Mise en oeuvre de l'ACP

Angelina Roche

Executive Master Statistique et Big Data

2018–2019

Plan du cours

Plan

Rappel dernier cours

- L'ACP est une méthode de réduction de la dimension.
- Objectifs :
  - visualiser les données,
  - pouvoir appliquer des méthodes normalement réservées à la faible dimension.
- Les axes principaux sont ceux maximisant la variance projetée.
- Ce sont les vecteurs propres associés aux plus grandes valeurs propres de la matrice de covariance/corrélation.

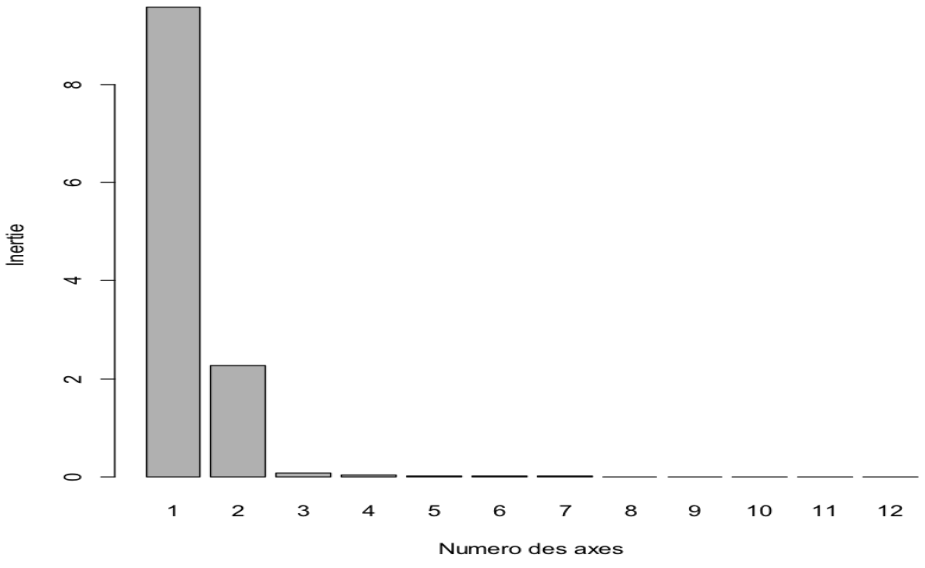
Rappel dernier cours

- L'ACP est une méthode de réduction de la dimension.
- Objectifs :
  - visualiser les données,
  - pouvoir appliquer des méthodes normalement réservées à la faible dimension.
- Les axes principaux sont ceux maximisant la variance projetée.
- Ce sont les vecteurs propres associés aux plus grandes valeurs propres de la matrice de covariance/corrélation.
- Objectifs du cours d'aujourd'hui : mettre en place et interpréter une ACP.

Exemple illustratif : température moyenne de différentes villes

	Janv	Fevr	Mars	Avril	Mai	Juin	Juil	Aout	Sept	Oct	N
Bordeaux	5.60	6.60	10.30	12.80	15.80	19.30	20.90	21.00	18.60	13.80	9.
Brest	6.10	5.80	7.80	9.20	11.60	14.40	15.60	16.00	14.70	12.00	9.
Clermont	2.60	3.70	7.50	10.30	13.80	17.30	19.40	19.10	16.20	11.20	6.
Grenoble	1.50	3.20	7.70	10.60	14.50	17.80	20.10	19.50	16.70	11.40	6.
Lille	2.40	2.90	6.00	8.90	12.40	15.30	17.10	17.10	14.70	10.40	6.
Lyon	2.10	3.30	7.70	10.90	14.90	18.50	20.70	20.10	16.90	11.40	6.
Marseille	5.50	6.60	10.00	13.00	16.80	20.80	23.30	22.80	19.90	15.00	10.
Montpellier	5.60	6.70	9.90	12.80	16.20	20.10	22.70	22.30	19.30	14.60	10.
Nantes	5.00	5.30	8.40	10.80	13.90	17.20	18.80	18.60	16.40	12.20	8.
Nice	7.50	8.50	10.80	13.30	16.70	20.10	22.70	22.50	20.30	16.00	11.
Paris	3.40	4.10	7.60	10.70	14.30	17.50	19.10	18.70	16.00	11.40	7.
Rennes	4.80	5.30	7.90	10.10	13.10	16.20	17.90	17.80	15.70	11.60	7.
Strasbourg	0.40	1.50	5.60	9.80	14.00	17.20	19.00	18.30	15.10	9.50	4.
Toulouse	4.70	5.60	9.20	11.60	14.90	18.70	20.90	20.90	18.30	13.30	8.
Vichy	2.40	3.40	7.10	9.90	13.60	17.10	19.30	18.80	16.00	11.00	6.

