# Application de l’intelligence artificielle à la chimie du sol africain

https://github.com/opsabarsec/African-soil-chemistry/

Marco Berta

# Table des Matiéres

# 1. Introduction

Les progrès dans l'analyse rapide et à faible coût d'échantillons de sol, le géo-référencement d'échantillons de sol et une plus grande disponibilité des données de télédétection terrestre offrent de nouvelles opportunités pour prédire les propriétés fonctionnelles du sol à des emplacements non échantillonnés. Par contre, pour la mesure de la concentration des éléments du sol liées à la fertilité, actuellement est nécessaire utiliser des test coûteux en termes d'argent et de temps. Les techniques principales sont l'analyse d'éléments organiques CHNSO et la chromatographie en phase liquide pour l'analyse des éléments à faible concentration (P, K, Ca)(Shamrikova *et al.*, 2020). L'analyse CHNS, analyse la plus courante en analyse élémentaire, est basée sur un procédé par combustion dont la température peut atteindre un maximum de 1150°C pour la mesure du carbone, de l'hydrogène et de l'azote(Fadeeva, Tikhova and Nikulicheva, 2008). La proportion de ces éléments détermine le degré d'évolution de la matière organique, c'est-à-dire de son aptitude à se décomposer plus ou moins rapidement dans le sol et à fertiliser les cultures. La chromatographie ionique (ICP) est une méthode particulièrement bien adaptée à l'analyse des anions et cations majeurs des solutions du sol, par contre demande une procédure longue de préparation des échantillons et des instruments coûteux(Papadakis and Papadopoulou-Mourkidou, 2002).

Cette contraintes sont encore plus importantes pour des pays en voie de développement telles que ces de l'Afrique subsaharienne.
Des analyses comme la mesure en fluorescence des rayons-X (XRF) (Kalnicky and Singhvi, 2001)et la spectroscopie infrarouge (FTIR)(Teong *et al.*, 2016) sont plus rapides et économiques mais ne donnent pas une mesure directe de la concentration des éléments importants comme le carbone, l'hydrogène et l'azote. D'ici l'intérêt de prédire ces variables par une modélisation mathématique et de trouver une corrélation entre les mesures obtenues par analyse conventionnelle (CHNS, ICP) et ces obtenues par des techniques plus rapides et économiques (spectroscopie XRF et FTIR).

Les résultats des mesures sur une large quantité échantillons de sol africain sont disponibles sur le site ( Amazon Web Services S3 bucket)  https://registry.opendata.aws/afsis/
La structure et taille des données sont présentées en Fig.1. Les mesures ont été faites par trois laboratoires. Parfois la mesure de la mème variable (ex. le pH) est faite par deux laboratoires différents. Ou la même quantité (ex. concentration de phosphore) est mesurée avec deux ou trois techniques différentes: electro-conductivité (M3_P), chromatographie (ICP_OES_P mg/Kg), XRF (P).
Des modélisations de cette données ont été réalise par des algorithmes classiques de régression (Ridge, Random Forest..) (*African-Soil-Analysis - FTIR spectroscopy*, no date)(Hengl *et al.*, 2017)pour prédire des variables chimique du terrain comme le pH ou  la concentration du phosphore (Fig.2).

