Projet - Sélection de modèle en régression linéaire

Code

Université de Bordeaux - M1 Master MAS - UE : Modèles de régression

Année 2024-2025

Nom : LE CAMUS
Prénom : Arthur

Nom : DOSSOU
Prénom : Félicia

Préambule

Dans ce projet, il est proposé de s'intéresser à l'analyse d'un jeu de données réelles disponible à l'URL : https://www.openlab.psu.edu/ansur2/.

Il s'agit du jeu de données `ANSUR II` (Anthropometric Survey of US Army Personnel) relatives à des caractéristiques de taille et forme du corps humain. Il s'agit d'un ensemble de 93 mesures effectuées sur plus de 6 000 militaires américains adultes, dont 4 082 hommes et 1 986 femmes. On va proposer de choisir la variable du **poids** et chercher à l'expliquer en fonction des $p = 92$ autres variables du jeu de données à l'aide d'un modèle linéaire. Le but étant de trouver un sous-ensemble de variables qui permettent de bien expliquer le poids d'individu.

Sommaire

1. Analyse descriptive

```
1.1 Prétraitement des données
1.2 Exploration initiale
1.3 Réduction de la dimension
1.4 Clustering des variables
1.5 Visualisation des relations
```

2. Sélection de modèle

```
2.1 Hommes
  2.1.1 Backward
    - Réduction initiale avec step
    - Génération de combinaisons de variables
    - Évaluation des critères
    - Résultats
    - Conclusion
  2.1.2 Forward
2.2 Femmes
  2.2.1 Backward
  2.2.2 Forward
```

---

1 / Analyse descriptive

1.1 Prétraitement des données

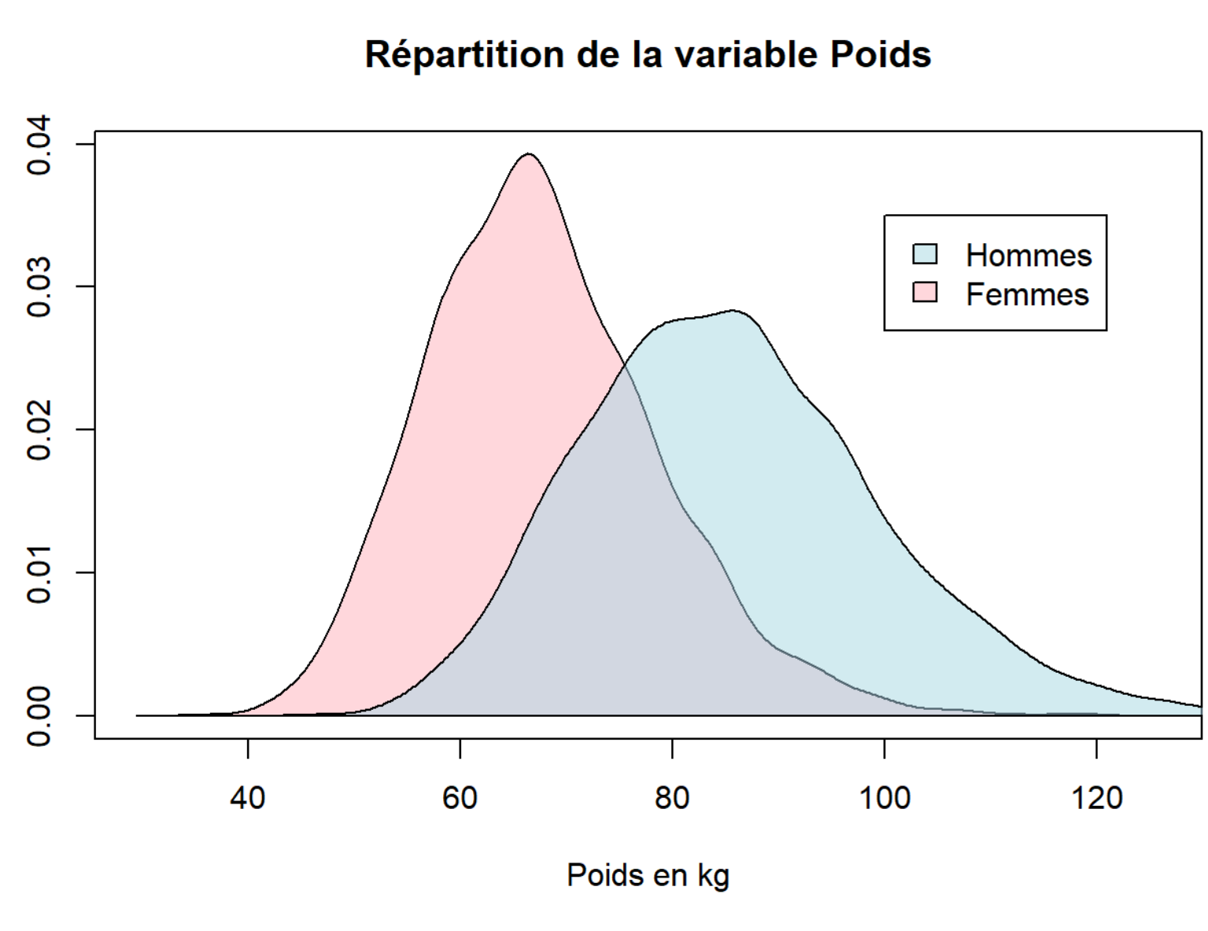
On charge les données anthropométriques ANSUR pour les hommes et les femmes, on sélectionne les variables explicatives, puis on isole le poids comme variable cible en la convertissant en kilogrammes.

Show

1.2 Exploration initiale

La répartition du poids est étudiée à l'aide de courbes de densité pour les hommes et les femmes, permettant de comparer leurs distributions respectives.