Choix du nombre d'axes



- Le premier axe explique à lui seul presque 80% de la variance.
- Les deux premiers axes expliquent plus de 98% de la variance.

  $\hookrightarrow$ bonne représentation des données dans le plan engendré par les deux premiers axes.

Plan

Projection du nuage des individus

- Les axes de l'ACP $v_1, ..., v_K$ sont des éléments de $\mathbb{R}^p$
- k-ème axe de l'ACP :

$$v_k = \begin{pmatrix} v_k^1 \\ \vdots \\ v_k^p \end{pmatrix}.$$

- $s_i^k = \tilde{x}_i v_k = \sum_{j=1}^p \tilde{x}_i^j v_k^j$: coordonnée du i-ème individu par rapport à l'axe k, où $\tilde{x}_i^j = (x_i^j - \bar{x}^j)/\sigma_j$ (ACP normée) ou $\tilde{x}_i^j = x_i^j - \bar{x}^j$ (ACP non normée).

Coordonnées des villes sur les premiers axes

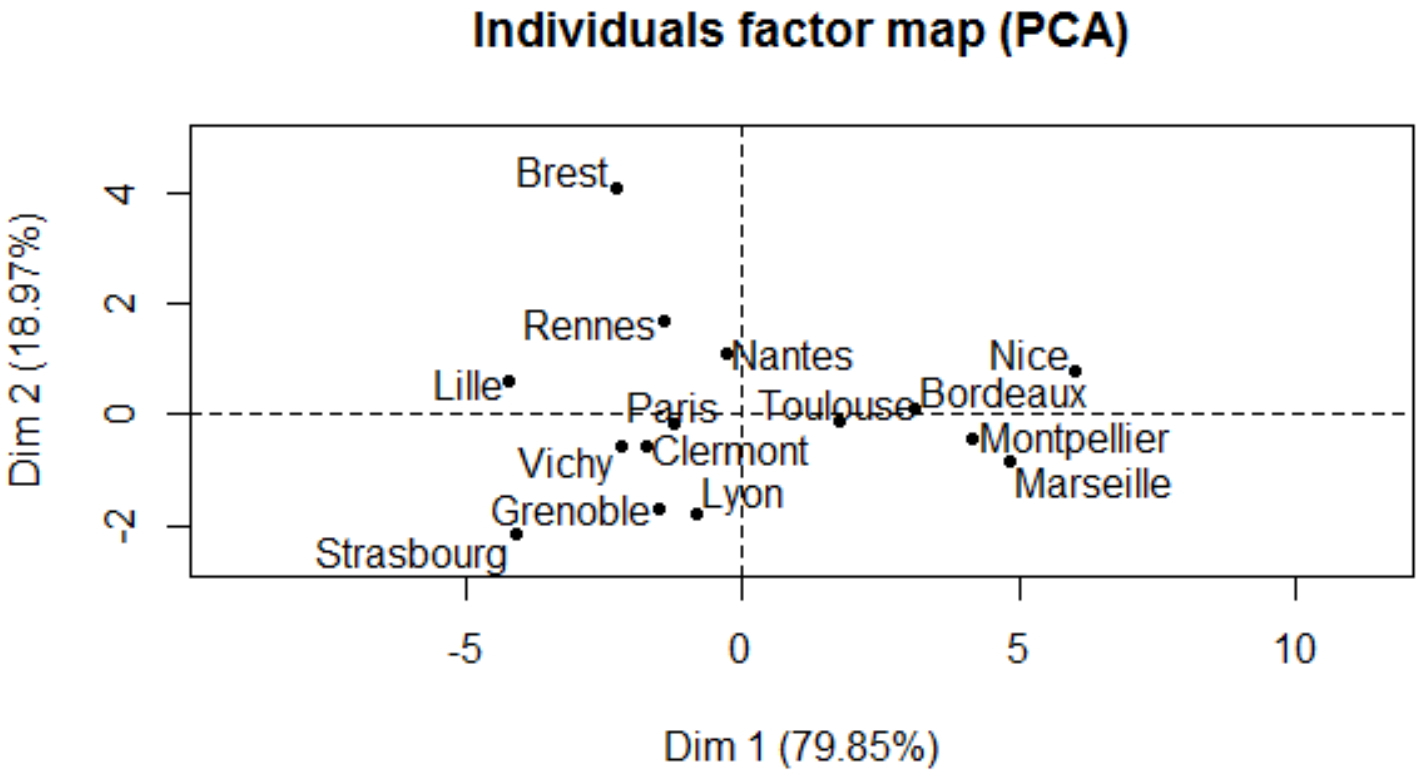


Figure – Représentation du nuage N_I projeté sur les deux premiers axes principaux (tracé des points (s_i^1, s_i^2), $i = 1, ..., n$).

Composantes principales

- $s^k = (s_1^k, ..., s_n^k)$: composante principale $\hookrightarrow$ assimilable à une variable.

$$\sum_{i=1}^{n} p_i s_i^k = \sum_{i=1}^{n} p_i \tilde{x}_i^t v_k = \left(\sum_{i=1}^{n} p_i \tilde{x}_i \right)^t v_k = 0$$

$\Rightarrow$ les composantes principales sont centrées.

Composantes principales (II)

- Soient

