vaut :

- 0 si  $distance\_ilot \leq Q1$
- 1 si  $Q1 < distance\_ilot \leq \text{médiane}$
- 2 si  $\text{médiane} < distance\_ilot \leq Q3$
- 3 si  $distance\_ilot > Q3$

#### **4.2.2.2. Définition des ordres de priorité**

L'ordre de priorité est calculé en prenant en compte, à la fois le score total sur l'ensemble des logements d'une part et les logements en dur uniquement d'autre part (tableau Table 1).

Table 1. – Ordres de priorité des îlots avec sous-couverture de logements supposée

Ordre de priorité	Score (ensemble des logements – logements durs)
1	5-5
2	4-5 ou 5-4
3	4-4
4	3-5 ou 5-3
5	3-4 ou 4-3
6	3-3
7	2-5 ou 5-2
8	2-4 ou 4-2

*Note de lecture : L'ordre de priorité 2 (score 5-4 ou 4-5) signifie que le score total vaut 5 pour l'ensemble des logements et 4 pour les logements durs ou inversement.*

Finalement, tous les îlots avec un ordre de priorité 1 à 8 ont pu bénéficier d'une cartographie complémentaire

#### **4.2.3. Un exemple d'îlot**

Dans cet îlot de la commune de Bandrele (voir Figure 24), les contours bleus représentent la prédiction des bâtis en 2021, les points verts ceux cartographiés initialement en 2021 et les points rouges représentent les logements identifiés dans une enquête cartographique complémentaire réalisée en 2024.

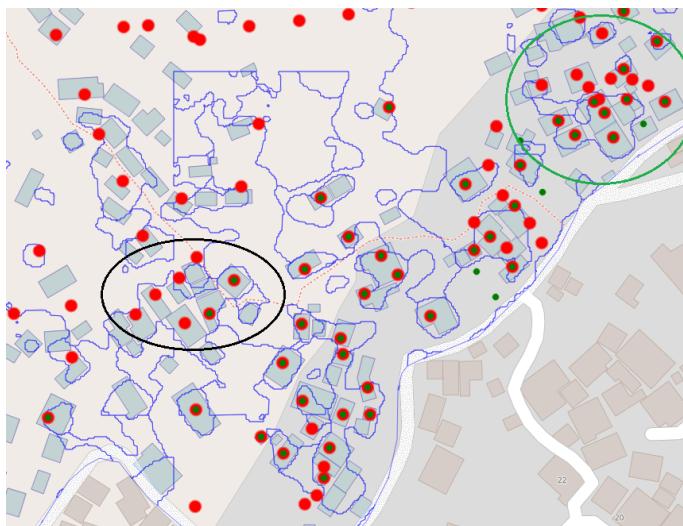


Figure 24. – Exemple d'un examen en 2024 d'un îlot de 2021

D'une part, on observe l'apparition de nouveaux logements en 2024, qui ne semblaient pas exister au moment de l'enquête cartographique initiale (en 2021). Ces ajouts sont particulièrement visibles sur la partie supérieure droite de l'image dans la zone circulaire verte : les points rouges apparaissent sans contours bleus.

Néanmoins, des logements semblent avoir été omis lors de la cartographie initiale de 2021 et ont pu être ajoutés lors du réexamen de 2024. Ceci est bien illustré par les logements présents par exemple dans la zone circulaire noire. On y remarque que plusieurs logements semblaient exister (présence d'un contour bleu), mais seuls deux avaient été cartographiés, matérialisés par les points verts. Ce constat confirme que l'utilisation d'images satellites traitées peut permettre de s'assurer de l'exhaustivité de la collecte des logements.

### 4.3. Application : bilan et perspectives

Dans les premières enquêtes annuelles du recensement à Mayotte, les données satellites ont permis de cibler des îlots pour lesquels les logements identifiés dans l'enquête cartographique initiale paraissent sous-estimés. Une enquête cartographie corrective a pu être réalisée dans ces îlots, améliorant la qualité des pondérations des logements dans les enquêtes annuelles de recensement et qui aurait pu être prise en compte avant la publication des populations si le premier cycle quinquennal avait été mené à son terme.