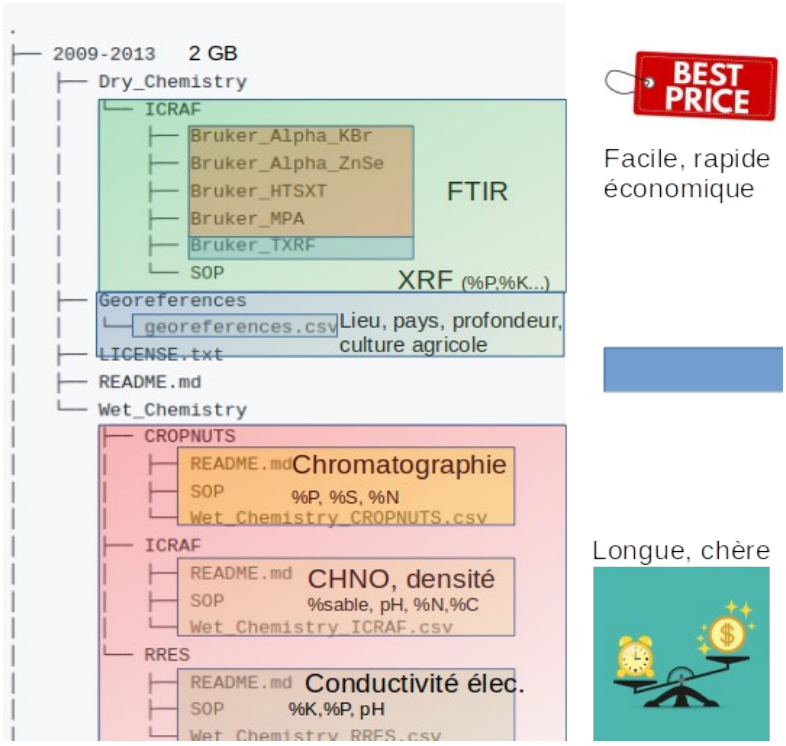


Fig.1 Structure des données AFSIS dans le AWS S3 bucket. Mesures des trois laboratoires différents. Parfois de la même quantité avec instruments/techniques differentes.

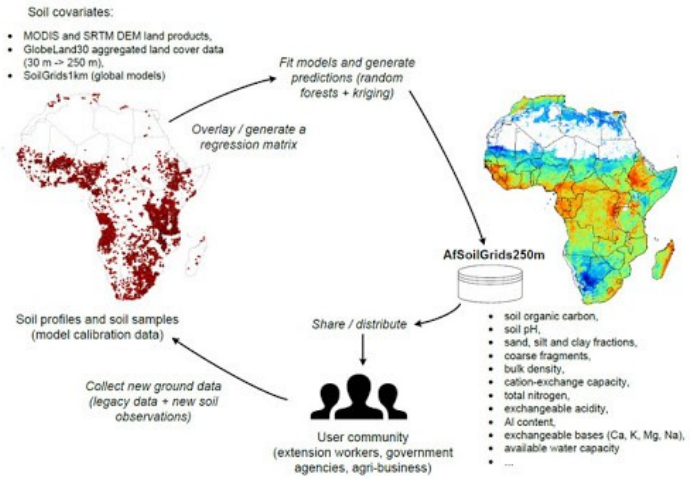


Fig.2 Modélisation des données chimiques du sol africain (Hengl *et al.*, 2017).

Dans cette étude on estime la possible corrélation entre les variables de composition chimique du sol, mesurées par XRF (X-ray fluorescence), et sa fertilité. En suite on essaye de prédire les variables de composition chimique par des mesures de spectroscopie infrarouge (FTIR), encore plus rapide et moins chère que l'XRF. Une publication académique récente (Xu *et al.*, 2019) indique que la modélisation des données (FTIR) avec l'algorithme ''Partial Least Square''(PLS) régression, utilisé depuis longtemps pour le développement des nouveaux médicaments(Deeb *et al.*, 2007), peut être appliqué avec succès à la chimie du sol. La modélisation classique avec la méthode des moindres carrés par contre peut être améliorée avec l'utilisation d'une réseaux des neurones convolutionelle (CNN)(Ng *et al.*, 2019; Padarian, Minasny and McBratney, 2019).

Dans cette étude les données de spectroscopie infrarouge sont corrélées par la méthode des moindres carrés partiels (PLS) aux variables chimiques du sol et on teste une version modifiée de une CNN réalisée dans un étude académique pour la prédiction des médicaments (drug discovery) par spectroscopie infrarouge(Bjerrum, Glahder and Skov, 2017) . Les codes Python et les données modélisées sont disponibles sur https://github.com/opsabarsec/African-soil-chemistry/ .

# 2. Exploration des données

La première étape d'exploration et traitement des données concerne la création d'un dataset unique pour la chimie du sol avec le donnés prises avec techniques différentes. Pour chaque mesure existe un identifiant unique de l'échantillon et il suffit de faire une jointure (inner join) des tableaux correspondant à chaque set de mesures avec la clé de l'identifiant d'échantillon. On élimine les valeurs aberrants (concentration négative) et on applique l'imputation du valeur moyen des mesures aux valeurs nuls. L'élimination des outliers par la méthode de Tukey (2X ecart interquartile) par contre supprimerait plus de la moitié des données, que présentent généralement une forte dispersion (Fig.3).

