# Projet de Sciences des Données

Explotation d'images satellites haute-résolution
pour la prevision d'indicateurs socio-économiques

Youcef - Kacer

25 Aout 2016

# Table des matières

# Introduction

Ce document présente le projet de sciences des données que je compte développer, dans le cadre de la validation du CES Data Scientist à Telecom-Paristech [TP14].
Ce projet consiste à exploiter des images satellitaires haute-résolution afin d'en extraire des indicateurs socio-économiques.
En effet, évaluer la densité démographique d'un pays, par exemple, peut représenter un coût non négligeable en terme de recensement. Or utiliser des images aériennes et leurs caractéristiques permettrait de prédire la population présente à moindre coût. En effet, les « edges » des routes et des batiments caractérisent les zones urbaines et donc les zones à forte population, alors que les champs et les forêts caractérisent des zones faiblement peuplées.

# Chapitre 1

# Images exploitées

Les images à exploiter proviennent du satellite Landsat 8 de la NASA et sont libres d'accès [Sur16]. Ce satellite scanne tout le globe terrestre tous les 16 jours, ce depuis 2013. Ces images permettent donc non seulement d'étudier une zone à un moment donnée mais aussi d'étudier son évolution sur une période donnée (entre 2013 et 2016, on aurait donc 91 couvertures d'une même zone).
Ces images sont très riches dans la mesure où elles présentent en tout 11 canaux, 9 dans le visible et 2 dans l'infra-rouge. Donc en plus des caractéristiques de formes, le niveau des images doit pouvoir nous renseigner sur la nature des materiaux et des objets présents au sein d'une zone (métal ou végétation par exemple), les canaux infra-rouges pourront très certainement quantifier la présence humaine.
Certaines combinaisons bien spécifiques de bandes peuvent apporter de l'information, comme le montre le tableau 1.1 tiré de [esr16] :

| | |
|---|---|
| Natural Color | 4 , 3 , 2 |
| False Color (urban) | 7 , 6 , 4 |
| Color Infrared (vegetation) | 5 , 4 , 3 |
| Agriculture | 6 , 5 , 2 |
| Atmospheric Peration | 7 , 6 , 5 |
| Healthy Vegetation | 5 , 6 , 2 |
| Land/Water | 5 , 6 , 4 |
| Natural With Atmospheric Removal | 7 , 5 , 3 |
| Shortwave Infrared | 7 , 5 , 4 |
| Vegetation Analysis | 6 , 5 , 4 |

TABLE 1.1 – Combinaisons possibles à 3 canaux

Ces bandes couvrent approximativement un périmètre de 185km dans la direction Nord-Sud et 185km dans la direction Est-Ouest, pour une résolution de 30 mètres (soit des images de 8000x8000 pixels). Cette résolution équivaut à la demi-longueur d'un terrain de football, ce qui ne sera pas suffisant pour récupérer les edges des batiments et autres structures, c'est pour cela qu'on envisagera de coupler nos images satellitaires à l'exploitation des images couleur de Google Earth.
Les figures 1.1,1.2,1.3 présentent respectivement les bandes 2,3 et 4 d'un dataset que nous avons pu télécharger depuis [Sur16], et pris autour de la ville de *Grenada (Espagne)*

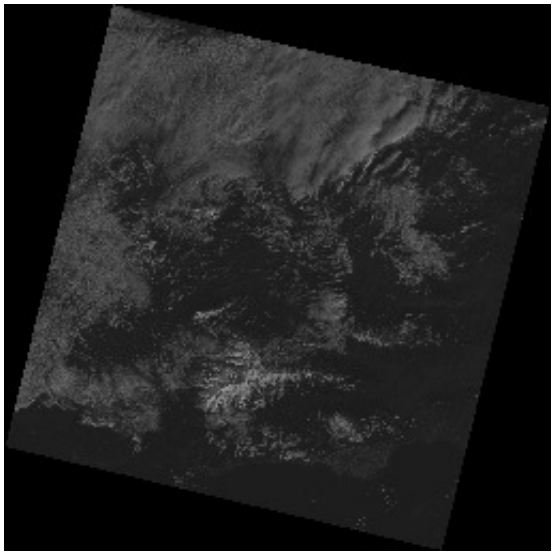

FIGURE 1.1 – image Lansat-8 (bande 2), région autour de *Grenada* (*Espagne*)

