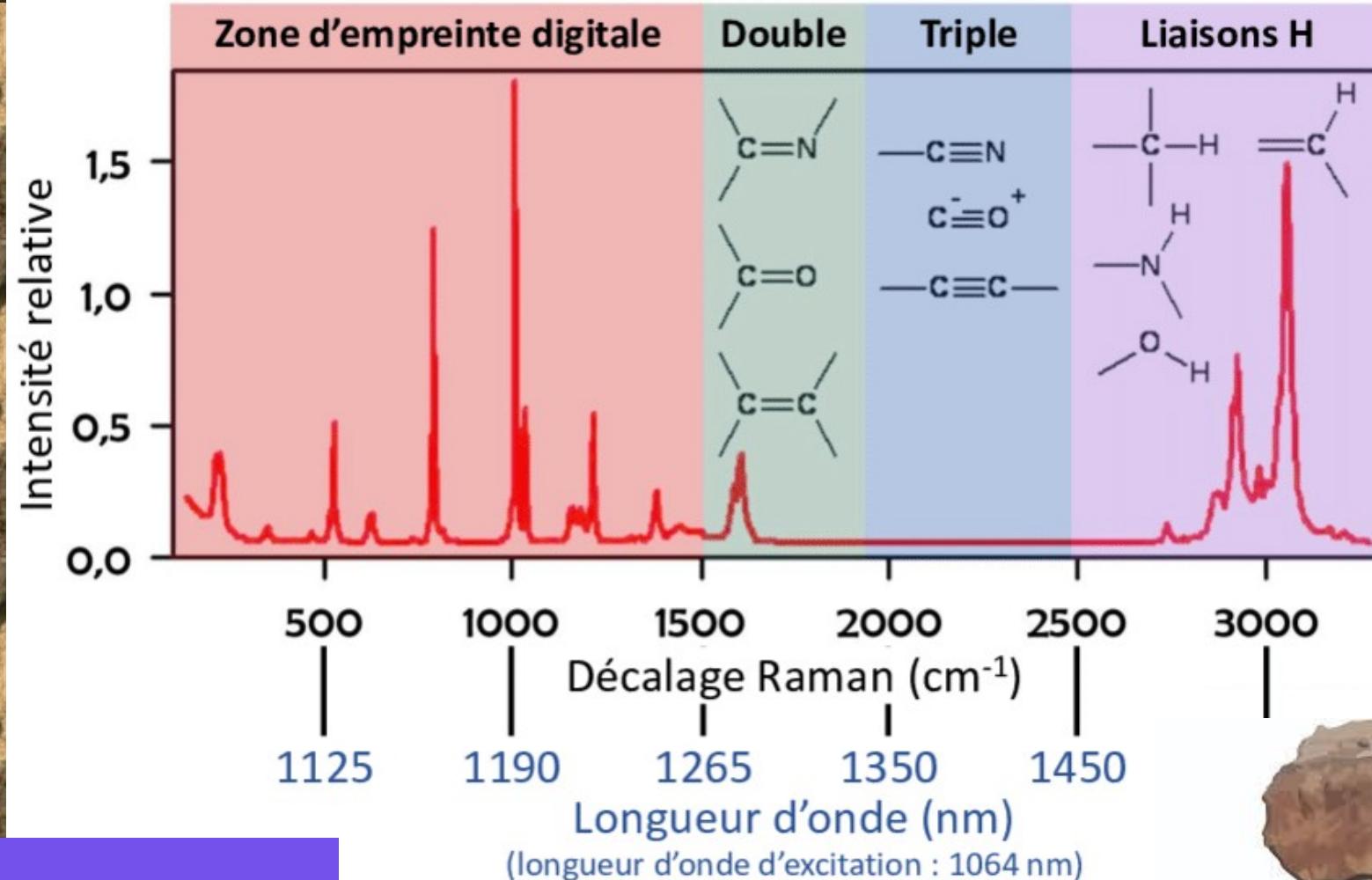




# Preuve de concept: Prédiction de la fertilité du sol par machine learning



OPENCLASSROOMS

[https://github.com/opsabarsec/  
African-soil-chemistry/](https://github.com/opsabarsec/African-soil-chemistry/)

AfSIS

# Outline

- **Introduction**
  - Objectif de l'étude
  - Jeux de données
- **Analyse exploratoire**
  - Les données AFSiS
  - Analyses de composition chimique
  - Sélection des spectres infrarouges
  - Distribution géographique
- **Modélisation**
  - Modèle de départ
  - Pistes de modélisation
  - Corrélation entre mesures et fertilité du sol
  - Corrélation entre spectroscopie et chromatographie
  - Régression multivariée par les moindres carrés partiels (PLS)
- **Conclusions**
  - Résultat et possible amélioration

## Modalités de la soutenance

5 min - Présentation de la problématique, de son interprétation et des pistes de recherche envisagées.

5 min - Présentation de l'exploration.

10 min - Présentation des différentes pistes de modélisation effectuées.

5 min - Présentation du modèle final sélectionné et résultats.

5 à 10 minutes de questions-réponses.

# Objectif du modèle

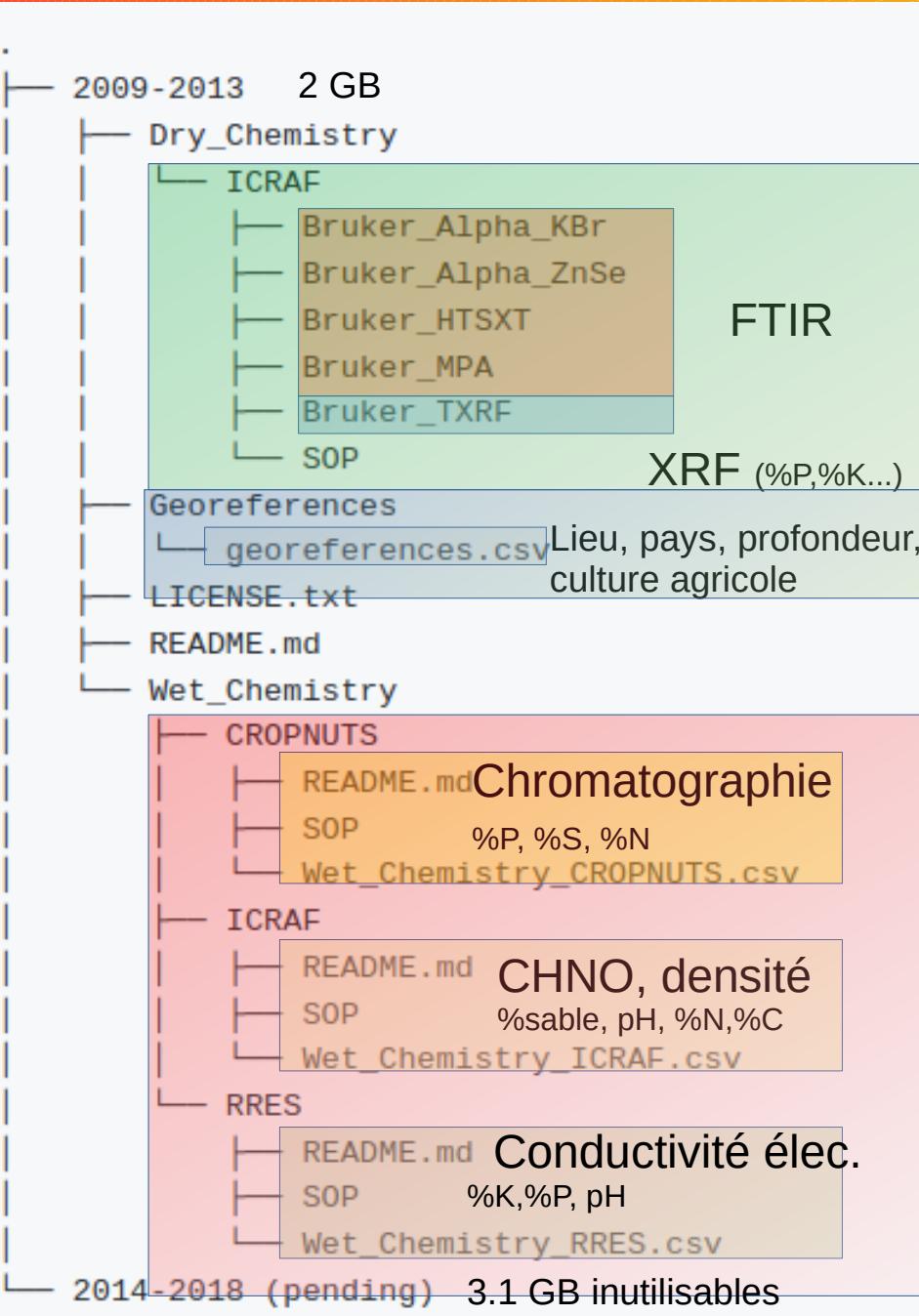
Dans cette étude avec un apprentissage supervisé, on essaye de trouver le meilleur modèle pour