$$S = \begin{pmatrix} s_1^1 & \dots & s_1^p \\ \vdots & \ddots & \vdots \\ s_n^1 & \dots & s_n^p \end{pmatrix}, P = \begin{pmatrix} v_1^1 & \dots & v_1^p \\ \vdots & \ddots & \vdots \\ v_n^1 & \dots & v_n^p \end{pmatrix}, \tilde{X} = \begin{pmatrix} \tilde{x}_1^1 & \dots & \tilde{x}_1^p \\ \vdots & \ddots & \vdots \\ \tilde{x}_n^1 & \dots & \tilde{x}_n^p \end{pmatrix},$$

  et $\Pi = \text{diag}(p_1, ..., p_n)$.

- Par définition : $S = \tilde{X}P$, d'où

$$S^t \Pi S = P^t \tilde{X}^t \Pi \tilde{X} P = P^t C P = \text{diag}(\lambda_1, ..., \lambda_p).$$

$$\Rightarrow \lambda_k = \sum_{i=1}^n p_i (s_i^k)^2, \quad \sum_{i=1}^n p_i s_i^j s_i^k = 0 \text{ si } j \neq k.$$

Composantes principales (II)

- Soient

$$S = \begin{pmatrix} s_1^1 & \dots & s_1^p \\ \vdots & \ddots & \vdots \\ s_n^1 & \dots & s_n^p \end{pmatrix}, P = \begin{pmatrix} v_1^1 & \dots & v_1^p \\ \vdots & \ddots & \vdots \\ v_n^1 & \dots & v_n^p \end{pmatrix}, \tilde{X} = \begin{pmatrix} \tilde{x}_1^1 & \dots & \tilde{x}_1^p \\ \vdots & \ddots & \vdots \\ \tilde{x}_n^1 & \dots & \tilde{x}_n^p \end{pmatrix},$$

  et $\Pi = \text{diag}(p_1, ..., p_n)$.

- Par définition : $S = \tilde{X}P$, d'où

$$S^t \Pi S = P^t \tilde{X}^t \Pi \tilde{X} P = P^t C P = \text{diag}(\lambda_1, ..., \lambda_p).$$

$\Rightarrow \lambda_k = \sum_{i=1}^n p_i (s_i^k)^2, \quad \sum_{i=1}^n p_i s_i^j s_i^k = 0$ si $j \neq k$.

  - La variance de la k-ème composante est égale à λ_k.
  - Les composantes principales sont décorrelées.