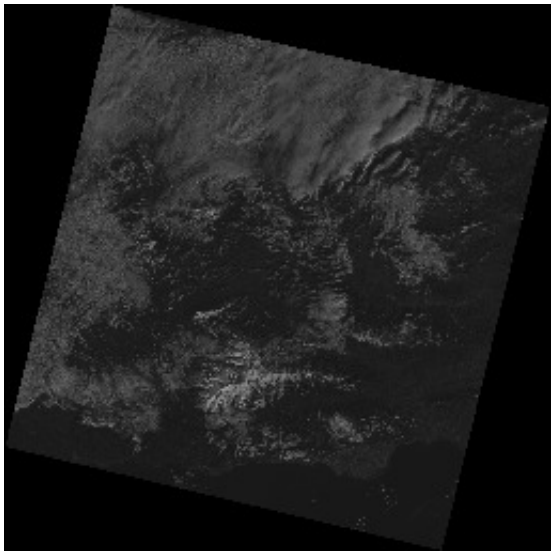

FIGURE 1.2 – image Lansat-8 (bande 3), région autour de *Grenada* (*Espagne*)

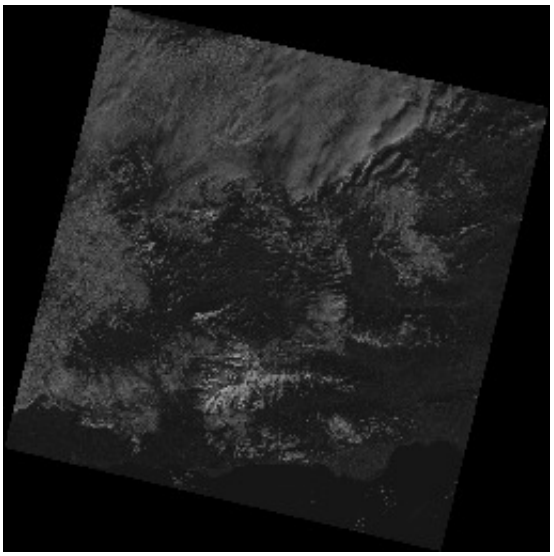

FIGURE 1.3 – image Lansat-8 (bande 4), région autour de *Grenada* (*Espagne*)

# Chapitre 2

# Méthode d'apprentissage

L'idée serait de s'intéresser à une certaine zone (un pays par exemple), dont on aurait l'indicateur de densité de population (valeur à prédire) pour un grand ensemble de communes du pays.
On pourra alors récupérer plusieurs images satellitaires quadrillant ce pays, et attribuer à chacune d'elles sa valeur de densité de population (on doit pouvoir utiliser la latitude et la longitude d'une image pour retrouver la commune concernée).
Ainsi, on récupère un ensemble classique d'images labelisées par sa densité de population.
Ensuite, on pourra extraire des descripteurs de ces images (histogramme orienté du gradient [DT05], entre autre) auquels on appliquera un algorithme de regression supervisé (la valeur à prédire, la densité de population, est plutôt continue que discrète).
On aurait donc un modèle de classification capable de prédire la densité de population d'une zone en fonction d'images satellites.
On pourra alors tester la généralisation du classifieur, en s'intéressant à d'autres pays.

# Chapitre 3

# Outils pour le stockage, le calcul et la visualisation

Si l'on s'intéresse à un territoire grand comme la France, on doit s'attendre à récupérer un total de 49 datasets pour quadriller cette zone [**?**]