- Prédire la fertilité du sol à partir des mesures de concentration rapides et économiques
- Prédire la composition du sol à partir des données des spectroscopie infrarouge

# Exploration des données

- Structure des données
- Nettoyage : imputation celles vides, outliers
- Sélection des données pour les modélisations

# Structure des données



Facile, rapide  
économique

Plusieurs laboratoires



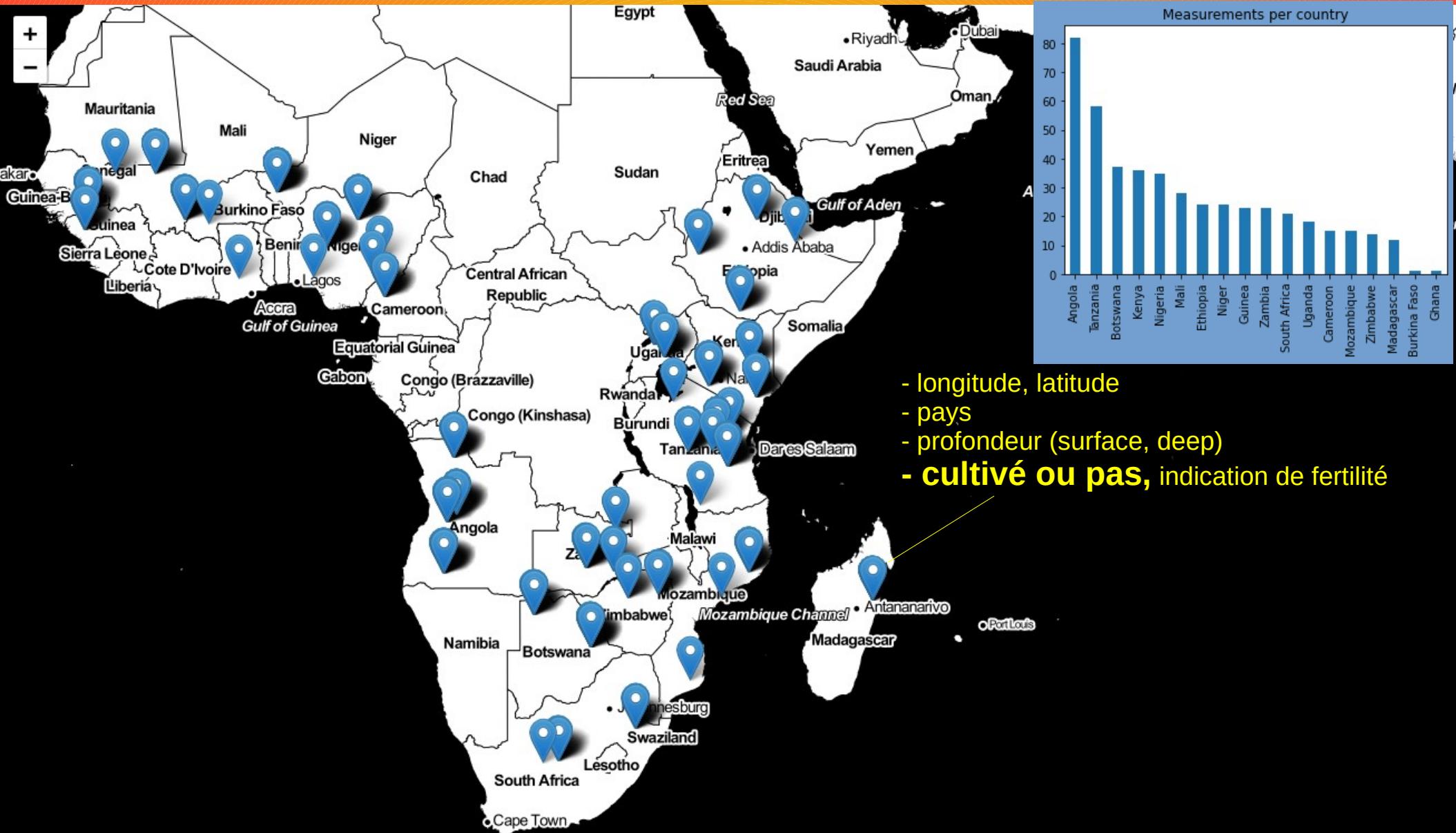
Spectres infrarouges  
mesurés avec  
instruments différents  
pour chaque  
échantillon

Longue, chère



Propriétés chimiques  
mesurées que pour  
une minorité des  
échantillons avec  
techniques différentes  
dans chaque  
laboratoire

# Distribution géographique



# Procédure de modélisation : composition et fertilité

	SSN	M3 Ca	M3 K	M3 Al	M3 P	M3 S	PH
0	icr006475	207.1	306.30	1095.0	4.495	18.960	4.682
1	icr006586	1665.0	1186.00	1165.0	12.510	13.600	7.062
2	icr007929	2518.0	72.57	727.6	21.090	14.810	7.114
3	icr008008	734.3	274.60	1458.0	109.200	11.400	5.650
4	icr010198	261.8	91.76	2166.0	3.958	5.281	5.501

Geo

Lab 1 → CSV1

Lab 2 → CSV2

Lab 3 → CSV3

Intégration des données composition chimique du sol et géoreferences  
(inner joins sur ID échantillon)

Nettoyage des données

Sélection des variables d'entrée

Corrélation avec la présence de cultures agricoles

	SSN	pH	Leco_N_ppm	C % Org	ICP OES K mg/kg	ICP OES P mg/kg
0	icr006454	7.85	800.0	0.94	8517.919223	96.575131
1	icr006455	8.03	600.0	0.70	10859.303780	117.423139
2	icr006474	5.01	500.0	0.57	1343.124117	87.040073
3	icr006475	4.57	500.0	0.47	1487.768795	83.555482
4	icr006492	6.78	900.0	0.98	2999.240760	150.936463

# Nettoyage des données

- 1) Suppression des doublons
- 2) Élimination des valeur négatives
- 3) Suppression des colonnes presque vides/variables inutiles
- 4) Imputation des valeurs manquantes avec la moyenne des valeurs

```
todrop = ['Soil material','Scientist', 'Site', 'Country', 'Region', 'Gid','RES ID']

df_geoelements_reduced = df_geoelements.drop(todrop, axis = 1)

print(df_geoelements_reduced.shape)
print(df_geoelements_reduced.isna().sum())

(467, 54)
SSN                 0
pH                  15
%N                  1

^def replace_missings(data):
    # this replaces missings with medians
    # NOTE: mixed string num columns it does not do anything with
    for cols in data.get_numeric_data().columns:
        data[cols].fillna(value=data[cols].median(), inplace=True)

replace_missings(df_geoelements_reduced)

print(df_geoelements_reduced['Cultivated'].unique())
df_geoelements_reduced['Cultivated'] = df_geoelements_reduced['Cultivated'].fillna('unknown')
```