Show



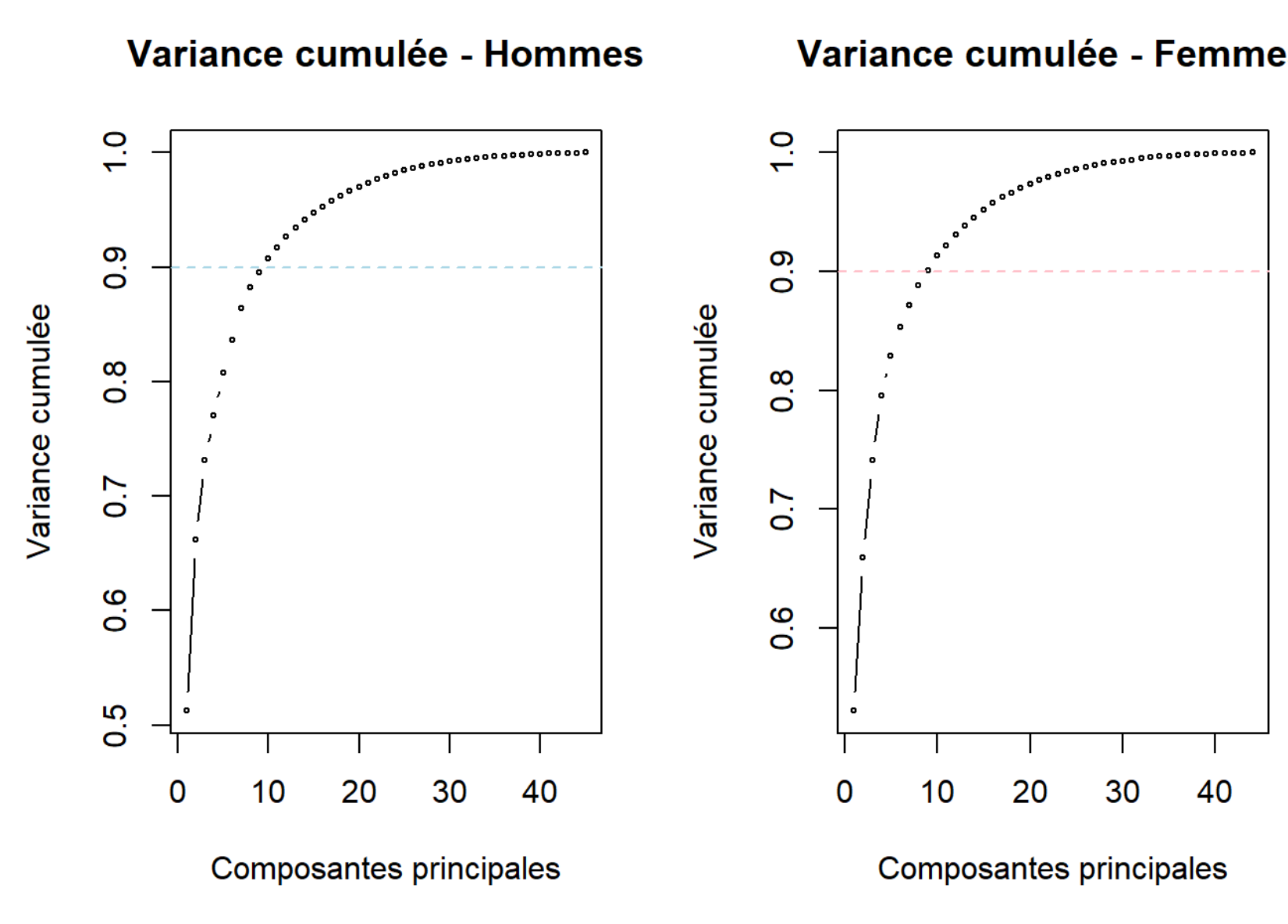
1.3 Réduction de la dimension

On élimine les variables fortement corrélées entre elles (corrélation > 0.8) pour obtenir des matrices réduites (`X_m_reduced` et `X_f_reduced`).Pour chaque paire de variables fortement corrélées, on conserve celle ayant la plus forte corrélation avec le poids et on élimine l'autre. La dimension passe ainsi de 92 variables à 45 pour les hommes et 44 pour les femmes. Ce sont ces matrices qu'on utilisera pour la sélection de modèle.

Show

Nous réalisons une analyse en composantes principales (ACP) sur les données centrées des hommes et des femmes afin de calculer la variance expliquée par chaque composante et déterminer le nombre minimal de composantes nécessaires pour conserver 90 % de la variance.

Show



Show

1.4 Clustering des variables

Il faut garder 11 composantes principales pour les hommes et 11 pour les femmes pour expliquer 90% de la variance.

