Final - AD50
UTBM – A2017
abderrahim.chariete@utbm.fr

Durée : 2 heures
Documents autorisés

Exercice 1 : (3 pts) Droites de régressions

Soit deux variables ***x*** et ***y*** correspondant respectivement au poids et à la taille de 10 personnes.

Q1. **(1 pt)** Déduire les colonnes de 3 à 7 du tableau ci-dessous :

	1	2	3	4	5	6	7
	Poids (x)	**Tailles (y)**	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})^2$	$(y - \bar{y})^2$	$(x - \bar{x})(y - \bar{y})$
1	48	155					
2	54	163					
3	57	160					
4	63	161					
5	65	157					
6	73	162					
7	85	165					
8	87	169					
9	96	187					
10	102	190					

Q2. **(1 pt)** Calculer les métriques descriptives suivantes :

Métriques descriptives		***x***	***y***
Moyenne	: $\bar{x}$		
Variance	: $Var(x)$		
Ecart-type	: S_x		
Covariance	: $Cov(x, y)$		
Cœfficient de corrélation	: r_{xy}		

Q3. **(1 pt)** Calculer la pente ***a*** et l'ordonnée à l'origine ***b*** des droites de régressions Dy/x et Dx/y.

Droites de régressions	***Pente : a***	***Ordonnée à l'origine : b***
Dy/x		
Dx/y		

Exercice 2 : (5 pts) Valeurs propres et vecteurs propres.

Soit M la matrice réelle 3 x 3 suivante : $M = \begin{bmatrix} 0 & 2 & -1 \\ 3 & -2 & 0 \\ -2 & 2 & 1 \end{bmatrix}$

Q1. **(1.5 pt)** Calculer le polynôme caractéristique de M.
Q2. **(1.5 pt)** Déterminer les valeurs propres de M.
Q3. **(2 pts)** Déterminer les espaces associés au valeurs propres trouvées (vecteurs propres).

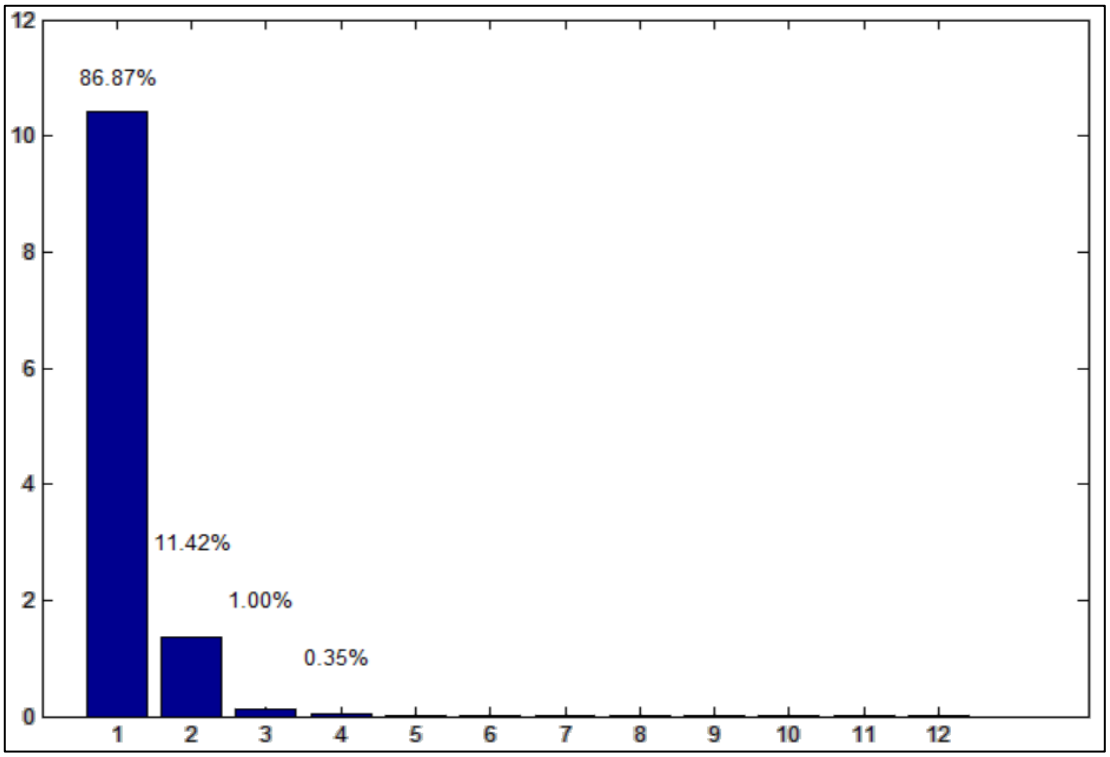
Problème : (12 pts) ACP, CAH, *k*-means et Silhouettes

