

Rapport Technique d'évaluation FirePy

Participants : Emmanuelle Cano, François Faupin, Thomas Gossart
Mentor : Pierre Adeikalam

Promotion : Data Scientist mars 2022



Contents

Contexte	2
Dans l'actualité	2
Dans la formation	2
Objectifs du projet	3
Business case	3
Méthodologie et environnement de travail	4
Déroulement du projet	5
Collecte des données	5
Description des données des satellites Sentinel 2	5
Description des données d'événements de feux	6
Sélection des incendies pour la base de données d'entraînement . .	9
Génération des images Sentinel 2	10
Génération des masques	11
Pre-processing des données	12
Qualité des données	12
Découpage en patchs	15
Normalisation	16
Augmentation des données	16
Dimensions du dataset d'entrée	16
Modélisation de type “segmentation sémantique”	18
U-Net	18
PSP-Net	21
Prédiction sur de nouvelles images	23
Bilan	25
Bibliographie	26

Contexte

Dans l'actualité

Les incendies de forêt augmentent en intensité et en fréquence à travers le monde en raison du changement climatique et de la hausse de la température mondiale.

L'observation de la Terre est un atout permettant de mieux comprendre et mesurer l'impact pour les populations et les infrastructures.

Les techniques de data science appliquées aux données satellitaires semblent pertinentes pour cet enjeu de surveillance de la surface terrestre.

L'enjeu est de pouvoir être capable de détecter des zones brûlées sur n'importe quelle partie du globe, surveiller la progression des incendies de forêt en temps quasi réel, ce qui est donc d'une importance cruciale pour les interventions d'urgence, mais aussi pour une estimation des enjeux économiques.

Dans la formation

Emmanuelle, de formation géomatique et télédétection, sans qui ce projet n'aurait pas vu le jour, a mis en œuvre la récupération des données de validation via QGIS,

François, Data Analyst, de par son expertise, avait déjà les connaissances théoriques et pratique en Deep Learning

Thomas, novice en Data Science au début du projet.

Ce projet était un puissant levier de montée en compétences sur les diverses problématiques et solutions en Data Science.

Objectifs du projet

Business case

Le relevé manuel des périmètres de feu est une tâche fastidieuse et sujette à erreur humaine. Le niveau de détails obtenu est limité car des poches épargnées par les flammes peuvent se trouver à l'intérieur des zones de feu.

L'automatisation de la détection de surfaces brûlées par un algorithme permettrait d'alléger la charge de travail humaine, d'apporter de la robustesse dans l'analyse, de passer à l'échelle la zone d'analyse et permettrait d'apporter une estimation préliminaire quant aux dégâts (naturels, les biens, les infrastructures). Le traitement mathématique des images satellites est aussi un moyen d'aller au-delà d'une information binaire brûlé / non brûlé et d'affiner le niveau de brûlure.

Les méthodes traditionnelles de détection de zones brûlées ont des performances limitées. En fonctionnant à l'aide de seuils, il est difficile pour ces techniques de détecter des petites zones brûlées ou des brûlures de faible intensité. Certains algorithmes basés sur la détection d'anomalie sur le voisinage de pixels ont un taux élevé de fausse détection. Enfin, les méthodes exploitant les différences dans les séquences d'image dans le temps ont l'inconvénient de nécessiter beaucoup de données.

Face à ce constat, les algorithmes de deep learning semblent particulièrement prometteurs et font l'objet de nombreuses recherches et publications.

La bonne disponibilité des données des satellites Sentinel 2 financés par le programme Copernicus de l'ESA est un atout essentiel pour le lancement du projet. En effet, les images sont disponibles gratuitement et couvrent depuis 2015 une grande partie de la surface terrestre.

En résumé:

L'objectif du projet est de mettre au point un algorithme capable de détecter les zones brûlées suite à un incendie à partir des images des satellites Sentinel 2 et avec une résolution très précise (au niveau du pixel).

Méthodologie et environnement de travail

Dans la cartographie des techniques de machine learning, le challenge du projet correspond à une tâche de classification supervisée et, plus particulièrement, de segmentation sémantique. Il s'agit de classifier chaque pixel d'une image (pixel brûlé / non brûlé).

Afin d'atteindre cet objectif, la méthodologie globale appliquée suivra les étapes ci-dessous:

- Recherche bibliographique sur les méthodes de détection de zones brûlées
- Constitution d'une base de données suffisamment propre et variée
- Entrainement d'un algorithme de Deep Learning à partir d'une région pour laquelle des données de vérité terrain (informations collectées par l'homme et non automatisées) sont disponibles,
- Application de l'algorithme à d'autres scènes et d'autres incendies pour produire une cartographie du contour de la zone brûlée une fois le feu maîtrisé par classification au pixel.

La nature des données d'entrée a nécessité une montée en compétences sur les images géo-référencées. En effet, des structures de données particulières (Fichiers vecteur Shapefile, Raster, GeoDataframe...) et des outils dédiés (QGIS, Geopandas, Rasterio¹) ont dû être mis en œuvre.

Enfin, afin de pouvoir travailler en collaboratif sur un environnement permettant les calculs de deep learning, nous avons utilisé les outils Google (Drive, Colab, Google Earth Engine) ainsi qu'un espace Github dédié.

¹<https://rasterio.readthedocs.io/en/latest/>

Déroulement du projet

Collecte des données

Pour réaliser ce projet, nous avons eu besoin de deux types distincts de données. Les images satellite, qui serviront pour l'entraînement du modèle, et les labels, pour la vérification.

Description des données des satellites Sentinel 2

Les 2 satellites Sentinel 2² (2A et 2B) ont été déployés en juin 2015 dans le cadre du programme Copernicus financé par l'Union Européenne et géré par l'ESA³(European Spatial Agency). L'objectif est de mettre à disposition des informations sur le sol, les océans, l'atmosphère et la sécurité.

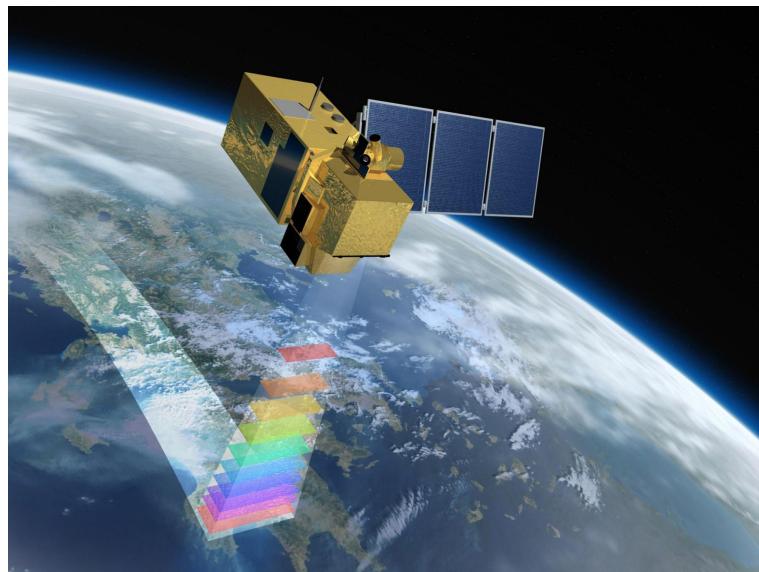


Figure 1: Représentation graphique

Le capteur multi-spectral de Sentinel 2 (MSI) permet de réaliser des acquisitions dans 13 bandes spectrales de résolutions spatiales différentes (de 10 à 60 m) dans les domaines du visible, du proche et du moyen infrarouge.

