

# Formation Data Analyst


Projet 5 : Produisez une étude de marché

17/04/2021

Aïssa MOUACHA



# SOMMAIRE

- ❑ Introduction
  - ❑ Data pre-processing
  - ❑ Génération dendrogramme CAH
  - ❑ ACP
  - ❑ Analyse avancée
  - ❑ Test d'adéquation
  - ❑ Test de comparaison
  - ❑ Conclusion
- 

Votre entreprise d'agroalimentaire de poulet souhaite se développer à l'international.



**Stratégie** : exportation dans le(s) nouveau(x) pays ciblé(s).

**Objectif** : cibler certains pays / "groupes" de pays dont on connaît les caractéristiques.

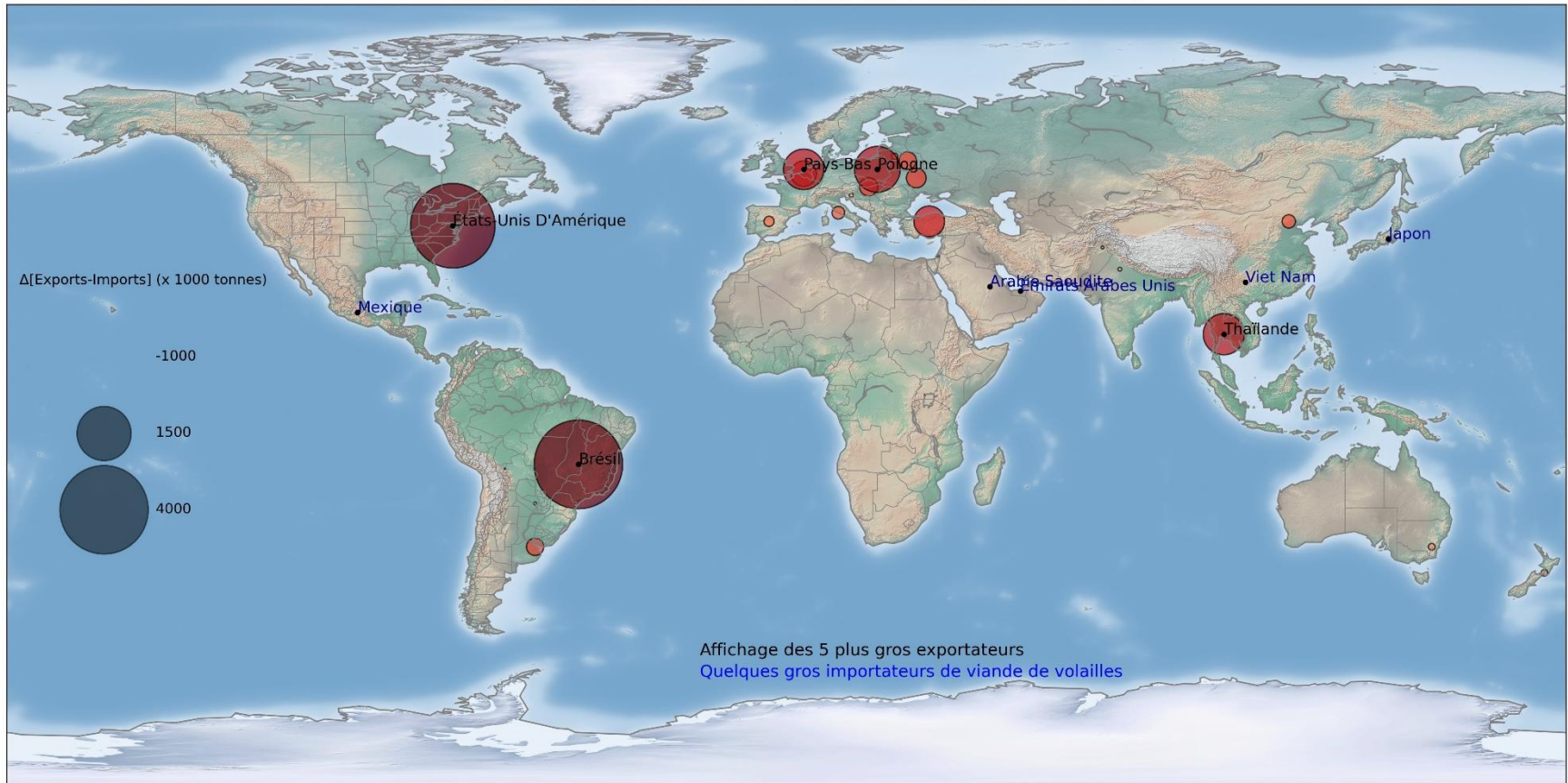
**Source** : site FAO

## Présentation des étapes de pre-processing

- Téléchargement des données FAO
- Calcul de l'évolution population (2014 vs 2018)
- Calcul du ratio protéines animales/total protéines
- Calcul des derniers indicateurs proposés (Dispo protéines/calories)
- Ajout de variables « additionnelles »
- Hypothèse d'indice de Stabilité Politique pour Nv-Calédonie & Poly-française:
  - imputation par la moyenne des îles voisines
- Hypothèse de revenu national brut pour Taïwan:
  - les données des variables retenues sont relativement proches pour les 3 RAS/province chinoises : RNB Taïwan = moy (Honk-Kong + Macao)