Nous souhaitons analyser des données de températures moyennes sur 40 ans relevées dans 35 grandes villes en Europe :

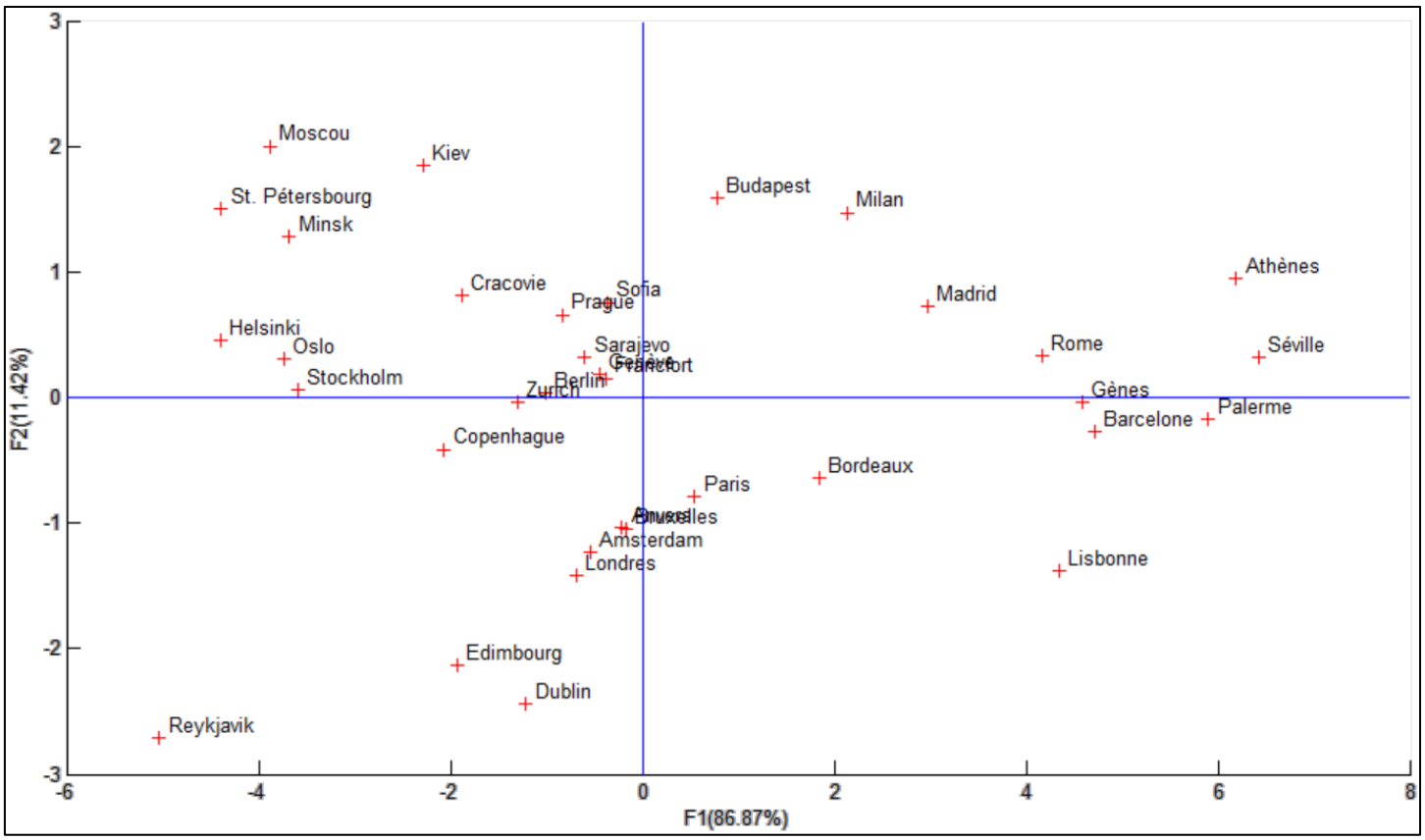
	Jan.	Fév.	Mars	Avr.	Mai	Juin	Juil.	Aout	Sep.	Oct.	Nov.	Déc.	Moyenne	Amplitude	Latitude	Longitude	Région
St. Pétersbourg	-8,2	-7,9	-3,7	3,2	10	15,4	18,4	16,9	11,5	5,2	-0,4	-5,3	4,5	26,6	59,6	30,2	Est
Moscou	-9,3	-7,6	-2	6	13	16,6	18,3	16,7	11,2	5,1	-1,1	-6	5,1	27,6	46,2	1,5	Est
Minsk	-6,9	-6,2	-1,9	5,4	12,4	15,9	17,4	16,3	11,6	5,8	0,1	-4,2	5,5	24,3	53,5	27,3	Est
Kiev	-5,9	-5	-0,3	7,4	14,3	17,8	19,4	18,5	13,7	7,5	1,2	-3,6	7,1	25,3	50,3	30,3	Est
Cracovie	-3,7	-2	1,9	7,9	13,2	16,9	18,4	17,6	13,7	8,6	2,6	-1,7	7,7	22,1	50	19,6	Est
Prague	-1,3	0,2	3,6	8,8	14,3	17,6	19,3	18,7	14,9	9,4	3,8	0,3	9,2	20,6	50	14,2	Est
Sofia	-1,7	0,2	4,3	9,7	14,3	17,7	20	19,5	15,8	10,7	5	0,6	9,6	21,7	42,4	23,2	Est
Budapest	-1,1	0,8	5,5	11,6	17	20,2	22	21,3	16,9	11,3	5,1	0,7	10,9	23,1	47,3	19	Est
Reykjavik	-0,3	0,1	0,8	2,9	6,5	9,3	11,1	10,6	7,9	4,5	1,7	0,2	4,6	11,4	64,1	21,6	Nord
Helsinki	-5,8	-6,2	-2,7	3,1	10,2	14	17,2	14,9	9,7	5,2	0,1	-2,3	4,8	23,4	60,1	25	Nord