L'indice permettant de détecter les zones brûlées fera appel aux bandes 7 (NIR) et 12 (SWIR), toutes les deux à 20 m de résolution spatiale.

²<https://sentinels.copernicus.eu/.../resolutions/spatial>

³<https://sentinel.esa.int/web/sentinel/missions/sentinel-2>

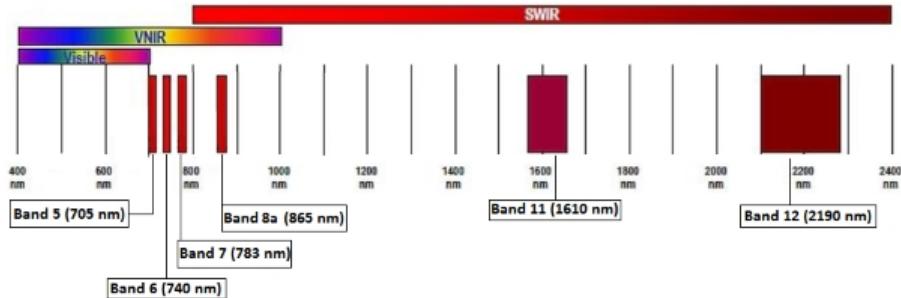


Figure 2: SENTINEL-2 20 m spatial resolution bands: B5 (705 nm), B6 (740 nm), B7 (783 nm), B8a (865 nm), B11 (1610 nm) and B12 (2190 nm)

On pourra également utiliser les bandes du visible (2,3,4) pour réaliser des compositions colorées permettant d'analyser visuellement les territoires à traiter ainsi que de calculer d'autres indices en lien avec la densité de végétation.

En effet, les surfaces brûlées présentent une réflectance plus faible que la végétation saine dans le moyen infrarouge, en raison de l'absorption des radiations par les cendres, ceci indépendamment des écosystèmes.

Toutefois, cette réflectance des surfaces brûlées dans le moyen infrarouge est voisine de celles des surfaces très humides et rend les confusions possibles. Ceci justifie la prise en compte d'un deuxième discriminant plus sévère, l'augmentation des températures de surface dans les zones brûlées pendant la journée en raison de la forte absorption des radiations solaires et de l'absence de l'évapotranspiration qui, dans les conditions normales, assure le transfert de l'énergie dans l'atmosphère sous forme de chaleur latente, à travers la vapeur d'eau. De plus, la présence des cendres et du charbon décroît l'albédo⁴ de surface et en augmente la température d'environ 7 à 8° Kelvin.

Description des données d'événements de feux

En tant que référence pour établir les datasets de training / validation / test, nous avons utilisé deux ressources principales :

- Pour la partie nord américaine: “Fire Perimeters in California Database⁵(CALFIRE), provided by the Fire and Resource Assessment Program⁶ (FRAP)”

Il est ainsi possible de récupérer la base de données du projet FRAP du gouvernement californien, qui met notamment à disposition, en complément des caractéristiques historiques des incendies survenus dans le pays depuis 1950,

⁴Grandeur caractérisant la proportion d'énergie lumineuse réfléchie ou diffusée par un corps éclairé.

⁵ <https://www.fire.ca.gov/incidents>

⁶<https://frap.fire.ca.gov/frap-projects/fire-perimeters/>

les périmètres des zones brûlées sous forme de données géographiques vecteurs (géodatabase ESRI).

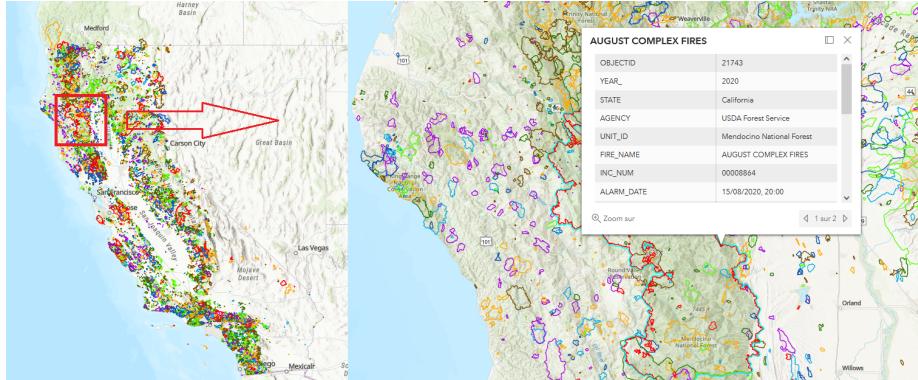


Figure 3: Shapefile contenant l’ensemble des incendies de la base CALFIRE

Pour produire le jeu de données d’entraînement, le choix s’est principalement porté sur les incendies de la Californie entre 2018 et 2020. Ces événements sont en effet richement documentés (voir ci-dessous des exemples d’événements récents).

FIRE NAME (CAUSE)	DATE	COUNTY	ACRES	STRUCTURES	DEATHS
1 AUGUST COMPLEX (<i>Under Investigation</i>)*	August 2020	Mendocino, Humboldt, Trinity, Tehama, Glenn, Lake, & Colusa	1,032,648	935	1
2 MENDOCINO COMPLEX (<i>Under Investigation</i>)	July 2018	Colusa, Lake, Mendocino & Glenn	459,123	280	1
3 SCU LIGHTNING COMPLEX (<i>Under Investigation</i>)*	August 2020	Stanislaus, Santa Clara, Alameda, Contra Costa, & San Joaquin	396,624	222	0
4 CREEK FIRE (<i>Under Investigation</i>)*	September 2020	Fresno & Madera	379,895	853	0
5 LNU LIGHTNING COMPLEX (<i>Under Investigation</i>)*	August 2020	Napa, Solano, Sonoma, Yolo, Lake, & Colusa	363,220	1,491	6

Figure 4: Top 5 largest California Wildfires

- Pour la partie Européenne, nous avons exploité la base de données des feux au Portugal de l’institut INCF⁷ (Instituto da Conservação da Natureza e das Florestas).

Pour chaque source, il était possible de télécharger un fichier shapefile contenant l’ensemble des incendies. L’outil QGIS⁸ a permis de contrôler les géométries puis de récupérer les géométries de chaque zone brûlée dans un fichier dédié.

Tout d’abord, qu’est ce qu’un shapefile ?

Un shapefile est un format de fichier pour les systèmes d’informations géographiques (SIG) et contient toute l’information liée à la géométrie des objets décrits, qui peuvent être des points, des lignes ou des polygones.