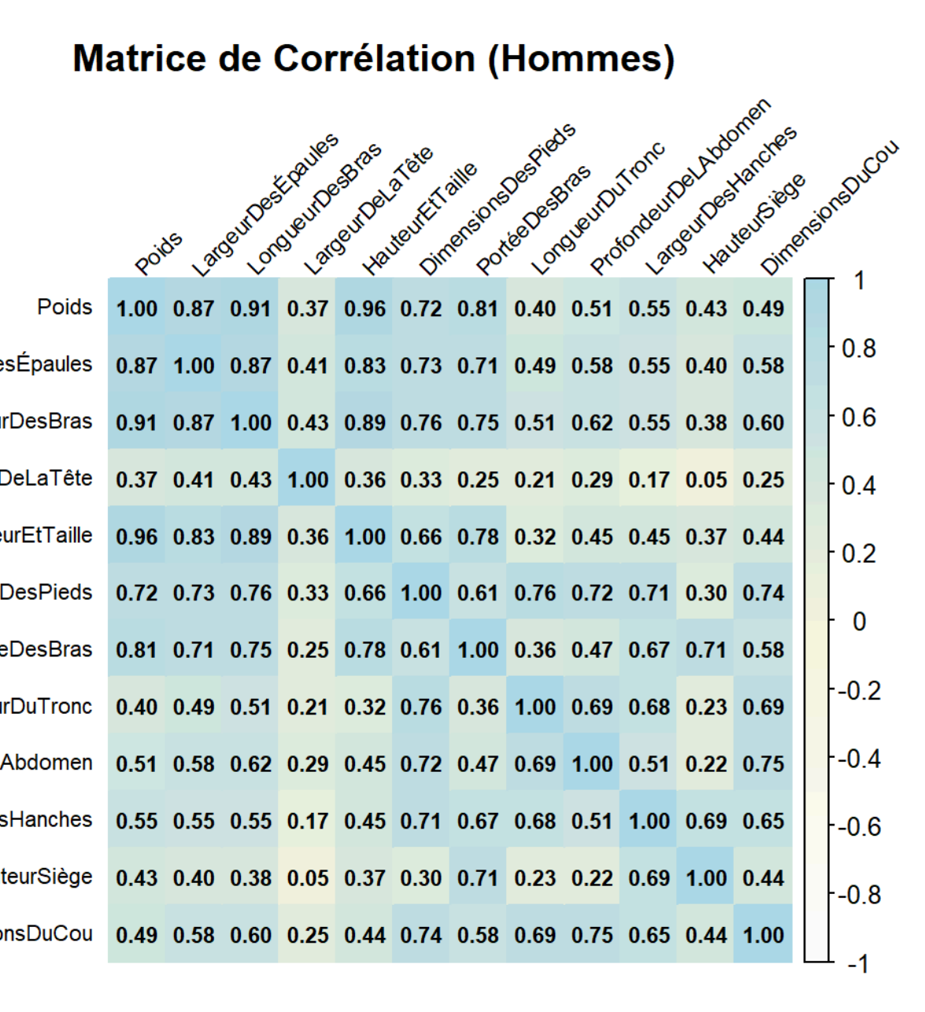
Nous regroupons les variables issues de l'ACP en clusters pour les hommes et les femmes en identifiant la composante principale qui représente le mieux chaque variable. Pour cela, nous calculons les distances de projection afin de prioriser les variables les mieux projetées dans le nouvel espace. Enfin, nous attribuons des noms significatifs à chaque cluster pour en faciliter l'interprétation.

Show

1.5 Visualisation des relations

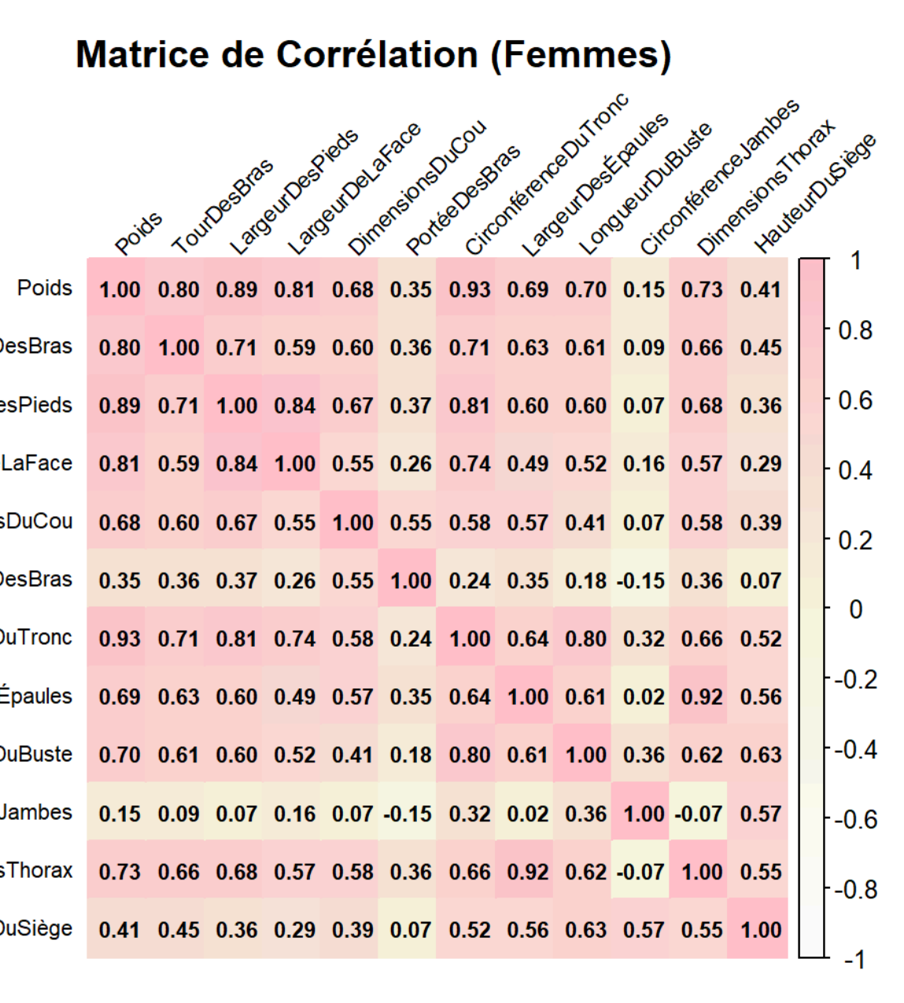
Les variables ont été regroupées par cluster, et leur moyenne a été calculée pour chaque ligne afin de créer un dataframe synthétique. Ce dataframe permet d'analyser les interactions entre clusters, notamment via une **matrice de corrélation**.