Oslo	-4,3	-3,8	-0,6	4,4	10,3	14,9	16,9	15,4	11,1	5,7	0,5	-2,9	5,6	21,2	59,5	10,5	Nord
Stockholm	-3,5	-3,5	-1,3	3,5	9,2	14,6	17,2	16	11,7	6,5	1,7	-1,6	5,8	20,7	59,2	18	Nord
Copenhague	-0,4	-0,4	1,3	5,8	11,1	15,4	17,1	16,6	13,3	8,8	4,1	1,3	7,8	17,5	55,4	12,3	Nord
Edimbourg	2,9	3,6	4,7	7,1	9,9	13	14,7	14,3	12,1	8,7	5,3	3,7	8,3	11,8	55	3	Nord
Dublin	4,8	5	5,9	7,8	10,4	13,3	15	14,6	12,7	9,7	6,7	5,4	9,3	10,2	53,2	6,1	Nord
Londres	3,4	4,2	5,5	8,3	11,9	15,1	16,9	16,5	14	10,2	6,3	4,4	9,7	13,5	51,4	0	Nord
Zurich	-0,7	0,7	4,3	8,5	12,9	16,2	18	17,2	14,1	8,9	3,9	0,3	8,7	18,7	47,2	8,3	Ouest
Berlin	-0,2	0,1	4,4	8,2	13,8	16	18,3	18	14,4	10	4,2	1,2	9,1	18,5	52,3	13,2	Ouest
Genève	0,1	1,9	5,1	9,4	13,8	17,3	19,4	18,5	15	9,8	4,9	1,4	9,7	19,3	46,1	6,1	Ouest
Francfort	0,2	1,8	5,4	9,7	14,3	17,5	19	18,3	14,8	9,8	4,9	1,7	9,8	18,8	50,1	8,4	Ouest
Amsterdam	2,9	2,5	5,7	8,2	12,5	14,8	17,1	17,1	14,5	11,4	7	4,4	9,9	14,6	52,2	4,5	Ouest
Bruxelles	3,3	3,3	6,7	8,9	12,8	15,6	17,8	17,8	15	11,1	6,7	4,4	10,3	14,4	50,5	4,2	Ouest