Cette expertise des données satellites et de leur traitement a toutefois de nouveau été mobilisée lors de l'enquête cartographique 2025, enquête exhaustive préalable au recensement exhaustif de la population prévu du 27 novembre 2025 au 10 janvier 2026. Les données satellites de Pléiades de mars 2025, postérieures au cyclone Chido, ont, en effet, également été utilisées dans le cadre de cette enquête cartographique préparatrice à la collecte exhaustive de fin 2025. En effet, d'une part pour orienter au mieux la collecte, ces fonds de cartes sont présents sur les tablettes des enquêteurs. D'autre part, des contrôles qualité ont été effectués en bureau au fur et à mesure de la collecte pour garantir son exhaustivité.

Bien que l'utilisation des données satellites permette de diminuer le risque de sous-couverture des logements à Mayotte, la prédiction de surface du bâti du modèle de segmentation sémantique ici mobilisé comporte quelques limites. Premièrement, la qualité des images avec notamment la présence de nuage peut nuire à la qualité des prédictions. Deuxièmement, les

images satellites utilisées ne sont livrées, en général qu'une fois par an. Comme le bâti peut évoluer rapidement, le décalage entre la prédiction réalisée à l'année n et la situation réelle sur le terrain peut entraîner des différences notables. Troisièmement, les données satellites fournissent une estimation de la surface bâtie au sol. Néanmoins, elles ne fournissent pas d'informations sur le nombre de logements. Par exemple, la surface au sol d'une maison peut être identique à celle d'un immeuble de plusieurs étages alors que leur nombre de logements peut être très différent. Par ailleurs, la prédiction s'appuie sur tous types de bâtis, y compris les locaux à vocation industrielle et commerciale. Or, la surface au sol de ces locaux peut être importante, notamment dans des îlots contenant des zones industrielles ou commerciales. Par exemple, dans 2 îlots de Kaweni (voir Figure 25), le bâti de la zone industrielle (surface rose) est important par rapport au bâti résidentiel. Pour un de ces 2 îlots, plusieurs établissements scolaires sont également présents. Ces bâtiments ont une influence sur l'évolution du bâti puisqu'une destruction ou une construction de ce type de bâti faussera l'évolution du bâti résidentiel que l'on cherche à approcher via les données satellites.

Les images satellites peuvent donc servir pour prédire les moyens à mobiliser zone par zone (il est plus long de cartographier une zone où le bâti a fortement évolué), pour contrôler/questionner en bureau le travail de repérage des enquêteurs sur le terrain mais elles ne le remplacent pas. Arpenter tout le territoire de façon systématique et méthodique reste indispensable à la qualité et à la précision du recensement.

