Produisez une étude de marché

Internationalisation des ventes

Frédéric Gainza – Avril 2021

Rappel de la mission

Contexte de l'étude : Volonté de se développer à l'international

→ exporter nos produits dans un premier temps

Objectif de l'étude : Réaliser une première recherche des pays susceptibles de devenir des marchés d'exportation

→ déterminer un groupe de pays candidats, aux caractéristiques propres

Sommaire

1. Déterminer les critères de choix des pays candidats → *Déterminer les variables*

2. Réaliser un partitionnement → *Classification hiérarchique*

3. Visualiser le partitionnement → *Analyse en Composantes Principales*

4. Vérification de la qualité du partitionnement → *Tests statistiques*

01

Détermination des variables

Critères de choix des pays candidats

1. Les variables actives

Variables actives : Variables qui vont servir à créer le partitionnement

Variables alimentaires *(année de référence 2018)* :

Source : FAOSTAT

- proportion de protéines d'origine animale dans la disponibilité alimentaire du pays : part_prot_anim
- disponibilité alimentaire en protéines par jour par habitant : dispo_prot_g_p_j
- disponibilité alimentaire en calories par jour par habitant : dispo_alim_kcal_j_pers

	code_pays	pays	annee	dispo_alim_kcal_j_pers	dispo_prot_g_p_j	part_prot_anim
0	1	Arménie	2018	2997.0	94.35	48.050021
1	2	Afghanistan	2018	2040.0	55.52	19.434438
2	3	Albanie	2018	3360.0	115.74	53.347732

On vérifie présence de valeurs nulles ou en doubles

```
1 df_mini.isna().sum()
2 df_mini.duplicated().sum()
```

```
code_pays              0
pays                   0
annee                  0
dispo_alim_kcal_j_pers 0
dispo_prot_g_p_j       0
part_prot_anim         0
dtype: int64

0
```

1. Les variables actives

Variable démographique :

Sources : FAOSTAT et Banque Mondiale

- évolution (en pourcentage) de la population entre 2008 et 2018 : evol_pop_pct

Vérification de la pop mondiale

```
1 pop_mondiale=pop['pop_2018'].sum()
2 pop_mondiale
```

```
9090746.135
```

Somme trop importante

	code_pays	pays	evol_pop_pct
35	41	Chine, continentale	5.472811
81	96	Chine - RAS de Hong-Kong	7.118232
112	128	Chine - RAS de Macao	22.590875
191	214	Chine, Taiwan Province de	3.073173
237	351	Chine	5.447454

Chine comptée 2 fois

On supprime *code_pays* 351

```
pop=pop[pop['code_pays'] != 351]
```

Dataframe de nos variables actives :

	code_pays	pays	dispo_kcal	dispo_prot	part_prot_anim	evol_pop_pct
40	60	El Salvador	2696.0000	77.0300	34.9344	4.7129
43	67	Finlande	3343.0000	117.9900	63.9237	3.8186
110	158	Niger	2601.0000	82.6000	14.1283	47.1573
95	136	Mauritanie	2877.0000	82.1200	36.1179	33.5860
19	32	Cameroun	2733.0000	71.8900	15.6628	30.9755

2. Les variables illustratives

Variables illustratives : Variables qui seront lues par le modèle mais sans participer à sa définition

↳ elles ne sont utilisées que pour la description des classes

- le continent auquel il appartient (variable catégorielle)
- le PIB/habitant
  - exprimé en $ PPA internationaux constants de 2011
- le niveau de sécurité (sécurité financière)
  - données issues du site de la COFACE
  - note qui rend compte du risque d'impayé (note transformée en donnée numérique)
- la part de consommation de volaille dans la consommation totale
  - exprimée en kg par personne et par an
- la part des importations nettes
  - part_imp_vol = (Importations / (Production + Exportations + Variation des Stocks)) * 100
- le niveau d'imposition
  - taux d'imposition des bénéfices commerciaux

3. Le dataframe global

Remarques :

- Suppression de la France de la liste des pays
- Absence de données pour certains pays :

  ↳ 8 pays pour le *PIB/hab* et 7 pays pour le *taux d'imposition* ⟹ Pays exclus de l'analyse

⇨ L'analyse va porter sur 156 pays au total

	code_pays	pays	code_cont	continent	pop_M	evol_pop	pib_hab	risque	taux_impot	dispo_kcal	dispo_prot	part_prot_anim	qte_volaille	vol_part_import
123	189	Sainte-Lucie	3	Ameriques	181.8890	6.9866	15261.5000	5	34.7000	2618.0000	85.8200	62.3747	60.4764	500.0000
77	118	Koweït	0	Asie	4137.3120	55.7718	50478.6000	3	13.0000	3471.0000	103.2100	49.3847	47.8572	210.0000
86	130	Malawi	2	Afrique	18143.2170	32.1632	1042.5000	6	34.5000	2392.0000	67.0700	13.8959	3.3621	0.0000
25	40	Chili	3	Ameriques	18729.1600	12.0952	24258.7000	3	34.0000	3029.0000	91.6400	54.1248	40.8454	17.4100
37	58	Équateur	3	Ameriques	17084.3580	17.5335	11561.8000	6	32.3000	2606.0000	65.9600	46.5818	20.4866	0.0000
32	53	Bénin	2	Afrique	11485.0440	32.0587	3160.8000	4	48.9000	2755.0000	63.9700	19.4622	9.9260	354.5500
27	44	Colombie	3	Ameriques	49661.0480	12.2157	14455.6000	4	71.9000	3114.0000	72.9200	51.1657	34.6147	5.5200
132	203	Espagne	1	Europe	46692.8580	1.3546	40328.9000	2	47.0000	3322.0000	107.5500	62.0456	30.4758	9.4700