# Balance commerciale mondiale

## Balance commerciale : Viande de Volailles en 2018

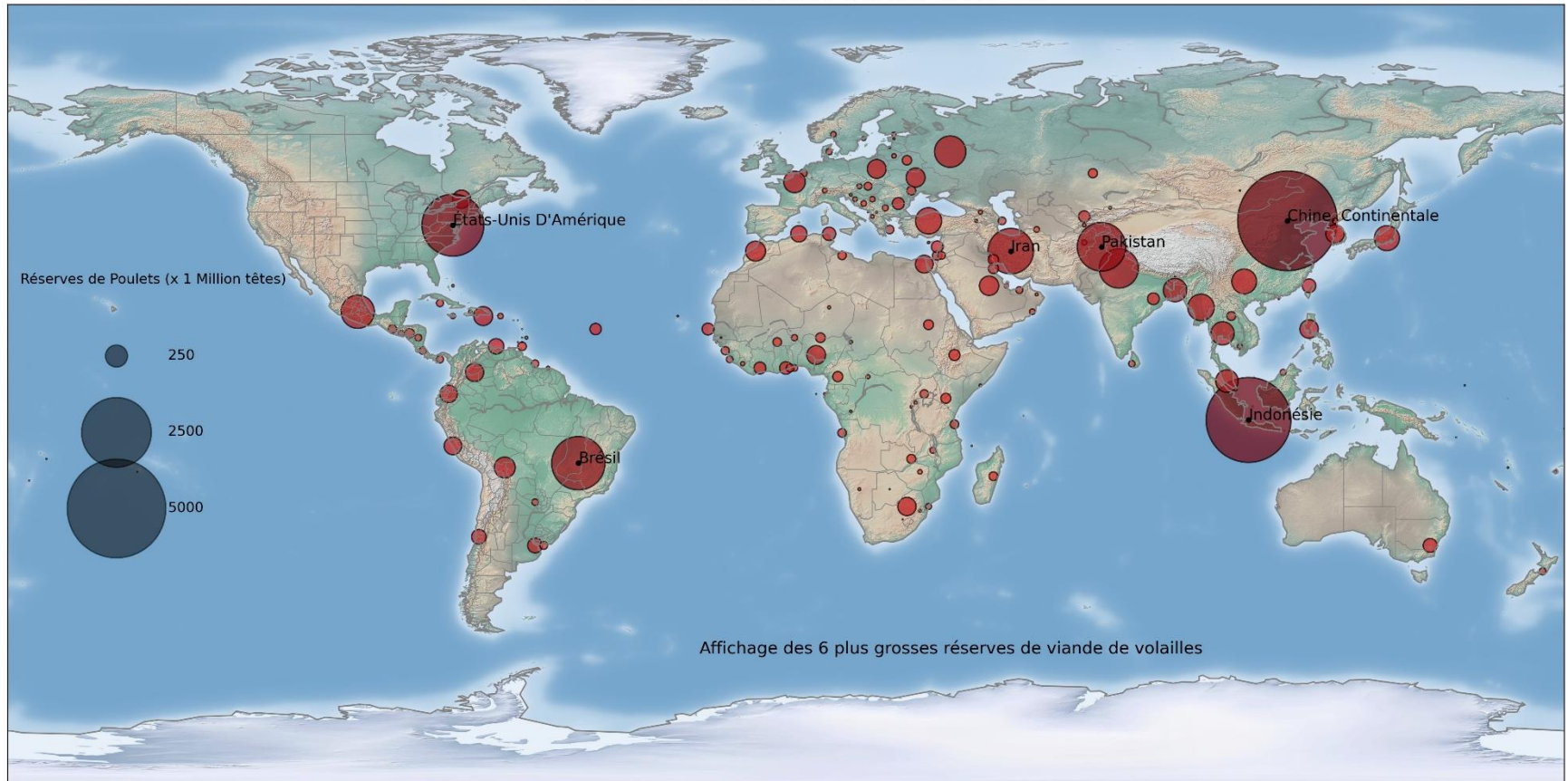


- Identification des 2 très/plus gros exportateurs
- Des pays comme le Mexique, l'Arabie Saoudite, le Japon présente une BC déficitaire (importations >> exportations)
- En Europe : Pays-Bas et Pologne sont de forts exportateurs



# Réserve mondiale de poulets

Réserve mondiale de Poulets en 2018



- Pour les réserves, on retrouve USA et Brésil, mais aussi Chine, Indonésie ...
- Peu de réserves en Afrique, Europe centrale et nordique

## Présentation des étapes de pre-processing

- Dans l'analyse qui va suivre, j'ai choisi de n'exclure aucun pays parmi ceux présentés précédemment, de sorte à dénaturer le moins possible le dataset et dérouler le process de clustering avec toutes les parties.
- Dans un autre contexte, la pré-étude que j'aurais menée aurait conduit à écarter le maximum de pays ne présentant pas les caractéristiques adéquates pour répondre à notre objectif.
- Éléments clés à prendre en compte pour la faisabilité de l'étude :
  - Tenir compte des **politiques commerciales** :
    - libéralisation des échanges (facilitée en Europe)
    - **taxes & quotas** (privilégier la proximité)
    - **impact environnemental** (réduire empreinte carbone)
  - Sonder le marché **concurrentiel** (gros exportateur/producteur de poulets à écarter)
  - Tenir compte de la richesse intérieure/**pouvoir d'achat** par habitant
  - Tenir compte de la **démographie**

# Présentation des variables : 1<sup>er</sup> clustering

- 4 variables proposées sont calculées
- Choix d'ajouter 3 autres variables (issues du site FAO)

