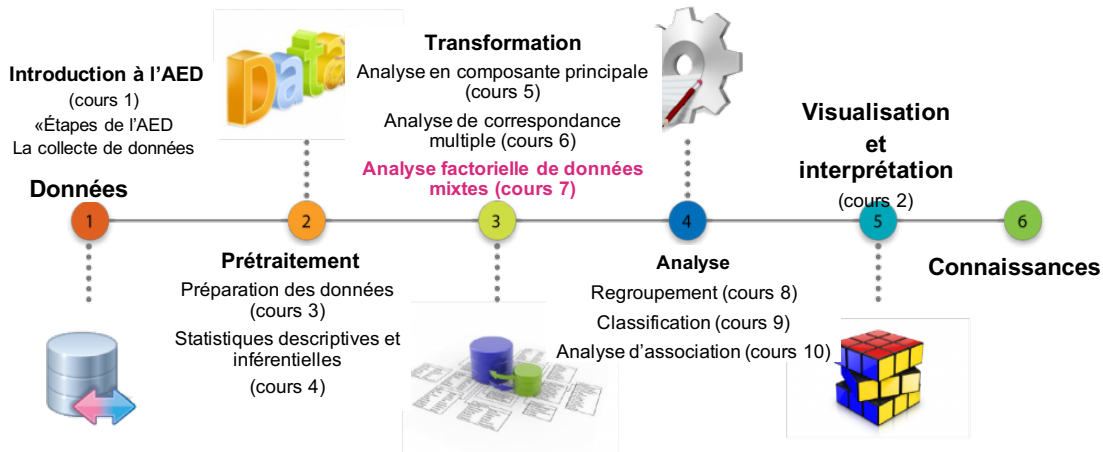


Cours 7 - Analyse Factorielle des Données Mixtes (AFDM)

Neila Mezghani

6 avril 2021



Plan du cours

- 1 Introduction
- 2 Représentation des données
- 3 Réalisation d'une AFDM
 - Principe de l'AFDM
 - Nombre d'axes à retenir
- 4 Analyse des variables et des individus
 - Représentation des individus
 - Analyse des variables

Plan du cours

- 1 Introduction
- 2 Représentation des données
- 3 Réalisation d'une AFDM
 - Principe de l'AFDM
 - Nombre d'axes à retenir
- 4 Analyse des variables et des individus
 - Représentation des individus
 - Analyse des variables

Définition de l'AFDM

- L'Analyse Factorielle des Données Mixtes (AFDM ou FAMD pour Factor Analysis of Mixed Data en anglais) est une méthode destinée à analyser un jeu de données contenant à la fois des variables quantitatives et qualitatives.
- Autrement dit, l'AFDM peut être considéré comme une analyse mixte entre l'analyse en composantes principales (ACP) et l'analyse des correspondances multiples (ACM).

Objectifs de l'AFDM

l'AFDM permet de faire des représentations graphiques du contenu d'un ensemble de données mixtes pour extraire des connaissances :

- Sur des **similitudes** ou **ressemblance** entre les individus en prenant en compte des variables mixtes.
- Sur des **relations** ou **liaisons** entre toutes les variables, tant quantitatives que qualitatives.

Questionnements

Les questions auxquelles permet de répondre une AFDM sont :

- Quels sont les individus qui se ressemblent ? (proximité entre les individus)
- Sur quelles variables sont fondées ces ressemblances avec une difficulté supplémentaire en relation avec le type des données ?
- Quelles sont les associations entre les variables ? Entre quantitatives, l'idée de la corrélation s'impose ; entre les qualitatives, on utilise le TDC ; mais comment faire entre quantitatives et qualitatives ?

Plan du cours

- 1 Introduction
- 2 Représentation des données
- 3 Réalisation d'une AFDM
 - Principe de l'AFDM
 - Nombre d'axes à retenir
- 4 Analyse des variables et des individus
 - Représentation des individus
 - Analyse des variables

Données

Les données se présentent sous la forme d'une matrice/tableau :

- n individus
- p_1 variables quantitatives stockées dans la matrice X_1
- p_2 variables qualitatives stockées dans la matrice X_2

	1	...	j	...	p_1	1	...	j	...	p_2
1										
i	x_i^j					x_i^j				
n										

Données quantitatives

Les données quantitatives sont centrées et réduites : $Z_1 = X_1$ centrée réduite avec $\dim(Z_1) = n \times p_1$