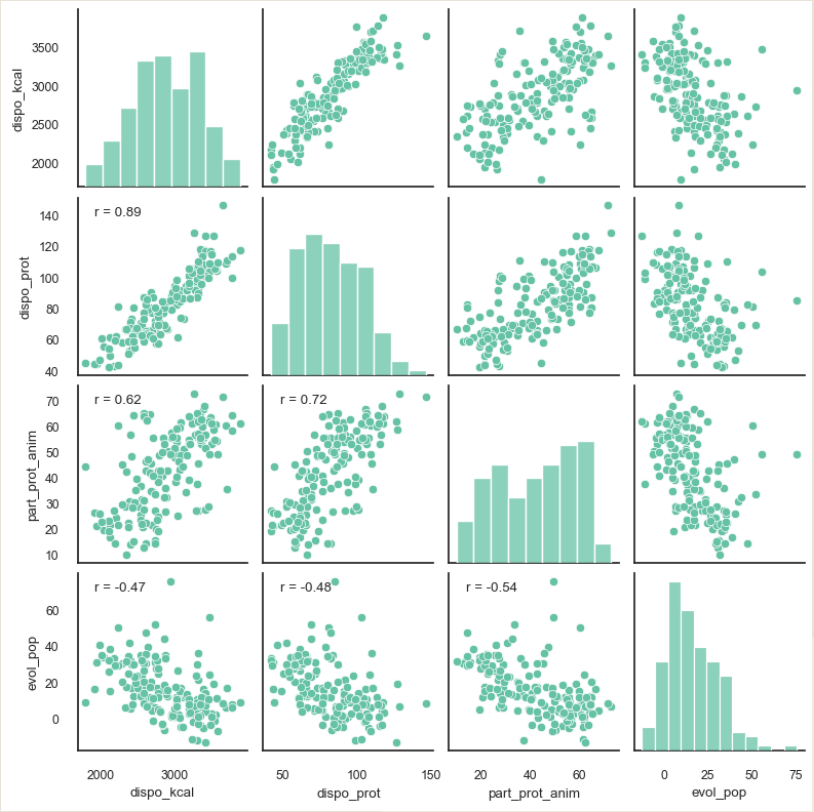
02

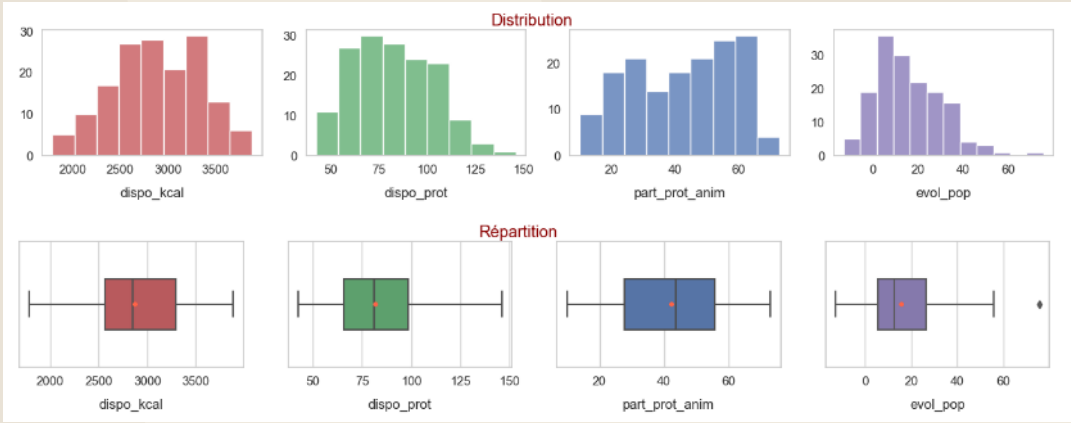
Partitionnement

Détermination des clusters

1. Approche graphique

Avant de réaliser la classification hiérarchique, on représente graphiquement les 4 variables actives





Graphiques de répartition :

- Globalement pas de valeurs aberrantes (1 seul outlier au total)

Graphique d'analyse des corrélations 2 à 2 :

- Corrélations positives entre la 'dispo_kcal' et la 'dispo_prot' (r = 0,89), et dans une moindre mesure entre la 'dispo_prot' et la 'part_prot_anim' (r = 0,72)

1. Classification Hiérarchique Ascendante (CHA)

On va réaliser une classification automatique

↳ cf. une classification d'individus décrits par des données quantitatives, à l'aide d'approches géométriques utilisant les distances

On se place dans un ensemble : - composé de n individus (ici n = 156)

- décrits par p variables quantitatives (ici p = 4)

Principe :

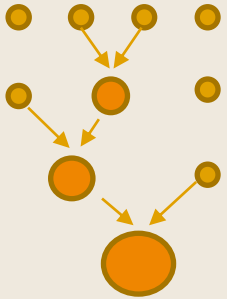
- on considère chaque pays comme un cluster
- puis à chaque itération du modèle, on regroupe les pays selon une certaine méthode
- on continue à itérer jusqu'à que tous les pays soient regroupés en un seul groupe

Étape 1 : Initialisation → 4 clusters

Étape 2 : 1er regroupement → 3 clusters

Étape 3 : 2eme regroupement → 2 clusters

Étape 4 : 3eme regroupement → 1 cluster



1. Classification Hiérarchique Ascendante (CHA)

source : Ricco Rakotomalala

Nécessité de déterminer 2 paramètres liés aux distances

La métrique utilisée

↳ cf. *la distance entre 2 individus*

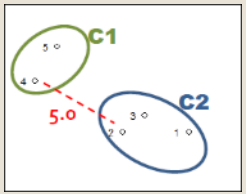
- la **distance euclidienne**, c'est-à-dire la longueur du segment qui relie 2 points, est la plus souvent utilisée

La méthodes de regroupement

↳ cf. *la distance entre 2 groupes*

Saut minimal (single linkage)

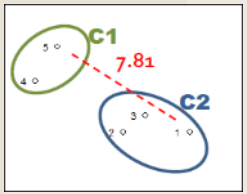
$$d(C1,C2) = \min_{a \in C1, b \in C2} d(a,b)$$



Distance entre 2 groupes = distance entre leurs *2 points les + proches*

Saut maximum (complete linkage)

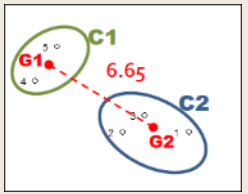
$$d(C1,C2) = \max_{a \in C1, b \in C2} d(a,b)$$



Distance entre 2 groupes = distance entre leurs *2 points les + éloignés*

Distance de Ward

$$d^2(C1,C2) = \frac{n_1 \times n_2}{n_1 + n_2} d^2(G1,G2)$$



Distance entre 2 groupes = distance pondérée entre les barycentres
=> Permet de *maximiser l'inertie interclasse*

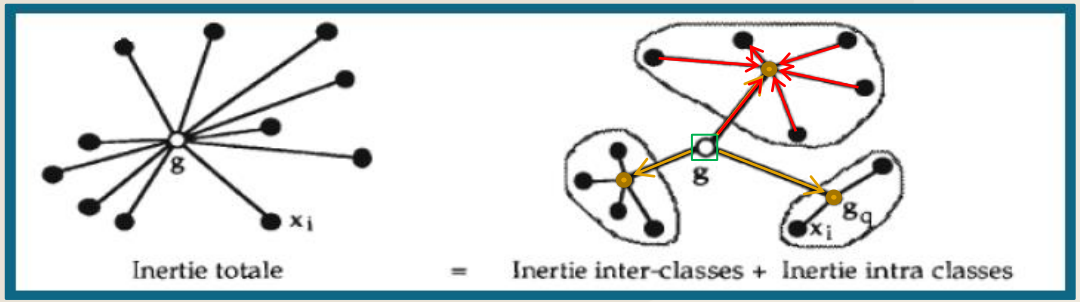
1. Classification Hiérarchique Ascendante (CHA)

Définition de l'inertie

L'inertie est un indicateur de dispersion ➔ comparable à la variance dans le cas multidimensionnel

Selon la relation de Huygens → suite à une partition, l'inertie totale se décompose en 2 éléments :

- L'inertie interclasse : dispersion entre les groupes
  - ↳ dispersion des centres de groupes autour du centre global
- L'inertie intraclasse : dispersion à l'intérieur des groupes