Représentation des variables

- Corrélation de la variable $\tilde{x}^j$ par rapport à la k-ème composante principale s^k :

$$\text{cor}(\tilde{x}^j, s^k) = \frac{1}{n} \sum_{i=1}^{n} p_i \tilde{x}_i^j \frac{s_i^k}{\sigma(s^k)} \text{ où } \sigma^2(s^k) = \sum_{i=1}^{n} p_i (s_i^k)^2 = \lambda_k.$$

- Rappels :
  - $-1 \leq \text{cor}(\tilde{x}^j, s^k) \leq 1$,
  - Plus $|\text{cor}(\tilde{x}^j, s^k)|$ proche de 1, plus on considèrera que la variable j est liée à l'axe k.
  - $|\text{cor}(\tilde{x}^j, s^k)| < 0$: corrélation négative,
  - $|\text{cor}(\tilde{x}^j, s^k)| > 0$: corrélation positive.

Cercles des corrélations

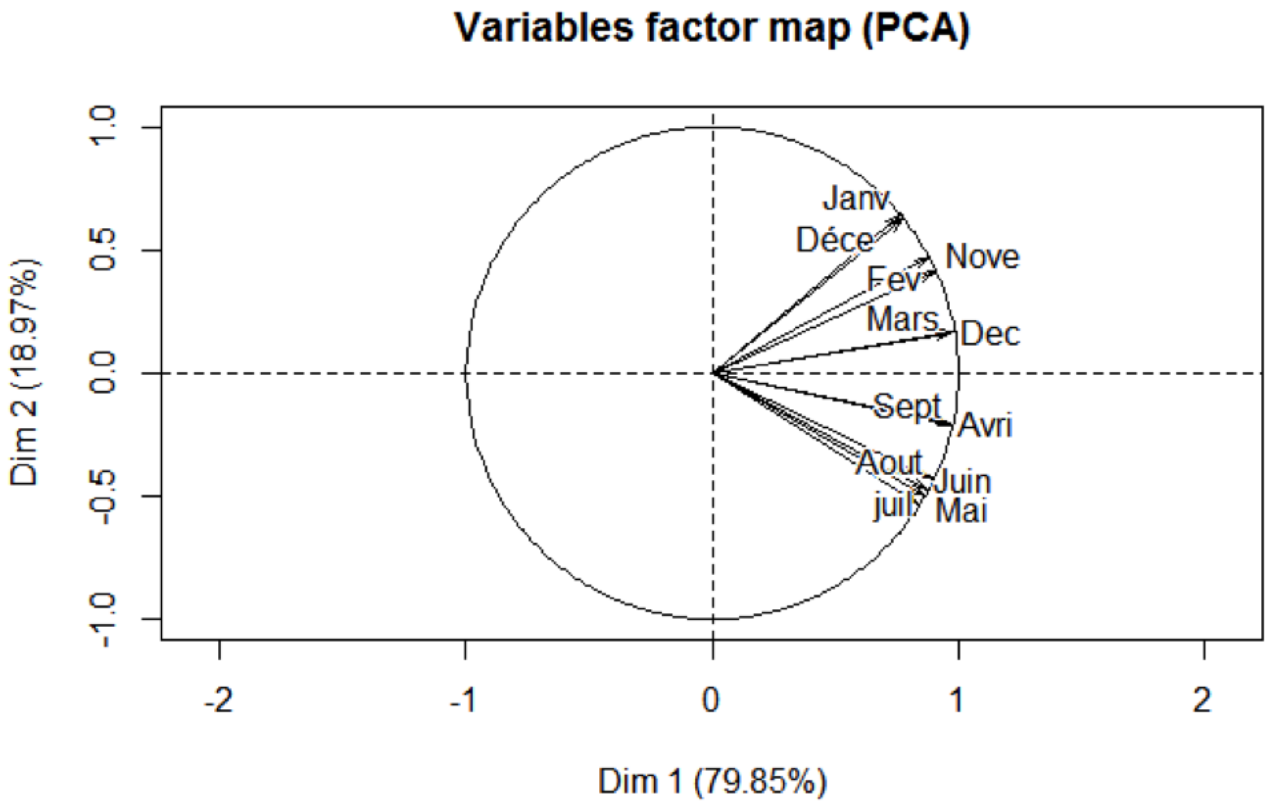


Figure – Représentation des corrélations sous la forme d'un cercle. Chaque flèche pointe sur le point de coordonnées $(\text{cor}(\tilde{x}^j, s^1), \text{cor}(\tilde{x}^j, s^2))$, $j = 1, ..., p$.