4 variables proposées + 3 variables

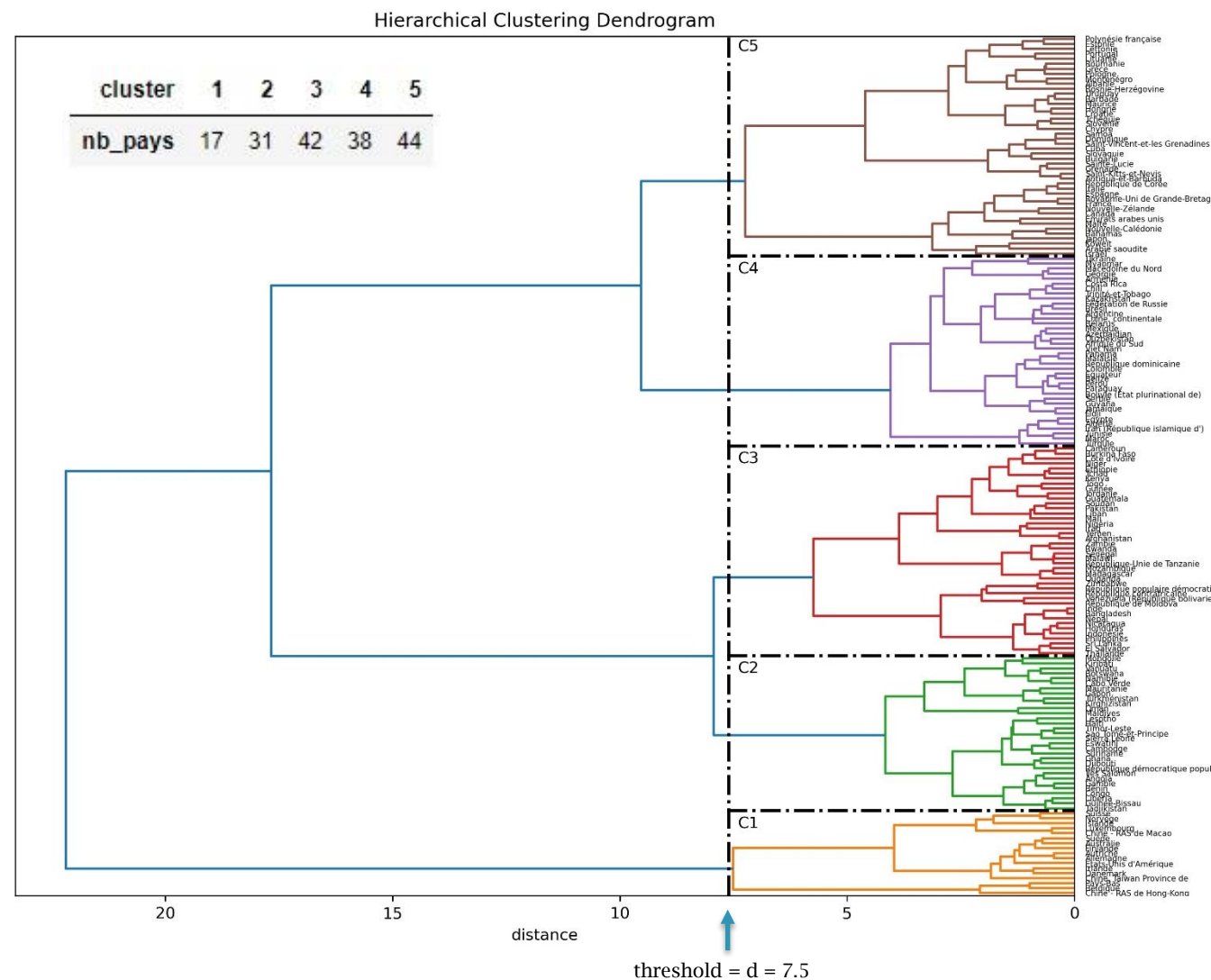
Zone	evo_pop_18_v_14 (%)	Proportion de protéines d'origine animale (%)	Disponibilité alimentaire en protéines (kg/hbt)	Disponibilité alimentaire en calories (Kcal/hbt)	Stabilité politique	Tx_depd_importations (%)	Revenu National Brut/hbt
Australie	5.5	66.7	38.7	1237715	1.0	1.3	56682.7
Belgique	2.3	58.7	36.4	1375685	0.4	226.1	47529.1
Chine - RAS de Hong-Kong	3.3	73.1	46.9	1192820	0.8	343.5	51252.6
Chine - RAS de Macao	7.1	64.8	40.0	1214720	1.3	86.7	79270.6
Chine, Taiwan Province de	1.0	50.6	31.5	1088795	0.8	25.1	65262.0

$$TDI = \frac{\text{Importations}}{\text{Production} + \text{Importations} - \text{Exportations}} \times 100$$