Distance de Ward

→ maximiser inertie interclasse

Groupes les plus différents les uns des autres

⇔ minimiser inertie intraclasse

Groupes d'individus les plus semblables

1. Classification Hiérarchique Ascendante (CHA)

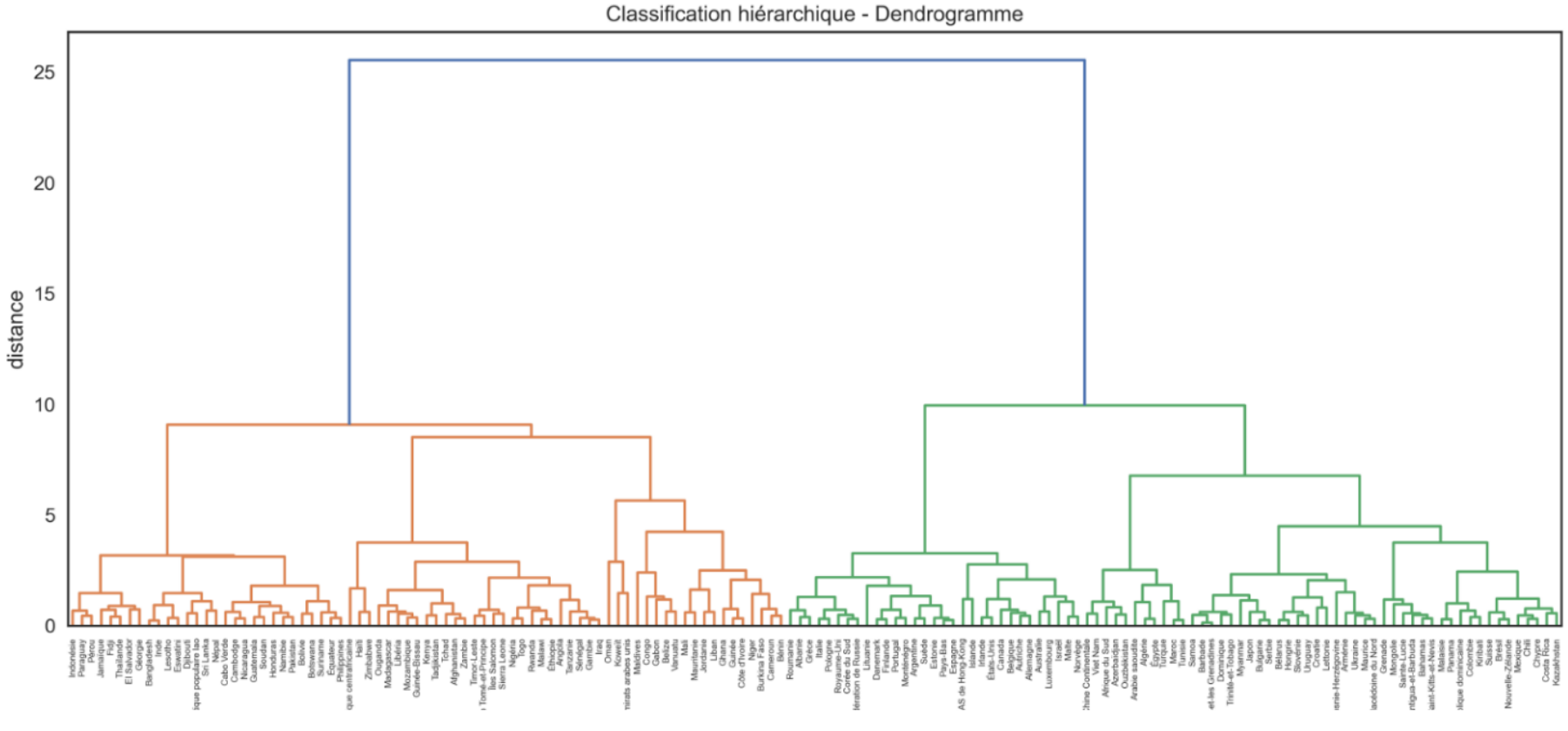
Réalisation du dendrogramme

Définition : Représentation graphique sous forme d'arborescence d'une CAH

Mise en œuvre : utilisation de la fonction scipy.cluster.hierarchy (module SciPy de Python)

Remarque : on va centrer-réduire nos variables → double intérêt :

- des données indépendantes de l'unité ou de l'échelle choisie ;
- des variables avec la même moyenne (0) et le même écart-type(1)


Classification hiérarchique - Dendrogramme
distance
25
20
15
10
5
0

1. Classification Hiérarchique Ascendante (CHA)

Détermination du nombre de groupes optimal

Graphiquement : intérêt du dendrogramme : permet de déterminer directement le nombre de groupes

↳ on coupe l'arbre lorsque que les distances entre les regroupements (et donc les différences entre les individus) deviennent importantes

Dans notre cas, on peut couper l'arbre à une distance de 8, et obtenir ainsi 5 groupes

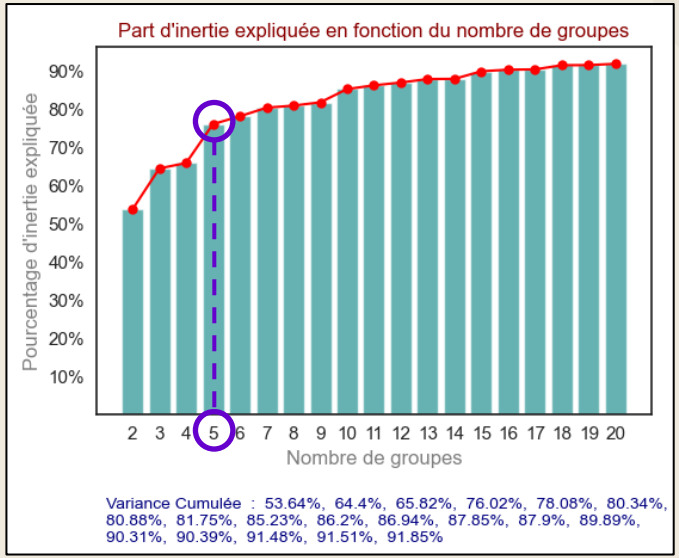
Confirmation par le calcul

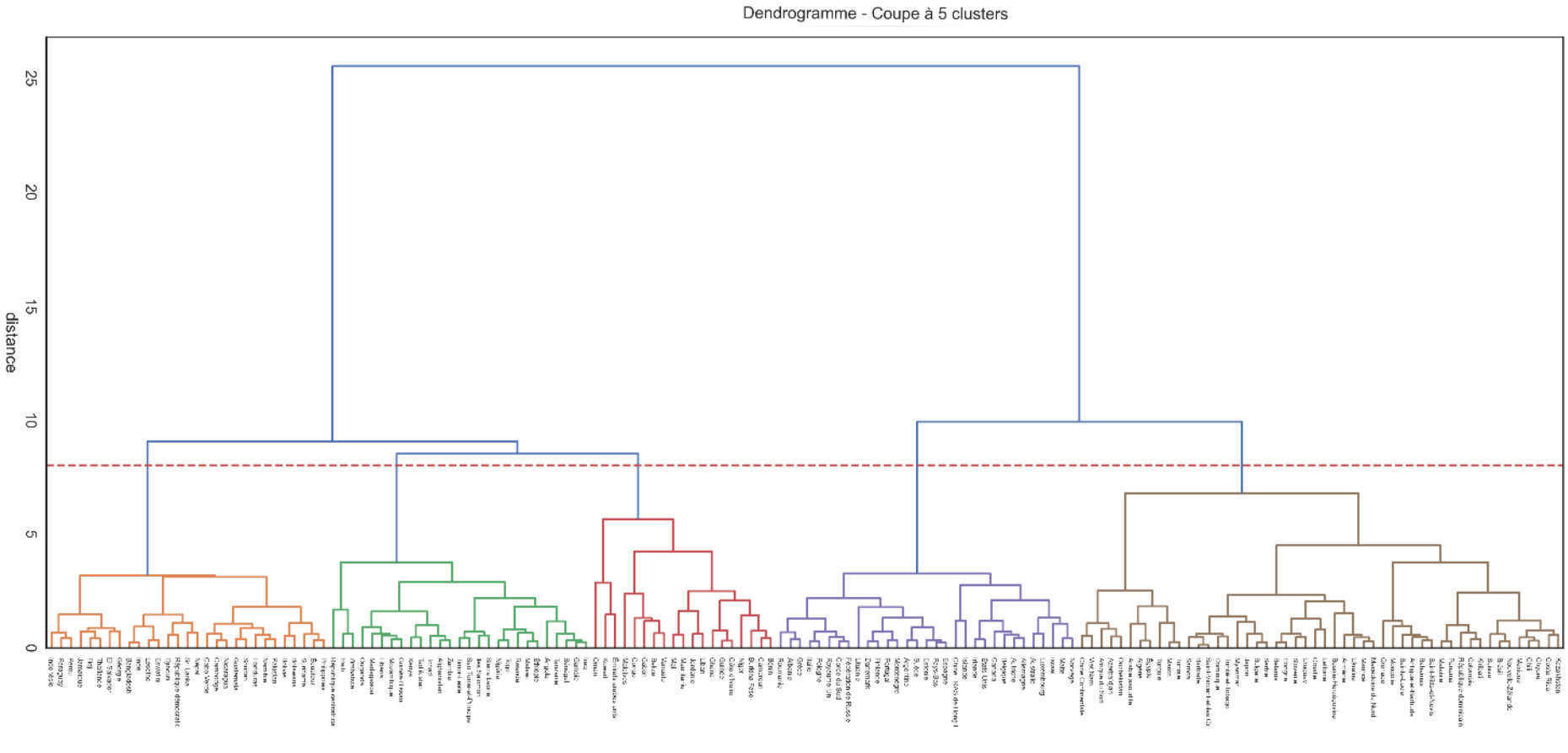
→ ex : la part d'inertie expliquée en fonction du nombre de groupes