⁷<http://www2.icnf.pt/portal/florestas/dfci/inc/cartografia/areas-ardidas>

⁸<https://www.qgis.org/fr/site/>



Figure 5: Shapefile contenant l'ensemble des incendies de la base INCF

D'après la description de la méthodologie sur ces sites, ces périmètres ne sont pas produits à partir de méthodes automatisées mais de digitalisation manuelle à partir de cartes ou de photointerprétation. La majorité des périmètres de feu est réalisée au travers de relevés GPS au sol.



Ces périmètres vont nous aider à récupérer les images correspondant aux événements dans des limites d'emprises pertinentes, mais aussi à construire le jeu d'entraînement et à contrôler nos résultats.

Sélection des incendies pour la base de données d'entraînement

La lecture des bases de données Shapefile de la Californie et du Portugal permet d'obtenir un “Geo-dataframe pandas” qui est une extension du dataframe Pandas avec un géo-référencement de chaque observation.

Pour chaque événement, nous avons donc une colonne avec le périmètre de zones brûlées sous la forme d'une liste de polygones définis par des coordonnées.

FIREALERT_HISTORICAL14D_DAY1_P1															
index	date	state	agency	unit_id	firm_name	inc_no	firm_alert_date	fire_cont_date	shape_length	shape_area	fire_id	fire_area	bbox	firr_geometry	fire_alert_year
0	250	2020	CA	USF	MNF	AUGUST COMPLEX FIRE	00008864	2020-08-16	2020-11-17 4.18704e+06	41.91870e+05	291	417918.70	Polygon ((-125.988 36.7056, -125.988 36.7056, -125.988 36.7056, -125.988 36.7056, -125.988 36.7056))	MULTIPOLYGON (((-125.988 36.7056, -125.988 36.7056, -125.988 36.7056, -125.988 36.7056, -125.988 36.7056)))	2020
1	1619	2020	CA	CDF	UNU	RANCH	00008545	2015-07-27	2018-09-13 3.48227e+05	16.93003e+09	9170	168005.03	Polygon ((-125.941 36.519227, -125.941 36.519227, -125.941 36.519227, -125.941 36.519227, -125.941 36.519227))	MULTIPOLYGON (((-125.941 36.519227, -125.941 36.519227, -125.941 36.519227, -125.941 36.519227, -125.941 36.519227)))	2018
2	175	2020	CA	CDF	SOU	SCCU	00005760	2020-08-16	2020-09-11 5.0337e+06	16.08920e+09	106589	165589.7	Polygon ((-125.941 36.400511, -125.941 36.400511, -125.941 36.400511, -125.941 36.400511, -125.941 36.400511))	MULTIPOLYGON (((-125.941 36.400511, -125.941 36.400511, -125.941 36.400511, -125.941 36.400511, -125.941 36.400511)))	2020
3	2474	2020	CA	USF	SFH	CREEK	00001394	2020-08-04	2020-12-24 8.16907e+05	1.53716e+09	248	153716.77	Polygon ((-125.91130 36.702775, -125.902 36.707166, -125.902 36.707166, -125.91130 36.702775, -125.91130 36.702775))	MULTIPOLYGON (((-125.91130 36.702775, -125.902 36.707166, -125.902 36.707166, -125.91130 36.702775, -125.91130 36.702775)))	2020

Figure 6: Geo-dataframe Pandas

Dans notre démarche, nous avons privilégié les gros incendies (filtrage sur la colonne “fire_area” > 2500 ha) afin de disposer d’images d’au moins 256 x 256 pixels (1 pixel équivaut à 20m).

A l'issue de la récupération des Shapefiles contenant plusieurs centaines d'incendies, nous avons scripté l'extraction des éléments suivant : longitude_1 et longitude_2, latitude_1 et latitude_2 (afin d'obtenir la bounding box en coordonnées GPS), date de début de l'incendie, date de fin de l'incendie (pour

dater et faciliter la recherche des images Sentinel 2 pré-fire et post-fire). Ces informations permettent de solliciter le service Google Earth Engine afin de télécharger les images Sentinel 2.

Génération des images Sentinel 2

Google Earth Engine⁹ est un service fournissant un catalogue d'images satellite et de dataset géo-référencés de plusieurs petabytes. Des capacités d'analyse de la surface de la Terre sont mises à la disposition des chercheurs et développeurs.

Ce service gratuit est un bon moyen de collecte des gros volumes de données d'imagerie satellitaire Sentinel 2.

L'extraction des données Sentinel 2 a été codée de la façon suivante:

```
# Collecting the Sentinel image
image = (ee.ImageCollection('COPERNICUS/S2_SR')
          .filterBounds(bounding_box)
          .filterDate(fire_date1, fire_date2)
          .filter(ee.Filter.lt('CLOUDY_PIXEL_PERCENTAGE', 10))
          .select(bands)
          .sort('system:index', opt_descending=False).mosaic()
          .clip(bounding_box))
```

Figure 7: Extrait du code d'extraction des données sur GEE

- Requête du catalogue Sentinel 2
- Filtrage sur la zone d'intérêt via les coordonnées GPS issues du géo-dataframe
- Filtrage sur la période d'intérêt issue du géo-dataframe
- Filtrage sur les images contenant moins de 10% de couverture nuageuse
- Filtrage sur les bandes du capteur (3 canaux RGB + 2 infra-rouge NIR, SWI). D'après certaines publications, le choix de ces 5 fréquences produit en effet les meilleures performances de classification.
- Reconstruction et fusion d'une mosaïque d'images (par empilement)
- Découpage sur la zone d'intérêt

Une fois l'objet image généré, il est possible de le télécharger avec la résolution souhaitée (20m) dans un fichier qui sera stocké dans le drive Google. Nous avons décidé que le fichier de sortie serait un fichier raster TIF car lui-seul nous permettait d'avoir à la fois les 5 canaux, aucune perte de qualité sur les fichiers volumineux, ainsi que les métadonnées associées.

⁹<https://earthengine.google.com/>

Qu'est-ce qu'une image raster?

Ce format est utilisé pour exploiter les images satellites, aériennes ou de plan. Cette représentation d'une photographie ou d'un plan est stockée sous la forme d'une grille de pixels en ligne et en colonne. Chaque cellule de cette matrice contient à la fois l'intensité des canaux et les coordonnées géographiques.