# Valeurs aberrants

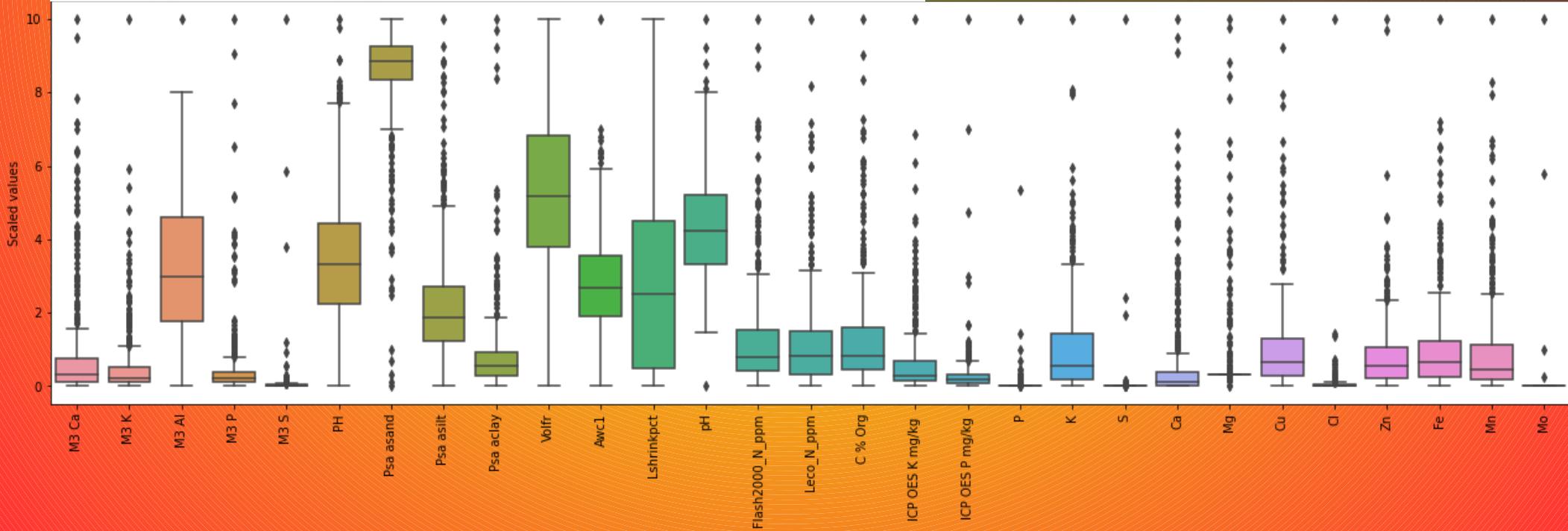
```
def detect_outliers(df, n, features):
    """
    Takes a dataframe df of features and returns a list of the indices
    corresponding to the observations containing more than n outliers according
    to the Tukey method.
    """
    outlier_indices = []

    # iterate over features(columns)
    for col in features:
        # 1st quartile (25%)
        Q1 = np.percentile(df[col], 25)
        # 3rd quartile (75%)
        Q3 = np.percentile(df[col], 75)
        # Interquartile range (IQR)
        IQR = Q3 - Q1

        # outlier step
        outlier_step = 3 * IQR
```

## Identification des valeurs aberrantes avec la règle $2,0 \times$ écart interquartile

209/476 outliers

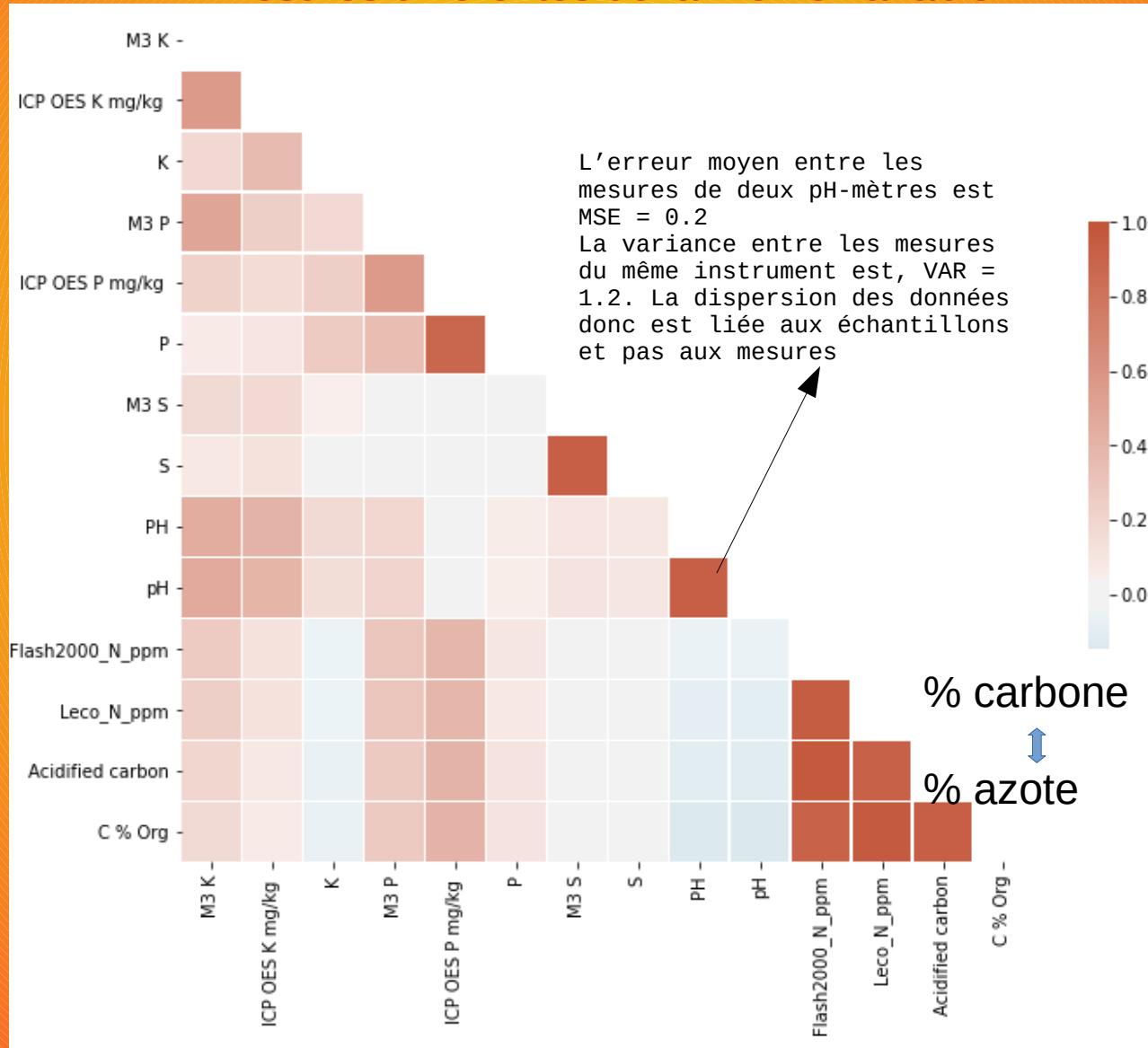


On a une variabilité liée à la mesure ou à l'échantillon ?

