Université de Technologie de Compiègne

Génie informatique

Rapport du projet de SY09

Authors :

Crauser Julien et Antoine Collas

15 avril 2018

1 Cuisine

Question 1

Le jeu de données présent dans le fichier recettes-pays.data contient 51 variables pour 26 individus. Nous pouvons donc noter que nous disposons de plus de variables que d'individus. Parmi ces 51 variables, une seule est qualitative (les origines des recettes) et 50 sont quantitaves. Ces dernières prennent leurs valeurs entre 0 et 1 (0 et 0.82 pour être plus précis). Dans ce jeux de données les recettes ont été agrégées par origine, il y a donc 26 origines (une par ligne de notre tableau individus-variables). De plus, le jeux de données ne présente aucunes valeurs manquantes. Cependant certaines origines sont à remarquer. En effet, il y a par exemple une distinction entre l'Afrique et le Maroc ou encore entre l'Asie et Chine, Japon, Vietnam, Thailande. De plus, les origines semblent être regroupées par région géographique mais des origines comme j̈uive,c̈adiens, sont présentes.

Question 2

Nous réalisons une ACP sur le jeu de données. Nous obtenons 26 axes principaux : autant que d'individus. En effet, comme il y a moins d'individus que de variables, il est suffisant de prendre 26 axes pour représenter tous les individus.

Nous obtenons les pourcentages d'inertie expliquée visibles sur la figure 1. Les 3 premiers axes représentent 72% de la variance du nuage de points.

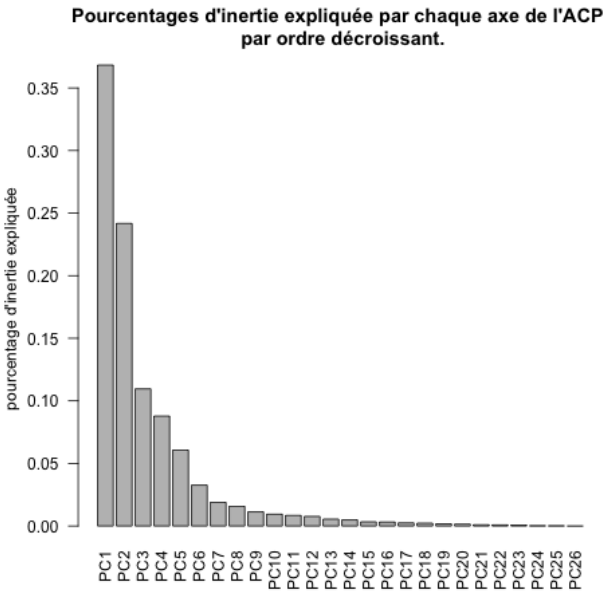
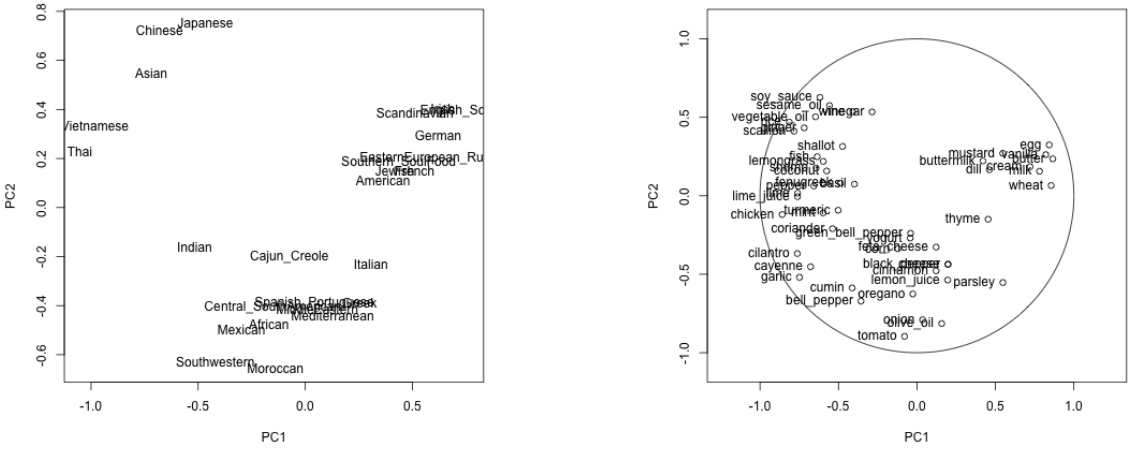


