SARAVANANE Sylvain BOUHIL Julian	L3 MIASHS BDD-SOCIOLOGIE PROJET

Base de donnée : Trafic de voyageurs et marchandises depuis 1841 — SNCF Open Data

- Description de la base de données choisie

Le projet que nous avons réalisé s'appuie sur la base de donnée concernant l'évolution des trafics de voyageurs et marchandises pour le transport ferroviaire en France depuis 1841 d'après la SNCF

→ Etudes de variables quantitatives : [Voyageurs, Voyageurs-km, Tonnes, Tonnes-km]
Nous avons quatre variables qui sont exprimés en chiffres (parmi lesquels, on considère la variable « année » comme étant une variable supplémentaire pour comprendre les résultats au cours du temps depuis 1841 jusqu'à 2019)

→ 179 observations avec 5 variables

Année (int) : De 1841 à 2019
Voyageurs (num) : Donnée exprimée en millions de voyageurs transportés en une année.
Voyageurs-km (num) : Donnée exprimée en milliard, la somme des kilomètres parcourus des voyageurs en une année.
Tonnes (num) : Donnée exprimée en millions de marchandises transportées en une année.
Tonnes-km (num) : Donnée exprimée en milliard, la somme des kilomètres parcourus des tonnes de marchandises transportées en une année.

- Idée de recherche à laquelle répond l'analyse et le choix de la méthode d'analyse factorielle

Cette base de donnée s'appuie donc de l'évolution du système ferroviaire en France, des comparaison de chaque année depuis 1841. On peut émettre des déductions statistiques. Nous allons donc évaluer et interpréter les trajets ferroviaires (des voyageurs et des marchandises) au cours du temps depuis 1841.

Nous avons donc retenu l'ACP (Analyse en composantes principales), puisque les informations contenues du tableau de données sont un ensemble d'observation mesurées sur un ensemble de variables quantitatives
On peut représenter la corrélation des données entre elles

Émettre une démarche mathématique pour interpréter ses résultats comme par exemple :
- la construction de la matrice des corrélations des variables,
- les valeurs propres/part d'inertie
- les contributions (en fonction des variables)
- les cosinus carrées

Le logiciel R a reconnu notre base de donnée avec des variables en tant que variable de type factor. Ainsi nous avons dû convertir chaque variables en variables numériques avec la ligne de code :
BDD$VAR ← as.numeric(BDD$VAR) : Avec BDD le nom de notre base de donnée (PROJECT) et VAR le nom de la variable à changer (Tonnes.km,Tonnes,Voyageurs,Voyageurs.km).

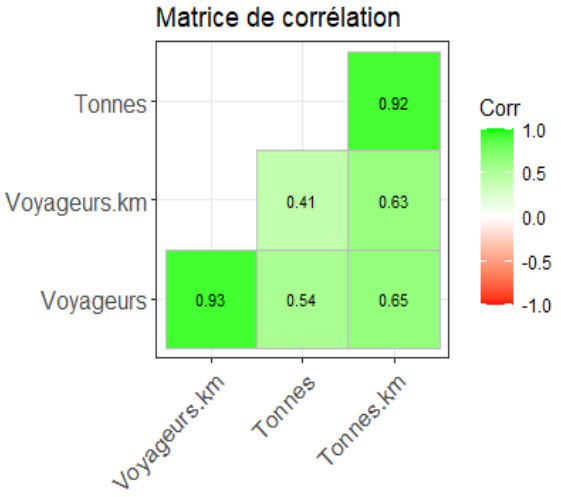
Ce changement en variable numérique permettra donc de réaliser convenablement une ACP.

<u>Réalisation de la matrice de corrélation :</u> corr ← cor(PROJECT[,-c(0,5)])
#(ou bien, ou bien cor (PROJECT[,1:4]))

Avec les variables numériques (donc nos variables quantitatives) de la base de donnée, nous avons réaliser une matrice de corrélation (que j'ai appelé : *corr*, dans le fichier R)

	Voyageurs	Voyageurs.km	Tonnes	Tonnes.km
Voyageurs	1.0000000	0.9317773	0.5428066	0.6512404
Voyageurs.km	0.9317773	1.0000000	0.4065103	0.6303175
Tonnes	0.5428066	0.4065103	1.0000000	0.9231754
Tonnes.km	0.6512404	0.6303175	0.9231754	1.0000000

A l'aide du package ggcarrplot, nous avons réaliser un graphique de cette matrice de corrélation, pour une meilleure interprétation visuelle. (Arrondi à 2 chiffres après la virgule)

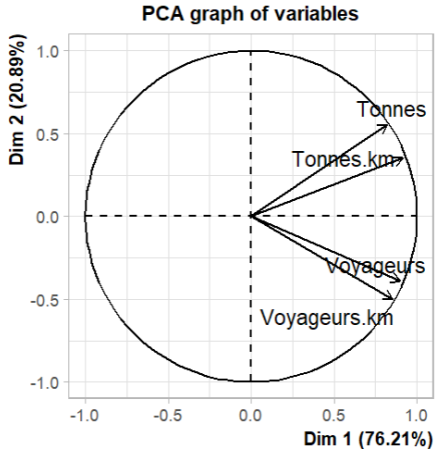


Nous pouvons faire une interprétation générale : Tout d'abord nous pouvons remarquer que chaque coefficient a une corrélation positive (supérieure à 0). Rappelons tout de même que la corrélation est le degré de liaison avec une composante, elle est comprise entre [-1,1].