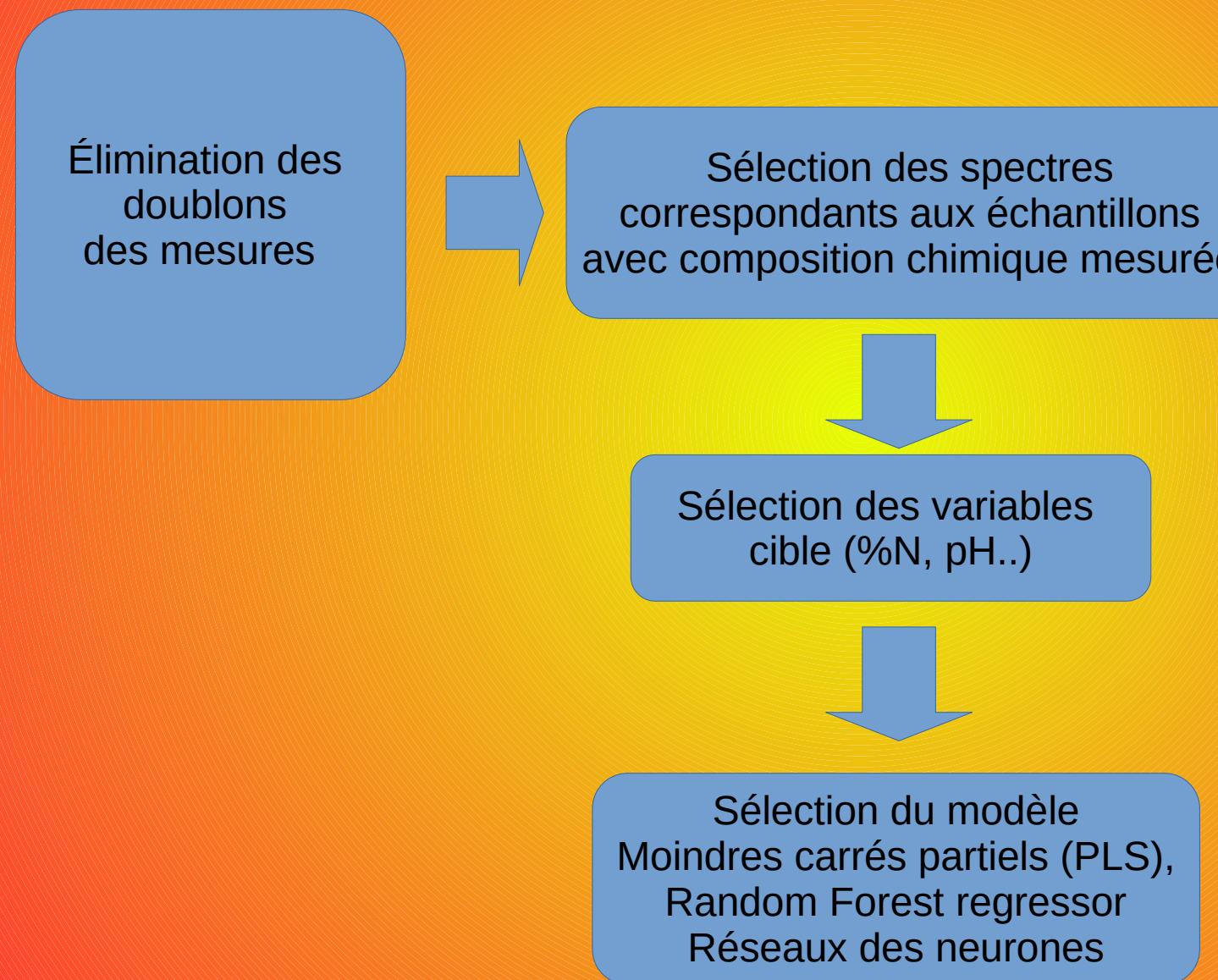
# Sélection variables composition

## Matrice des corrélations

Mesures différentes de la même variable



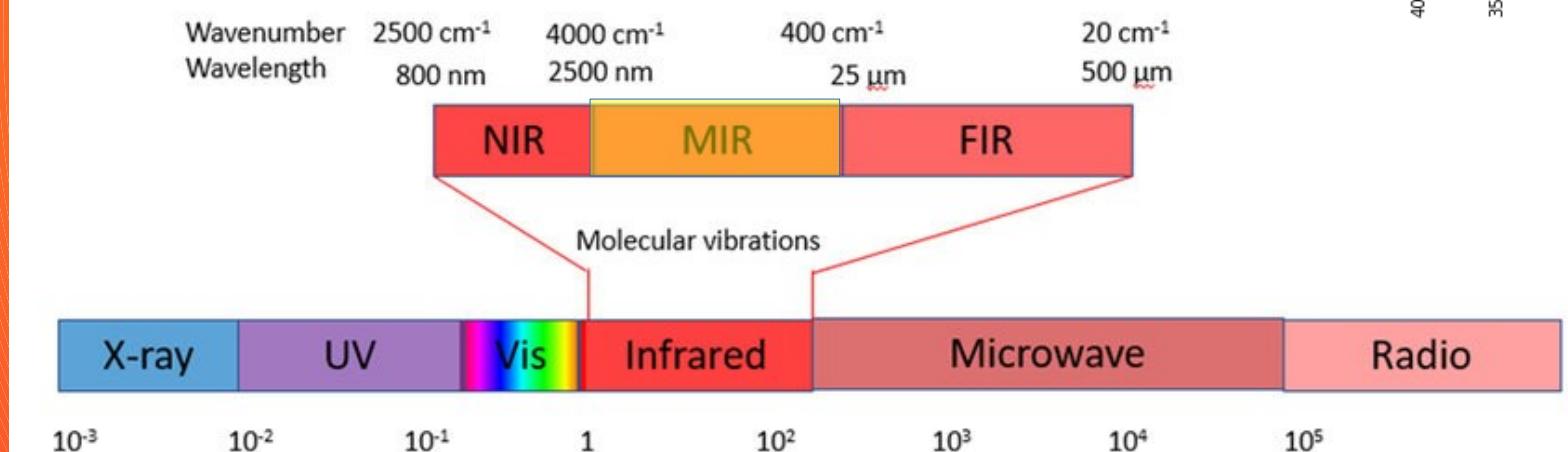
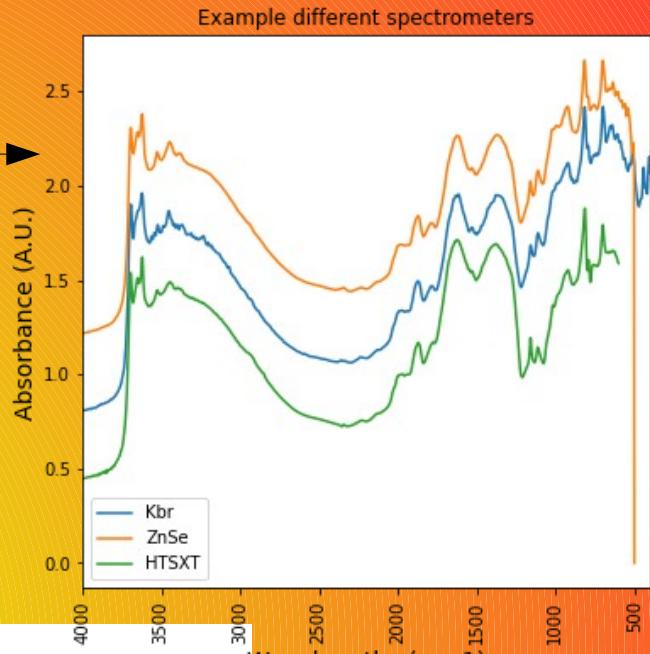
# Procédure de modélisation : spectroscopie et composition



# Spectres infrarouges : sélection

- Élimination des doublons
- Que les spectres dans la région IR moyen

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.



- Que les spectres correspondants aux échantillons avec caractérisation chimique

5 GB



30 MB

# Modélisation 1 : composition et fertilité

- Sélection des données “moins chères” (XRF)
- Encodage des variables catégoriques
- Évaluation du modèle