On utilise alors la *méthode "du coude"*

→ on choisit le coude de la courbe comme le nombre de groupes à constituer

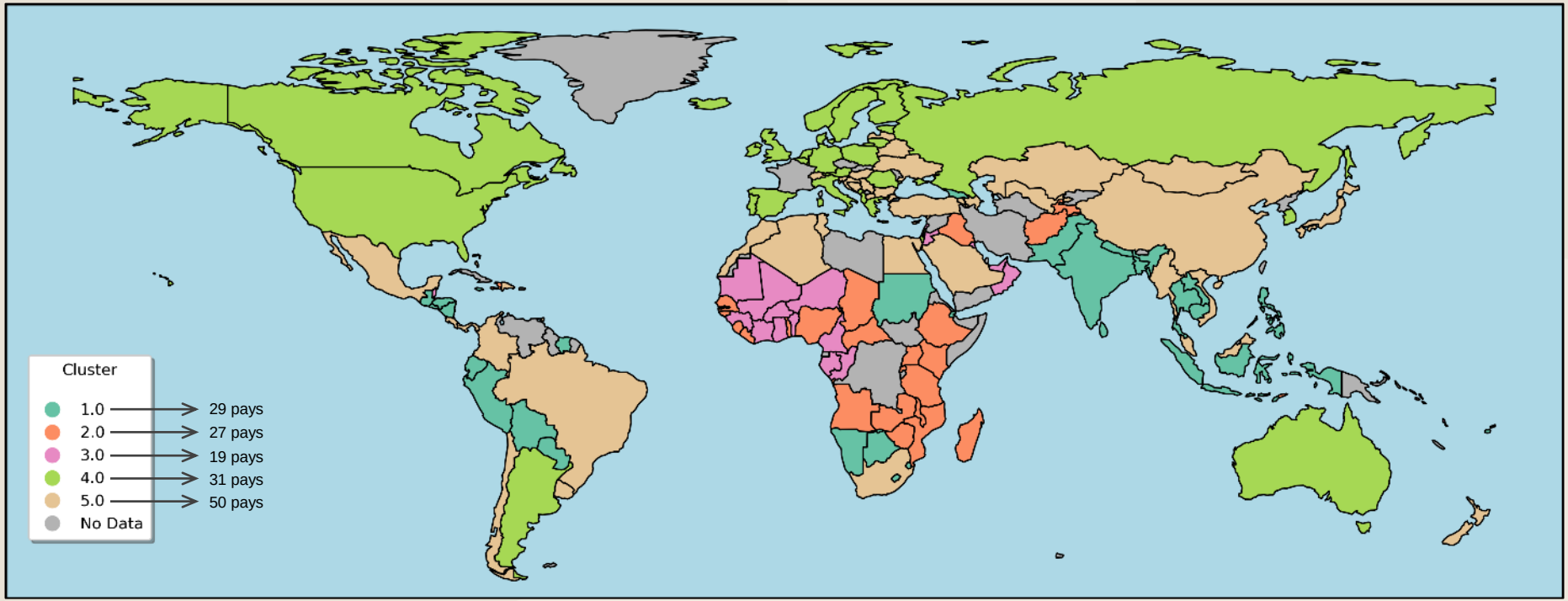
→ ici, c'est bien à partir de 5 groupes que l'augmentation marginale de la variance cumulée diminue.




Dendrogramme - Coupe à 5 clusters
distance
0
5
10
15
20
25

1. Classification Hiérarchique Ascendante (CHA)

Visualisation des clusters



2. Analyse des groupes obtenus

Caractérisation des groupes

Analyse des centroïdes

Obj : caractériser nos groupes pour savoir s'ils correspondent au profil que l'on recherche

cluster	evol_pop	pib_hab	risque	taux_impot	dispo_kcal	dispo_prot	part_prot_anim	qte_volaille	vol_part_import
1	13.05	9150.84	4.93	35.88	2628.62	68.29	34.34	14.06	157.50
2	29.68	2876.51	5.59	41.73	2244.11	54.94	22.62	4.27	38.83
3	38.62	13082.18	5.05	39.04	2788.63	75.40	33.95	15.75	362.22
4	4.99	46593.94	2.55	40.86	3461.65	111.99	60.49	27.34	74.67
5	8.21	20984.85	4.24	38.45	3034.96	88.68	50.05	29.76	311.55
mean	15.89	19777.28	4.37	39.09	2877.33	82.07	42.50	20.24	194.81

Groupes 1 et 2 :

- faible PIB/hab
- faibles valeurs "alimentaires"

Groupe 3 :

- forte croissance démographique et pays très importateurs de volaille
- mais mêmes faiblesses que les groupes 1 et 2

Groupes 4 et 5 → les 2 groupes les plus intéressants

- groupe 4 : les valeurs les plus élevées partout
  Mais pays exportateurs de volaille → est-on compétitif ?
- groupe 5 : même profil que groupe 4 mais avec de plus faibles valeurs
  Avantage : pays largement importateurs

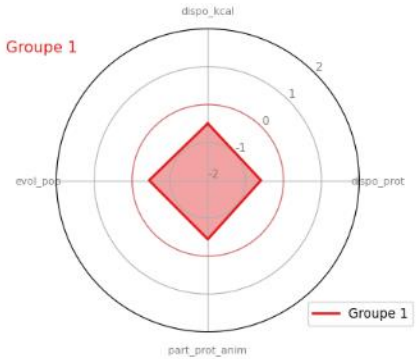
En se limitant aux variables actives :

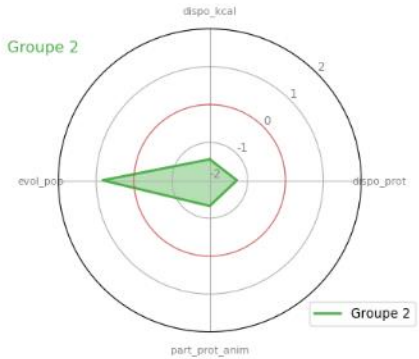
→ On retient le groupe 4

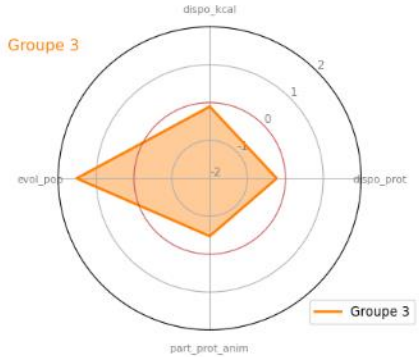
2. Analyse des groupes obtenus

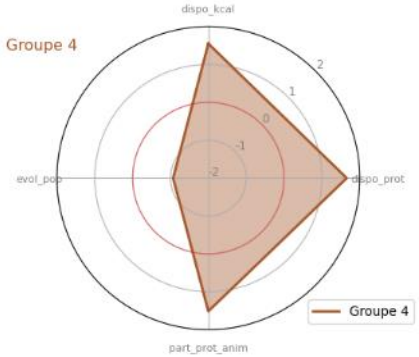
Représentation graphique du profil de chaque groupe