Show


Matrice de Corrélation (Hommes)

	Poids	LargeurDesÉpaules	LongueurDesBras	LargeurDeLaTête	HauteurEtTaille	DimensionsDesPieds	PortéeDesBras	LongueurDuTronc	ProfondeurDeLAbdomen	LargeurDesHanches	HauteurSiège	DimensionsDuCou
Poids	1.00	0.87	0.91	0.37	0.96	0.72	0.81	0.40	0.51	0.55	0.43	0.49
LargeurDesÉpaules	0.87	1.00	0.87	0.41	0.83	0.73	0.71	0.49	0.58	0.55	0.40	0.58
LongueurDesBras	0.91	0.87	1.00	0.43	0.89	0.76	0.75	0.51	0.62	0.55	0.38	0.60
LargeurDeLaTête	0.37	0.41	0.43	1.00	0.36	0.33	0.25	0.21	0.29	0.17	0.05	0.25
HauteurEtTaille	0.96	0.83	0.89	0.36	1.00	0.66	0.78	0.32	0.45	0.45	0.37	0.44
DimensionsDesPieds	0.72	0.73	0.76	0.33	0.66	1.00	0.61	0.76	0.72	0.71	0.30	0.74
PortéeDesBras	0.81	0.71	0.75	0.25	0.78	0.61	1.00	0.36	0.47	0.67	0.71	0.58
LongueurDuTronc	0.40	0.49	0.51	0.21	0.32	0.76	0.36	1.00	0.69	0.68	0.23	0.69
ProfondeurDeLAbdomen	0.51	0.58	0.62	0.29	0.45	0.72	0.47	0.69	1.00	0.51	0.22	0.75
LargeurDesHanches	0.55	0.55	0.55	0.17	0.45	0.71	0.67	0.68	0.51	1.00	0.69	0.65
HauteurSiège	0.43	0.40	0.38	0.05	0.37	0.30	0.71	0.23	0.22	0.69	1.00	0.44
DimensionsDuCou	0.49	0.58	0.60	0.25	0.44	0.74	0.58	0.69	0.75	0.65	0.44	1.00

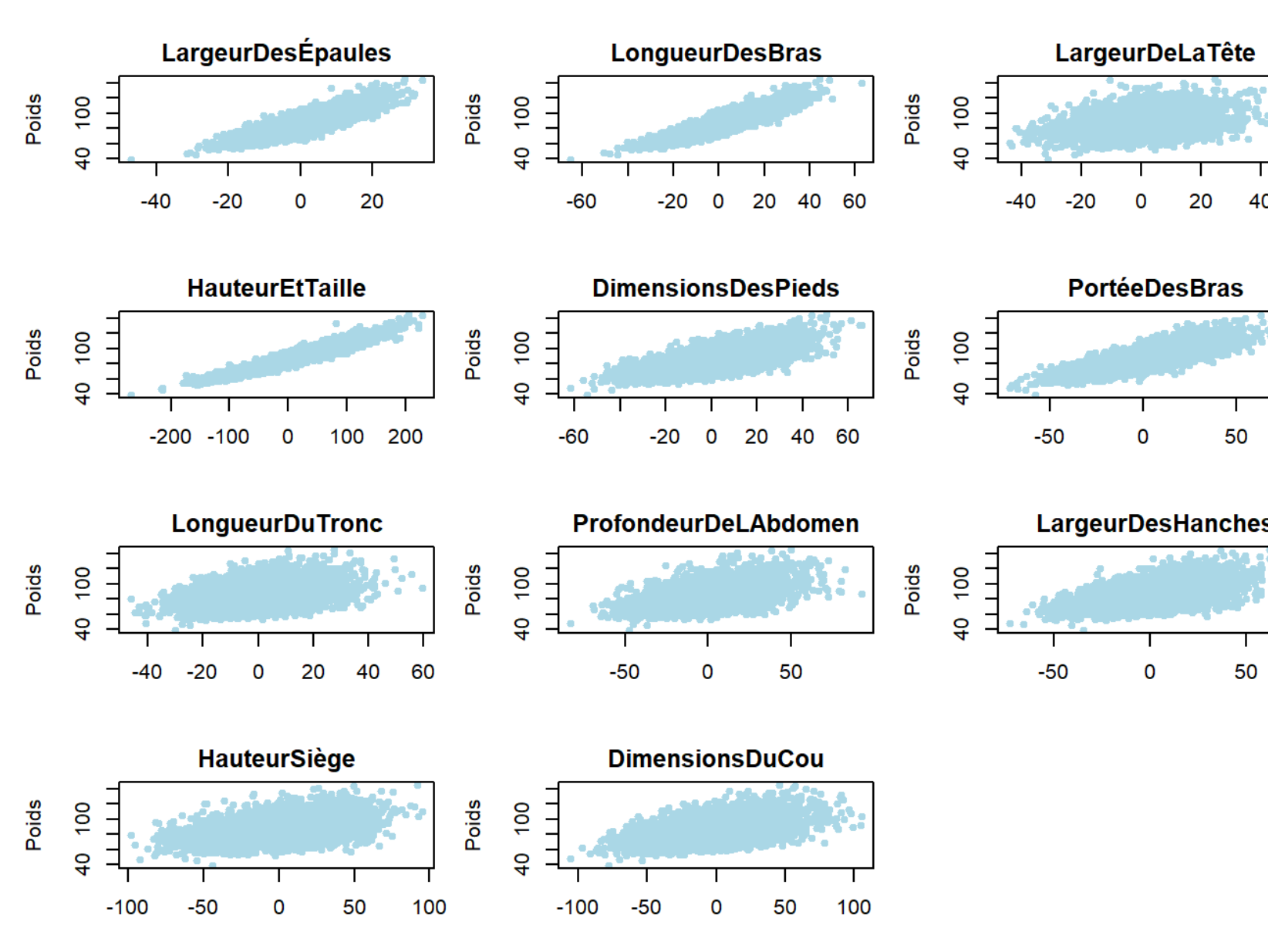
Show


Matrice de Corrélation (Femmes)