# Sélection et modélisation des variables

(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

X

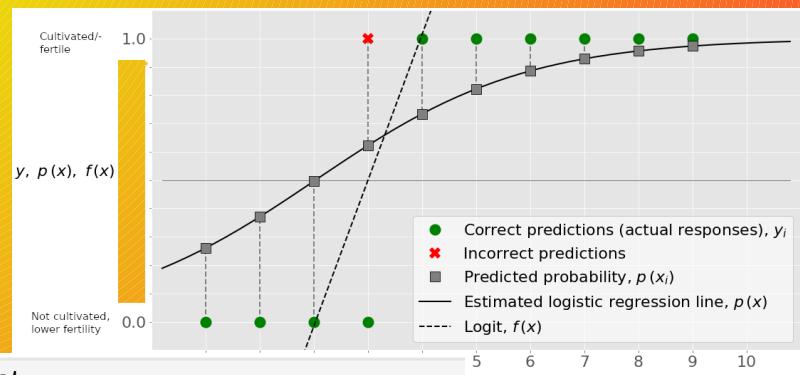
% fertile land samples 22.3

## Régression logistique

# Cross-validation Que 30 échantillons de test

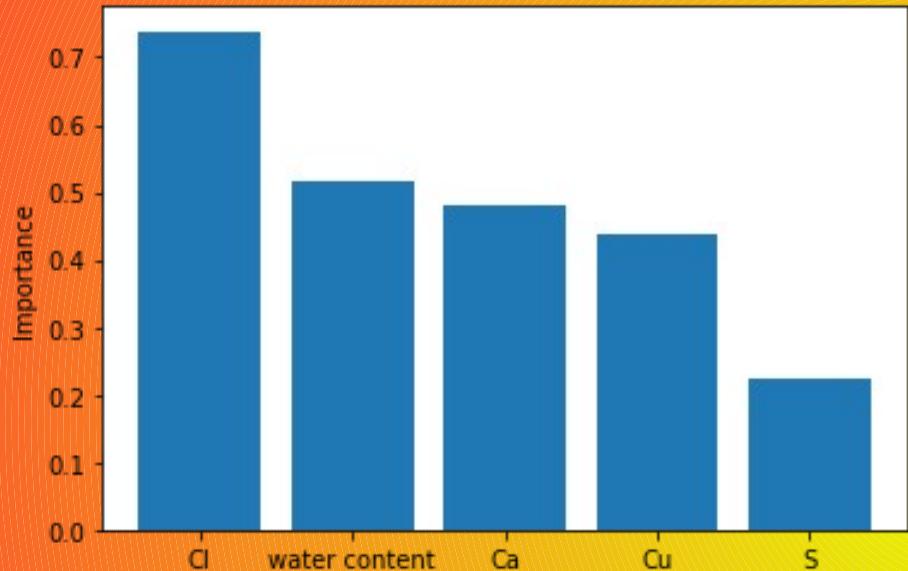
```
y_cv = cross_val_predict(lr_cultivated, X1_test, y1_test, cv=5)
scores = cross_val_score(lr_cultivated, X1_test, y1_test, cv=5)
scores
```

```
array([0.83333333, 0.5 , 0.83333333, 0.83333333, 0.83333333])
```



# Les variables plus importantes

Coefficients de la régression logistique



```
n_components = 5

pca = PCA(n_components= n_components)
principalComponents = pca.fit_transform(X_cult_scaled)

df_PCA_composition = pd.DataFrame(data = principalComponents , columns = ['compositional_component1','compositional_component2','compositional_component3','compositional_component4','compositional_component5'])

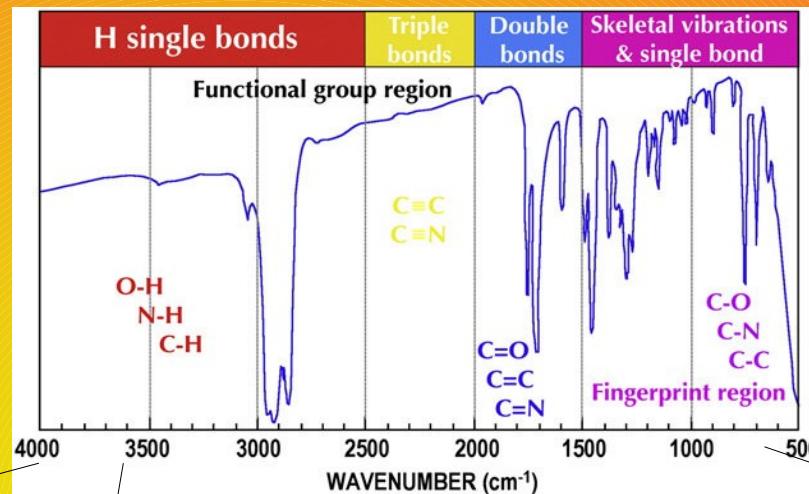
print('the new component contains',pca.explained_variance_ratio_ * 100, '% of the information from the', n_components , 'variables', 'for a total of',sum(pca.explained_variance_ratio_ * 100), '%')
#
the new component contains [30.6097796  18.49971864  15.10061963  10.03959827  7.97674046] % of the information from the 5 variables for a total of 82.22645659713287 %
```

# Modélisation 2 : spectroscopie IR et composition

- Sélection des variables cible
- Modèles classiques
- Réseaux des neurones convolutionelle

# Corrélation FTIR : jeux données

échantillon



	3998	3997	3996	3994	3993	3991	3990	3988	3987	3986	...	412	4
icr033603	0.908592	0.908981	0.909435	0.909956	0.910562	0.911266	0.912069	0.912953	0.913888	0.914843	...	1.849416	1
icr042897	0.810133	0.809940	0.809745	0.809616	0.809566	0.809563	0.809558	0.809516	0.809437	0.809359	...	2.069042	2
icr049675	0.711836	0.712355	0.713021	0.713747	0.714438	0.715037	0.715535	0.715951	0.716298	0.716546	...	2.042677	2
icr034693	0.788686	0.789201	0.789494	0.789611	0.789634	0.789656	0.789744	0.789912	0.790109	0.790249	...	2.083768	2
icr033950	0.752478	0.753251	0.753915	0.754423	0.754733	0.754816	0.754683	0.754402	0.754084	0.753866	...	1.925999	1

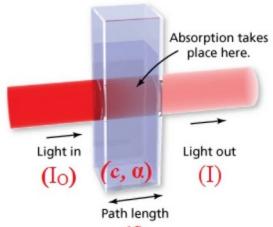
	SSN	M3 Ca	M3 K	M3 Al	M3 P	M3 S	PH	Psa asand	Psa asilt	Psa aclay	...	K	S	Ca	Mg	Cu	Cl	Zn	Fe
0	icr006475	207.1	306.30	1095.000	4.495	18.960	4.682	97.848667	1.845333	0.306000	...	12991.3	45.7	944.1	5575.0	13.0	210.2	22.0	12501.3
1	icr006586	1665.0	1186.00	1165.000	12.510	13.600	7.062	89.520000	9.553667	0.926333	...	15173.5	45.7	9301.0	5519.1	18.4	152.6	38.5	24094.6
2	icr021104	258.7	35.25	441.400	4.424	3.608	5.522	89.950000	5.205000	4.845000	...	6838.9	45.7	884.3	5575.0	2.8	229.5	2.3	2213.4
3	icr033622	11858.3	1156.00	108.286	31.233	25.460	8.583	91.445000	6.310000	2.245000	...	15845.8	45.7	46529.1	33771.2	13.0	58.2	29.8	24135.0
4	icr006570	896.2	607.30	1151.000	5.986	20.080	6.661	97.789667	1.885000	0.325667	...	12201.6	45.7	1790.1	5575.0	7.7	122.5	13.4	9135.1

X<sub>FTIR</sub>

Y<sub>C</sub>

# FTIR et corrélation multilinéaire

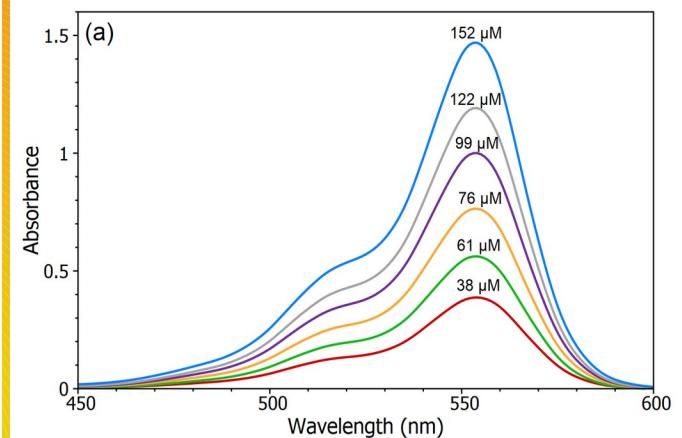
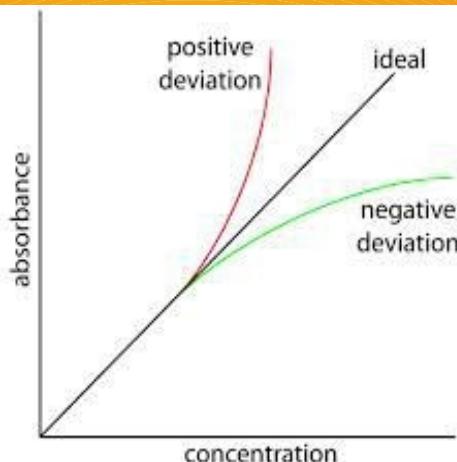
## Beer-Lambert Equation