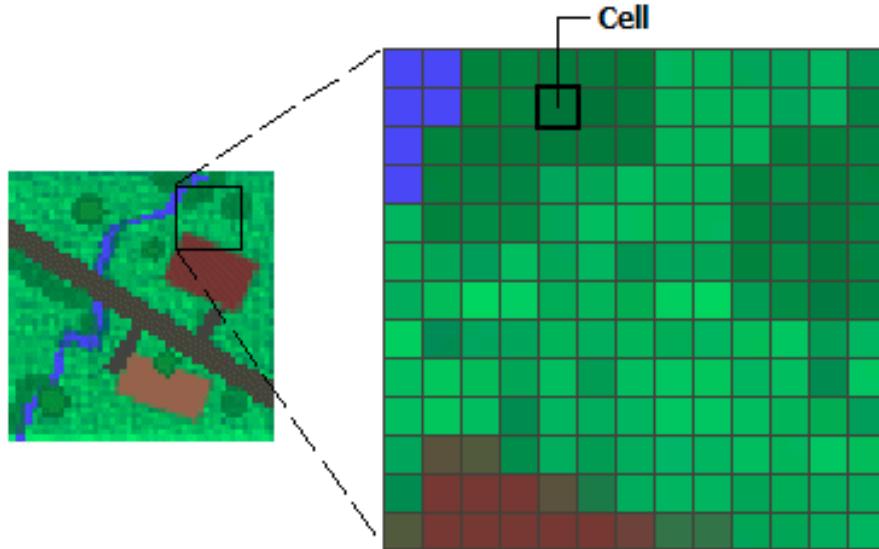


Figure 8: Représentation d'un Raster

Génération des masques

La génération du masque se déroule en 4 étapes:

- Préparation d'une image raster de masque construite sur les caractéristiques de l'image Sentinel 2 correspondante et contenant des valeurs nulles. La librairie GDAL¹⁰ contenant le driver “GTiff” a été utilisée.
- Récupération de la géométrie issue du shapefile correspondant. La librairie OGR¹¹ contenant le driver “ESRI shapefile” a été utilisée.
- Ajout de la géométrie de la zone brûlée à l'image raster de masque.
- Export de l'image raster de masque en fichier tiff.

Voici un exemple de masque et la vue en couleurs naturelles de l'image satellite associée:

¹⁰<https://gdal.org/>

¹¹<https://gdal.org/>



Figure 9: Vue RGB et mask associé

Pre-processing des données

Qualité des données

La visualisation des images en couleurs naturelles RGB a mis en évidence des problèmes de qualité de données. Certaines images doivent être écartées du dataset d'entraînement afin de ne pas dégrader les performances de classification. Malgré le faible nombre d'images à disposition, il a été décidé de privilégier la qualité des données.

Les facteurs de non qualité sont liés à la présence d'éléments entre la zone brûlée et le capteur du satellite. Contrairement à d'autres satellites disposant d'un capteur actif (radar envoyant une onde électromagnétique traversant les éléments), les images Sentinel 2 ne font que recevoir les émissions de la Terre dans différentes fréquences. Tout élément non désiré entre le capteur et la zone brûlée est donc perturbateur pour l'entraînement du modèle.

Les phénomènes conduisant à exclure certaines images sont illustrés ci-dessous :

- Présence de fumées

Pour certains événements, les dates de départ et de fin de feu sont peut-être erronées de quelques jours. Nous obtenons ainsi quelques images avec la présence massive de fumée qui semble donner un effet de flou.

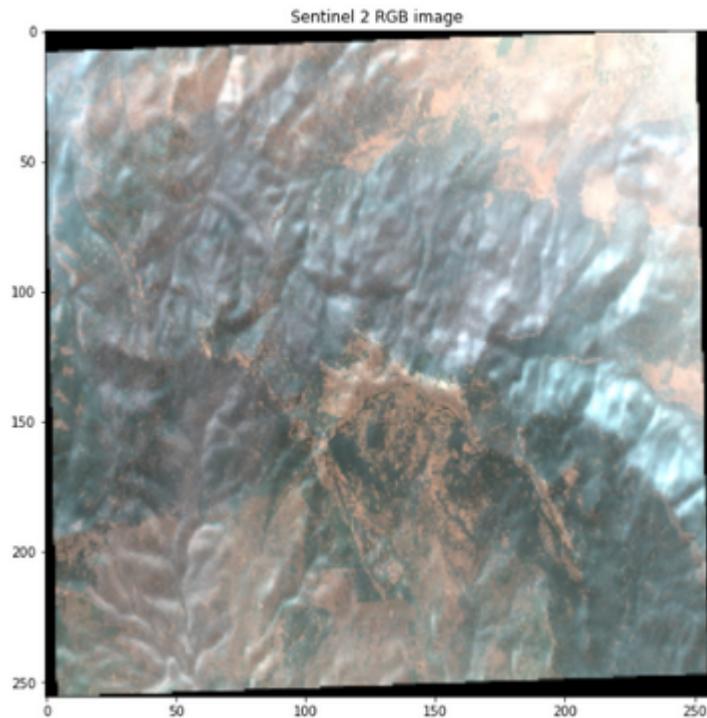


Figure 10: Présence de fumée

- Présence de neige

La plupart des feux sont situés dans des zones montagneuses. Il est donc possible de visualiser de la neige sur les reliefs et cela masque les zones brûlées.

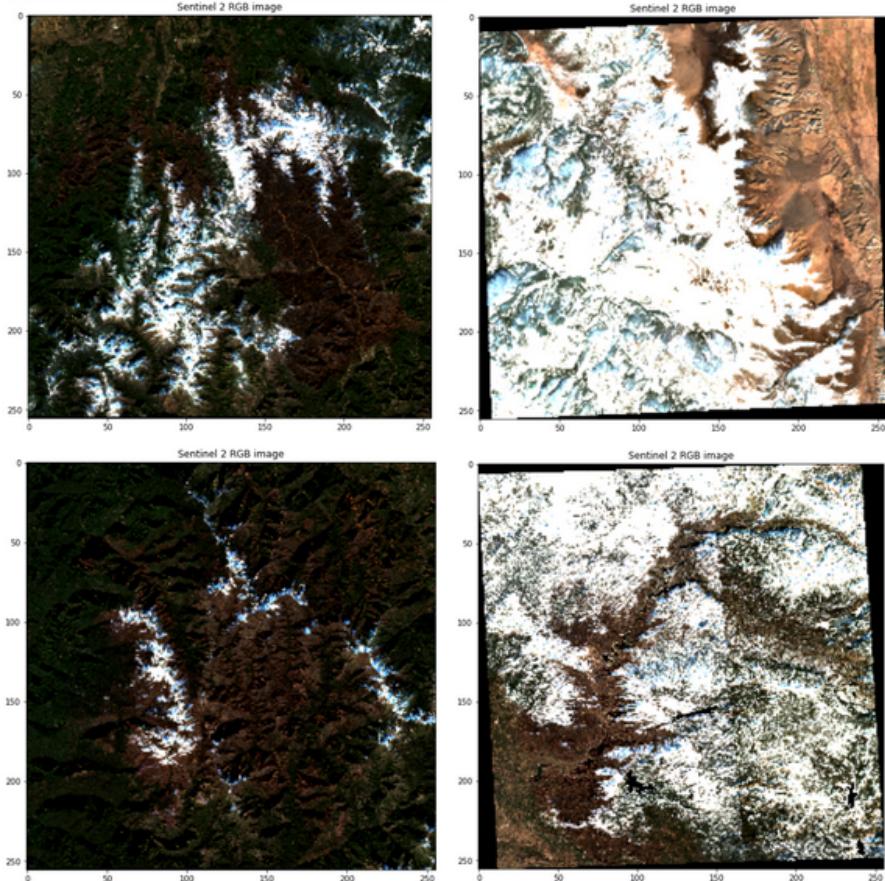


Figure 11: Présence de neige

Enfin, il faut noter une limitation au niveau de la qualité des données de labellisation issues des relevés topographiques. En zoomant sur certaines images, certaines zones ont été déclarées "non brûlées" alors qu'on peut penser qu'il s'agit d'une erreur humaine.