	1	...	j	...	p_1
1					
i	z_{ij}				
n					
Moyenne	μ_j				
Écart-type	σ_j				

avec : μ la moyenne de la variable
et σ l'écart type.

$$z_{ij} = \frac{x_i^j - \mu_j}{\sigma_j}$$

Données qualitatives (1)

Les données quantitatives se présentent selon :

	1	...	j	...	p_2
$X_2 =$					
i			h_{ij}		
n					

- n est le nombre d'individus, p_2 le nombre de variables qualitatives
- $h_{ij} \in M_j$ avec M_j l'ensemble des modalités de la j ème variable
- $m_j = \text{card}(M_j)$ le nombre de modalités de la variable j et $m = m_1 + \dots + m_{p_2}$ le nombre total de modalités.

Données qualitatives (1)

Données qualitatives sont transformées en un tableau disjonctif complet \mathbf{K} :

$$\mathbf{K} = \begin{array}{c|ccccc} & 1 & \dots & s & \dots & m \\ \hline 1 & & & & & \\ \vdots & & & \vdots & & \\ i & \dots & & k_{is} & \dots & \\ \vdots & & & \vdots & & \\ n & & & & & \\ \hline \end{array}$$

Chaque colonne s est l'indicatrice de la modalité s

- $k_{is} = 1$ si l'individu i possède la modalité s
- $k_{is} = 0$ sinon

Données qualitatives (2)

Les données quantitatives sont ensuite pondérée par $\sqrt{p_j}$: ce que nous appelons le codage-ACP de la variable qualitative.

	1	...	j	...	m
1	k_{ij}				
i					
n					
Effectif	n_j				
Proportion	p_j				

avec : n_j l'effectif de la modalité j
et p_j la proportion de la modalité.

$$z_{ij} = \frac{h_{ij}}{\sqrt{p_j}}$$

$$p_j = \frac{n_j}{n}$$

Données préparées

On obtient un tableau sommaire des données $Z = (Z_1|Z_2)$ tel que :

- Les variables quantitatives sont centrées et réduites.
- Les indicatrices des modalités sont normalisées par la racine carrée de la proportion.

$Z =$

	1	...	j	...	p_1	1	...	j	...	m	
1	z_{ij}					z_{ij}					
i											
n											
	Z_1					Z_2					

Exemple : Données

Modele	puissance	longueur	hauteur	poids	CO2	origine	carburant	4X4
GOLF	75	421	149	1217	143	Europe	Diesel	non
CITRONC4	138	426	146	1381	142	France	Diesel	non
P607	204	491	145	1723	223	France	Diesel	non
VELSATIS	150	486	158	1735	188	France	Diesel	non
CITRONC2	61	367	147	932	141	France	Essence	non
CHRY300	340	502	148	1835	291	Autres	Essence	non
AUDIA3	102	421	143	1205	168	Europe	Essence	non
OUTLAND	202	455	167	1595	237	Autres	Diesel	oui
PTCRUISER	223	429	154	1595	235	Autres	Essence	non
SANTA_FE	125	450	173	1757	197	Autres	Diesel	oui

- Variables quantitatives : $p_1 = 5$
- Variables qualitatives : $p_2 = 3$

Exemple : Données quantitatives centrées réduire

Modele	puissance	longueur	hauteur	poids	CO2
GOLF	-1,103	-0,614	-0,419	-0,989	-1,127
CITRONC4	-0,304	-0,485	-0,733	-0,411	-1,148
P607	0,532	1,192	-0,838	0,795	0,558
VELSATIS	-0,152	1,063	0,524	0,837	-0,179
CITRONC2	-1,280	-2,007	-0,628	-1,993	-1,169
CHRY300	2,256	1,476	-0,524	1,189	1,990
AUDIA3	-0,761	-0,614	-1,047	-1,031	-0,600
OUTLAND	0,507	0,263	1,466	0,344	0,853
PTCRUISER	0,773	-0,408	0,105	0,344	0,811
SANTA_FE	-0,469	0,134	2,094	0,915	0,011

Exemple : Données qualitatives normalisées (1)