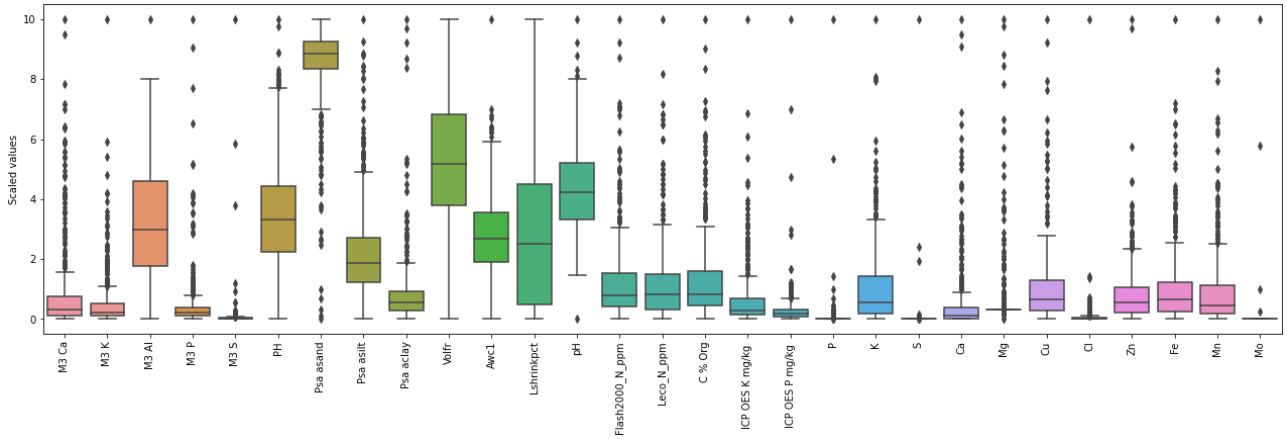


Fig. 3 Boxplots avec données normalisées des variables de composition chimique du sol.

Par contre le fait d'avoir l'estimation de la même variable faite par deux laboratoires différents mous permet de déterminer si cette dispersion est due à la mesure ou aux échantillons.Deux mesures de la même quantité (ex. le pH) son sensées être idéalement identiques. Ceci pratiquement n'est jamais le cas, mais on peut mesurer la fiabilité de la mesure par l'écart et le taux de corrélation des données prises par deux instruments différents. En Fig.4 on présente le matrice de corrélation des mesures différentes pour les mêmes quantités

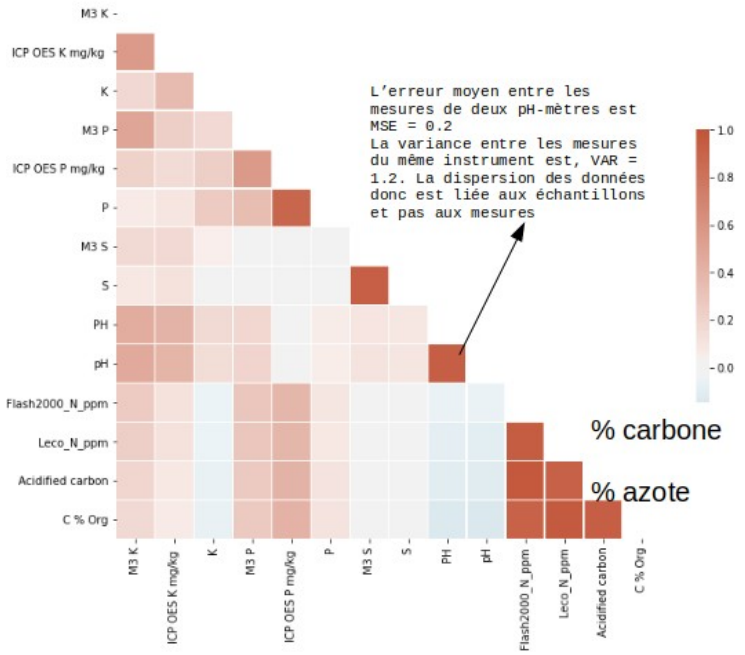


Fig.4 Matrice de corrélation des mesures des mêmes quantités.