Plan

Contribution d'un individu à l'inertie d'un axe

- Rappel :

$$\lambda_k = \sum_{i=1}^{n} p_i (s_k^i)^2.$$

- Contribution de l'individu i à l'inertie de l'axe k :

$$\text{ctr}(i,k) = \frac{p_i (s_k^i)^2}{\lambda_k}.$$

- Lorsque les individus ne sont pas anonymes, ceux ayant une contribution importante (par exemple $> 1/n$) peuvent aider à l'interprétation des axes.
- Attention aux individus ayant une contribution trop importante ($> 25\%$).

Indices de contribution – données températures

```
res.pca = PCA(temp,quanti.sup=13:16)
res.pca$ind$contrib[,1:2]
```

	Dim.1	Dim.2
Bordeaux	6.78	0.03
Brest	3.58	49.07
Clermont	2.07	1.03
Grenoble	1.63	8.34
Lille	12.37	1.04
Lyon	0.49	9.36
Marseille	16.25	2.01
Montpellier	11.97	0.56
Nantes	0.06	3.64
Nice	25.11	1.82
Paris	1.07	0.07
Rennes	1.44	8.18
Strasbourg	11.73	13.82
Toulouse	2.10	0.05
Vichy	3.37	0.97

Qualité de représentation d'un individu

- Nous avons : $\text{dist}(0, \tilde{x}_i)^2 = \sum_{j=1}^{p} (s_i^j)^2$.
- Qualité de représentation de l'individu i sur l'axe j :

$$Q(i,k) = \frac{(s_i^j)^2}{\text{dist}(0, \tilde{x}_i)^2}.$$

- On appelle parfois cet indice *cosinus carré*.

Qualité de représentation – données température

```
res.pca$ind$cos2[,1:2]
```

	Dim.1	Dim.2
Bordeaux	0.95	0.00
Brest	0.23	0.76
Clermont	0.88	0.10
Grenoble	0.43	0.52
Lille	0.97	0.02
Lyon	0.18	0.82
Marseille	0.96	0.03
Montpellier	0.99	0.01
Nantes	0.06	0.89
Nice	0.98	0.02
Paris	0.89	0.01
Rennes	0.42	0.57
Strasbourg	0.78	0.22
Toulouse	0.95	0.01
Vichy	0.92	0.06

Contribution et qualité de représentation d'une variable

- Contribution de la variable j à l'inertie de l'axe k :

$$\text{ctr}(j,k) = \frac{\text{cor}(\tilde{x}^j, s^k)^2}{\sum_{\ell=1}^{p} \text{cor}(\tilde{x}^\ell, s^k)^2}.$$

- Qualité de représentation de la variable j sur l'axe k :

$$Q(j,k) = \frac{\text{cor}(\tilde{x}^j, s^k)^2}{\sum_{\ell=1}^{p} \text{cor}(\tilde{x}^j, s^\ell)^2}.$$

Contribution et qualité de représentation d'une variable – données températures

`res.pca$var$contrib[,1:2]`

	Dim.1	Dim.2
Janv	6.05	18.24
Fevr	8.09	9.67
Mars	9.79	1.07
Avril	9.81	1.82
Mai	7.95	9.90
Juin	7.78	10.95
Juil	7.39	12.41
Aout	8.43	8.12
Sept	9.90	1.90
Oct	10.03	1.28
Nov	8.52	7.53
Dec	6.26	17.12

`res.pca$var$cos2[,1:2]`

	Dim.1	Dim.2
Janv	0.58	0.42
Fevr	0.78	0.22
Mars	0.94	0.02
Avril	0.94	0.04
Mai	0.76	0.23
Juin	0.75	0.25
Juil	0.71	0.28
Aout	0.81	0.18
Sept	0.95	0.04
Oct	0.96	0.03
Nov	0.82	0.17
Dec	0.60	0.39