Ceci est un point de vigilance sur les métriques de performance du modèle de classification qui ne pourra pas atteindre 100% à juste titre.

Vous trouverez ci-dessous des exemples d'erreurs de labellisation.

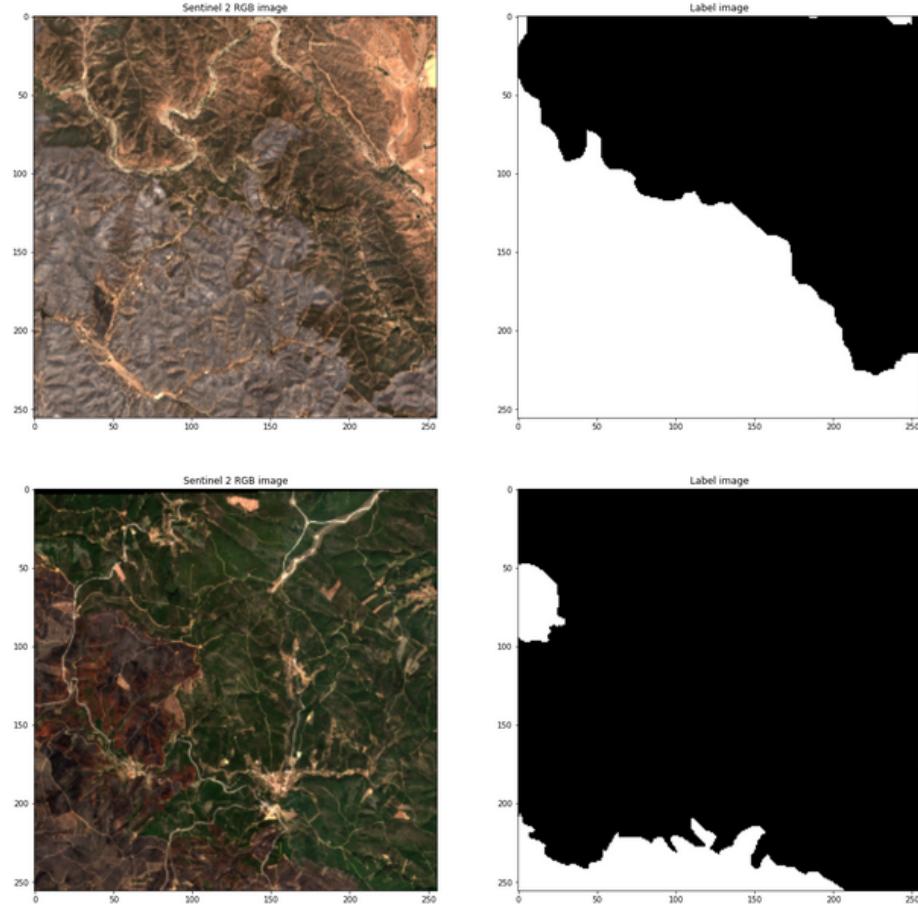


Figure 12: Erreur de labellisation

Découpage en patchs

La taille des images dépend de l'étendue des feux. Il n'est pas conseillé d'effectuer un "resize" de chaque image car le modèle risque d'apprendre des interpolations d'image très différentes. Ainsi, la bonne pratique est de découper des patchs de 256 par 256 pixels que l'on pourra soumettre à l'algorithme de segmentation sémantique.

L'extraction de patchs a été réalisée via le package "Patchify¹²".

¹²<https://pypi.org/project/patchify/>

L'exploitation des patchs permet de démultiplier le nombre d'images pour l'apprentissage. Ainsi, une trentaine de feux génère environ 1200 patchs de 256 par 256 pixels.

Ce même package est capable de reconstituer une grande image initiale à partir de patchs. Cette fonctionnalité ("unpatchify") est très utile pour réaliser des prédictions sur des grandes images

Normalisation

Les données ont été normalisées afin de contenir que des valeurs comprises entre 0 et 1.

Les facteurs multiplicatifs suivants ont été appliqués:

- 1 / 10000 pour les images Sentinel (la réflectance mesurée par le capteur du satellite est en effet multipliée par 10000 pour des raisons de stockage en entier)
- 1/ 255 pour les masques de label

Ce traitement a été intégré au sein du générateur fournissant au modèle les données à la volée.

Augmentation des données

Afin d'augmenter le nombre d'images pour l'apprentissage, une stratégie d'augmentation de données a été mise en place.

Le package "Albumentations¹³" a été sélectionné pour cette tâche car il permet de gérer les images possédant 5 canaux.

Ce traitement a été intégré au sein du générateur fournissant au modèle les données à la volée.

Les augmentations d'image implémentées sont les suivantes:

- Probabilité de 30% de flip horizontal
- Probabilité de 30% de flip vertical
- Probabilité de 30% de rotation de 90°

Dimensions du dataset d'entrée

A l'issue des différents traitements de pré-processing, le dataset se présente sous la forme d'un batch d'images à 5 canaux.

Le nettoyage des images Sentinel a conduit à la constitution d'une base de données de 1200 patchs issus des feux de Californie et Portugal. Ces images ne tiennent pas toutes en mémoire RAM. Elles sont donc fournies au modèle via des batchs mis à disposition par un générateur personnalisé.

¹³<https://albumentations.ai/>

Les dimensions du dataset en entrée du modèle sont:

Pour une image Sentinel 2:

$(nb\ batch, nb\ pixels\ hauteur, nb\ pixels\ largeur, nb\ canaux)$

$(32, 256, 256, 5)$

Pour un masque labellisant les zones brûlées:

$(nb\ batch, nb\ pixels\ hauteur, nb\ pixels\ largeur, nb\ canaux)$

$(32, 256, 256, 1)$

Enfin, notre base de données est constituée d'environ 1200 patchs avec le masque associé.

Voici un exemple ci-dessous:

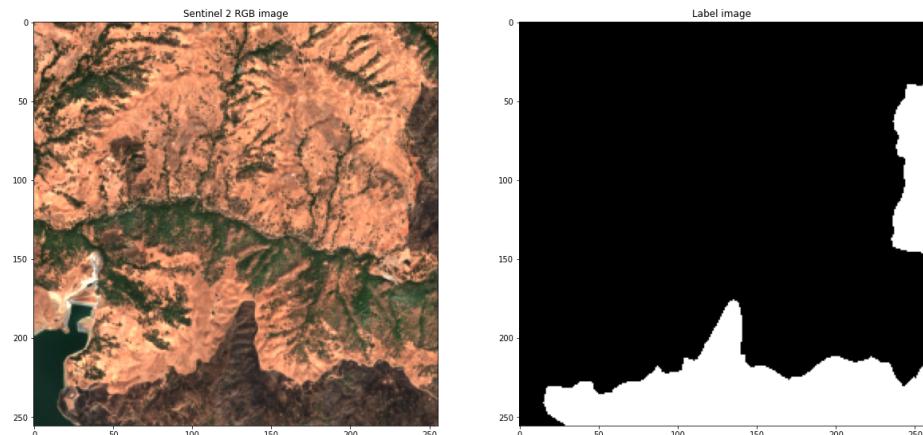


Figure 13: Exemple de Patch + mask associé

Modélisation de type “segmentation sémantique”

U-Net

La modélisation U-Net¹⁴ a été proposée dans la publication “U-Net: Convolutional Networks for Biomedical Image Segmentation”. Cette méthode de classification est de type FCNN “Fully Convolutional Neural Network”, c'est-à-dire sans couches fully connected.