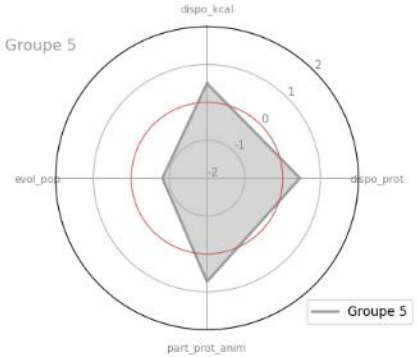
(valeurs centrées–réduites des centroïdes)











Confirmation graphique :

→ Groupe 4 le plus intéressant

3. Choix des pays

On va sélectionner les 10 pays du groupe 4 (sur un total de 31) qui correspondent le plus à notre besoin

1. Suppression des pays au marché trop étroit

On supprime les pays de moins de 1 million d'habitants → Il reste 27 pays

2. Notation des pays restants

Pour chaque variable (active et illustrative), on attribue une note à chaque pays en fonction de son classement pour cette variable

Exemple : variable "dispo_kcal" :

- on tri les valeurs par ordre croissant et on associe une note de 1 à 27
(1 pour le pays avec la valeur la plus faible ; 27 pour le pays avec la valeur la plus élevée)
- puis on additionne les notes obtenues par chaque pays, en les pondérant
- on retient les 10 meilleurs scores

pays	dispo_kcal	dispo_kcal_note
Suède	3184.0000	1
Estonie	3247.0000	2
Chine - RAS de Hong-Kong	3267.0000	3
Pays-Bas	3297.0000	4
Argentine	3307.0000	5
Espagne	3322.0000	6
...	...	...

On obtient les 10 pays suivants :

Chine - RAS de Hong-Kong	368.5
Irlande	366.5
Danemark	351.5
États-Unis	348.5
Autriche	291.0
Israël	288.0
Norvège	287.5
Australie	278.5
Canada	277.5
Portugal	263.0

4. Visualisation du partitionnement via l'ACP

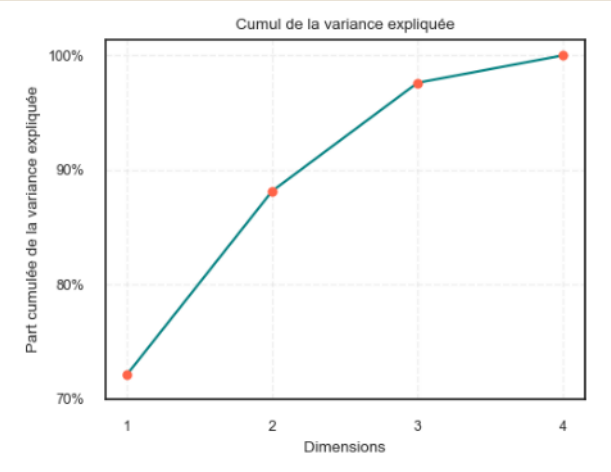
- Intérêt pratique de l'ACP :

  → permettre de représenter graphiquement un ensemble de données multidimensionnelles dans un plan à 2 dimensions,

  → grâce à la création de variables synthétiques qui minimisent la perte d'information

- Détermination du nombre de facteurs à retenir

  → ex: graphique du cumul de la variance restituée selon le nombre de facteurs



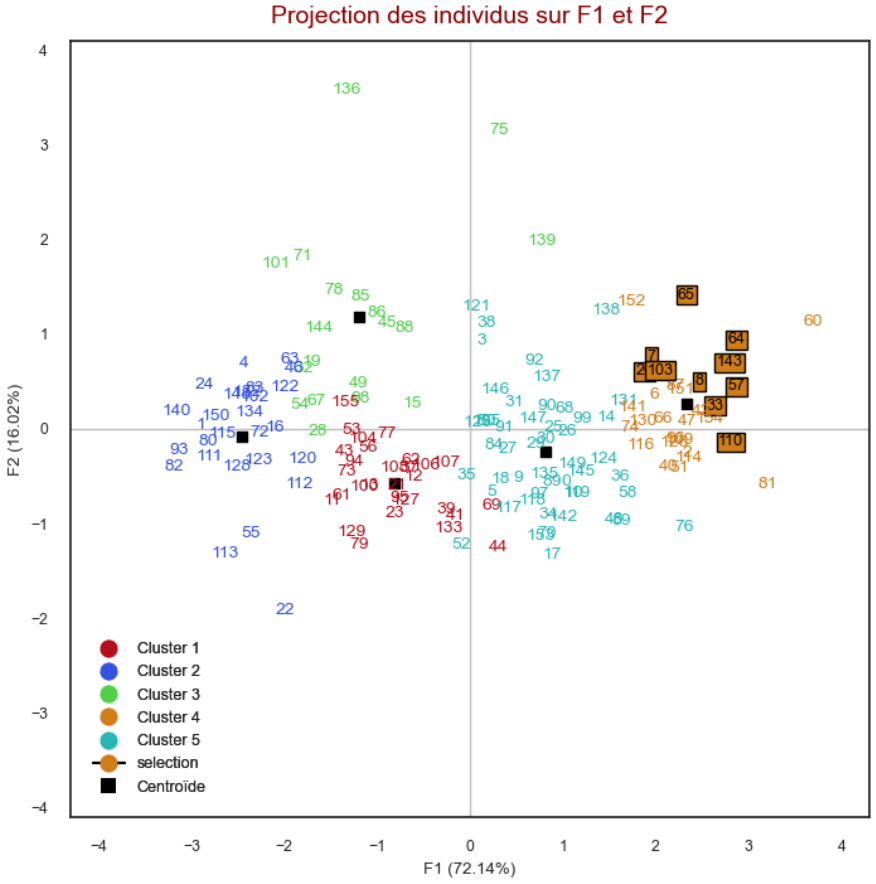
En se limitant à 2 dimensions, on récupère près de 90% de l'information initiale