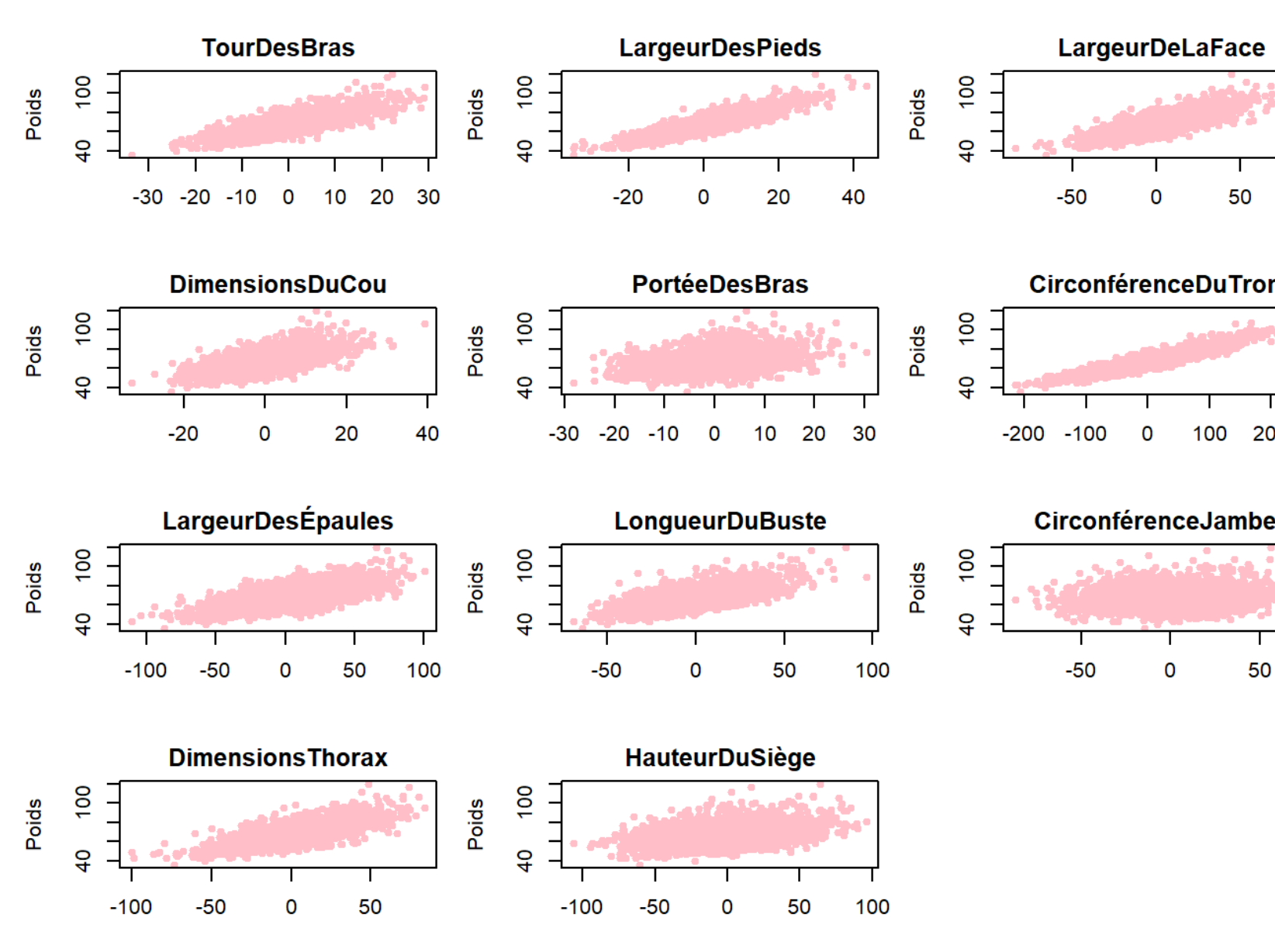
	Poids	TourDesBras	LargeurDesPieds	LargeurDeLaFace	DimensionsDuCou	PortéeDesBras	CirconférenceDuTronc	LargeurDesÉpaules	LongueurDuBuste	CirconférenceJambes	DimensionsThorax	HauteurDuSiège
Poids	1.00	0.80	0.89	0.81	0.68	0.35	0.93	0.69	0.70	0.15	0.73	0.41
TourDesBras	0.80	1.00	0.71	0.59	0.60	0.36	0.71	0.63	0.61	0.09	0.66	0.45
LargeurDesPieds	0.89	0.71	1.00	0.84	0.67	0.37	0.81	0.60	0.60	0.07	0.68	0.36
LargeurDeLaFace	0.81	0.59	0.84	1.00	0.55	0.26	0.74	0.49	0.52	0.16	0.57	0.29
DimensionsDuCou	0.68	0.60	0.67	0.55	1.00	0.55	0.58	0.57	0.41	0.07	0.58	0.39
PortéeDesBras	0.35	0.36	0.37	0.26	0.55	1.00	0.24	0.35	0.18	-0.15	0.36	0.07
CirconférenceDuTronc	0.93	0.71	0.81	0.74	0.58	0.24	1.00	0.64	0.80	0.32	0.66	0.52
LargeurDesÉpaules	0.69	0.63	0.60	0.49	0.57	0.35	0.64	1.00	0.61	0.02	0.92	0.56
LongueurDuBuste	0.70	0.61	0.60	0.52	0.41	0.18	0.80	0.61	1.00	0.36	0.62	0.63
CirconférenceJambes	0.15	0.09	0.07	0.16	0.07	-0.15	0.32	0.02	0.36	1.00	-0.07	0.57
DimensionsThorax	0.73	0.66	0.68	0.57	0.58	0.36	0.66	0.92	0.62	-0.07	1.00	0.55
HauteurDuSiège	0.41	0.45	0.36	0.29	0.39	0.07	0.52	0.56	0.63	0.57	0.55	1.00

Pour obtenir une visualisation plus détaillée de la relation entre chaque cluster synthétique et le poids , nous avons créé une grille de graphiques de scatter plots. Cela permet d'identifier des tendances spécifiques ou des outliers.

Show



Show



---

2 / Sélection de modèle

Pour sélectionner les variables explicatives les plus pertinentes, nous avons commencé par appliquer une réduction automatique via la fonction STEP.

Dans le modèle réduit obtenu, toutes les p-values associées aux coefficients étaient très faibles, indiquant que les variables restantes étaient significatives. Par conséquent, il n'était pas possible d'éliminer d'autres variables sur la base de leur manque de significativité.

Nous avons donc généré toutes les combinaisons possibles de 4 variables issues du modèle réduit et évalué chaque combinaison selon différents critères : SCR, R² ajusté, Cp de Mallow, AIC, BIC et PRESS. Pour analyser ces résultats, nous avons visualisé les valeurs obtenues pour chaque critère sur l'ensemble des combinaisons à l'aide d'histogrammes.