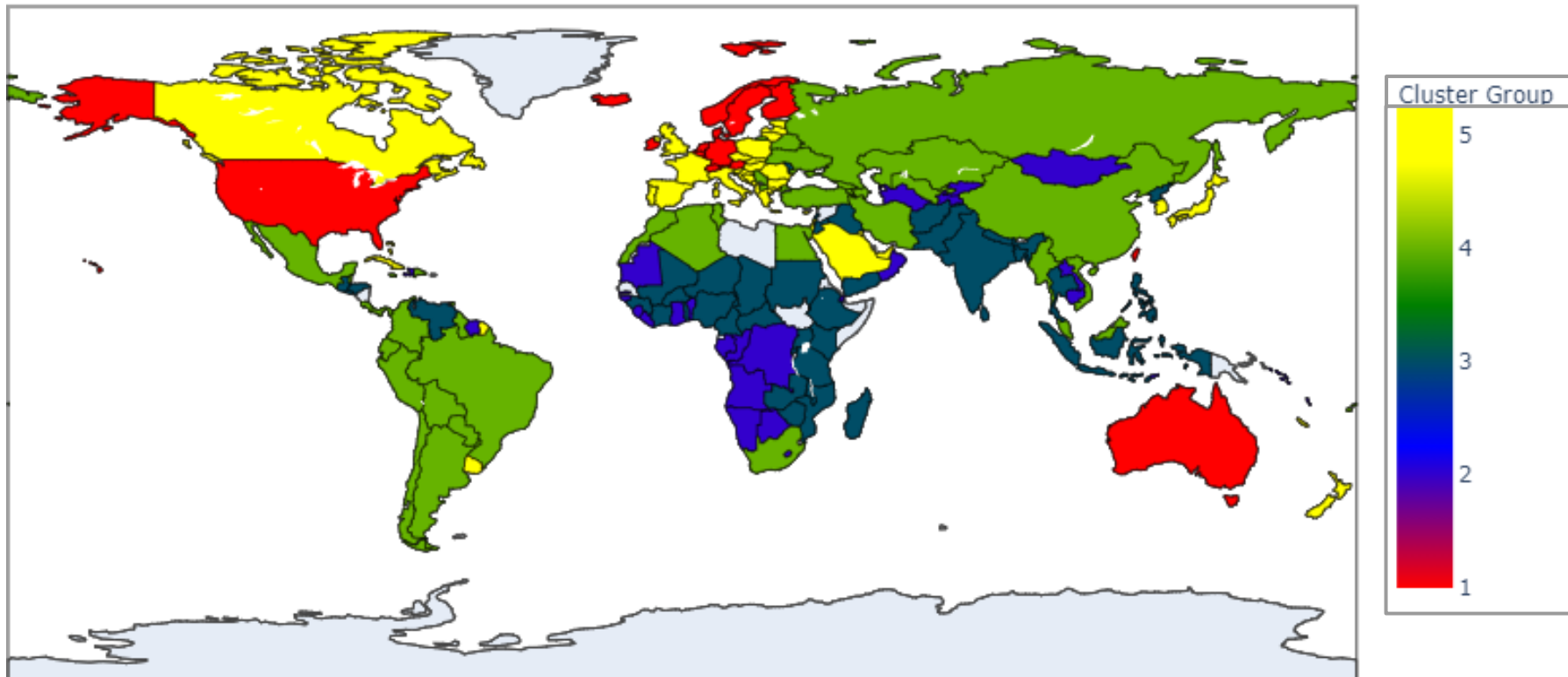
# Génération du dendrogramme

- Découpage en 5 clusters et association cluster/pays



## Visualisation géographique des 5 clusters

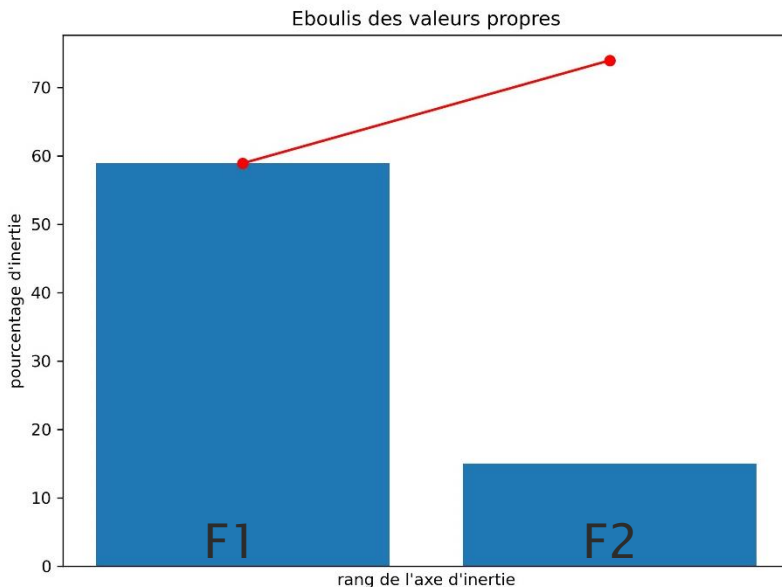
- Choropleth : groupement des pays basé sur le clustering



```
data = [dict(type='choropleth', locations = Pays['Area'], text= Pays['Area'],
            locationmode = 'country names', # 'ISO-3', 'USA-states', 'country names', 'geojson-id'
            z = Pays['Cluster'], colorbar = {'title': 'Cluster Group'}, marker_line_color='black', marker_line_width=0.5,
            zauto=False, autocolorscale=False, reversescale=False, colorscale = my_colors, zmin=1, zmax=6, showscale=True)]
layout = dict(title='Grouping of Countries based on Clustering',
            geo=dict({'scope': 'world'}, showframe = True, projection = {'type': 'equiangular'}, showcountries = True))
map3 = dict(data=data, layout=layout)
```