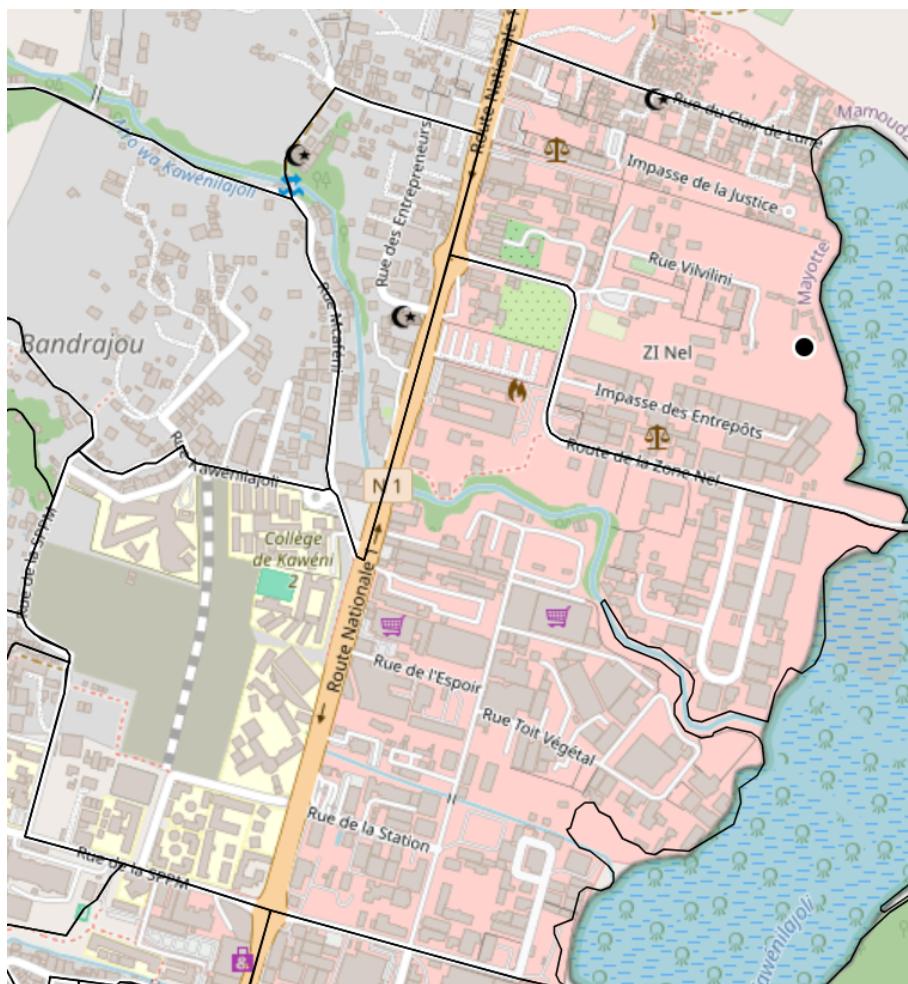


Figure 25. – Présence d'une zone industrielle dans 2 îlots de Kaweni

## 5. Conclusion et perspective

L'objectif initial du projet mené autour des images satellites était de développer une méthode fiable pour repérer automatiquement les logements dans les DROM. Ce projet visait à aider les travaux de contrôle de la qualité des enquêtes cartographiques à Mayotte ainsi qu'à appuyer les enquêtes cartographiques réalisées dans les DROM en amont de la collecte du recensement. L'idée était de suivre les évolutions du bâti, en identifiant des nouvelles constructions et des démolitions, dans des environnements où les données administratives souvent incomplètes ou de qualité insuffisante.

Pour détecter ces évolutions, l'utilisation d'images satellitaires s'avérait particulièrement adaptée. La segmentation sémantique d'images permet d'identifier et de délimiter les objets d'intérêt présents sur une image (ici les bâtiments). L'entraînement d'un modèle de deep learning constitue ainsi une approche pertinente pour extraire automatiquement les informations relatives aux bâtiments et à leur localisation, afin de comparer l'évolution du bâti sur un même territoire observé à plusieurs dates.

En explorant différentes architectures de deep learning, notamment UNet, DeepLabV3+, et SegFormer, nous avons démontré la capacité de ces modèles à segmenter automatiquement les images satellites et à identifier les changements du bâti. Le modèle SegFormer s'est distingué par sa précision, bien qu'il demande plus de ressources computationnelles que DeepLabV3+, son principal concurrent ici.

Les recherches menées sur les données d'entraînement, portant à la fois sur le choix de la source satellitaire et sur les annotations, ont conduit à retenir les images Pléiades, en raison de leur très haute résolution spatiale, bien qu'elles soient plus difficiles à acquérir que les images Sentinel-2. Par ailleurs, la couverture du sol COSIA a été sélectionnée pour la richesse de ses classes thématiques et la finesse de sa précision, offrant ainsi de nombreuses possibilités d'exploitation, notamment au-delà de la seule détection du bâti.

Ces travaux de prédiction du bâti ont été utilisés afin de contrôler la qualité des enquêtes cartographiques à Mayotte. Ils ont permis de repérer des îlots où le nombre de logements semblait sous-estimé lors de l'enquête cartographique initiale, conduisant à la réalisation d'une enquête cartographique corrective, dans ces îlots. Cette enquête corrective a contribué à améliorer la qualité des pondérations de logements utilisées dans les enquêtes annuelles de recensement et auraient pu être intégrés avant la publication des populations si le premier cycle quinquennal avait été mené à son terme.

Toutefois, il reste plusieurs axes d'amélioration à exploiter, sur chaque partie de la chaîne de traitement produisant les prédictions de bâtiments. Les usages finaux des travaux présentés (estimations de population et planification de charge) dépendent entièrement de la qualité des prédictions réalisées par l'algorithme. Il faut noter que les prédictions ne distinguent pas les bâtiments des logements, et encore moins le nombre de logements à son sein. Cette distinction est difficilement réalisable via une image satellite, car même un œil humain ne saurait trancher. Également, un travail manuel sur les données peut être effectué pour optimiser l'entraînement en assurant une cohérence temporelle entre les images et les annotations. Il est cependant difficile d'estimer le rapport entre ce coût et les gains qui seront enregistrés sur les prédictions.