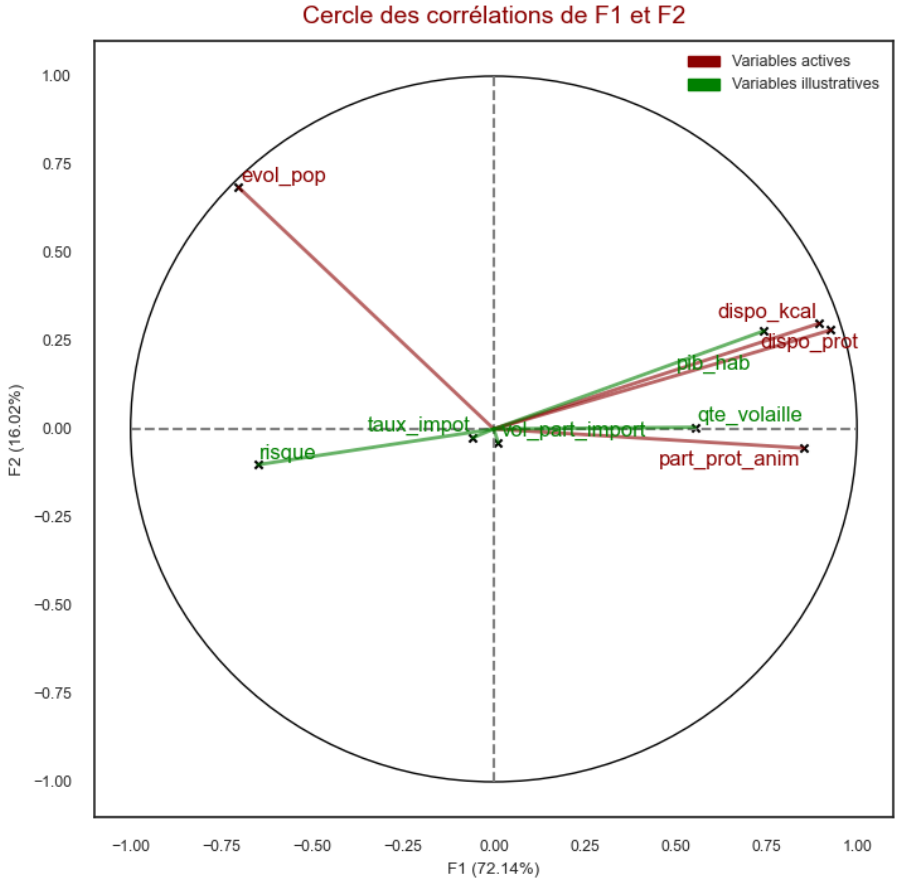
→ On se limite donc au premier plan factoriel

➔ on affiche maintenant le résultat de l'ACP, c'est-à-dire :

- la projection des individus
- le cercle des corrélations

4. Visualisation du partitionnement via l'ACP





5. Analyse de l'ACP

- Qualité de représentation des variables

  - Analyse graphique

    - variable d'autant mieux représentée que l'extrémité du vecteur qui la représente est proche du cercle de corrélations

      ↳nos 4 variables actives sont donc bien représentées

  - Confirmation par le calcul

    - on additionne le COS^2 des facteurs 1 et 2 pour chaque variable

      → plus la valeur est proche de 1, plus la variable est bien représentée

	id	COS²_var_F1	COS²_var_F2
0	evol_pop	0.49	0.47
1	dispo_kcal	0.80	0.09
2	dispo_prot	0.86	0.08
3	part_prot_anim	0.73	0.00

	id	qualite_1er_plan
0	evol_pop	0.96
1	dispo_kcal	0.89
2	dispo_prot	0.94
3	part_prot_anim	0.73

5. Analyse de l'ACP

- Caractérisation des axes

  - Analyse graphique

    Axe 1 : 72% de l'info initiale

    - la projection sur l'axe factoriel de l'extrémité de la flèche représentant une variable correspond au coefficient de corrélation entre la variable et l'axe factoriel

      - les 3 variables "alimentaires" sont corrélées à l'axe 1 et de façon positive
      - la variable "démographique" est corrélée négativement à l'axe 1

    Remarque :

    - pour les variables bien représentées, plus l'angle entre 2 variables est faible, plus la corrélation entre ces 2 variables est élevée

      ↳ ainsi, les variables "dispo_kcal" et "dispo_prot" ont une corrélation proche de 1

5. Analyse de l'ACP

- Caractérisation des axes

Axe 2 : 16% de l'info initiale

- seule la variable "évolution de la population" est corrélée de façon significative à l'axe 2 (corrélation positive)

→ *On peut caractériser nos 2 premiers axes factoriels de la façon suivante :*

* L'axe 1 peut se définir comme le niveau de la disponibilité alimentaire et de l'importance relative des protéines d'origine animale.

* L'axe 2 rend compte de l'importance de la croissance démographique

- Confirmation analytique

Les corrélations par axes factoriels

	id	COR_F1	COR_F2
0	evol_pop	-0.70	0.69
1	dispo_kcal	0.90	0.30
2	dispo_prot	0.93	0.28
3	part_prot_anim	0.85	-0.05

5. Analyse de l'ACP

- Analyse des individus

Il faut analyser le graphique de projection des individus en fonction :

de notre objectif en terme de pays cible :

- Croissance démographique,
- Disponibilité alimentaire élevée,
- Proportion de protéines animales élevée

du cercle de corrélations :

- Axe 1 : Importance alimentaire
- Axe 2 : Importance démographique

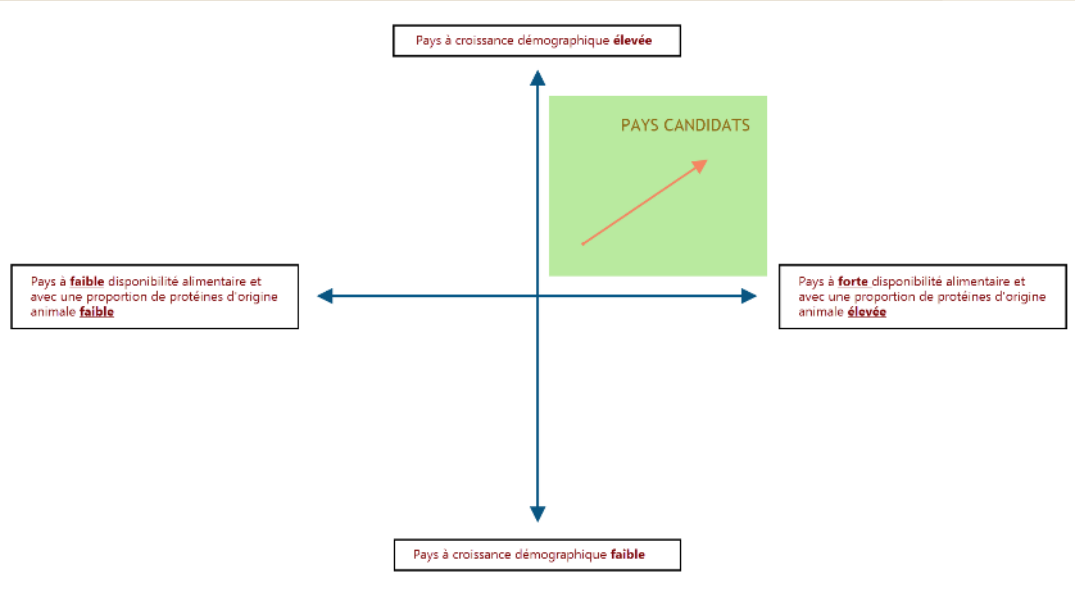
Pays recherchés :

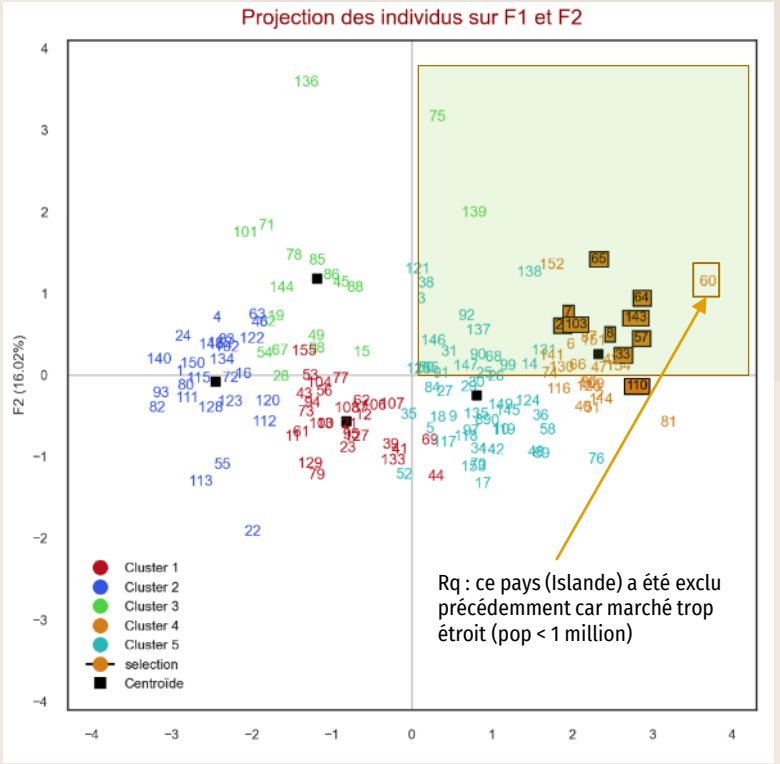
→ Pays avec des coordonnées F1 et F2 élevées et positives