Dans cette matrice de corrélation, on peut distinguer une forte corrélation entre Voyageurs et Voyageurs.km (à 0,93) et entre Tonnes et Tonnes.km (à 0,92)
La matrice présente également un coefficient de corrélation faible (la plus faible de la matrice) entre Voyageurs.km et Tonnes (à 0,41)

<u>Réalisation de l'ACP :</u> Je nomme ' res.ACPROJECT ', le nom de la réalisation de l'ACP

Pour interpréter notre base de donnée et de présenter les informations et les observations de nos variables quantitatives et des données corrélées, nous allons donc réaliser une ACP sur R afin d'avoir une représentation graphique mais également l'allure des nuages de points, en posant la variable supplémentaire « année » : **PCA(PROJECT[,-c(0,5)],scale.unit=TRUE,quali.sup=1**



Les variables mesurant les transports de marchandises (Tonnes et Tonnes.km) sont donc corrélés.

De même, les variables mesurant les transports de voyageurs (Voyageurs et Voyageurs.km) sont également corrélés.

Les quatre vecteurs sont corrélés plus ou moins vers l'axe 1.

Nous ajoutons, avec l'aide de cette formule : **res.ACPROJECT$eig,** les valeurs propres de nos quatre composantes de notre ACP : $\lambda_1 = 3.049$; $\lambda_2 = 0.836$; $\lambda_3 = 0.109$: $\lambda_4 = 0.007$

Ainsi nous pouvons établir la part d'inertie avec les valeurs propres de chacune de nos composantes.

Choix des axes :

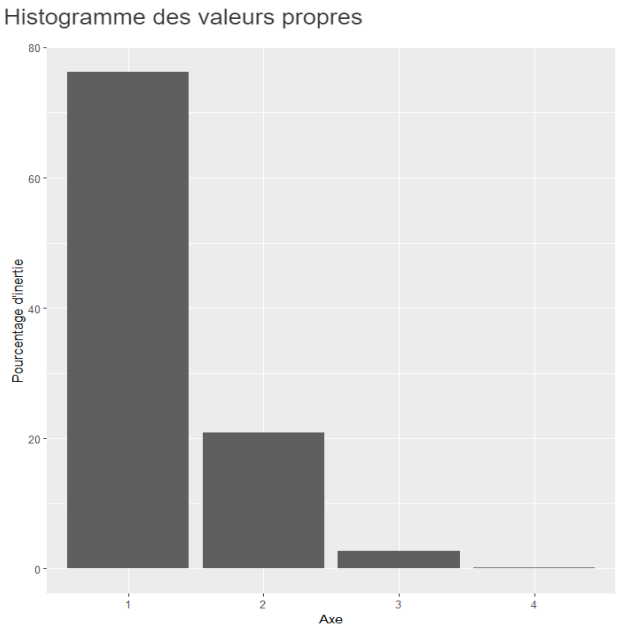


Tableau des valeurs propres

Axe	%	Cum. %
1	76.2	76.2
2	20.9	97.1
3	2.7	99.8
4	0.2	100.0

En utilisant la règle du coude, les axes 3 et 4 n'apporteront peu d'informations (une interprétation plus ou moins faible)
La règle de Kaiser, quand à lui, stipule que nous devons sélectionner les composantes dont la valeur propre λ_k est supérieure à 1. Ainsi, seule la composante 1 se réfère au critère de Kaiser ($\lambda_1 = 3{,}048$)
En se référençant des différents critères, nous avons deux résultats contradictoires, nous avons donc préféré à l'interprétabilité en se reposant sur les deux premiers axes.