Les réflexions sur l'algorithme sélectionné sont tout aussi importantes. En effet, la littérature scientifique est foisonnante sur les modèles de Segmentation et il est donc nécessaire de

réaliser une veille technique permanente sur le sujet. Certains modèles peuvent en théorie s'adapter à des images de résolutions différentes. Une veille sur les pratiques des autres instituts sur l'utilisation de l'imagerie satellitaire est indispensable. Les contacts avec l'IGN devraient être également renforcés dans la mesure où la constitution même de COSIA réside dans des travaux sur l'imagerie aérienne. Enfin, des échanges plus fréquents avec des acteurs du monde académique, voire la mise en place de projets de recherche visant exclusivement à répondre aux cas d'usages mentionnés en introduction, profiteraient grandement à l'avancée du projet.

Du point de vue opérationnel seulement, partant du principe que la méthode de constitution des données et le choix de l'algorithme sont arrêtés, l'intégration de l'outil dans le processus de production Insee n'est pas aisée. En amont déjà, l'obtention des images devrait être internalisée à l'Insee avec la création d'un service responsable de cette acquisition. De même, les temps moyens d'acquisition i.e. le décalage entre la date de commande de l'image satellite et son obtention effective devraient être mesurés afin de prendre la mesure du décalage potentiel entre la vérité du terrain et celle photographiée. Des entraînements devraient aussi être réalisés assez fréquemment afin d'actualiser l'algorithme avec de nouvelles données et améliorer ses capacités de prédiction. Ensuite, les résultats des algorithmes pourraient être expertisés par les agents en bureau qui vérifieront la pertinence des prédictions en les superposant aux images dont elles sont issues. Un travail autour du calcul des différences de bâti entre deux millésimes sur un même territoire doit aussi être approfondi, pour cibler les réelles constructions/destructions et non les décalages de prédictions du modèle. Cette phase d'expertise permettrait également aux agents d'émettre des propositions d'amélioration de l'outil de mise à disposition des résultats.

## Bibliographie

- [1] European Statistical System Committee (ESSC), « Warsaw Memorandum on the use of Earth Observation data for official statistics ». [En ligne]. Disponible sur: <https://dgins2021.stat.gov.pl/warsaw-memorandum>
- [2] J. Long, E. Shelhamer, et T. Darrell, « Fully Convolutional Networks for Semantic Segmentation ». [En ligne]. Disponible sur: <https://arxiv.org/abs/1411.4038>
- [3] Y. LeCun *et al.*, « Backpropagation Applied to Handwritten Zip Code Recognition », *Neural Computation*, vol. 1, n° 4, p. 541-551, 1989, doi: 10.1162/neco.1989.1.4.541.
- [4] P. Kim, « Convolutional Neural Network », in *MATLAB Deep Learning: With Machine Learning, Neural Networks and Artificial Intelligence*, P. Kim, Éd., Apress, 2017, p. 121-147. doi: 10.1007/978-1-4842-2845-6\_6.
- [5] O. Ronneberger, P. Fischer, et T. Brox, « U-Net: Convolutional Networks for Biomedical Image Segmentation », *arXiv:1505.04597 [cs]*, mai 2015, [En ligne]. Disponible sur: <http://arxiv.org/abs/1505.04597>
- [6] L.-C. Chen *et al.*, « Rethinking Atrous Convolution for Semantic Image Segmentation », *arXiv*, déc. 2017, [En ligne]. Disponible sur: <http://arxiv.org/abs/1706.05587>
- [7] A. Vaswani *et al.*, « Attention Is All You Need », in *Advances in Neural Information Processing Systems*, 2017. [En ligne]. Disponible sur: <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>
- [8] A. Dosovitskiy *et al.*, « An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale », *arXiv preprint arXiv:2010.11929*, juin 2021, [En ligne]. Disponible sur: <http://arxiv.org/abs/2010.11929>
- [9] E. Xie, W. Wang, A. L. Yuille, A. Anandkumar, et J. M. Alvarez, « SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers », *arXiv preprint arXiv:2105.15203*, oct. 2021, [En ligne]. Disponible sur: <http://arxiv.org/abs/2105.15203>
- [10] S. Zheng *et al.*, « Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers ». [En ligne]. Disponible sur: <https://arxiv.org/abs/2012.15840>
- [11] Z. Liu *et al.*, « Swin Transformer: Hierarchical Vision Transformer using Shifted Windows ». [En ligne]. Disponible sur: <https://arxiv.org/abs/2103.14030>
- [12] N. Ravi *et al.*, « SAM 2: Segment Anything in Images and Videos ». [En ligne]. Disponible sur: <https://arxiv.org/abs/2408.00714>