Sauf que pour le potassium les mesures ont des coefficients de corrélation  linéaire élevés pour la même variable. On observe aussi que le de matière organiques (%C Org) et le %N sont fortement corrélés et on choisis que une des deux pour la modélisation.  Les spectres FTIR sont sélectionnés pour le longueur d'onde du moyen infrarouge (400-4000 $cm^{-1}$)(Xu *et al.*, 2019), chaque spectre correspondant à un échantillon avec la composition chimique mesurée. Après la sélection de la plage de longueur d'onde et l'élimination des doublons des mesures FTIR, la taille totale des donnée à traiter est réduite de 5GB à 30MB.

# 3. Modélisation et résultats

## 3.1 Corrélation entre fertilité et composition chimique (XRF)

Pour modéliser la corrélation entre variables chimiques/physiques et fertilité du sol, la présence (ou absence) des cultures agricoles dans le terrain de chaque échantillon était numérisée avec un code binaire 1, cultivé ou 0 pas cultivé. La majorité des variables explicatives était choisi parmi les mesures de concentration des éléments chimiques par fluorescence des rayons X (XRF). L'XRF est une technique rapide et relativement économique. Les variables du modèle sont présentées dans le tableau 1.

(148, 17)

| | pH | Nitrogen (ppm) | Water % | P | K | S | Ca | Mg | Cu | Cl | Zn | Fe | depth_sub | depth_top | Cultivated_n |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 4.57 | 392.82719 | 0.034962 | 50.6 | 12991.3 | 45.7 | 944.1 | 5575.0 | 13.0 | 210.2 | 22.0 | 12501.3 | 1.0 | 0.0 | 0 |
| 1 | 7.06 | 859.46908 | 0.042649 | 50.6 | 15173.5 | 45.7 | 9301.0 | 5519.1 | 18.4 | 152.6 | 38.5 | 24094.6 | 0.0 | 1.0 | 0 |
| 2 | 5.27 | 186.29957 | 0.091913 | 50.6 | 6838.9 | 45.7 | 884.3 | 5575.0 | 2.8 | 229.5 | 2.3 | 2213.4 | 1.0 | 0.0 | 0 |
| 4 | 6.50 | 377.73812 | 0.033089 | 50.6 | 12201.6 | 45.7 | 1790.1 | 5575.0 | 7.7 | 122.5 | 13.4 | 9135.1 | 0.0 | 1.0 | 0 |
| 5 | 5.41 | 2520.21961 | 0.112162 | 50.6 | 4731.1 | 45.7 | 4924.8 | 5575.0 | 56.1 | 145.1 | 40.6 | 67075.9 | 0.0 | 1.0 | 0 |

y[0,1]

X

Tableau 1. Variables liées à la composition et fertilité du sol pour chaque échantillon.

Le modèle plus pertinent pour la modélisation de la corrélation multivariée avec une variable cible y binaire est la régression logistique. Les résultats pour le jeux des données choisi sont présentes dans la tableaux 2a et 2b, et ils montrent que l'algorithme arrive à prédire avec une bonne précision les valeurs négatives (pas fertile), moins bien les valeurs positives.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.89 | 0.93 | 0.91 | 27 |
| 1 | 0.00 | 0.00 | 0.00 | 3 |
| accuracy | | | 0.83 | 30 |
| macro avg | 0.45 | 0.46 | 0.45 | 30 |
| weighted avg | 0.80 | 0.83 | 0.82 | 30 |

| | predicted negative | predicted positive |
|---|---|---|
| actual negative | 25 | 2 |
| actual positive | 3 | 0 |

Tableau 2. Résultats de la régression logistique entre composition et fertilité du sol.

Ceci peut être l'effet d'une pourcentage relativement faible (22%) échantillons pris dans des terrains cultivés.

## 3.2 Corrélation spectroscopie IR - composition chimique

La spectroscopie infrarouge (fourier-transform-infrared, FTIR en anglais) est une technique plus économique et rapide que l'XRF et peut donner des informations supplémentaires liées à la fertilité du sol. Dans la littérature académique on trouve que on arrive à prédire avec l'FTIR la concentration de nitrogen (%N), de sable, ou de matière organique (%C org) (Du and Zhou, 2009; Teong *et al.*, 2016; Ludwig *et al.*, 2019). La spectroscopie donne une mesure indirecte de ces quantités. Avec la spectroscopie on mesure l'intensité de la lumière pour chaque longueur d'onde dans une certaine plage de valeurs. Dans cette étude dans la région du infrarouge moyen : 2500nm-25μm ou 400-4000 $cm^{-1}$(Fig.5).