Cependant, le processus devient très long lorsque le nombre de variables augmente. Par exemple, pour sélectionner 4 variables parmi 34, cela implique de tester $\binom{34}{4}$ = 46 376 combinaisons possibles. De plus, le calcul du critère PRESS prend trop de temps car il nécessite n ajustements de modèles en laissant une observation de côté à chaque itération. Du fait de son coût, on a cherché une alternative sur internet, basée sur les leviers de la matrice de projection pour éviter de recalculer le modèle à chaque itération. Voici le site sur lequel on a trouvé la formule : Statology

2.1 Hommes

Show

```
Le modèle initial contient 45 variables.
```

Le modèle initial contient 45 variables.

2.1.1 Backward

Show

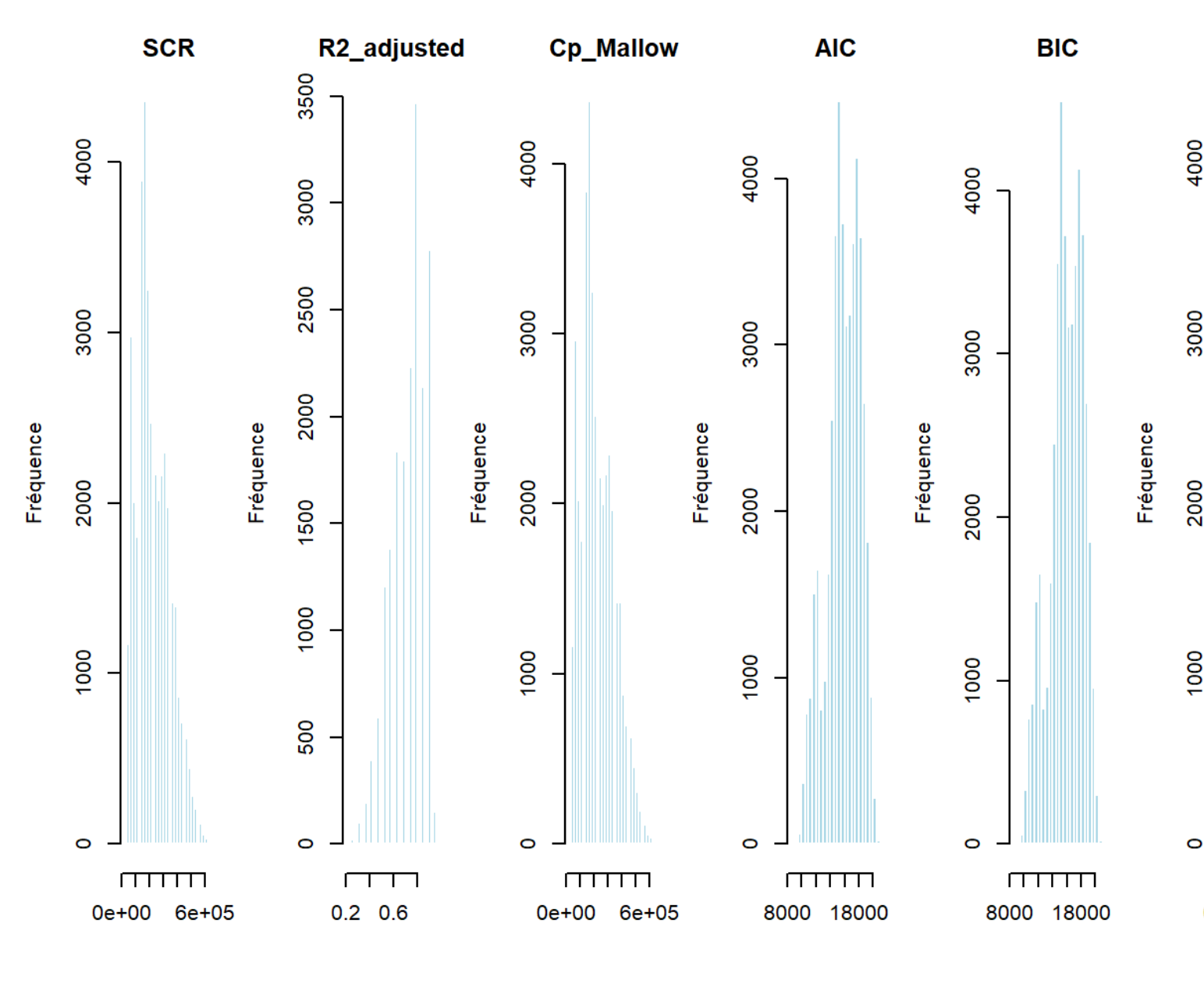
Show

```
Le modèle après STEP BACKWARD contient 34 variables.
```

Show

```
Meilleur modèle selon SCR = 35445.93 avec les variables : buttockcircumference, chestdepth, shouldercircumferenc
e, wristheight
 Meilleur modèle selon R² ajusté = 0.9570084 avec les variables : buttockcircumference, chestdepth, shouldercircu
mference, wristheight
 Meilleur modèle selon Cp de Mallow = 35532.87 avec les variables : buttockcircumference, chestdepth, shouldercir
cumference, wristheight
 Meilleur modèle selon AIC = 8832.922 avec les variables : buttockcircumference, chestdepth, shouldercircumferenc
e, wristheight
 Meilleur modèle selon BIC = 8864.493 avec les variables : buttockcircumference, chestdepth, shouldercircumferenc
e, wristheight
 Meilleur modèle selon PRESS = 35550.22 avec les variables : buttockcircumference, chestdepth, shouldercircumfere
nce, wristheight
```