L'architecture U-Net semble pertinente pour la segmentation d'images satellites. En effet, ce type de modélisation a démontré de bons résultats pour des tâches similaires et pour des bases de données d'entraînement peu fournies.

La structure originelle du U-Net est illustrée ci-dessous.

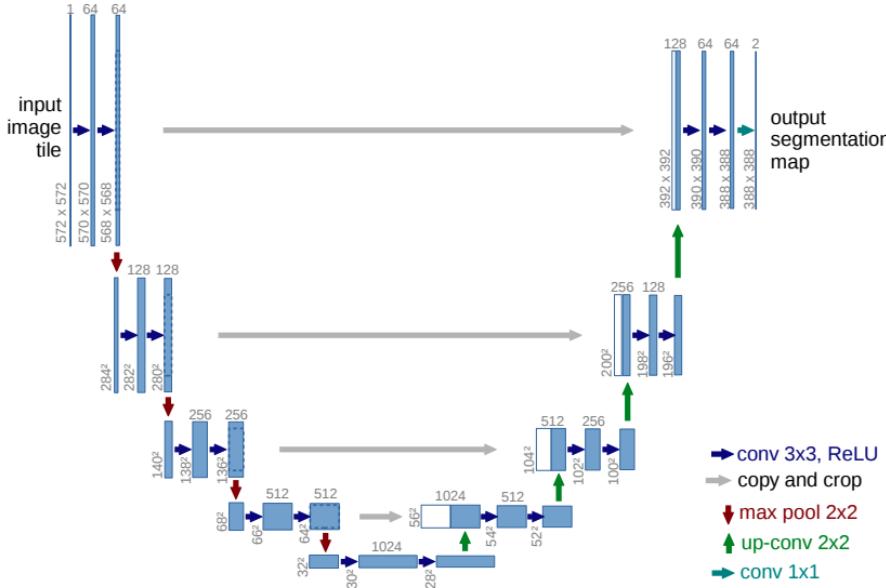


Figure 14: Schéma de la structure U-Net de notre modèle

Cette architecture comprend une partie d'encodage à l'aide une séquence de couches de convolution, de pooling et de dropout. A chaque étape de convolution de la phase d'encodage, l'information spatiale perdue est récupérée (flèches grises) pour être exploitée lors de la phase de décodage.

La reconstruction de l'image prédictive contenant la zone d'intérêt s'effectue au travers de couches de dé-convolutions appliquées aux features encodées et à l'information spatiale stockée pendant l'encodage.

¹⁴Convolutional Networks for Biomedical Image Segmentation

Afin d'obtenir une probabilité de classification binaire dans l'image de sortie, une fonction d'activation de type sigmoïde a été utilisée.

Le modèle codé dans le cadre du projet est illustré ci-dessous:

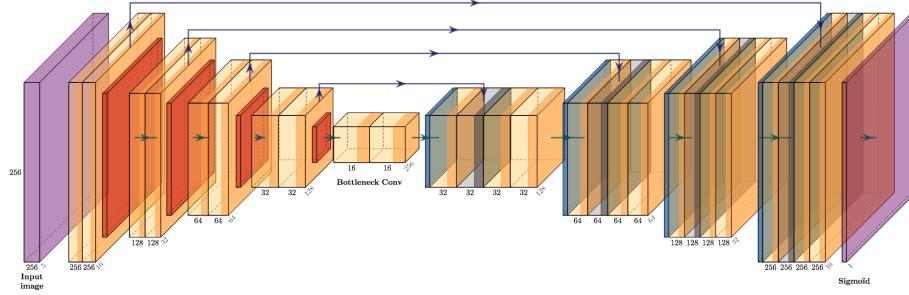


Figure 15: Schéma de notre modèle U-Net

Après plusieurs essais, la fonction de perte la plus performante correspond à un mix (50/50) entre la fonction de binary cross entropy et du coefficient DICE qui correspond à une version exploitable car différentiable de l'IOU (Intersection Over Union). Le choix de cette fonction coût produit les meilleurs résultats de prédiction sur l'échantillon de validation (voir tableau ci-dessous).

Choix	loss	acc	f1	precision	recall	iou	val_loss	val_acc	val_f1	val_precision	val_recall	val_iou
50% BCE	0.1854	0.9165	0.8908	0.8556	0.9326	0.7478	0.1800	0.9153	0.8788	0.8104	0.9644	0.7383
100% BCE	0.2505	0.9006	0.8640	0.8305	0.9083	0.6378	0.2907	0.8843	0.8589	0.7959	0.9372	0.6303
10% BCE	0.2373	0.9086	0.8763	0.8347	0.9295	0.6556	0.2896	0.8916	0.8640	0.8221	0.9149	0.6341

Figure 16: Résultats de prédictions (metrics)

La compilation du modèle a été réalisée avec l'optimizer Adam (et son paramétrage par défaut) et les métriques: "Accuracy", "F1 score", "Precision", "Recall" et "IOU".

L'entraînement du modèle U-Net a été réalisé sur 20 epochs avec le callback "early stopping" afin de stopper l'entraînement lorsque la fonction coût augmente et de restaurer les meilleurs poids.

Lors de la phase d'apprentissage, un échantillon de validation (20% du dataset) est mis de côté dans le but de contrôler la performance de prédiction du modèle.

Nous obtenons les résultats suivants:

Les métriques de performance du modèle atteignent un plateau correspondant à une accuracy sur les données de validation autour de 91,5% et un f1 score de 88%.

L'indicateur IOU est de 74%, ce qui est satisfaisant (car supérieur à 50%).