Donc des pays situés dans la partie "en haut à droite" du graphique de projection des individus

5. Analyse de l'ACP

- Analyse des individus





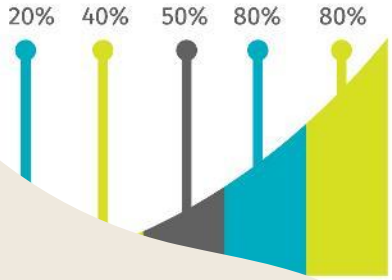
les 10 pays candidats obtenus à partir de la CHA correspondent aux "meilleurs" pays obtenus avec l'ACP.

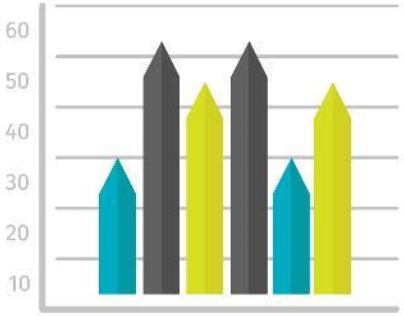
Vérifions maintenant la qualité de notre partition

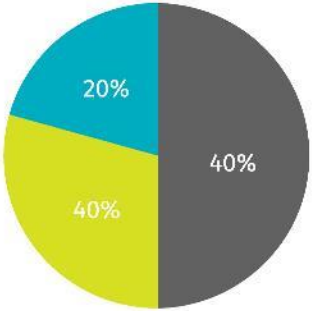
04

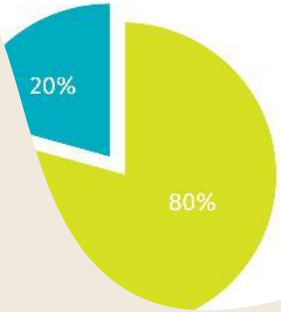
TESTS STATISTIQUES

Normalité des variables et comparaison des clusters









1. Objectif et principe

Objectif :

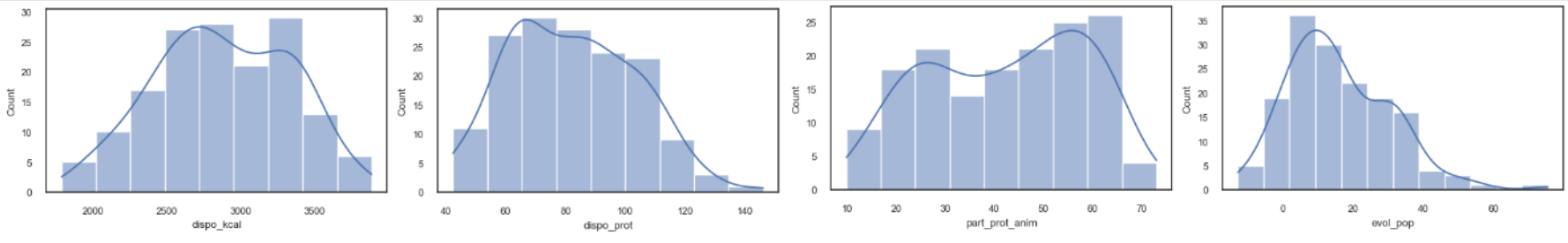
→ Vérifier que les groupes obtenus suite à la CHA sont significativement différents

Or les tests de comparaison supposent que la variable suive une loi normale

→ on commence donc par tester la normalité de la distribution de nos variables

2. Test de normalité

- Approche graphique
  - Distribution des 4 variables actives



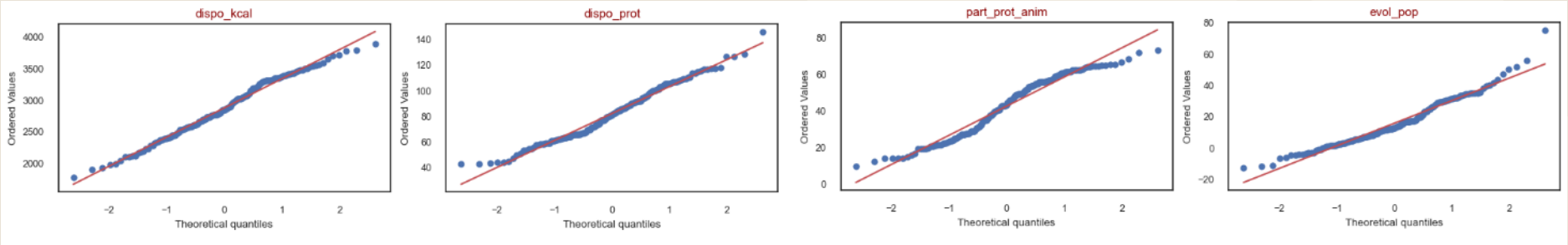
2. Test de normalité

- Approche graphique

  - Q-Q plot (Quantile-Quantile Plot)

On compare la distribution de nos variables avec la loi normale comme référentiel

→ si les données suivent une loi normale, les points obtenus forment une droite, ils sont alignés sur la diagonale principale



Graphiquement, nos 4 variables semblent suivre approximativement une loi normale :

- en particulier vrai pour la disponibilité en kcal,
- moins le cas pour la part de protéines animales (profil plus en courbes)

➔ Approche graphique reste approximative et subjective → intérêt de réaliser des tests statistiques

2. Test de normalité

- Test statistique

Test retenu → Test de Shapiro-Wilk

Le principe :

2 valeurs sont renvoyées par le test :

- **statistique** : valeur du test
- **p-value** : valeur d'interprétation du test, à comparer au seuil choisi → on choisi ici $\alpha = 0{,}05$

Interprétation :

Hypothèse nulle H0 : "La variable dont provient notre échantillon suit une loi normale"
Hypothèse alternative H1 : "La variable dont provient notre échantillon ne suit pas une loi normale"

si $p <= \alpha$: on rejette H0 au profit de l'alternative H1 => on rejette l'hypothèse de normalité
si $p > \alpha$: on ne peut pas rejeter H0 => on accepte l'hypothèse de normalité

2. Test de normalité

Résultat test statistique - Valeur pvalue

	dispo_kcal	dispo_prot	part_prot_anim	evol_pop
shapiro	0.1244	0.0264	0.0	0.0002

En vert, HO acceptée (la variable suit une loi normale)
En rouge, HO rejettée (la variable ne suit pas une loi normale)

Les tests confirment l'analyse graphique :

→ la variable disponibilité alimentaire en kcal est la plus compatible avec une distribution gaussienne