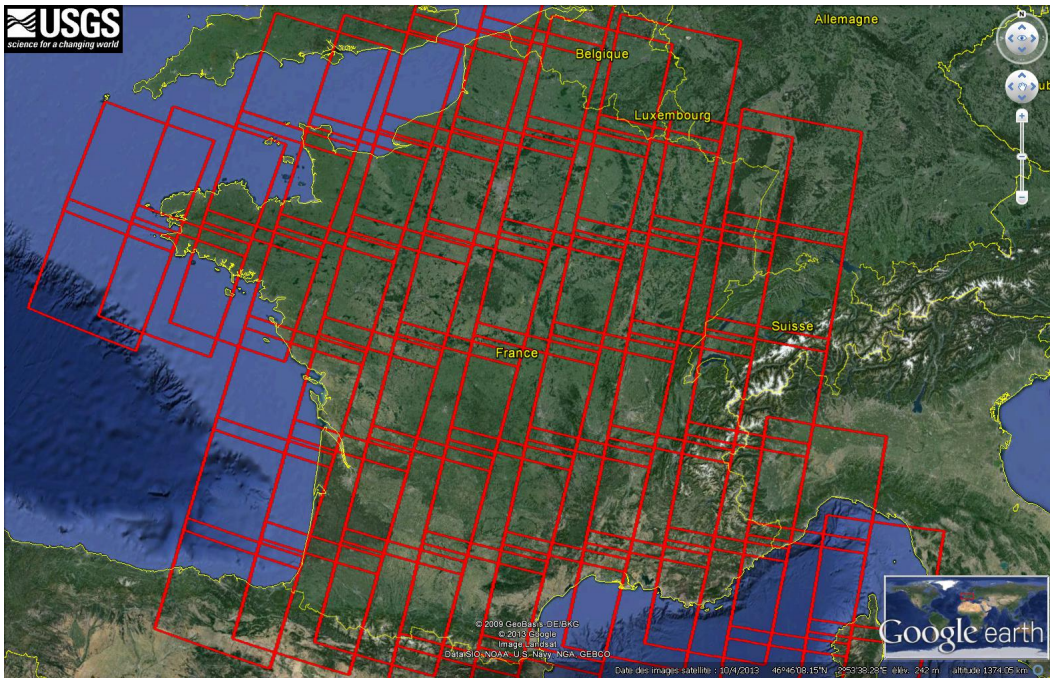


FIGURE 3.1 – Quadrillage de la France par Landsat-8 [geo16]

Chaque dataset compte les 11 bandes évoquées plus haut, mais aussi une image d'information de qualité renseignant pour chacun des pixels, la présence ou non de nuage, la présence ou non de mers ainsi que la présence ou non de neige.
Ainsi, chaque dataset mesure approximativement 1Go, on donc aurait une quantité totale de **49Go** . Des descripteurs comme les hitogrammes orientés du gradient (pour une taille de block 16 pixels , une taille de cellule de 4, un recouvrement de 4) pourraient multiplier cette quantité par un facteur 8, soit **392Go**, ce qui ne tiendrait pas en mémoire vive. Les images seront donc stockées sur HDFS [Whi09] (soit en mode « single-node cluster », soit en mode « multi-node cluster »via le cluster de Telecom ParisTech), cela permettra d'extraire les descripteurs par Map/Reduce. Les descripteurs seront aussi stockés de manière distribuée, via une table HBASE afin de pouvoir effectuer des requêtes et vérifier les valeurs calculées.
Une fois les descripteurs calculés, on utilisera la librairie de Machine Learning MLlib de Spark [MBY+16], dédiée à l'apprentissage sur données distribuées dans HDFS. Pour la regression, cette librairie ne permet cependant que les regressions linéaires (soit par moindres carrés, par Ridge ou par Lasso) et les regressions logistiques (pas de SVM regressif).
On testera toutes ces méthodes pour en analyser les erreurs en cross-validation.
Enfin, pour la visualisation des resultats, on proposera une page html utilisant la librairie javascript D3 [Jai14] pour l'affichage de la carte de densité de population. Cette interface web sera aussi intéractive que possible afin de permettre des zooms mais aussi l'observation de la densité de population sur une année antérieure.

# Liste des tableaux

# Table des figures

# Bibliographie


[DT05] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. pages 886–893, 2005.

[esr16] `https://blogs.esri.com/esri/arcgis/2013/07/24/band-combinations-for-landsat-8/`, 2016.

[geo16] `http://ids.equipex-geosud.fr/images-landsat-8/`, 2016.

[Jai14] Abhijit Jain. Data visualization with the d3.js javascript library. *J. Comput. Sci. Coll.*, 30(2) :139–141, December 2014.

[MBY+16] Xiangrui Meng, Joseph Bradley, Burak Yavuz, Evan Sparks, Shivaram Venkataraman, Davies Liu, Jeremy Freeman, DB Tsai, Manish Amde, Sean Owen, Doris Xin, Reynold Xin, Michael J. Franklin, Reza Zadeh, Matei Zaharia, and Ameet Talwalkar. Mllib : Machine learning in apache spark. *J. Mach. Learn. Res.*, 17(1) :1235–1241, January 2016.

[Sur16] U.S. Geological Survey. `http://eros.usgs.gov/`, 2016.

[TP14] Telecom-ParisTech. `http://www.telecom-evolution.fr/fr/formations-certifiantes/ces-data-scientist`, 2014.

[Whi09] Tom White. *Hadoop : The Definitive Guide*. O'Reilly Media, Inc., 1st edition, 2009.