Anvers	3,1	2,9	6,2	8,9	12,9	15,5	17,9	17,6	14,7	11,5	6,8	4,7	10,3	15	51,1	4,2	Ouest
Paris	3,7	3,7	7,3	9,7	13,7	16,5	19	18,7	16,1	12,5	7,3	5,2	11,2	15,3	48,5	2,2	Ouest
Bordeaux	5,6	6,7	9	11,9	15	18,3	20,4	20	17,6	13,5	8,5	6,1	12,7	14,8	44,5	0,3	Ouest
Sarajevo	-1,4	0,8	4,9	9,3	13,8	17	18,9	18,7	15,2	10,5	5,1	0,8	9,4	20,3	43,5	18,3	Sud
Milan	1,1	3,6	8	12,6	17,3	21,3	23,8	22,8	18,9	13,1	6,9	2,6	12,6	22,7	45,3	9,2	Sud
Madrid	5	6,6	9,4	12,2	16	20,8	24,7	24,3	19,8	13,9	8,7	5,4	13,9	19,7	40,2	3,4	Sud
Rome	7,1	8,2	10,5	13,7	17,8	21,7	24,4	24,1	20,9	16,5	11,7	8,3	15,4	17,3	41,5	12,3	Sud
Lisbonne	10,5	11,3	12,8	14,5	16,7	19,4	21,5	21,9	20,4	17,4	13,7	11,1	15,9	11,4	38,4	9,1	Sud
Gènes	8,7	8,7	11,4	13,8	17,5	21	24,5	24,6	21,8	17,8	12,2	10	16,1	15,9	44,3	9,4	Sud
Barcelone	9,1	10,3	11,8	14,1	17,4	21,2	24,2	24,1	21,7	17,5	13,1	10	16,2	15,1	41,2	2,2	Sud
Palerme	10,5	11,5	13,3	16,9	20,9	23,8	24,5	22,3	22,3	18,4	14,9	12	16,6	14	38,1	13,1	Sud
Athènes	9,1	9,7	11,7	15,4	20,1	24,5	27,4	27,2	23,8	19,2	14,6	11	17,8	18,3	37,6	23,5	Sud
Séville	10,7	11,8	14,1	16,1	19,7	23,4	26,7	26,7	24,3	19,4	14,5	11,2	18,2	16	37,2	5,6	Sud