Show



2.1.2 Forward

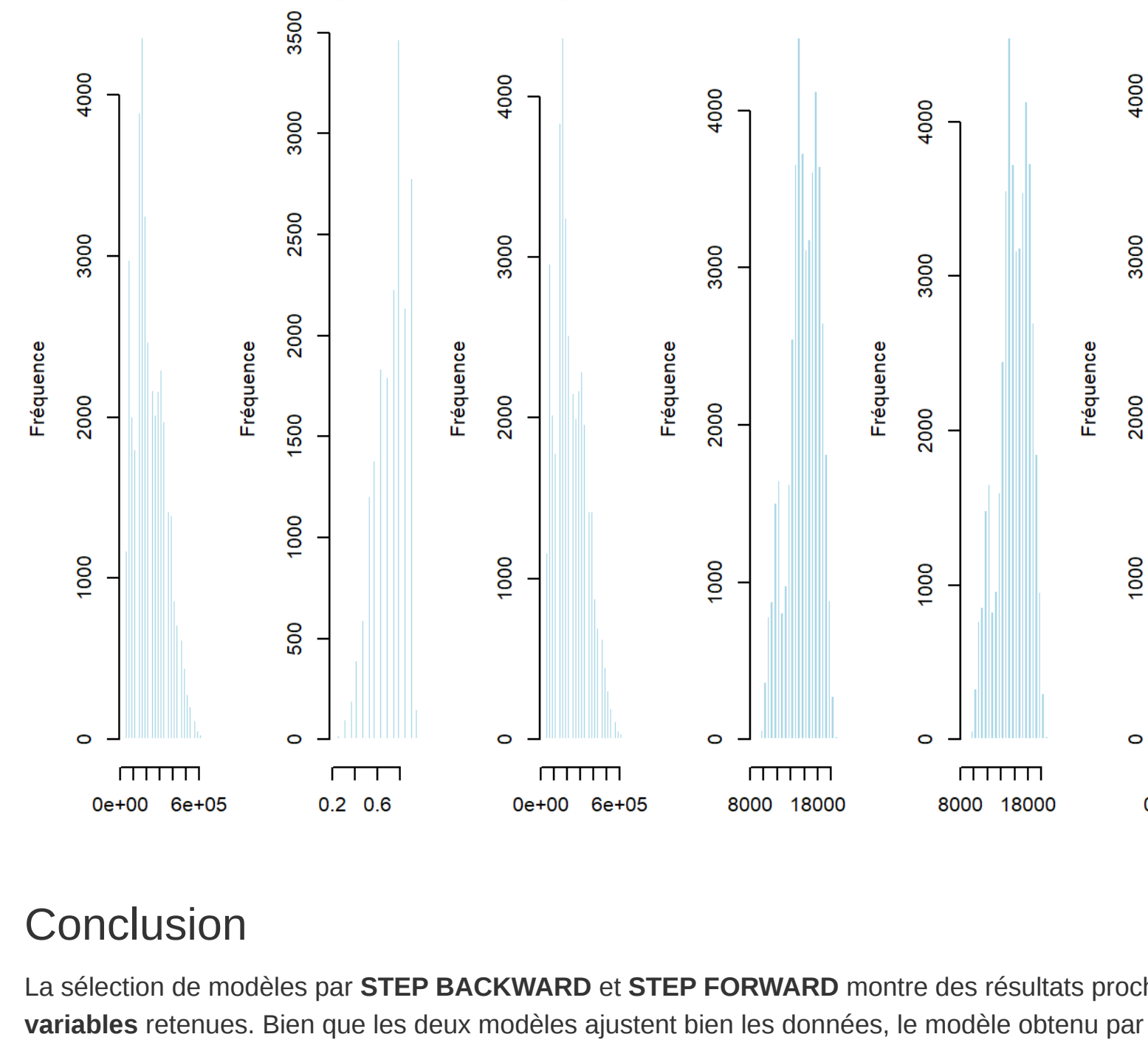
Show

```
Le modèle après STEP FORWARD contient 34 variables.
```

Show

```
Meilleur modèle selon SCR = 35445.93 avec les variables : buttockcircumference, shouldercircumference, wristheigh
t, chestdepth
 Meilleur modèle selon R² ajusté = 0.9570084 avec les variables : buttockcircumference, shouldercircumference, wr
istheight, chestdepth
 Meilleur modèle selon Cp de Mallow = 35532.87 avec les variables : buttockcircumference, shouldercircumference,
wristheight, chestdepth
 Meilleur modèle selon AIC = 8832.922 avec les variables : buttockcircumference, shouldercircumference, wristheig
ht, chestdepth
 Meilleur modèle selon BIC = 8864.493 avec les variables : buttockcircumference, shouldercircumference, wristheig
ht, chestdepth
 Meilleur modèle selon PRESS = 35550.22 avec les variables : buttockcircumference, shouldercircumference, wristhe
ight, chestdepth
```

Show



Conclusion

La sélection de modèles par **STEP BACKWARD** et **STEP FORWARD** montre des résultats proches, avec respectivement **34 variables** et **34 variables** retenues. Bien que les deux modèles ajustent bien les données, le modèle obtenu par STEP FORWARD présente des performances légèrement meilleures en termes de SCR, R² ajusté, Cp de Mallows, AIC, BIC et PRESS. Les variables clés identifiées dans les deux approches incluent principalement **circonférence des fesses**, **profondeur de la poitrine**, **circonférence des épaules** et **hauteur de poignet**. Le modèle réduit obtenu constitue ainsi un bon compromis entre la précision et la complexité.