Contributions des années et cosinus² :

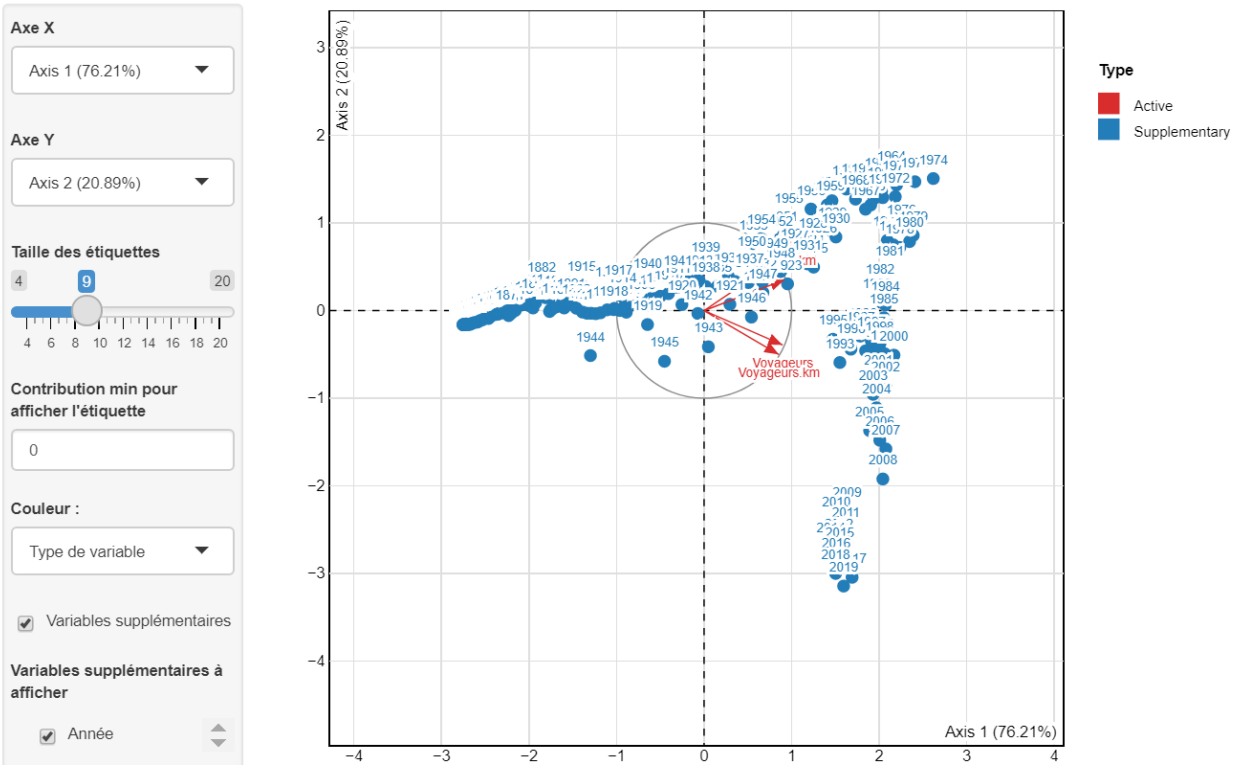
Pour savoir si les années ont été contributives, il faut que sa contribution au sein des différentes composantes soit au moins de l'ordre 1 % (contrib ≥ 0.01). Comprise entre 0 et 1, Cela permet de mesurer l'influence de la variable dans la définition de la composante.

Nous pouvons voir que les points 80, 99, 163, 24, 103, et 161 ont les seuls contributions nulles et donc ses points ne sont pas contributifs.
Les points 8 et 23 sont les points ont la contribution est la plus forte (estimée à 13,9 %)

Concernant le cosinus², les années 1885,1886,1887 sont les années les mieux représentées avec cos² égal à 1.

Concernant le V-test, il est particulièrement faible en valeur absolu (puisque |VT| = |-1,580| et donc inférieure à 2) et donc les V-test des différentes années sont très faibles.

Interprétation des axes :



Axe 1 (76,21%) :

On remarque que plus les points sont situés à gauche, plus le trafic est faible (moins de voyageurs et de marchandises transportés) et inversement, plus les points sont situés à droite, plus le trafic est élevé.

En effet, le trafic était beaucoup moins dense en 1841 qu'en 1974 par exemple, cela est dû à l'évolution du trafic ferroviaire entre ces deux dates. On observe un déplacement des points partant de la gauche jusqu'à la droite entre 1841 et 1974, puis une stabilisation de ces points à partir de 1974 jusqu'en 2019.

Axe 2 (20,89%) :

En observant la direction des vecteurs, on se rend compte que plus les points sont situés en haut du graphe, plus ce sont des marchandises qui furent transportées cette année-là. A l'inverse, plus les points sont situés en bas du graphe, plus ce sont des voyageurs qui ont été transportés. On observe dans un premier temps que les trains étaient surtout utilisés pour transporter de la marchandise jusque dans les années 1970, puis le transport de voyageurs s'est démocratisé entre ces années jusqu'à maintenant, où les trains sont majoritairement utilisés pour transporter des voyageurs.

Cependant, on observe quelques exceptions, en 1944, 1945, 1919, 1943 qui sont des années où les guerres mondiales ont eu lieu, ce qui explique cette tendance à transporter plus de passagers que de marchandises au cours de ces périodes.