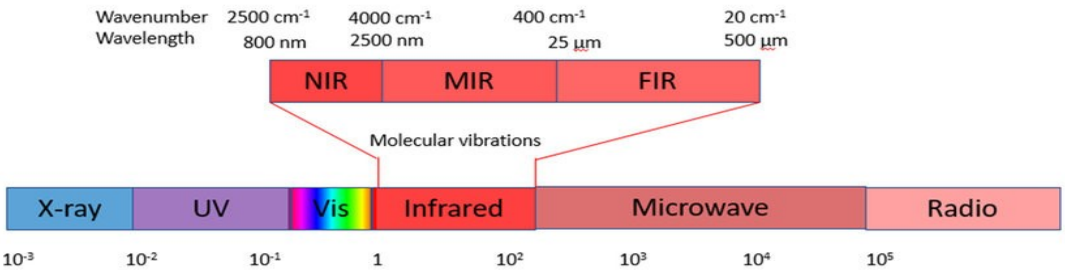


Fig.5 région spectrale infrarouge proche (near infrarouge, NIR), moyen (MIR) et loin (far infrared FIR)

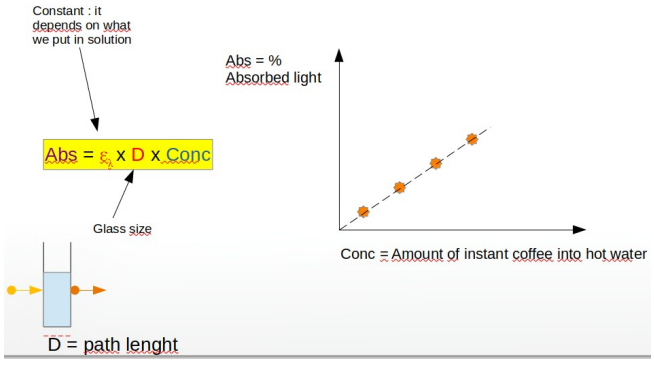


Fig.6 Loi de Lambert-Beer

L'intensité transmise quand un rayon de lumiére passe par une solution est inversement proportionnelle à l'absorbance de cette solution. L'absorbance est une fonction linéaire de la concentration de la solution, selon la loi de Lambert-Beer (Fig.6).

Si on arrive à comprendre dans quelle région du spectre un certain élément en solution absorbe l'intensité de la lumière on peut estimer indirectement sa concentration. On peut donc traiter chaque longueur d'onde comme une variable explicative dans une matrice $X_{mXn}$, et on choisis une ou plusieurs variables cible (pH,%sable,%N..). La méthode classique d'une prédiction linéaire multivariée avec plusieurs variables cible est avec l'algorithme des moindres carrés partiels (PLS). Similaire à l'ACP (analyse composants principaux), la régression par les moindres carrés partiels (PLS) est une technique qui réduit les prédicteurs à un plus petit ensemble de composantes non corrélées et qui effectue la régression par les moindres carrés sur ces composantes, plutôt que sur les données initiales. La fonctionnalité PLS est particulièrement utile lorsque les prédicteurs sont fortement colinéaires, comme les intensités à chaque longeur d'onde. Si avec l'ACP on choisi les composantes que expliquent mieux l'ensemble des variables explicatives, avec la PLS on essaye de trouver les plus corrélés/explicatives des variables cible. Une modélisation plus récente et performante était faite avec des CNN appliqués aux spectres(Zhang *et al.*, 2019). Par contre les même procédures ou une version modifiée de la CNN ne donnent pas des bons résultats avec le jeux de donnée sélectionné pour cette étude (Fig.7).