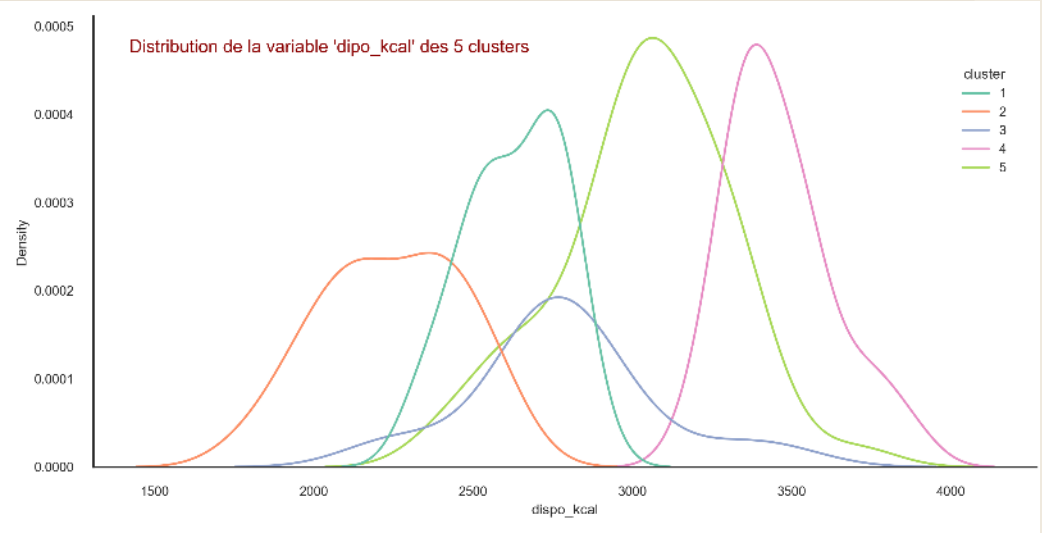
On choisit cette variable, la "dispo_kcal", pour réaliser les tests de comparaison entre les clusters

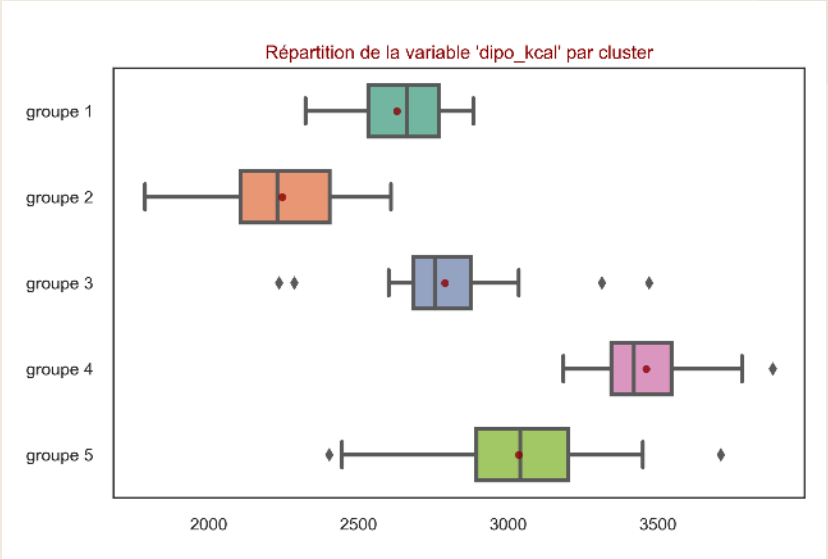
3. Test de comparaison dans le cas gaussien

- Choix des clusters

  On a choisi le groupe 4 comme groupe candidat
  → intérêt à savoir si ce groupe est bien spécifique
  → on va donc le comparer aux autres clusters

- Approche graphique



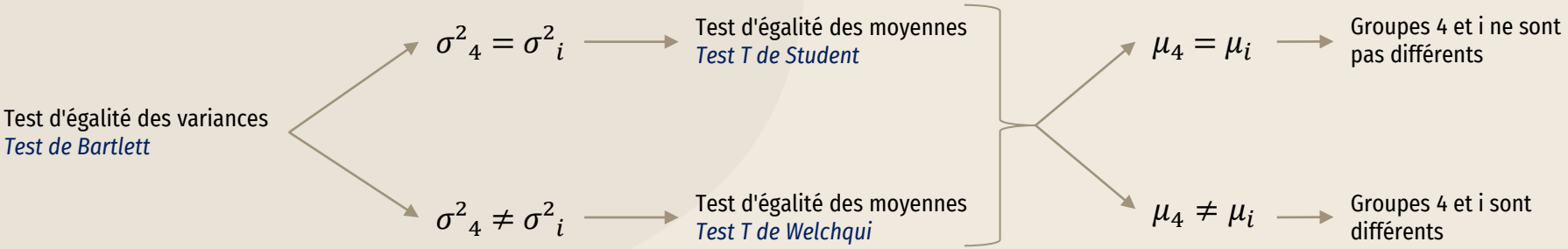


3. Test de comparaison dans le cas gaussien

- Tests statistiques - Méthodologie

Le principe → on procède en 2 étapes : 1. test d'égalité des variances
2. test d'égalité des moyennes

Soit : la variable testée "dispo_kcal"
$\sigma^2{}_4$ la variance du groupe 4 et $\sigma^2{}_i$ la variance du groupe i
μ_4 la moyenne du groupe 4 et μ_i la moyenne du groupe i
} avec i ∈ [1, 2, 3, 5]



3. Test de comparaison dans le cas gaussien

- Test d'égalité des variances

Hypothèse nulle H0 : "Les variances des 2 groupes sont égales"
Hypothèse alternative H1 : "Les variances des 2 groupes ne sont pas égales"

si $p <= \alpha$: on rejette H0 au profit de l'alternative H1 => on rejette l'hypothèse d'égalité des variances
si $p > \alpha$: on ne peut pas rejeter H0 => on accepte l'hypothèse d'égalité des variances

Groupe 4 et groupe 1 : pvalue = 0.659106
Groupe 4 et groupe 2 : pvalue = 0.132292

pvalue > 0,05 → on ne peut rejeter H0 → on considère les variances comme égales

Groupe 4 et groupe 3 : pvalue = 0.007728
Groupe 4 et groupe 5 : pvalue = 0.006489

pvalue < 0,05 → on rejeter H0 → on considère les variances comme différentes

3. Test de comparaison dans le cas gaussien

- Test d'égalité des moyennes

Hypothèse nulle H0 : "Les moyennes des 2 groupes sont égales"
Hypothèse alternative H1 : "Les moyennes des 2 groupes ne sont pas égales"

si $p <= \alpha$: on rejette H0 au profit de l'alternative H1 => on rejette l'hypothèse d'égalité des moyennes
si $p > \alpha$: on ne peut pas rejeter H0 => on accepte l'hypothèse d'égalité des moyennes

Groupe 4 et groupe 1 : pvalue ≃ 0
Groupe 4 et groupe 2 : pvalue ≃ 0
Groupe 4 et groupe 3 : pvalue ≃ 0
Groupe 4 et groupe 5 : pvalue ≃ 0

pvalue < 0,05 → on rejette H0 → on considère les moyennes significativement différentes

➔ Pour la variable "dispo_kcal", le groupe de pays retenu comme candidat potentiel est bien différent des autres groupes

Conclusion

Classification hiérarchique → déterminer des groupes aux caractéristiques propres

→ Sélectionner un groupe de pays candidats potentiels en tant que marchés d'exportation

ACP → Visualiser notre partitionnement sur plan à 2 dimensions grâce à la détermination de variables synthétiques

Tests Stats → Confirmer la qualité de notre partition : groupes différents

---

➔ 1ère étude qui analyse les pays susceptibles de devenir des marchés d'exportation

Mais étude qui repose sur le choix arbitraire et non exhaustif de 4 variables actives

↳ Nécessité d'analyses complémentaires

MERCI POUR VOTRE ATTENTION

QUESTIONS ?