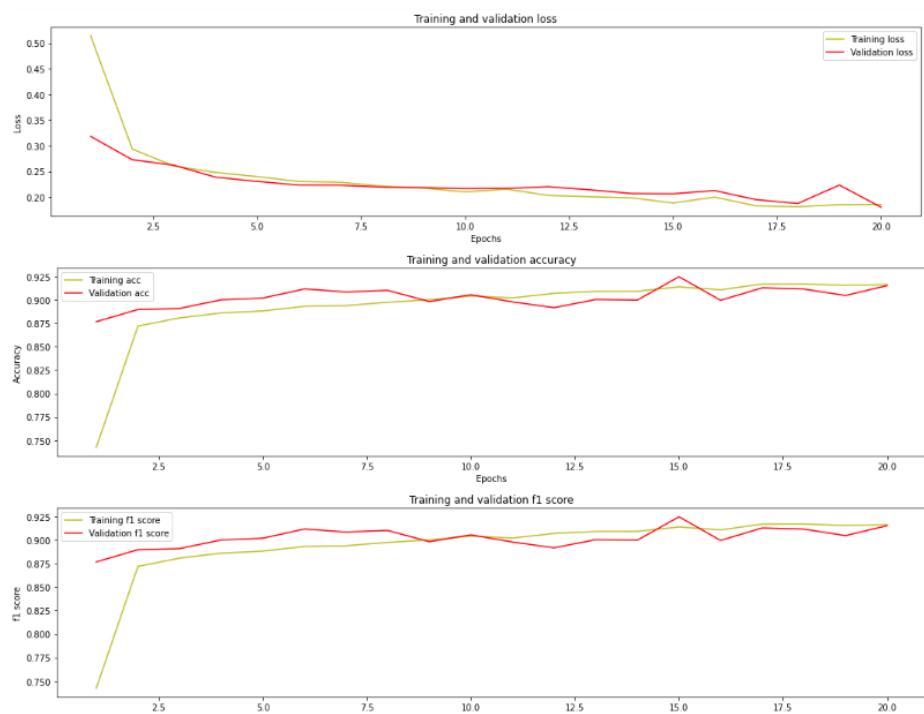


Figure 17: Courbes de training et validation

L'architecture U-Net affiche de bonnes performances de prédiction. Avant d'exploiter cet algorithme en production, les tuiles de test mises de côté peuvent être analysées afin de contrôler visuellement la pertinence du modèle.



Figure 18: Prédiction sur patch



Figure 19: Prédiction sur patch



Figure 20: Prédiction sur patch

Dans l'image ci-dessous, on remarque que le lac est visible sur la cartographie de probabilité de zone brûlée mais pas sur la prédiction brûlé/non brûlé. Un réglage du seuil de décision semble nécessaire.

L'application du modèle sur les patchs ci-dessous confirme la bonne performance de l'algorithme sur des régions non brûlées (champs agricoles, zones urbaines).

PSP-Net

Afin de confronter le modèle U-Net, il a été proposé (suite à plusieurs articles et Githubs mettant en avant ce modèle dans la segmentation d'images satellite) d'utiliser le modèle PSPNet¹⁵ (Pyramid Scene Parsing Network).

¹⁵<https://arxiv.org/abs/1612.01105>



Figure 21: Prédiction sur patch

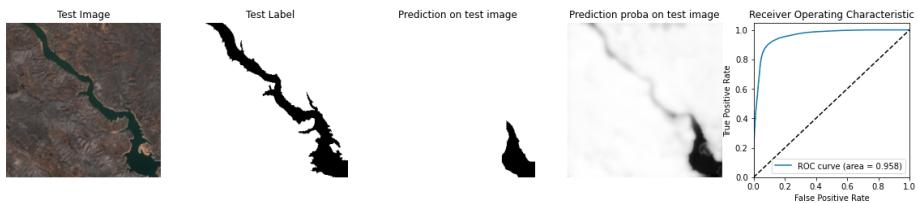


Figure 22: Prédiction avec erreur sur un lac (réflectance)



Figure 23: Prédiction sur zone agricoles non brûlées

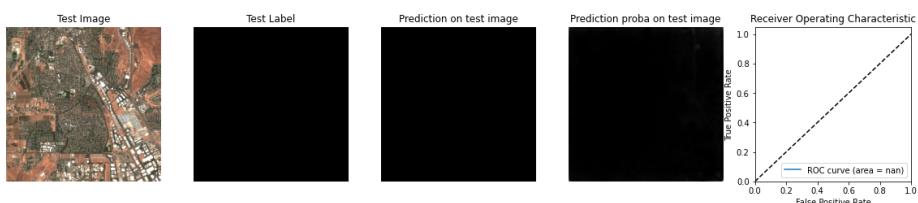


Figure 24: Prédiction sur zone urbaines non brûlée

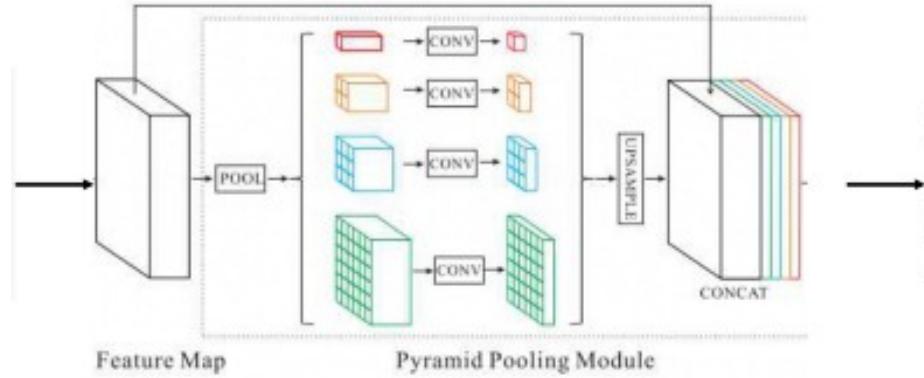


Figure 25: Architecture PSPNet

La difficulté réside dans le fait que ce modèle peut s'avérer puissant mais uniquement sur des images constituées de 3 bandes (RGB) et devient trop gourmand en ressources sur nos fichiers en 5 bandes car là où le modèle U-Net nécessite 1,941,393 paramètres, le PSP requiert 31,203,073 paramètres.

La modélisation a été effectuée sur l'ensemble des données en Patch, mais s'est vite soldée par un échec car la RAM disponible sur les serveurs Colab n'était pas suffisante.

Afin de tout de même lui laisser ses chances, nous avons procédé à un redimensionnement des images afin d'alléger la charge mais les résultats n'étaient pas concluants non plus (nous pouvons constater que la prédiction a des artefacts et n'est pas aussi précise que pour le modèle U-Net)

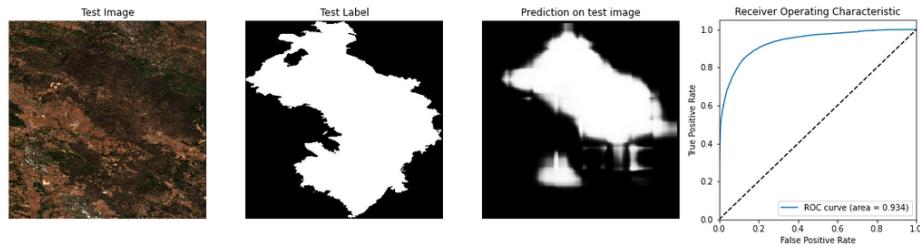


Figure 26: Prédiction du modèle PSPNet

Prédiction sur de nouvelles images

Le modèle appris sur le dataset d'entraînement peut être appliqué à de nouvelles

images.