On décide de synthétiser l'information disponible avec l'analyse en composantes principales ACP :

Q1. **(0,5 pt)** Combien de facteurs (composantes principales) sont nécessaires pour l'analyse. Justifier votre réponse.



Le nuage de points des individus sur le premier plan factoriel (F1-F2) est le suivant :

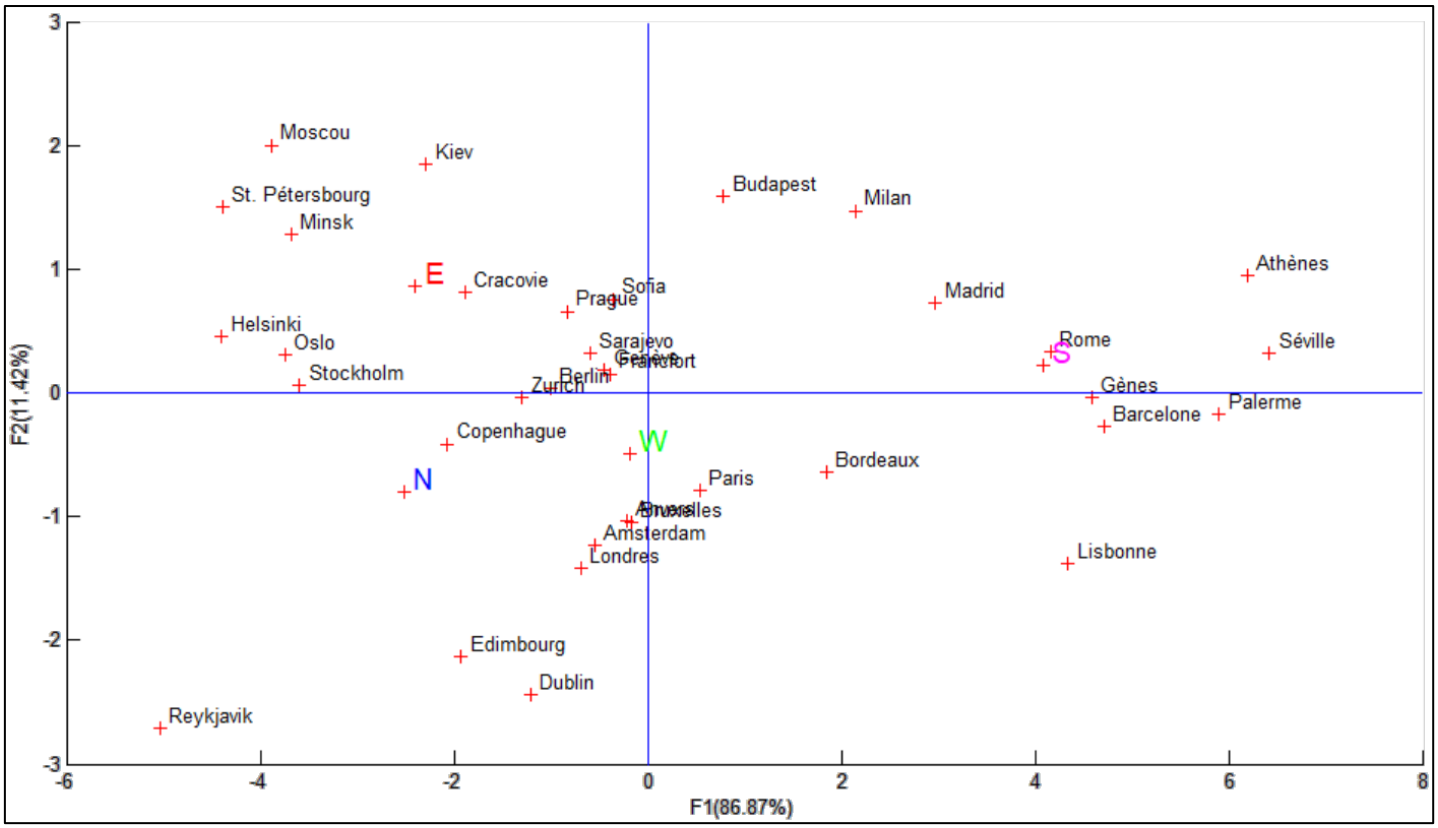


Q2. **(1 pt)** Au vu du premier plan factoriel ci-dessus, quelle interprétation pouvez-vous donner du 1er et 2ème facteur ?