Modele	origine_Autres	origine_Europe	origine_France	carburant_Diesel	burant_Esser	4X4_non	4X4_oui
GOLF	0	1	0	1	0	1	0
CITRONC4	0	0	1	1	0	1	0
P607	0	0	1	1	0	1	0
VELSATIS	0	0	1	1	0	1	0
CITRONC2	0	0	1	0	1	1	0
CHRY300	1	0	0	0	1	1	0
AUDIA3	0	1	0	0	1	1	0
OUTLAND	1	0	0	1	0	0	1
PTCRUISER	1	0	0	0	1	1	0
SANTA_FE	1	0	0	1	0	0	1
Nj	4	2	4	6	4	8	2
Pj	0,4	0,2	0,4	0,6	0,4	0,8	0,2


Exemple : Données qualitatives normalisées (2)

Modele	origine_Autres	origine_Europe	origine_France	carburant_Diesel	burant_Esso	4X4_non	4X4_oui
GOLF	0	1	0	1	0	1	0
CITRONC4	0	0	1	1	0	1	0
P607	0	0	1	1	0	1	0
VELSATIS	0	0	1	1	0	1	0
CITRONC2	0	0	1	0	1	1	0
CHRY300	1	0	0	0	1	1	0
AUDIA3	0	1	0	0	1	1	0
OUTLAND	1	0	0	1	0	0	1
PTCRUISER	1	0	0	0	1	1	0
SANTA_FE	1	0	0	1	0	0	1
Nj	4	2	4	6	4	8	2
Pj	0,4	0,2	0,4	0,6	0,4	0,8	0,2



Modele	origine_Autres	origine_Europe	origine_France	carburant_Diesel	burant_Esso	4X4_non	4X4_oui
GOLF	0,000	2,236	0,000	1,291	0,000	1,118	0,000
CITRONC4	0,000	0,000	1,581	1,291	0,000	1,118	0,000
P607	0,000	0,000	1,581	1,291	0,000	1,118	0,000
VELSATIS	0,000	0,000	1,581	1,291	0,000	1,118	0,000
CITRONC2	0,000	0,000	1,581	0,000	1,581	1,118	0,000
CHRY300	1,581	0,000	0,000	0,000	1,581	1,118	0,000
AUDIA3	0,000	2,236	0,000	0,000	1,581	1,118	0,000
OUTLAND	1,581	0,000	0,000	1,291	0,000	0,000	2,236
PTCRUISER	1,581	0,000	0,000	0,000	1,581	1,118	0,000
SANTA_FE	1,581	0,000	0,000	1,291	0,000	0,000	2,236

Exemple : Données préparées



The diagram above the table shows a red double-headed arrow labeled p_1 spanning the first 6 columns (quantitative variables). Another red double-headed arrow labeled $m = m_1 + m_2 + m_3$ spans the last 8 columns (qualitative variables).

Modele	puissance	longueur	hauteur	poids	CO2	origine_Autres	origine_Europe	origine_France	carburant_Diesel	carburant_Essence	4X4_non	4X4_oui
GOLF	-1,103	-0,614	-0,419	-0,989	-1,127	0,000	2,236	0,000	1,291	0,000	1,118	0,000
CITRONC4	-0,304	-0,485	-0,733	-0,411	-1,148	0,000	0,000	1,581	1,291	0,000	1,118	0,000
P607	0,532	1,192	-0,838	0,795	0,558	0,000	0,000	1,581	1,291	0,000	1,118	0,000
VELSATIS	-0,152	1,063	0,524	0,837	-0,179	0,000	0,000	1,581	1,291	0,000	1,118	0,000
CITRONC2	-1,280	-2,007	-0,628	-1,993	-1,169	0,000	0,000	1,581	0,000	1,581	1,118	0,000
CHRY300	2,256	1,476	-0,524	1,189	1,990	1,581	0,000	0,000	0,000	1,581	1,118	0,000
AUDIA3	-0,761	-0,614	-1,047	-1,031	-0,600	0,000	2,236	0,000	0,000	1,581	1,118	0,000
OUTLAND	0,507	0,263	1,466	0,344	0,853	1,581	0,000	0,000	1,291	0,000	0,000	2,236
PTCRUISER	0,773	-0,408	0,105	0,344	0,811	1,581	0,000	0,000	0,000	1,581	1,118	0,000
SANTA_FE	-0,469	0,134	2,094	0,915	0,011	1,581	0,000	0,000	1,291	0,000	0,000	2,236

- Z_1 = Variables quantitatives : $p_1 = 5$
- Z_2 = Variables qualitatives : $p_2 = 3$ avec $m_1 = 3$, $m_2 = 2$ et $m_3 = 2$
 $\implies m = 7$
- La matrice Z est de dimension $p = p_1 + m = 12$