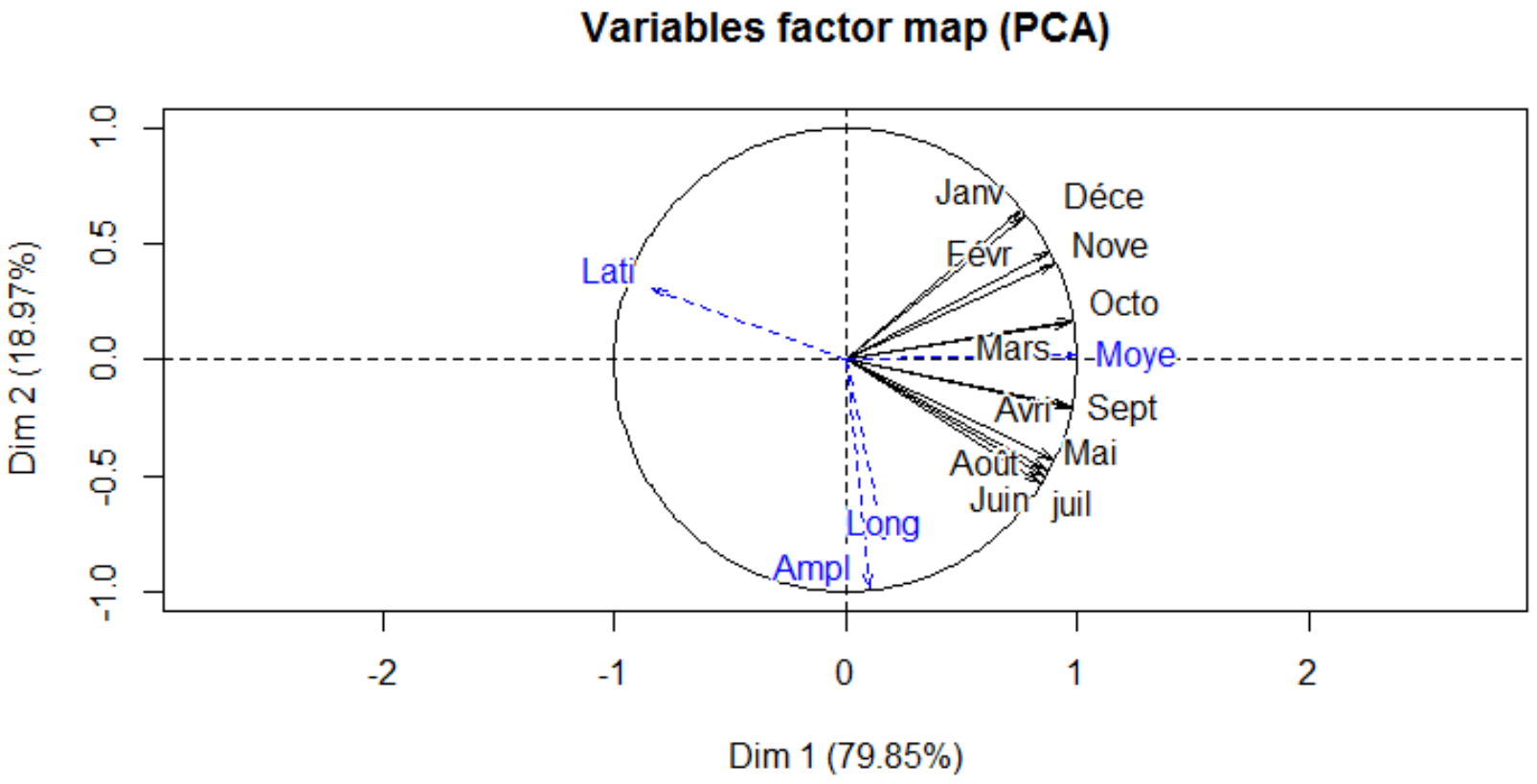
Règles empiriques

- Un individu ayant une mauvaise représentation tout en ayant une contribution importante sera écarté de l'analyse (individu supplémentaire).
- On ne représente que les points ayant une qualité de représentation *acceptable*.
- Deux individus à la fois proches et bien représentés sur un même axe sont réellement proches.

Variables supplémentaires

- Utilité : variables construite à partir d'autres variables mais pouvant aider à l'interprétation ou variables quantitatives supplémentaires.
- Variables quantitatives : ajout sur le cercle des corrélations.
- Variables qualitatives : ajout dans le nuage des individus (coloration des individus en fonction des modalités par exemple).

Données température



Individus supplémentaires

- Utilité : individus ayant une contribution trop importante, ou dont on doute de la fiabilité, nouvelle étude,....
- Ajout dans le nuage des individus.

Individu supplémentaire

On ajoute à l'analyse les données concernant la ville de Dijon.