Q3. **(1 pt)** Quelle interprétation pouvez-vous donner aux individus (groupes, cas isolés, corrélations, décorrélation, etc.) ?

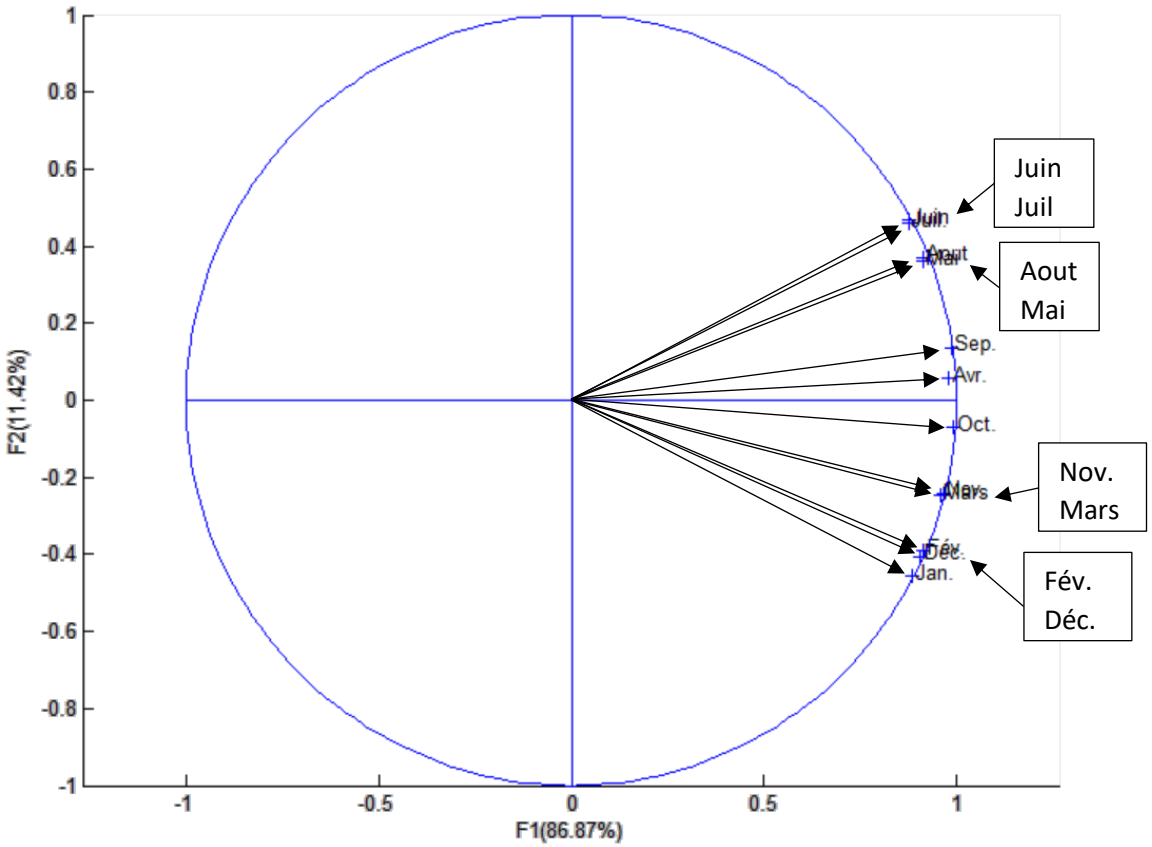
Q4. **(1 pt)** Quel est l'individu le moins bien représenté par le premier plan factoriel ? Quel est l'individu le mieux représenté ?

Nous allons utiliser une information supplémentaire concernant la région de chaque ville. Quarte modalités, ***E***, ***N***, ***W*** et ***S*** pour respectivement les régions ***Est***, ***Nord***, ***Ouest*** et ***Sud***, sont projetées sur le premier plan. Chaque modalité est projetée au centroïde des villes de la région correspondante.