$$A = \log_{10}(\frac{I_0}{I}) = \epsilon cl$$

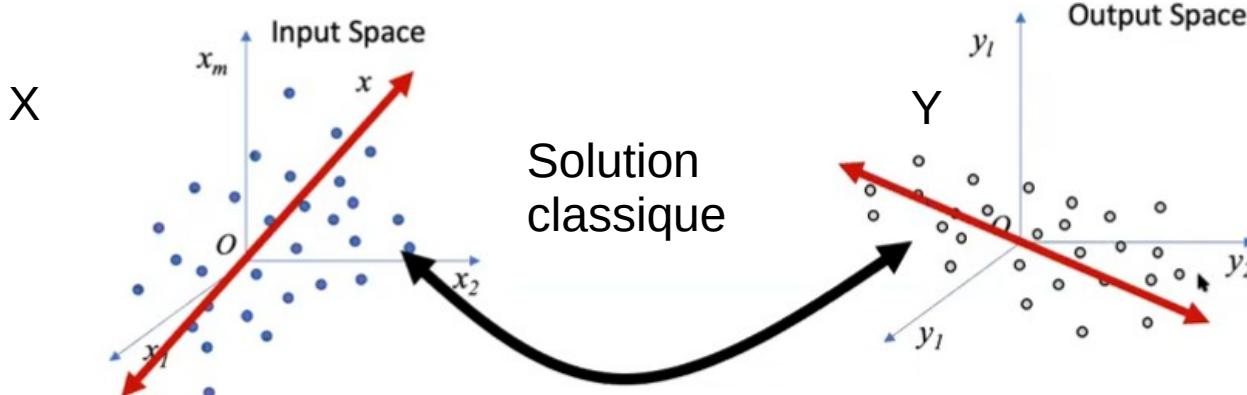
$$A = \epsilon cl$$

La théorie indique un modélisation linéaire pour la spectroscopie IR

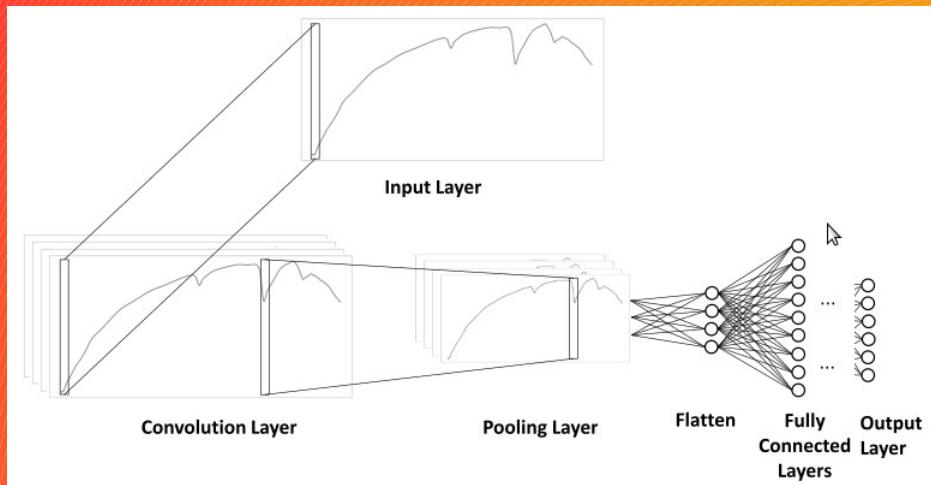


## Régression multivariable avec les moindres carrés partiels

Partial Least Squares Regression is a latent modeling method for predicting a set of outputs in relation to a reduced order inputs. The basic idea is to find a low-dimensional set of input space variables that is most correlated with a given set of output data. It is to analyze data in both input space and output space.

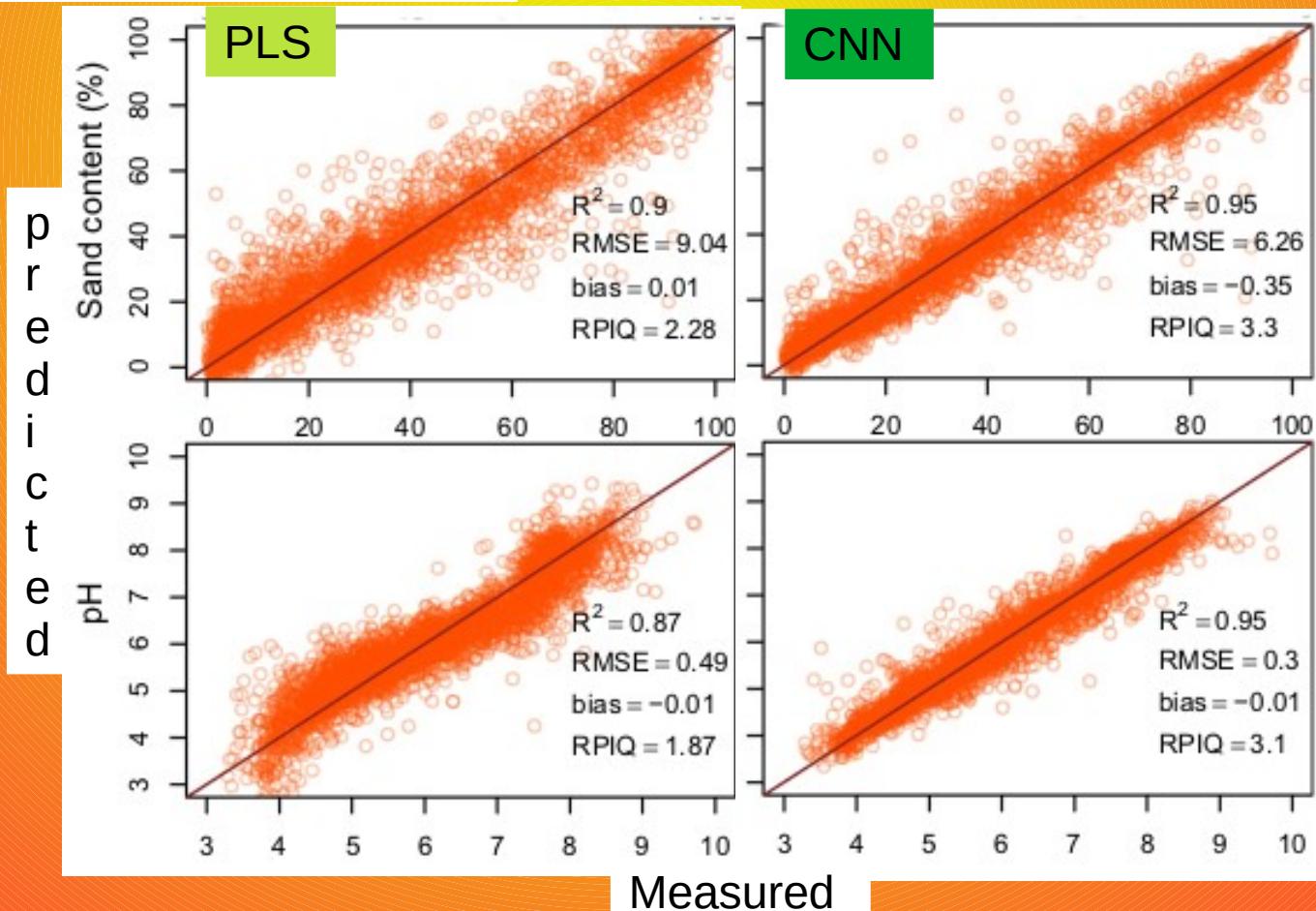


# Modèle amélioré par CNN



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.