# Eboulis des valeurs propres

- Variance totale 1<sup>er</sup> plan factoriel : 75%



```
# Choix du nombre de composantes à calculer
n_comp = 2 # Nbmaxi = Min(p,n-1) = Min(4,172-1)

# selection des colonnes à prendre en compte dans l'ACP
data_pca = table_vf

# préparation des données pour l'ACP
X = data_pca.values
names = list(data_pca.index)
features = data_pca.columns

# Centrage et Réduction
std_scale = preprocessing.StandardScaler().fit(X) #Compute the mean and std to be used for later scaling
std_scale = preprocessing.RobustScaler().fit(X) #Removes the median and scales the data according to the quantile range
X_scaled = std_scale.transform(X)

# Calcul des composantes principales
pca = decomposition.PCA(n_components=n_comp)
pca.fit(X_scaled)

# Projection des individus
X_projected = pca.transform(X_scaled)

# Eboulis des valeurs propres
display_scree_plot(pca)

# Cercle des corrélations
pcs = pca.components_

# Adaptation de la fonction pr obtention 2 infos sur le même tracé (ajout scale factor pour les corrélations)
display_circles(X_projected, pcs, n_comp, pca, [(0,1)], labels = np.array(features), illustrative_var=clusters, alpha=1)

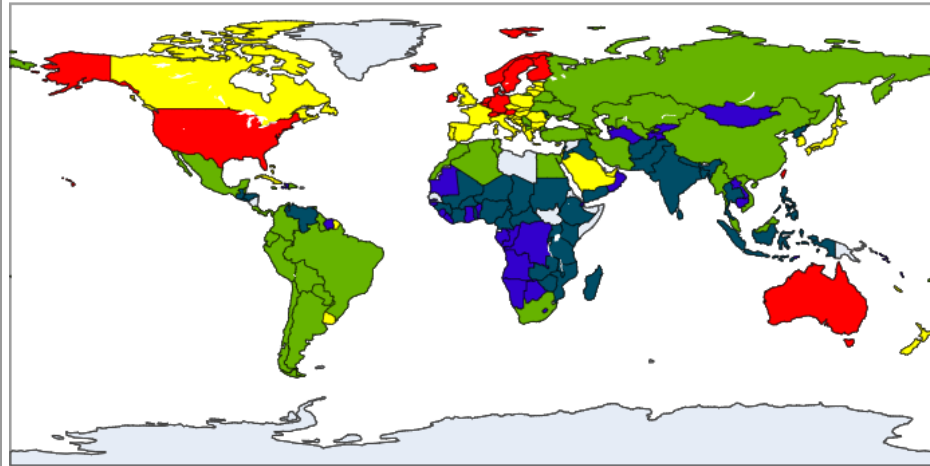
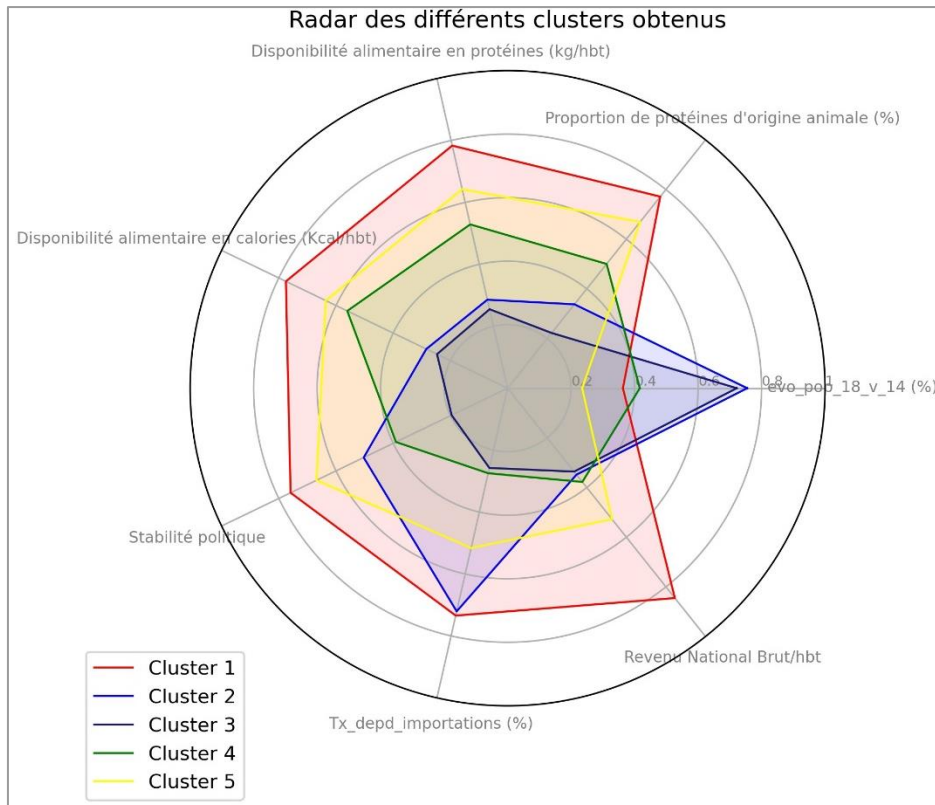
# Projection des individus
display_factorial_planes(X_projected, n_comp, pca, [(0,1)], labels=np.array(names), alpha=1, illustrative_var=clusters)
```

Définies dans le fichier **functions.py**



# Radar Clustering

- La caractérisation **moyenne** des différents clusters obtenus est présentée ci-dessous



- «Bons et mauvais clients» apparaissent ici justifiant le choix du cluster qui a été fait.