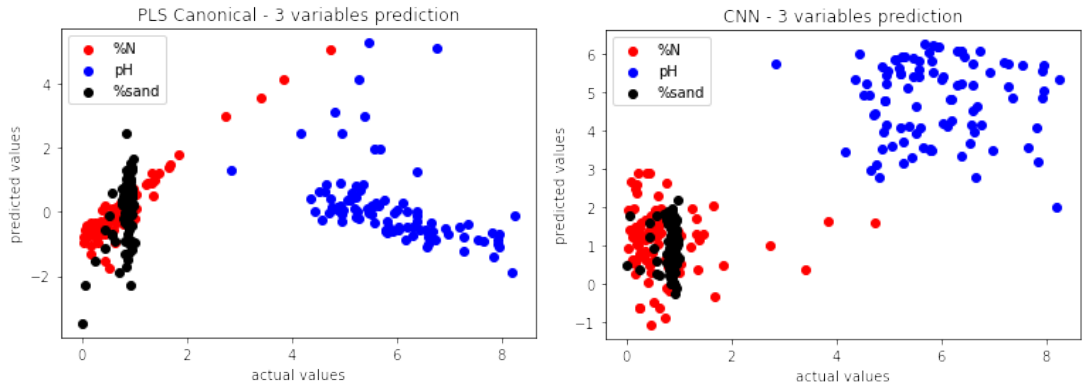


Fig.7 Prédictions par réseaux des neurones et PLS vs. valeurs mesurés.

Les faibles corrélations sont le résultat de la dispersion des valeurs comme on peut observer en Fig.8 (‘‘garbage in – garbage out’’ en anglais).

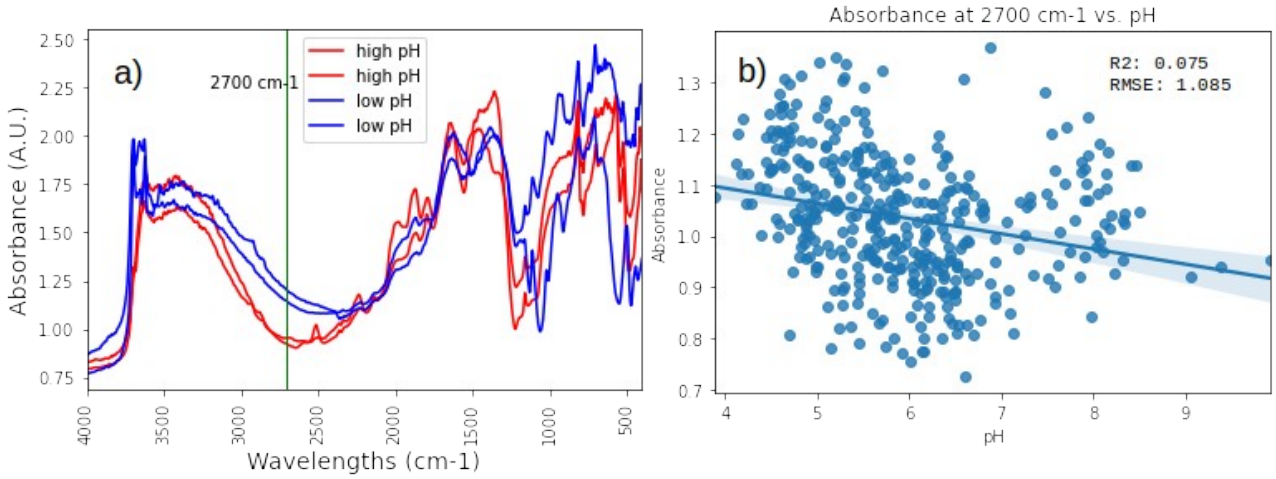


Fig.8 a) Spectres infrarouges des échantillons avec le plus faible et le plus haut pH. On observe une différence dans la région 2200-3200 $cm^{-1}$. La régression linéaire b) de l’intensité à 2700 $cm^{-1}$ a un coefficient $R^2$ très réduit, due à la forte dispersion des valeurs.

Si on prends que une variable cible, le pH, les spectres des échantillons avec le valeurs extrêmes présentent l’écart plus important à 2700 $cm^{-1}$ (Fig.8a). Par contre en Fig.8 b la dispersion des valeurs d’intensité vs. PH à 2700 $cm^{-1}$ donne un coefficient de régression linéaire proche au zero. Il n’es pas possible donc d’obtenir un bon modèle surtout avec l’hypothèse de linéarité comme pour la PLS.

# 4. Conclusion

Les échantillons avec une analyse complète de la composition sont une minorité du jeu de données (30 MB sur 5GB). Par contre une régression logistique entre des variables de composition chimique et la présence de cultures agricoles donne une bonne performance. Ceci n’es pas le cas pour la modélisation des spectres FTIR par PLS ou avec des réseaux des neurones. La théorie de spectroscopie infrarouge implique une corrélation linéaire entre absorbance et concentration mais une simple régression linéaire pour la variable pH montre que le coefficient est proche au zero et cette condition initiale n’est pas respectée pour le jeu des données examiné.