Plan du cours

- 1 Introduction
- 2 Représentation des données
- 3 Réalisation d'une AFDM**
 - Principe de l'AFDM
 - Nombre d'axes à retenir
- 4 Analyse des variables et des individus
 - Représentation des individus
 - Analyse des variables

Rappel des objectifs de l'ACP et de l'ACM

- Objectif : de produire un premier axe factoriel F1 qui soit le plus lié possible avec les variables.
- Si la liaison n'est pas parfaite, un second axe factoriel est déterminé pour expliquer l'information résiduelle, non prise en compte par le premier et ainsi de suite...
- On garde les premiers axes factoriels selon certains critères (critère du coude, de Kaiser, ...) qui permettent d'avoir suffisamment d'information sur les données.

ACP

Avec l'analyse en composantes principales lorsque les variables sont toutes quantitatives, la variance restituée par le premier facteur est :

$$\lambda_1 = \sum_j r^2(F_1, X_j) \quad (1)$$

Où $r^2()$ est le carré du coefficient de corrélation linéaire entre la variable X_j et le facteur F_1

ACM

Avec l'analyse en correspondance multiples lorsque les variables sont toutes qualitatives, la variance restituée par le premier facteur est :

$$\lambda_1 = \sum_j \eta^2(F_1, X_j) \quad (2)$$

Où $\eta^2()$ est le carré du rapport de corrélation entre la variable X_j et le facteur F_1

Plan du cours

- 1 Introduction
- 2 Représentation des données
- 3 Réalisation d'une AFDM
 - Principe de l'AFDM
 - Nombre d'axes à retenir
- 4 Analyse des variables et des individus
 - Représentation des individus
 - Analyse des variables

AFDM

- L'AFDM peut être traitée de plusieurs manières.
- L'AFDM est une généralisation pour l'analyse factorielle des données mixtes
- Nous allons nous baser sur l'Approche de Jérôme Pagès (« Analyse factorielle des données mixtes », Revue de Statistique Appliquée.
- Consiste à utiliser une ACP normée

AFDM

- Consiste à définir un critère qui réunit ces spécifications pour traiter simultanément les deux types de variables :

$$\lambda_1 = \sum_j r^2(F_1, X_j) + \sum_j \eta^2(F_1, X_j) \quad (3)$$

puisque $(0 < r^2 < 1)$ et $(0 < \eta^2 < 1)$, les deux types de variables présent de manière équilibrée dans l'analyse, cet aspect est primordial pour la praticabilité de l'approche

AFDM

L'AFDM = la généralisation pour l'analyse factorielle des données mixtes

$$\lambda_1 = \underbrace{\sum_j r^2(F_1, X_j)}_{\text{Si}=0 \Rightarrow \text{ACM}} + \underbrace{\sum_j \eta^2(F_1, X_j)}_{\text{Si}=0 \Rightarrow \text{ACP}} \quad (4)$$

λ_1 on parle du premier axe, mais on peut généraliser sur tout les autres axes.

Plan du cours

- 1 Introduction
- 2 Représentation des données
- 3 Réalisation d'une AFDM
 - Principe de l'AFDM
 - Nombre d'axes à retenir
- 4 Analyse des variables et des individus
 - Représentation des individus
 - Analyse des variables

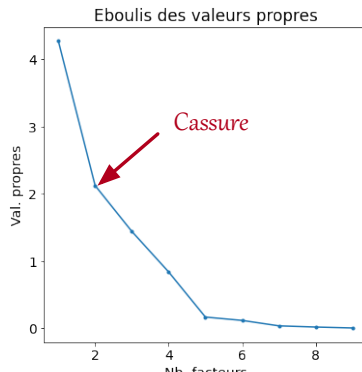
Nombre d'axes à retenir

- La transformation en TDC entraîne une redondance des données
 - ⇒ on garde au plus $p - p_2$ axes factoriels
 - ⇒ on garde $p - p_2$ valeurs propres non nulles
- Le critère de Kaiser utilisé dans l'ACP, ne peut pas être appliqué ici parce que certaines variables ne sont pas nativement quantitatives
- On utilise souvent le diagramme de coude.

Exemple : Nombre d'axes à retenir

- Dans notre cas : On garde aux plus $12 - 3 = 9$ axes factoriels.
- Avec les deux premiers axes factoriels, on préserve 71% de la variance cumulée.