# Raffinement de l'analyse sur la base des 17 pays choisis

- Rappel : réduction empreinte carbone et libéralisation échanges facilitée
- Considération d'une nouvelle variable « Distance »

```
# ajout de variables supplémentaires
url_distance = 'INPUTS_FORMATION/P5_csv_Distances.csv'
dist = pd.read_csv(url_distance, sep=';', encoding='utf-8')
distance = dist.copy()
distance = distance.dropna()
table_vf = pd.merge(table_vf,distance, how="left", on="Zone")
```

Zone	Distance (km)
Afghanistan	5590
Afrique du Sud	9354
Albanie	1604
Algérie	1340
Allemagne	440

4 variables proposées

+

3

+

1

Zone	evo_pop_18_v_14 (%)	Proportion de protéines d'origine animale (%)	Disponibilité alimentaire en protéines (kg/hbt)	Disponibilité alimentaire en calories (Kcal/hbt)	Stabilité politique	Tx_dep_d_importations (%)	Revenu National Brut/hbt	Distance (km)
Australie	5.5	66.7	38.7	1237715	1.0	1.3	56682.7	16975
Belgique	2.3	58.7	36.4	1375685	0.4	226.1	47529.1	262
Chine - RAS de Hong-Kong	3.3	73.1	46.9	1192820	0.8	343.5	51252.6	9639
Chine - RAS de Macao	7.1	64.8	40.0	1214720	1.3	86.7	79270.6	9602
Chine, Taiwan Province de	1.0	50.6	31.5	1088795	0.8	25.1	65262.0	9911

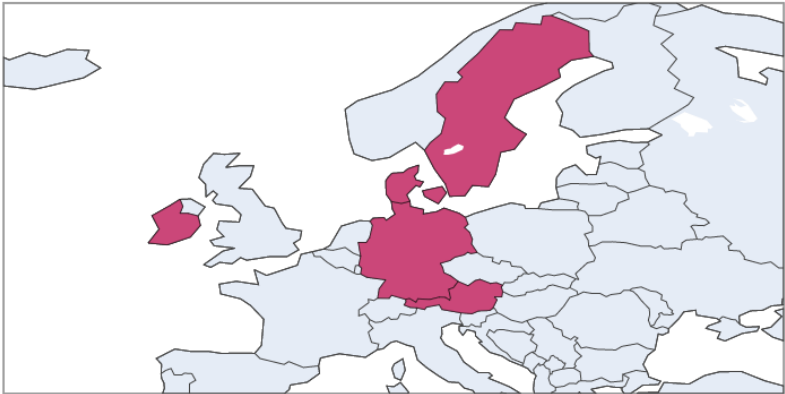


# Priorisation et stratégie finale

- nlargest (10, [«TDI (%)»])
  - nsmallest ( 8, [«Distance (km)»])
  - nsmallest ( 6, [«D Exports-Imports»])
  - marché > 4 Millions

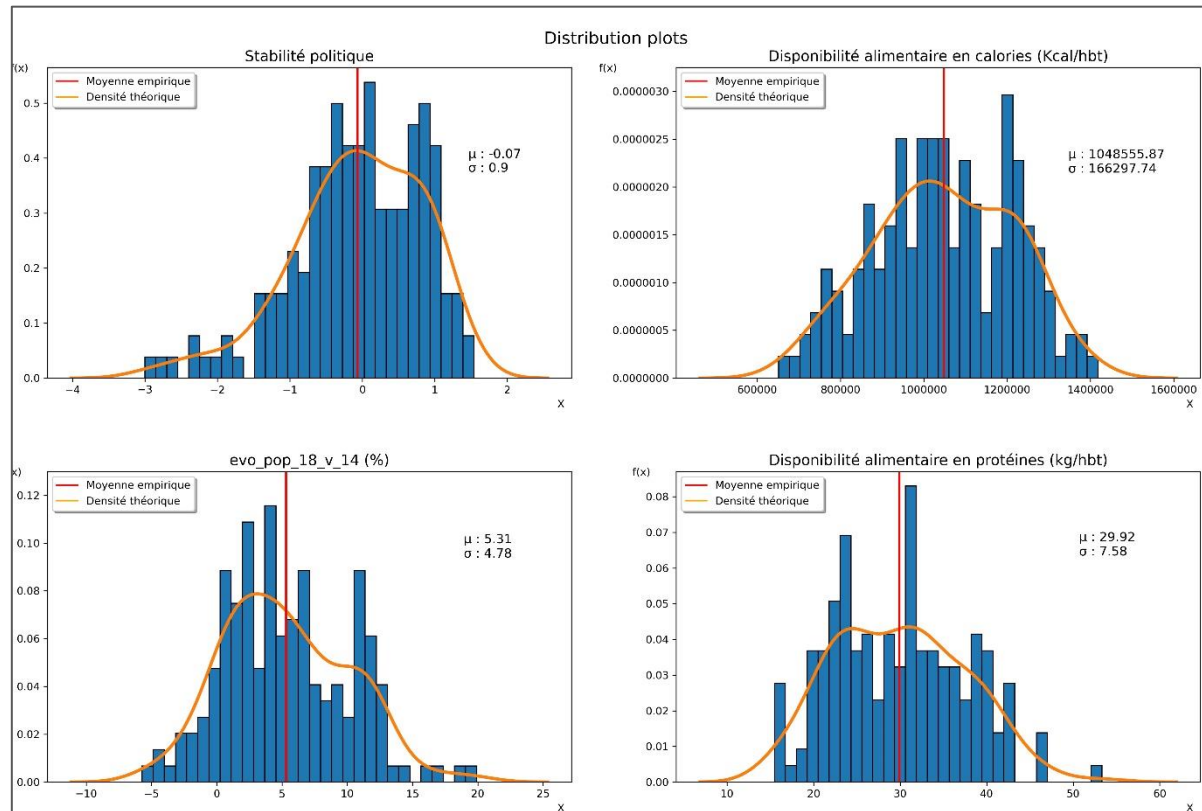


	Cluster	evo_pop_18_v_14 (%)	Proportion de protéines d'origine animale (%)	Disponibilité alimentaire en protéines (kg/hbt)	Disponibilité alimentaire en calories (Kcal/hbt)	Stabilité politique	Tx_depd_importations (%)	Revenu National Brut/hbt	Distance (km)	D Exports-Imports	Population_18 (x1000)
Zone											
Allemagne	5	2.06	60.75	38.47	1297210	0.60	48.34	49335.79	440	-231.0	83124.0
Suède	5	2.89	63.99	38.53	1162160	0.91	37.50	56569.57	1546	-55.0	9972.0
Autriche	5	3.20	60.33	39.83	1348675	0.92	62.30	51904.41	1035	-31.0	8891.0
Irlande	5	4.15	61.24	42.92	1418025	1.03	63.64	62488.32	778	-15.0	4819.0
Danemark	5	1.55	68.10	42.73	1241365	0.96	90.97	62635.09	1028	1.0	5752.0