2.2 Femmes

Show

```
Le modèle initial contient 44 variables.
```

2.1.1 Backward

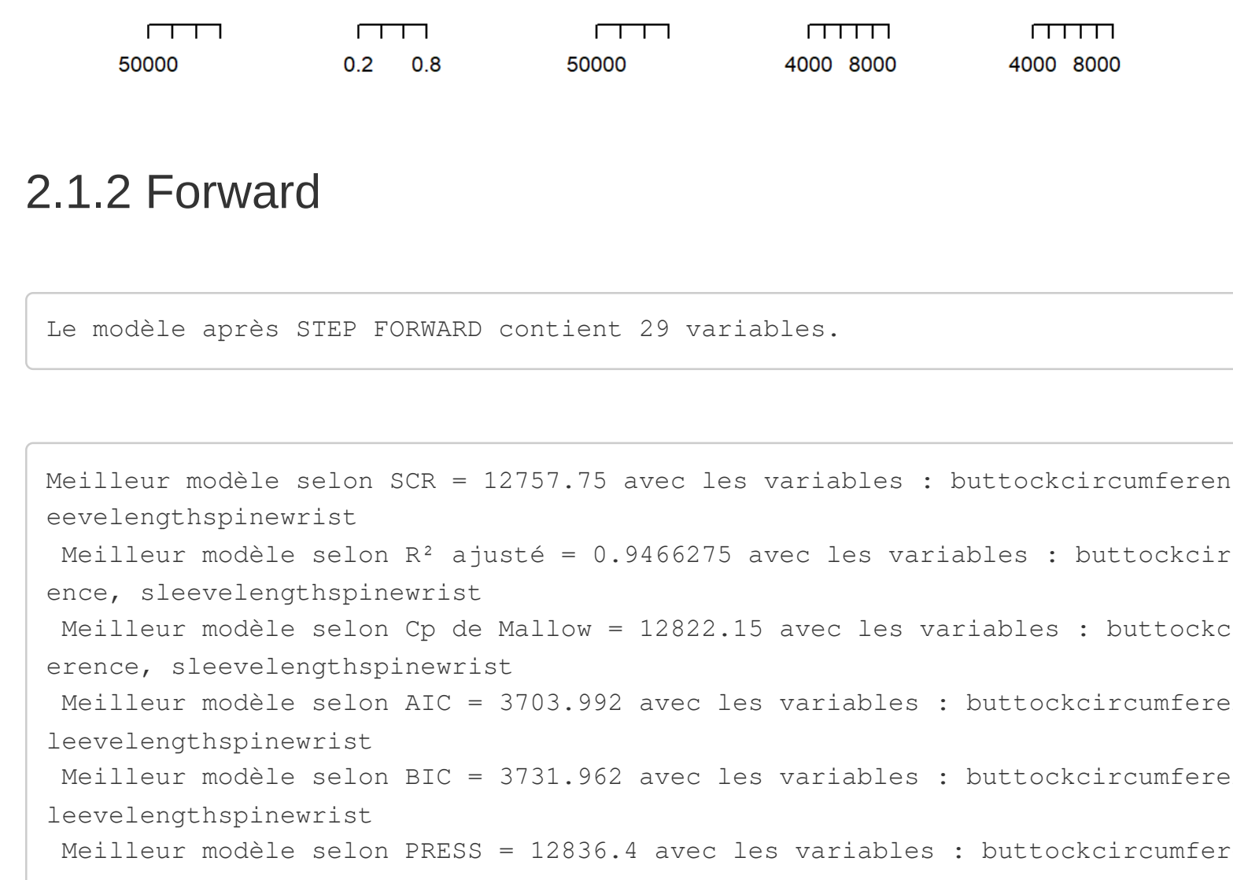
Show

```
Le modèle après STEP BACKWARD contient 31 variables.
```

Show

```
Meilleur modèle selon SCR = 12757.75 avec les variables : buttockcircumference, calfcircumference, chestdepth, sl
eevelengthspinewrist
 Meilleur modèle selon R² ajusté = 0.9466275 avec les variables : buttockcircumference, calfcircumference, chestd
epth, sleevelengthspinewrist
 Meilleur modèle selon Cp de Mallow = 12822.15 avec les variables : buttockcircumference, calfcircumference, ches
tdepth, sleevelengthspinewrist
 Meilleur modèle selon AIC = 3703.992 avec les variables : buttockcircumference, calfcircumference, chestdepth, s
leevelengthspinewrist
 Meilleur modèle selon BIC = 3731.962 avec les variables : buttockcircumference, calfcircumference, chestdepth, s
leevelengthspinewrist
 Meilleur modèle selon PRESS = 12836.4 avec les variables : buttockcircumference, calfcircumference, chestdepth,
sleevelengthspinewrist
```

Show



2.1.2 Forward

Show

```
Le modèle après STEP FORWARD contient 29 variables.
```

Show

```
Meilleur modèle selon SCR = 12757.75 avec les variables : buttockcircumference, chestdepth, calfcircumference, sl
eevelengthspinewrist
 Meilleur modèle selon R² ajusté = 0.9466275 avec les variables : buttockcircumference, chestdepth, calfcircumfer
ence, sleevelengthspinewrist
 Meilleur modèle selon Cp de Mallow = 12822.15 avec les variables : buttockcircumference, chestdepth, calfcircumf
erence, sleevelengthspinewrist
 Meilleur modèle selon AIC = 3703.992 avec les variables : buttockcircumference, chestdepth, calfcircumference, s
leevelengthspinewrist
 Meilleur modèle selon BIC = 3731.962 avec les variables : buttockcircumference, chestdepth, calfcircumference, s
leevelengthspinewrist
 Meilleur modèle selon PRESS = 12836.4 avec les variables : buttockcircumference, chestdepth, calfcircumference,
sleevelengthspinewrist
```

Show

Conclusion

La sélection de modèles par **STEP BACKWARD** et **STEP FORWARD** montre des résultats proches, avec respectivement **31 variables** et **29 variables** retenues. Bien que les deux modèles ajustent bien les données, le modèle obtenu par **STEP FORWARD** présente des performances légèrement meilleures en termes de SCR, R² ajusté, Cp de Mallows, AIC, BIC et PRESS. Les variables clés identifiées dans les deux approches incluent principalement **circonférence des fesses**, **profondeur de la poitrine**, **circonférence du mollet** et **longueur de l'avant bras**.

Pour aller plus loin, on pourrait utiliser une méthode heuristique, comme vu l'année dernière en cours de Techniques Algorithmiques, pour réduire le nombre de combinaisons à tester. L'algo, en s'approchant séquentiellement de la solution optimale, permettrait d'éviter l'exploration exhaustive des combinaisons inutiles?