Axes	Val.propre	% expliqué	% cumulé
1	4,2731	47,4793	47,4793
2	2,1219	23,5765	71,0558
3	1,4387	15,9858	87,0416
4	0,8364	9,2930	96,3345
5	0,1640	1,8226	98,1571
6	0,1145	1,2719	99,4290
7	0,0336	0,3736	99,8026
8	0,0158	0,1756	99,9782
9	0,0020	0,0218	100,0000



Plan du cours

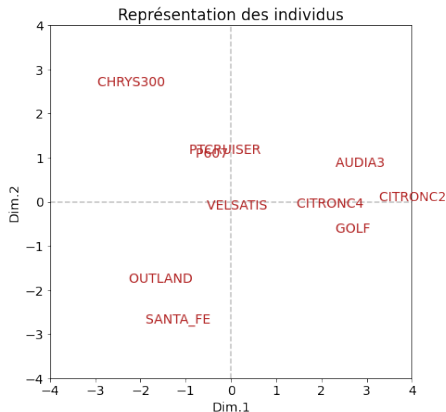
- 1 Introduction
- 2 Représentation des données
- 3 Réalisation d'une AFDM
 - Principe de l'AFDM
 - Nombre d'axes à retenir
- 4 Analyse des variables et des individus
 - Représentation des individus
 - Analyse des variables

Plan du cours

- 1 Introduction
- 2 Représentation des données
- 3 Réalisation d'une AFDM
 - Principe de l'AFDM
 - Nombre d'axes à retenir
- 4 Analyse des variables et des individus
 - Représentaion des individus
 - Analyse des variables

Exemple : Représentation factorielle des individus

Modele	Coord.F1	Coord.F2
GOLF	2.317805	-0.687271
CITRONC4	1.445368	-0.122286
P607	-0.779731	1.018563
VELSATIS	-0.541063	-0.160059
CITRONC2	3.275864	0.025117
CHRY300	-2.957705	2.628112
AUDIA3	2.316133	0.795341
OUTLAND	-2.255919	-1.840426
PTCRUISER	-0.931809	1.090852
SANTA_FE	-1.888943	-2.747943



Plan du cours

- 1 Introduction
- 2 Représentation des données
- 3 Réalisation d'une AFDM
 - Principe de l'AFDM
 - Nombre d'axes à retenir
- 4 Analyse des variables et des individus
 - Représentation des individus
 - Analyse des variables

Analyse simultanée des variables quantitatives et qualitatives

- La représentation des variables quantitatives et qualitatives dans le même repère est un vrai défi et pose une complexité de calcul.
- Par contre, la représentation des corrélations entre les différentes variables et les axes factoriels peut être une solution : les carrés des corrélations (variables quantitatives) et des rapports de corrélation (qualitatives) avec les facteurs.

Analyse simultanée des variables quantitatives et qualitatives

- L'AFDM est la généralisation pour l'analyse factorielle des deux types de variables :

$$\lambda_1 = \sum_j r^2(F_1, X_j) + \sum_j \eta^2(F_1, X_j) \quad (5)$$

avec $(0 < r^2 < 1)$ et $(0 < \eta^2 < 1)$

- Les deux indicateurs d'intensité de relation varient entre 0 et 1
 - ⇒ Leurs valeurs sont comparables
 - ⇒ On peut évaluer les contributions (Ctr) des variables et leurs qualité de représentation (Cos^2)

Le carré du rapport de corrélation (1)

- Le carré du rapport de corrélation caractérise, pour chaque variable, la dispersion relative de ses modalités.
- Il se définit par le ratio entre la variance inter-modalités et la variance totale.
- Pour la variable j , le carré du rapport de corrélation sur le facteur h est donné par :

$$\eta^2(F_h, X_j) = \sum_{k \in X_j} G_{kh}^2 \quad (6)$$

avec G_{kh} les coordonnées de la modalité k sur l'axe factoriel h

Le carré du rapport de corrélation (2)

- G_{kh} les coordonnées de la modalité k sur l'axe factoriel h est donné par :

$$G_{kh}^2 = \mu_{kh}^2 \times \frac{p_k}{\lambda_h} \quad (7)$$

avec λ_h la variance associée à l'axe factoriel h
 p_k la proportion de la modalité k .