# 5. Bibliographie


*African-Soil-Analysis - FTIR spectroscopy* (no date). Available at: https://github.com/pcohen89/African-Soil-Analysis/blob/master/PC soil analysis.py (Accessed: 15 November 2020).

Bjerrum, E. J., Glahder, M. and Skov, T. (2017) 'Data Augmentation of Spectral Data for Convolutional Neural Network (CNN) Based Deep Chemometrics'. Available at: http://arxiv.org/abs/1710.01927 (Accessed: 18 November 2020).

Deeb, O. *et al.* (2007) 'Effect of the electronic and physicochemical parameters on the carcinogenesis activity of some sulfa drugs using QSAR analysis based on genetic-MLR and genetic-PLS', *Chemosphere*, 67(11), pp. 2122–2130. doi: 10.1016/j.chemosphere.2006.12.098.

Du, C. and Zhou, J. (2009) 'Evaluation of soil fertility using infrared spectroscopy: A review', *Environmental Chemistry Letters*, pp. 97–113. doi: 10.1007/s10311-008-0166-x.

Fadeeva, V. P., Tikhova, V. D. and Nikulicheva, O. N. (2008) 'Elemental analysis of organic compounds with the use of automated CHNS analyzers', *Journal of Analytical Chemistry*, 63(11), pp. 1094–1106. doi: 10.1134/S1061934808110142.

Hengl, T. *et al.* (2017) 'Soil nutrient maps of Sub-Saharan Africa: assessment of soil nutrient content at 250 m spatial resolution using machine learning', *Nutrient Cycling in Agroecosystems*, 109(1), pp. 77–102. doi: 10.1007/s10705-017-9870-x.

Kalnicky, D. J. and Singhvi, R. (2001) 'Field portable XRF analysis of environmental samples', *Journal of Hazardous Materials*, 83(1–2), pp. 93–122. doi: 10.1016/S0304-3894(00)00330-7.

Ludwig, B. *et al.* (2019) 'Accuracy of Estimating Soil Properties with Mid-Infrared Spectroscopy: Implications of Different Chemometric Approaches and Software Packages Related to Calibration Sample Size', *Soil Science Society of America Journal*, 83(5), pp. 1542–1552. doi: 10.2136/sssaj2018.11.0413.

Ng, W. *et al.* (2019) 'Convolutional neural network for simultaneous prediction of several soil properties using visible/near-infrared, mid-infrared, and their combined spectra', *Geoderma*, 352, pp. 251–267. doi: 10.1016/j.geoderma.2019.06.016.

Padarian, J., Minasny, B. and McBratney, A. B. (2019) 'Using deep learning to predict soil properties from regional spectral data', *Geoderma Regional*, 16, p. e00198. doi: 10.1016/j.geodrs.2018.e00198.

Papadakis, E. N. and Papadopoulou-Mourkidou, E. (2002) 'Determination of metribuzin and major conversion products in soils by microwave-assisted water extraction followed by liquid chromatographic analysis of extracts', *Journal of Chromatography A*, 962(1–2), pp. 9–20. doi: 10.1016/S0021-9673(02)00537-X.

Shamrikova, E. V. *et al.* (2020) 'Nitrogen Compounds in the Soil of the Continental Margins of the European Russian Arctic', *Eurasian Soil Science*, 53(7), pp. 870–881. doi: 10.1134/S1064229320070133.

Teong, I. T. *et al.* (2016) 'Characterization of Soil Organic Matter in Peat Soil with Different Humification Levels using FTIR', in *IOP Conference Series: Materials Science and Engineering*. Institute of Physics Publishing, p. 012010. doi: 10.1088/1757-899X/136/1/012010.

Xu, X. *et al.* (2019) 'Detection of soil organic matter from laser-induced breakdown spectroscopy (LIBS) and mid-infrared spectroscopy (FTIR-ATR) coupled with multivariate techniques', *Geoderma*, 355. doi: 10.1016/j.geoderma.2019.113905.

Zhang, X. *et al.* (2019) 'DeepSpectra: An end-to-end deep learning approach for quantitative spectral analysis', *Analytica Chimica Acta*, 1058, pp. 48–57. doi: 10.1016/j.aca.2019.01.002.