Journal de bord DHS/Landsat

Ce projet a pour rôle de créer le dataset d’entraînement pour l’étude de la pauvreté en Afrique de 1990 à aujourd’hui à partir d’images satellites.

Les sorties attendues sont :

* Un fichier csv regroupant tous les clusters d’Afrique ayant fait l’objet d’une enquête DHS, leurs coordonnées géographiques, la boundingbox centrée autour du cluster, si le cluster est urbain (U) ou rural et l’IWI (inclusive wealth index) calculé à partir des résultats de l’enquête
* Une image Landsat pour chaque zone d’intérêt répertoriée dans le fichier csv (triées selon le pays et l’année)

# Construction du fichier csv

Pour la construction d’un fichier de données le plus exhaustif possible, nous disposons de trois sources :

* Les enquêtes indexées provenant directement de Global Data Lab
* Le fichier dataset de Petterson et al (soi-disant utilisant les mêmes sources que nous mais les plages sont différentes)
* Le fichier dataset de Yeh et al créé dans son projet Sustain Bench où il recalcule directement les IWI

Nous choisissons d’utiliser en priorité les données de Global Data Lab car elles sont au plus proche de la source et paraissent donc plus fiables. Néanmoins, les deux autres sources complètent le premier dataset de façon non négligeable (65 000 clusters au lieu de 40 000).

Pour s’assurer de la cohérence entre les trois jeux de données, nous avons calculé les coefficients de corrélation sur les données communes et ils sont assez bons.

La stratégie à adopter à partir de là reste à définir. La plus naturelle serait sûrement de calculer les coefficients de régression et appliquer une transformation linéaire aux données pour les mettre à l’échelle de Global Data Lab.

A ce jour, nous avons construit un fichier areas\_of\_interest.csv qui regroupe tous les clusters des trois sources et leurs boundingbox à 5km mais pas les IWI. Il permet de télécharger les images satellites qui seront utiles à notre modèle de machine learning.

Les différents défis rencontrés concernant l’homogénéisation des données ont été :

* La transcription des noms de pays de Sustain Bench (AGO => Angola)
* La correspondance enter Global Data et Petterson qui ne répertorie pas les cluster\_id (jointure par les coordonnées satellites)
* Attribuer des cluster\_id artificiels aux clusters présents uniquement dans le dataset Petterson (simple indexation avec remise à zéro dès qu’un cluster dispose déjà d’un id)

## Téléchargement des images satellites

## Microsoft Planetary Computer

Pour le téléchargement des images satellites (Landsat), nous nous sommes tournés vers MPC plutôt que Google Earth Engine car il a l’avantage de ne pas nécessiter d’inscription et permet de stocker les données directement en local (et pas sur drive) grâce à la fonction to\_raster de rasterio.

## Cloudless Mosaic

La documentation de MPC explique comment réaliser une image composite sans nuages. Le composite permet également de corriger les défauts de capteur du Landsat 7 s’il y a suffisamment d’items à superposer

Il est nécessaire de bien définir la zone d’intérêt car le traitement d’une tuile entière est extrêmement long. Un patch de 10km de côté prend quant à lui une dizaine de seconde à être généré.

Le filtrage selon la quantité de nuages et la fenêtre temporelles sont des éléments à paramétrer.

Par la suite, d’autres stratégies pourront être envisagée comme l’utilisation de données brutes et non composites pour éviter les biais, ou le composite par saison. Ce sont les perspectives d’amélioration pour le moment.