# Normalité des distribution : Distributions de Gauss

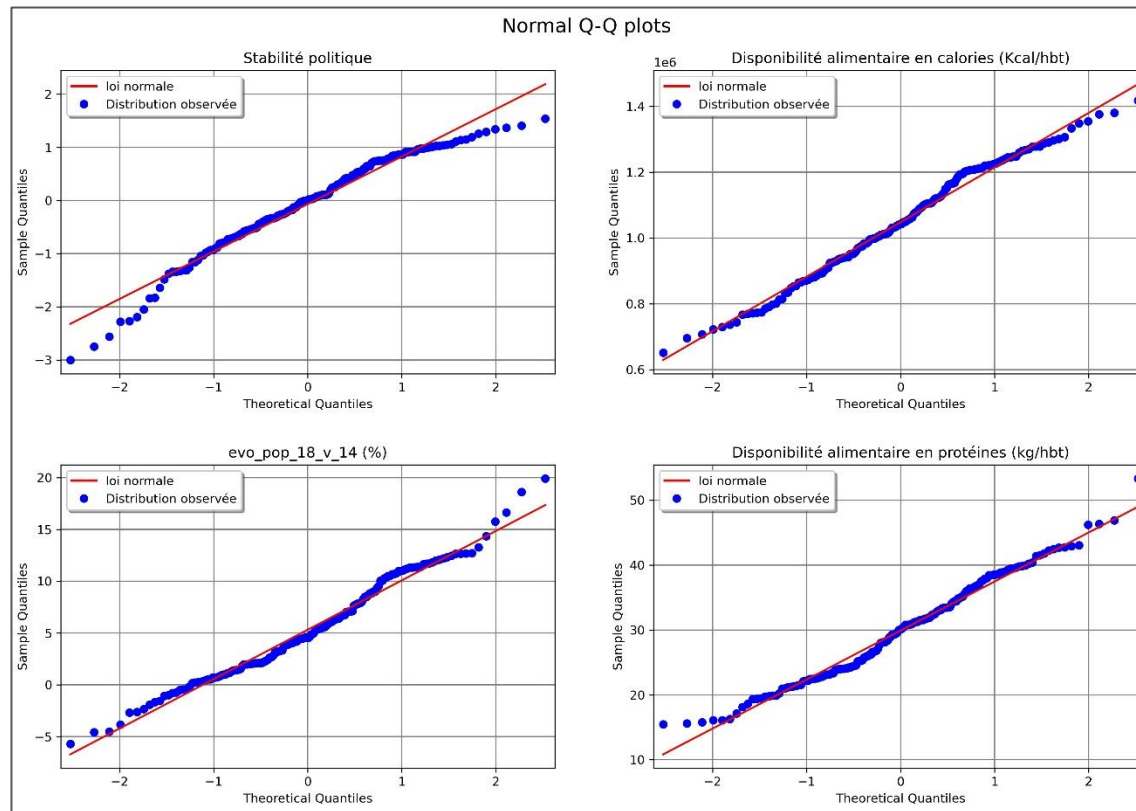
- Pour les graphiques et tests qui vont suivre, conservation de 3 variables (proposées) et 1 de mon choix



- Les distributions paraissent normales, à l'exception peut-être de la variable « Stabilité politique » ➔ **A priori à confirmer/infirmar**

# Normalité des distribution : Quantile-Quantile plots

- On retrouve cet « a priori » dans les tracés de quantiles



# Normalité des distribution : tests statistiques

- Utilisation des 2 tests les plus communs

```
from scipy.stats import ks_2samp
```

	Stabilité politique	Disponibilité alimentaire en calories (Kcal/hbt)	evo_pop_18_v_14 (%)	Disponibilité alimentaire en protéines (kg/hbt)
D_statistic	0.0788372	0.0654419	0.066814	0.0937674
p_value	0.30323	0.531472	0.504903	0.141031
Normality	normal	normal	normal	normal

La variable a une distribution normale (Ho)

But : ne pas rejeter Ho

On cherche à limiter le risque de se tromper en obtenant une  $p\_value >> 0.05$

➡ Le test de Kolmogorov-Smirnov valide pour les 4 variables le caractère normal de leur distribution

```
from scipy import stats as st
```

	Stabilité politique	Disponibilité alimentaire en calories (Kcal/hbt)	evo_pop_18_v_14 (%)	Disponibilité alimentaire en protéines (kg/hbt)
W_statistic	0.96219	0.985949	0.982063	0.982375
p_value	0.000130426	0.0820675	0.0256593	0.0281526
Normality	#	normal	#	#

➡ Le test de Shapiro-Wilk le réfute pour 3 variables

- On retiendra donc que seule la variable « Dispo alim en calorie » suit une loi normale.