FIGURE 1 – Pourcentages d'inertie expliquée

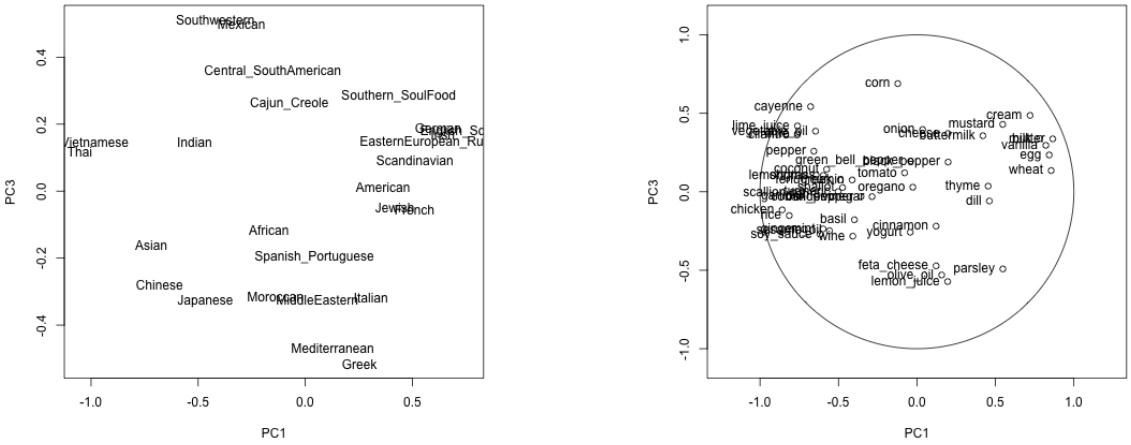
Comme les 3 premiers axes factoriels représentent 72% de la variance totale, nous analysons le nuage de points seulement avec les 3 premiers plans factoriels. Nous obtenons les plans factoriels des

figures 3, **??** et 4. Nous avons associé à chaque plan son cercle des corrélations.



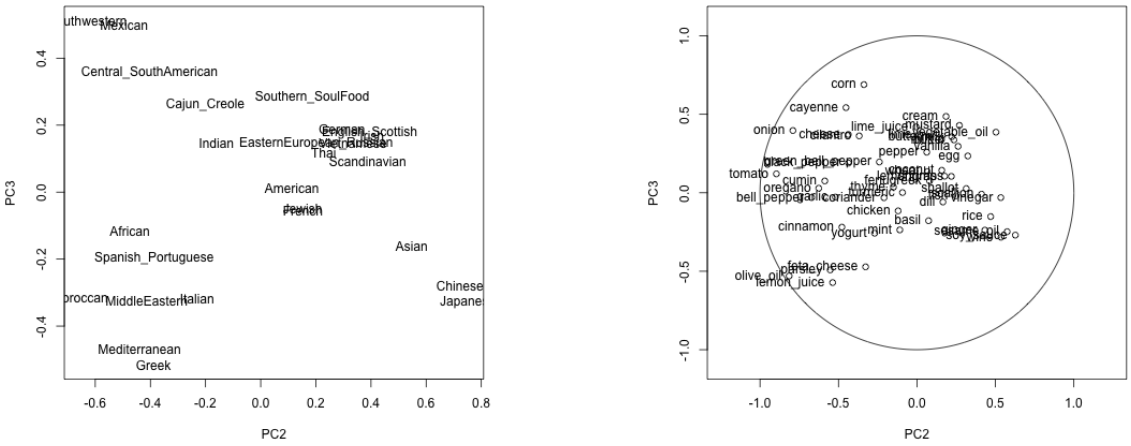
(a) Plan 1,2 (b) Cercle des corrélations

FIGURE 2 – Plan factoriel 1,2 avec son cercle associé



(a) Plan 1,3 (b) Cercle des corrélations

FIGURE 3 – Plan factoriel 1,3 avec son cercle associé



(a) Plan 2,3

(b) Cercle des corrélations

FIGURE 4 – Plan factoriel 2,3 avec son cercle associé

Dans le premier plan factoriel, nous observons que plusieurs groupes se forment. Les pays asiatiques se distinguent avec des ingrédients comme l'huile de sésame, l'huile de soja, le poisson, les échalottes... Un deuxième groupe se distingue : la cuisine occidentale avec des ingrédients comme le blé, les oeufs et le lait. Aussi, les pays consommant de la tomate et de l'huile d'olive sont exposés dans ce premier plan : Maroc, Mexique et Afrique.

Dans le deuxième plan factoriel, la Grèce et la méditerranée sont liées par leur consommation de féta, d'huile d'olive et de jus de citron. Ce plan permet aussi de voir la formation de deux sous groupes parmi les pays asiatiques. Le Vietnam et la Thaïlande consomment du poulet et du piment, ce qui est moins le cas de la Chine et du Japon.

Enfin le troisième plan factoriel met en évidence des pays consommant du maïs comme le Mexique.

Question 3

Nous obtenons le dendrogramme de la figure 5.