Cette tâche de prédiction se déroule en 3 étapes:

1. Découpage d'une image Sentinel 2 en patchs de 256 par 256 pixels
2. Prédiction des zones brûlées sur chaque patch à l'aide du modèle
3. Reconstitution d'une prédiction globale à partir des patchs prédits

Pour la dernière étape, 2 méthodes ont été mise en oeuvre:

1. Juxtaposition des patchs prédits. Le modèle est appliqué sur chaque patch et les résultats sont simplement collés en respectant l'ordre du découpage. Cette technique est disponible via la méthode unpatchify du package patchify. Bien que rapide, cette méthode fait apparaître des artefacts sur les bords de chaque bord. Le modèle a en effet plus de difficultés à apprendre près des bords. De plus, les prédictions ne sont possibles que sur un nombre fini de patchs, ce qui peut exclure une partie de l'image. Voir image en bas à gauche.
2. "Smoothed blending" des patchs prédits. Dans cet algorithme, le modèle est appliquée sur une fenêtre glissante de l'image satellite avec un overlap entre chaque tuile. Ensuite, les résultats de prédiction sont recombinés ensemble avec une interpolation de type spline. L'algorithme utilisé est issu d'un projet open source "Smoothly-Blend-Image-Patches"¹⁶. Bien que nécessitant plus de temps de calcul, cette méthode est plus performante et les images produites sont très réaliste. Voir image en bas à droite.

On remarque que le masque de zone brûlée issu de la prédiction est très proche de la vérité terrain. L'affichage de la probabilité de zone brûlée permet de faire apparaître des nuances dans le niveau de brûlure ainsi que des éléments internes (cours d'eau, routes...). Cette nouvelle connaissance est intéressante pour évaluer l'impact du feu sur les points stratégiques de la zone étudiée. Une meilleure gestion des conséquences de l'incendie est alors possible.

¹⁶<https://github.com/Vooban/Smoothly-Blend-Image-Patches>

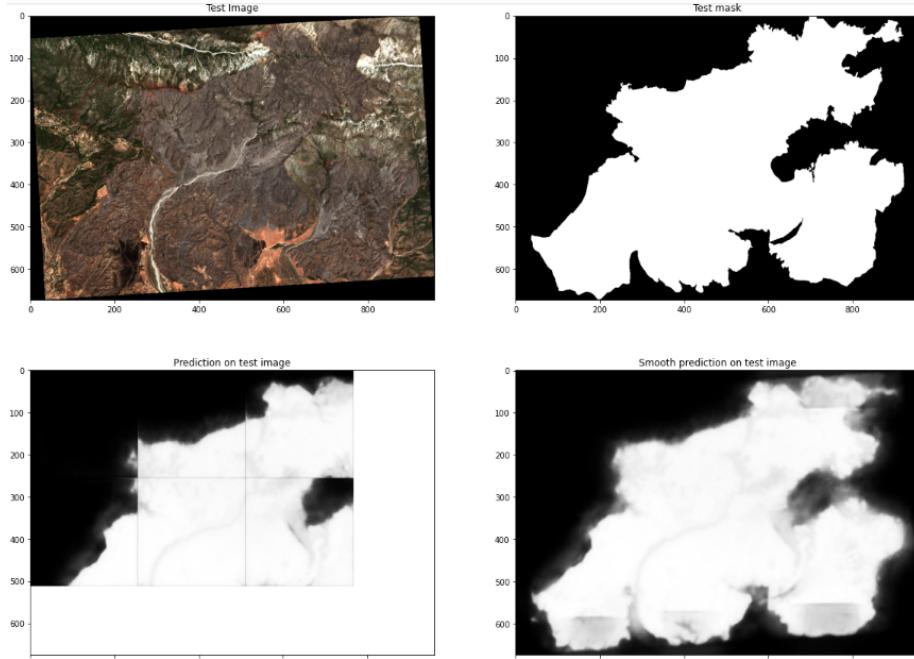


Figure 27: Exemple de prédiction après Smoothblending

Bilan

La réalisation du projet a permis de mettre en pratique les compétences acquises lors de la formation dans le cadre d'un challenge particulièrement stimulant. Le développement de cette application de machine learning a été l'opportunité de se familiariser avec un nouveau type de données: les images géo-référencées.

Le code du projet peut être consulté sur le Github suivant:

<https://github.com/DataScientest-Studio/firepy>

L'application mise au point offre des perspectives d'applications concrètes prometteuses (surveillance de la surface brûlée, impact des incendies sur les infrastructures...).

Cependant, avant d'être mis en production, le modèle de détection de zones brûlées nécessiterait d'être entraîné sur davantage de données. L'entraînement a principalement été réalisé sur des images de Californie, donc la généralisation aux feux d'autres régions du monde serait limitée. En effet, la diversité des environnements (forêt, type de végétation, volcan) n'est pas assez représentée dans le training dataset.

Bibliographie

1. U-Net: Convolutional Networks for Biomedical Image Segmentation
<https://arxiv.org/pdf/1505.04597.pdf>
2. CAL FIRE Incidents <https://www.fire.ca.gov/incidents>
3. Portugal Cartographia de area ardida <http://www2.icnf.pt/portal/florestas/dfci/inc/cartografia/areas-ardidas>
4. FRAP <https://frap.fire.ca.gov/frap-projects/fire-perimeters/>
5. USA : <https://data-nifc.opendata.arcgis.com/>
6. Satellite Sentinel 2 : <https://sentinel.esa.int/web/sentinel/missions/sentinel-2>
7. Spatial Resolution of SENTINEL-2 : <https://sentinels.copernicus.eu/web/sentinel/user-guides/sentinel-2-msi/resolutions/spatial>
8. QGIS : <https://www.qgis.org/fr/site/>
9. Google Earth Engine : <https://earthengine.google.com/>
10. Patchify : <https://pypi.org/project/patchify/>
11. Albumentation : <https://albumentations.ai/>
12. PSPNet : <https://arxiv.org/abs/1612.01105>
13. GDAL/OGR : <https://gdal.org/>
14. Rasterio : <https://rasterio.readthedocs.io/en/latest/>

List of Figures

1	Représentation graphique	5
2	SENTINEL-2 20 m spatial resolution bands: B5 (705 nm), B6 (740 nm), B7 (783 nm), B8a (865 nm), B11 (1610 nm) and B12 (2190 nm)	6
3	Shapefile contenant l'ensemble des incendies de la base CALFIRE	7
4	Top 5 largest California Wildfires	7
5	Shapefile contenant l'ensemble des incendies de la base INCF	8
6	Geo-dataframe Pandas	9
7	Extrait du code d'extraction des données sur GEE	10
8	Représentation d'un Raster	11
9	Vue RGB et mask associé	12
10	Présence de fumée	13
11	Présence de neige	14
12	Erreur de labellisation	15
13	Exemple de Patch + mask associé	17
14	Schema de la structure U-Net de notre modèle	18

15	Schéma de notre modèle U-Net	19
16	Résultats de prédictions (metrics)	19
17	Courbes de training et validation	20
18	Prédictions sur patch	21
19	Prédictions sur patch	21
20	Prédictions sur patch	21
21	Prédictions sur patch	22
22	Prediction avec erreur sur un lac (réflectance)	22
23	Prédictions sur zone agricoles non brûlées	22
24	Prédiction sur zone urbaines non brûlée	22
25	Architecture PSPNet	23
26	Prédiction du modèle PSPNet	23
27	Exemple de prédiction après Smoothblending	25

Made with Google doc, build with L^AT_EX