## Homoscédasticité (test d'égalité des variances)

- Utilisation des 2 tests les plus communs sur les populations :

cluster	1	2	3	4	5
nb_pays	17	31	42	38	44

```
from scipy.stats import bartlett
```

	Stabilité politique	Disponibilité alimentaire en calories (Kcal/hbt)	evo_pop_18_v_14 (%)	Disponibilité alimentaire en protéines (kg/hbt)
Statistic	8.2647	0.186381	1.26172	0.45788
p_value	0.00404234	0.665946	0.261326	0.498616
Variance	diff	equal	equal	equal

➡ Le test de Bartlett montre que l'homoscédasticité n'est pas vérifiée pour la variable « stabilité politique »

```
from scipy.stats import levene
```

	Stabilité politique	Disponibilité alimentaire en calories (Kcal/hbt)	evo_pop_18_v_14 (%)	Disponibilité alimentaire en protéines (kg/hbt)
Statistic	5.24943	0.236566	2.04429	0.0655585
p_value	0.0259578	0.628703	0.158647	0.798908
Variance	diff	equal	equal	equal

➡ Le test de Levene montre que l'homoscédasticité n'est pas vérifiée pour la variable « stabilité politique »

- On retiendra donc que l'homoscédasticité entre ces 2 populations n'est pas validée pour la variable « stabilité politique ».

Invalidation des Hypothèses de :

- 1) Normalité (*plusieurs variables dont la loi n'est pas normale*)
- 2) Homoscédasticité (*au moins 1 variable montrant une hétéroscédasticité*)

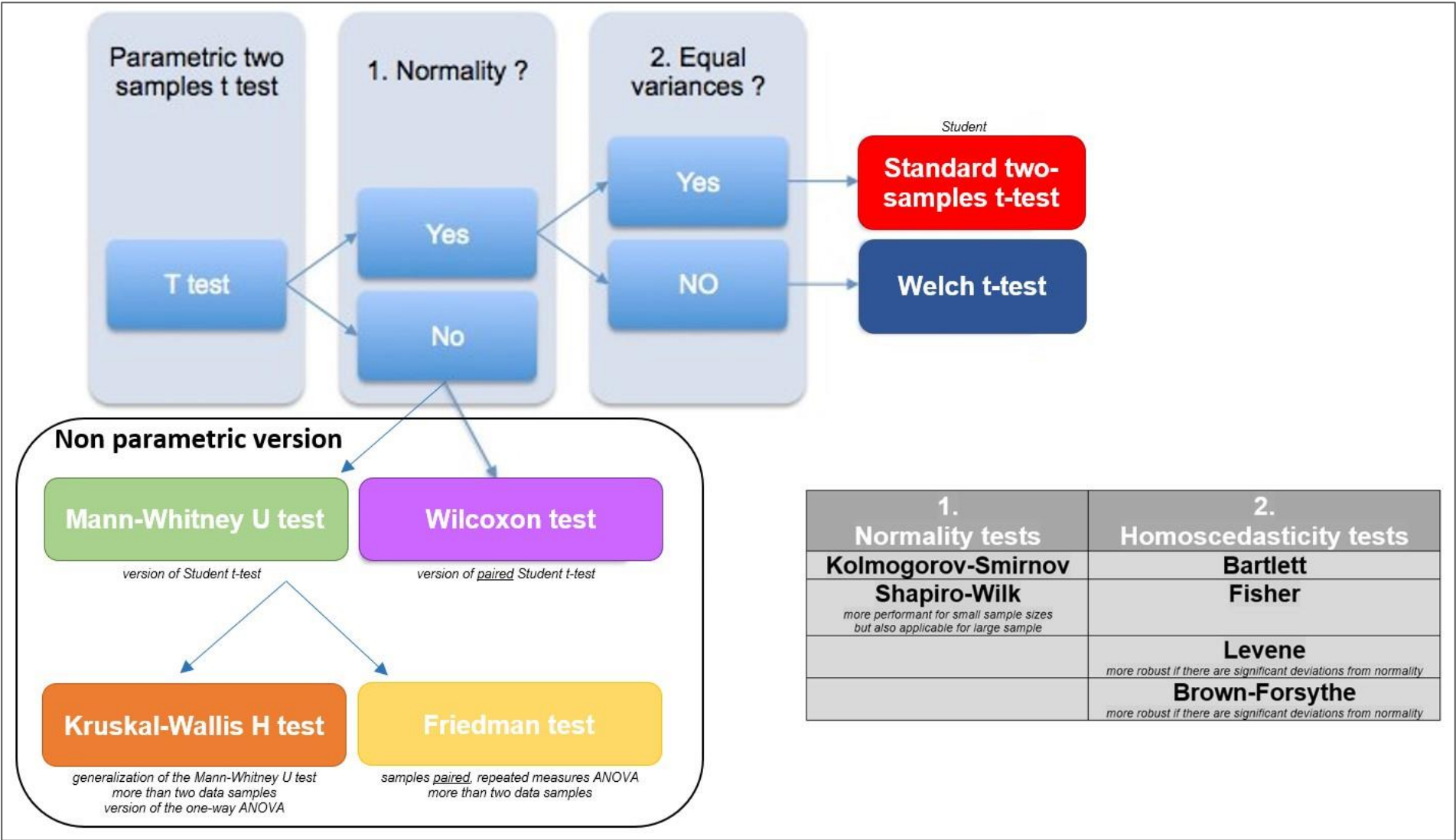
➔ On ne peut donc conclure quant à la différence entre les clusters

Des tests statistiques non-paramétriques tels que les tests de :

- Mann-Whitney U
- et plus généralement Kruskal Wallis H,

seraient plus à même de tester les différences significatives entre ces 2 clusters.





Merci