Les bons résultats publiés

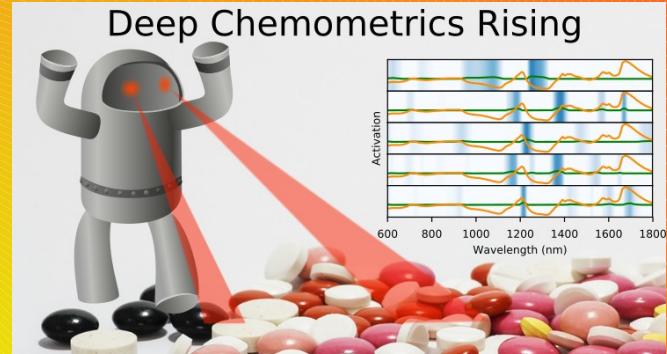


# Réseaux des neurones modifiée

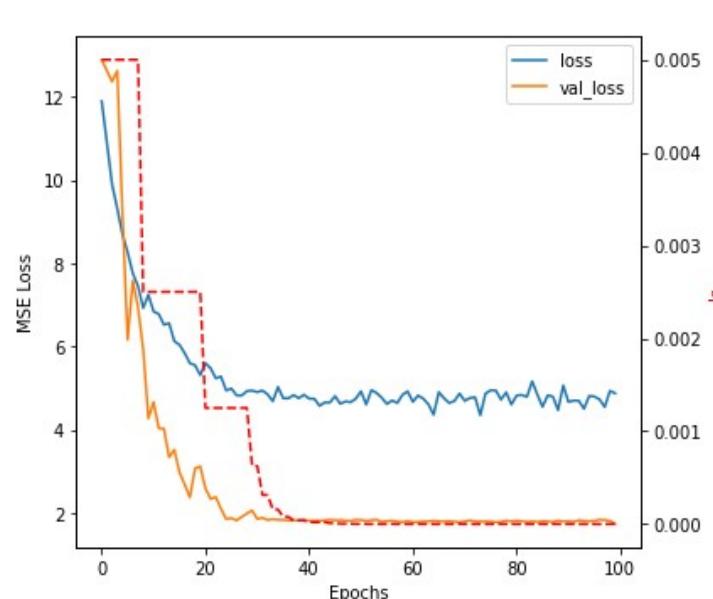
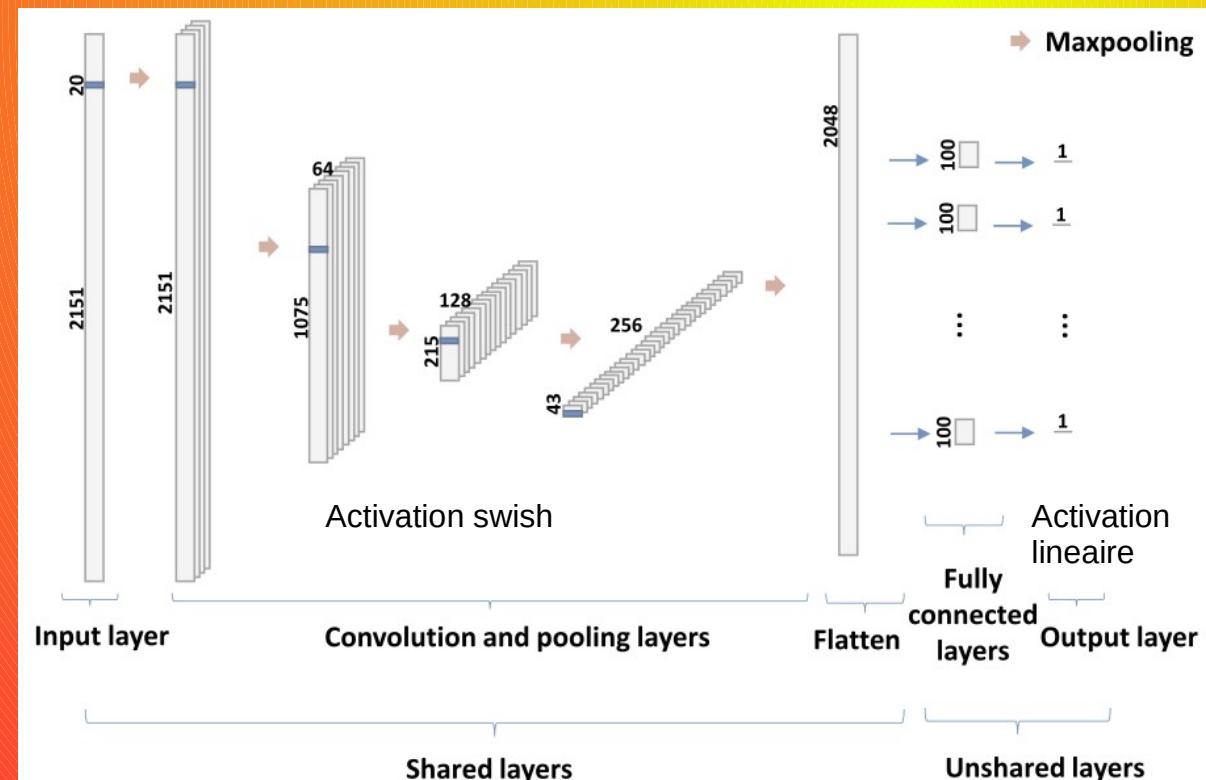
- Augmentation des données selon le modèle

<https://github.com/EBjerrum/Deep-Chemometrics>

- Modification de la fonction d'activation
- Batch size plus petite

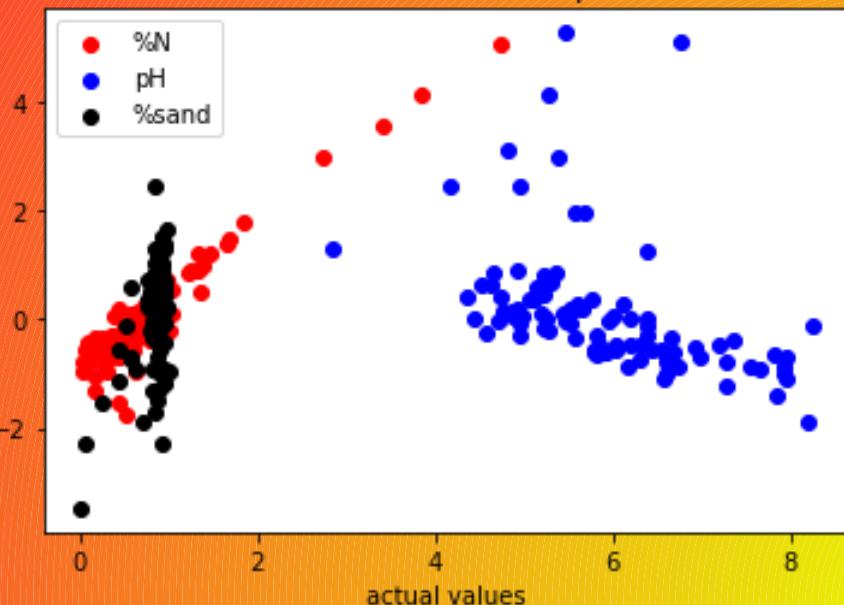


<https://www.wildcardconsulting.dk/useful-information/deep-chemometrics-deep-learning-for-spectroscopy/>



# Modèles et résultats

PLS Canonical - 3 variables prediction

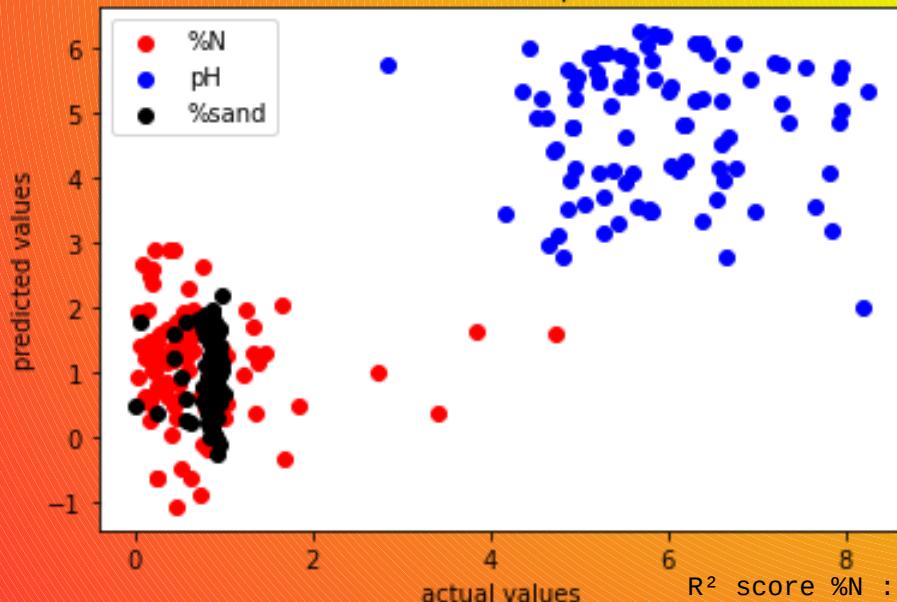


$R^2$  score %N : -0.18514679707524984  
 $R^2$  score pH : -33.07396764367303  
 $R^2$  score %sand : -48.144788339965494