Contributions des variables

- La contribution d'une variable quantitative est donnée :

$$Ctr_j(F_h) = \frac{r^2(F_h, X_j)}{\lambda_h} \quad (8)$$

avec λ_h la variance associée à l'axe factoriel h

- La contribution d'une variable qualitative est donnée :

$$Ctr_j(F_h) = \frac{\eta^2(F_h, X_j)}{\lambda_h} \quad (9)$$

Qualité de représentations

- La qualité de représentations (Cos^2 ou bien Qlt) d'une variable quantitative est donnée :

$$Cos_j^2(F_h) = r^2(F_h, X_j) \quad (10)$$

avec λ_h la variance associée à l'axe factoriel h

- La qualité de représentations d'une variable qualitative est donnée :

$$Cos_j^2(F_h) = \frac{\eta^2(F_h, X_j)}{m_j - 1} \quad (11)$$

Exemple : Analyse des variables

- $\lambda_1 = 4,27$ et $\lambda_2 = 2,12$

	Axe factoriel 1			Axe factoriel 2		
Variable	Cor	CTR(%)	Qlt (%)	Cor	CTR(%)	Qlt (%)
puissance	0,6713	15,7	67	0,2907	13,7	29
longueur	0,6339	14,8	63	0,0600	2,8	6
hauteur	0,3345	7,8	33	0,5831	27,5	58
poids	0,8640	20,2	86	0,0037	0,2	0
CO2	0,7933	18,6	79	0,1436	6,8	14
origine	0,6966	16,3	70	0,0160	0,8	1
carburant	0,0283	0,7	3	0,4046	19,1	40
4X4	0,2513	5,9	25	0,6201	29,2	62

Quantitatives: r^2

Qualitatives: η^2

- Vérification :

$$\lambda_1 = \sum_j r^2(F_1, X_j) + \sum_j \eta^2(F_1, X_j) = 4,27 \quad (12)$$

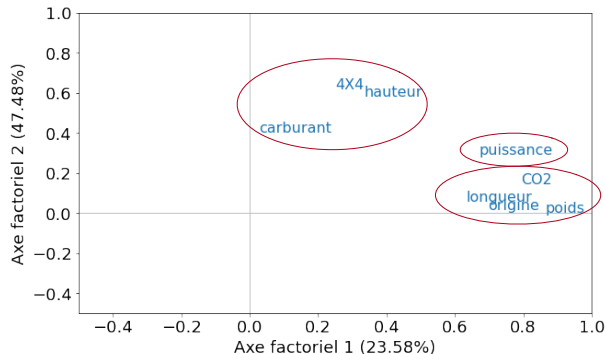
Exemple : Analyse des variables

- Les variables (Poids, CO2, Origine, Puissance, Longueur) présentent une corrélation élevée avec le 1er axe factoriel (Inertie expliquée 47,4% diapo 27).
- Pour le 2ème axe (4X4, Hauteur, Carburant, Puissance) sont les plus déterminants.

	Axe factoriel 1			Axe factoriel 2		
Variable	Cor	CTR(%)	Qlt (%)	Cor	CTR(%)	Qlt (%)
puissance	0,6713	15,7	67	0,2907	13,7	29
longueur	0,6339	14,8	63	0,0600	2,8	6
hauteur	0,3345	7,8	33	0,5831	27,5	58
poids	0,8640	20,2	86	0,0037	0,2	0
CO2	0,7933	18,6	79	0,1436	6,8	14
origine	0,6966	16,3	70	0,0160	0,8	1
carburant	0,0283	0,7	3	0,4046	19,1	40
4X4	0,2513	5,9	25	0,6201	29,2	62

Exemple : Analyse des variables

- Les variables (Poids, CO2, Origine, Longueur) présentent une forte liaison
- Les variables (4×4, Hauteur, Carburant) sont aussi liées et forment un cluster



Exemple : Analyse des variables et des individus

Modele	puissance	longueur	hauteur	poids	CO2	origine	carburant	4X4
GOLF	75	421	149	1217	143	Europe	Diesel	non
CITRONC4	138	426	146	1381	142	France	Diesel	non
P607	204	491	145	1723	223	France	Diesel	non
VELSATIS	150	486	158	1735	188	France	Diesel	non
CITRONC2	61	367	147	932	141	France	Essence	non
CHRY300	340	502	148	1835	291	Autres	Essence	non
AUDIA3	102	421	143	1205	168	Europe	Essence	non
OUTLAND	202	455	167	1595	237	Autres	Diesel	oui
PTCRUISER	223	429	154	1595	235	Autres	Essence	non
SANTA_FE	125	450	173	1757	197	Autres	Diesel	oui