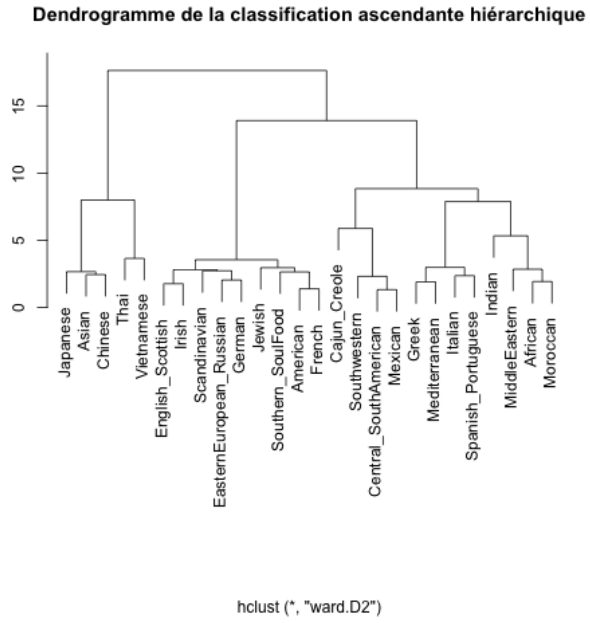


FIGURE 5 – Dendrogramme de la classification ascendante hiérarchique

Ce dendrogramme confirme plusieurs observations que nous avions déjà indiquées dans la deuxième question :

— les cuisines des pays asiatiques sont ensemble avec deux sous groupes : Japon-Chine, Thailande-Vietnam.

— les recettes de deux pays occidentaux, les Etats-Unis et la France, sont ensemble

— la cuisine méditerranée est regroupée avec les pays suivants : Grèce, Italie, Espagne, Portugal.

D'autres groupes que nous n'avions pas identifiés sont présents :

— les recettes des pays de l'europe de l'est avec l'Allemagne, la Russie et les pays de l'est.

— la cuisine des régions dans la partie inférieure au Texas : les Cadiens (présents au Texas et en Louisiane), Mexique, Amérique centrale, Amérique du sud.

— la cuisine d'Afrique et du Moyen-Orient (Afrique, Moyen-Orient et Maroc).

A un niveau plus macro nous avons 3 groupes :

— l'asie avec un indice de 7.5

— l'occident avec l'europe de l'est : un indice de 4 ce qui traduit une utilisation relativement homogène des ingrédients

— Amérique centrale et du sud, méditerranée, l'Inde, l'Afrique, avec un indice de 8 qui traduit une certaine hétérogénéité par rapport à l'occident par exemple

Question 4

Nous commençons par chercher quel serait un b̈onïombre de classes à choisir :

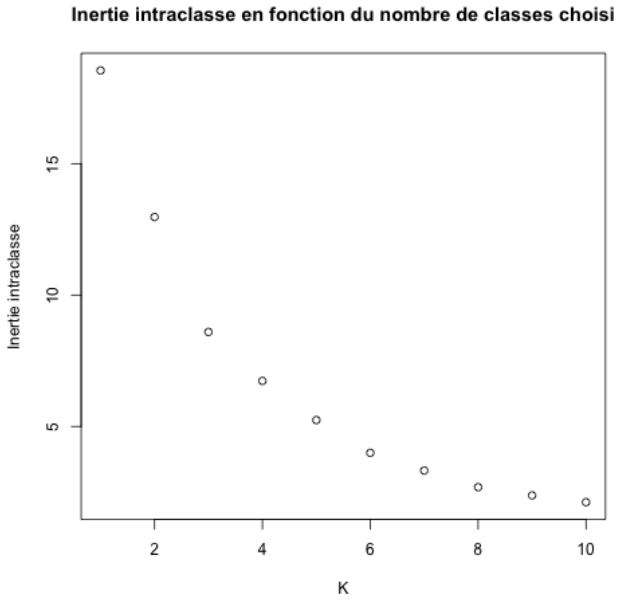


FIGURE 6 – Inertie intraclasse en fonction du nombre de classes choisi

En utilisant la règle du coude un nombre de classes est 3. Ce nombre est en adéquation avec la classification hiérarchique ascendante que nous avions obtenue dans la question précédente. En effet nous avions déjà identifié 3 groupes.

Nous obtenons la classification suivante :

1	African, Cajun_Creole, Central_SouthAmerican, Greek, Indian, Italian, Mediterranean, Mexican, MiddleEastern, Moroccan, Southwestern, Spanish_Portuguese
2	Asian, Chinese, Japanese, Thai, Vietnamese
3	American, EasternEuropean_Russian, English_Scottish, French, German, Irish, Jewish, Scandinavian, Southern_SoulFood

Cette classification est identique aux trois groupes que nous avions identifiés précedemment avec la CAH.

Question 5

Question 6

Le jeux de données qui est fourni contient un échantillon avec 2000 recettes (2000 lignes). Il y a 26 régions (les mêmes que dans le jeu de données précédent) et 51 variables. Les valeurs sont binaires, 0 pour l'absence et 1 pour la présence d'un ingrédient. Il n'y a aucune valeur manquante.

Question 7