RMSE PLS regression average: 3.610



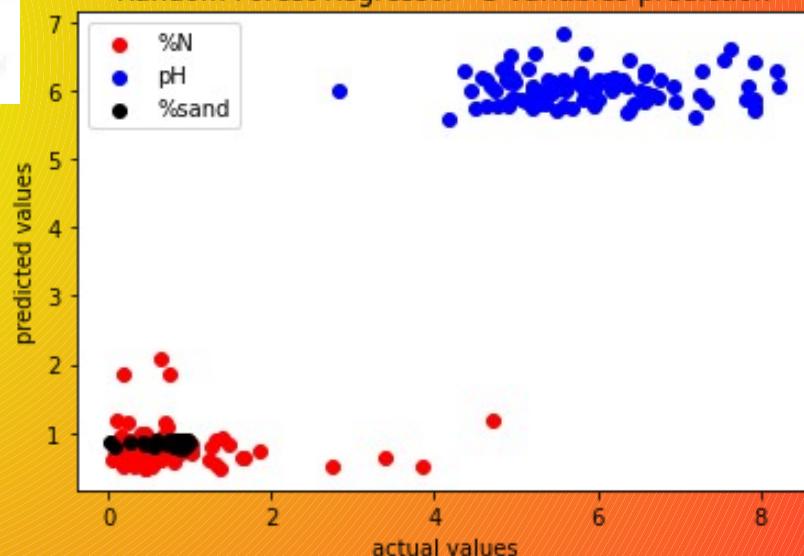
CNN - 3 variables prediction



$R^2$  score %N : -1.613130923301938  
 $R^2$  score pH : -2.108750706582395  
 $R^2$  score %sand : -12.153247920532396

RMSE CNN average: 1.330

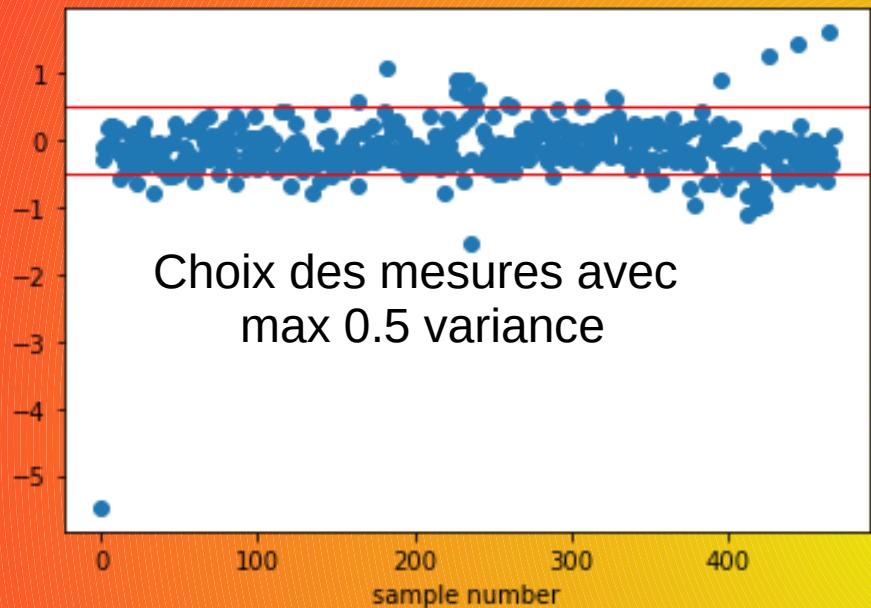
Random Forest Regressor - 3 variables prediction



$R^2$  score %N : -0.14604758535058426  
 $R^2$  score pH : -0.050488510389144814  
 $R^2$  score %sand : -0.01440851542187449

RMSE Random Forest average: 0.780

# Corrélation linéaire ? Vérifier

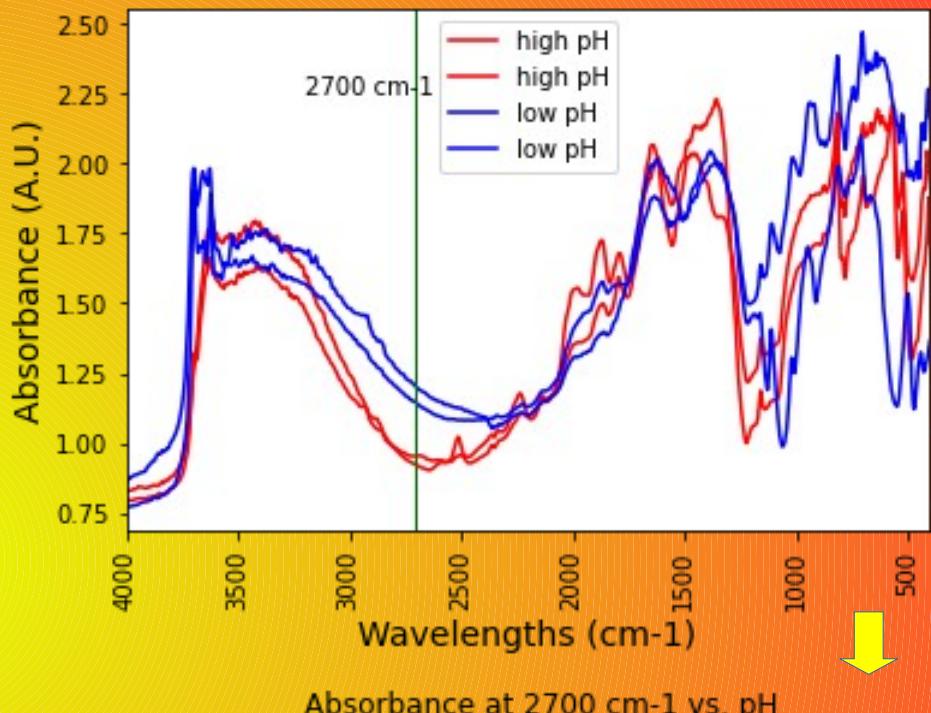


R<sup>2</sup>: 0.075  
RMSE: 1.085

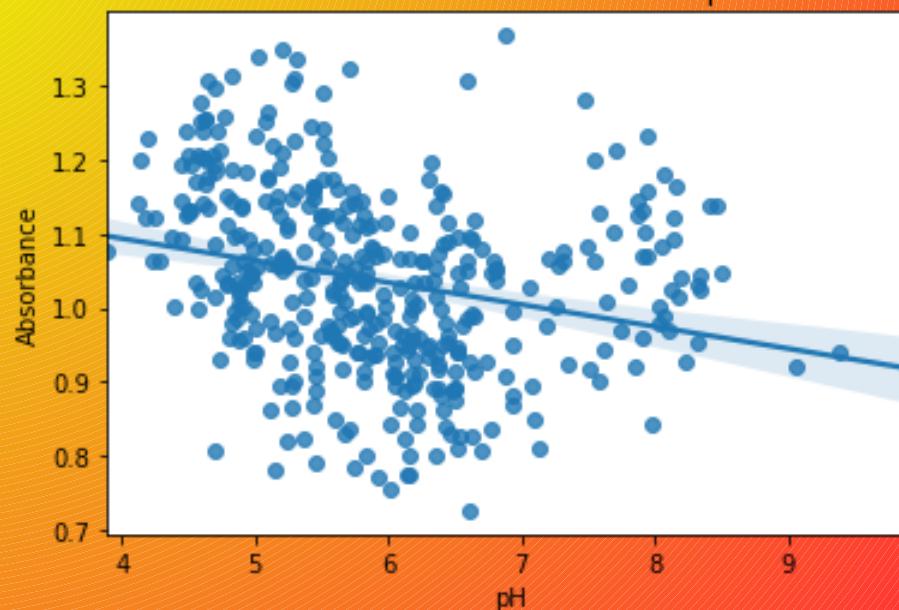
Les données ne satisfaisant pas la condition initiale



FTIR spectra pour échantillons avec valeurs extrêmes de pH



Absorbance at 2700 cm<sup>-1</sup> vs. pH



# Conclusions

- Peu d'échantillons avec analyse complète
- La régression logistique donne une bonne prédiction de la fertilité du sol à partir de 5 variables chimiques importantes
- La modélisation de la chimie du sol par réseaux des neurones ne donne pas des bonnes résultats
- Les données sont trop dispersées pour une corrélation linéaire entre absorbance dans infrarouge et concentration chimique (loi de Lambert Beer)