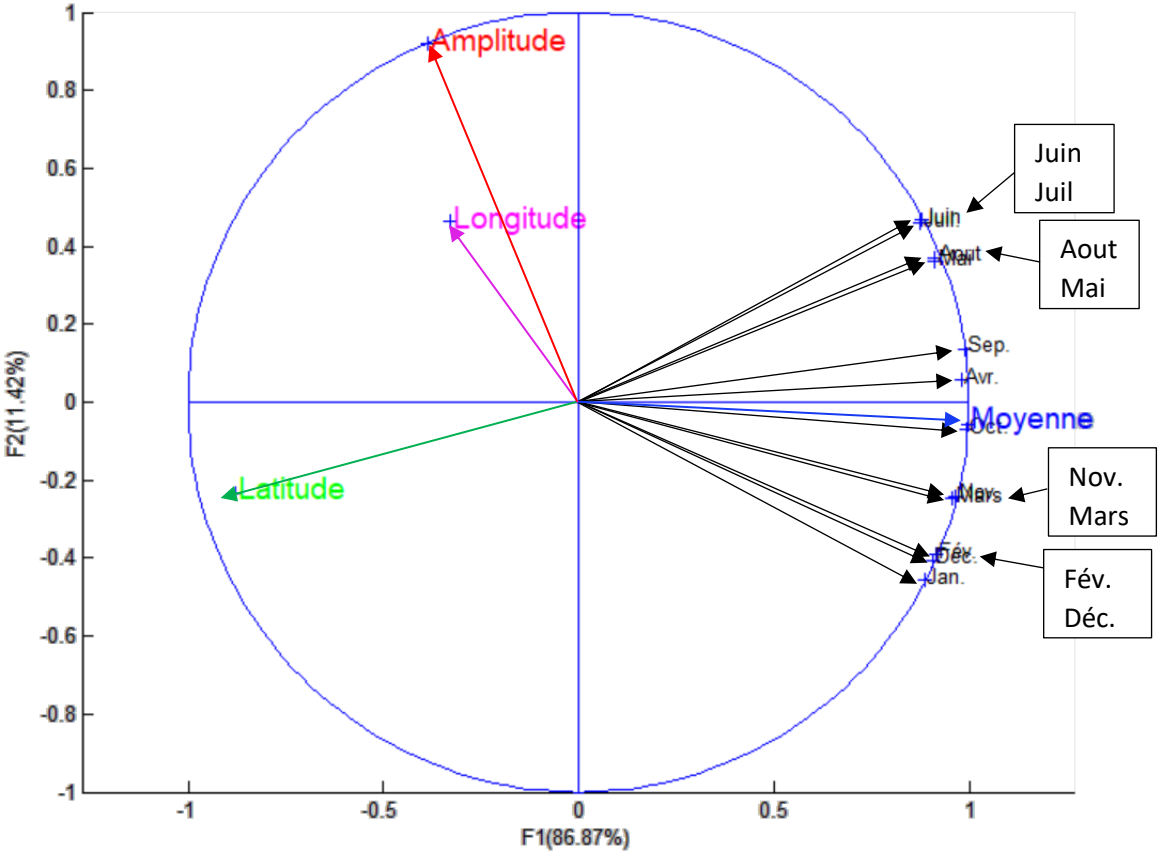
Q5. **(1 pt)** Comment interprétez-vous maintenant le premier plan factoriel ? Qu'est-ce que cette information a apporté à l'analyse ?

Le nuage de points de variables (le cercle de corrélation) de cette analyse est le suivant :



Q6. **(1 pt)** Comment interprétez-vous la participation de variables (les mois) à la construction des axes factoriels F1 et F2 ?

Nous allons utiliser des informations supplémentaires ici aussi. Il s'agit de la moyenne et de l'amplitude annuelles des températures, ainsi que la latitude et la longitude de chaque individu (les villes).



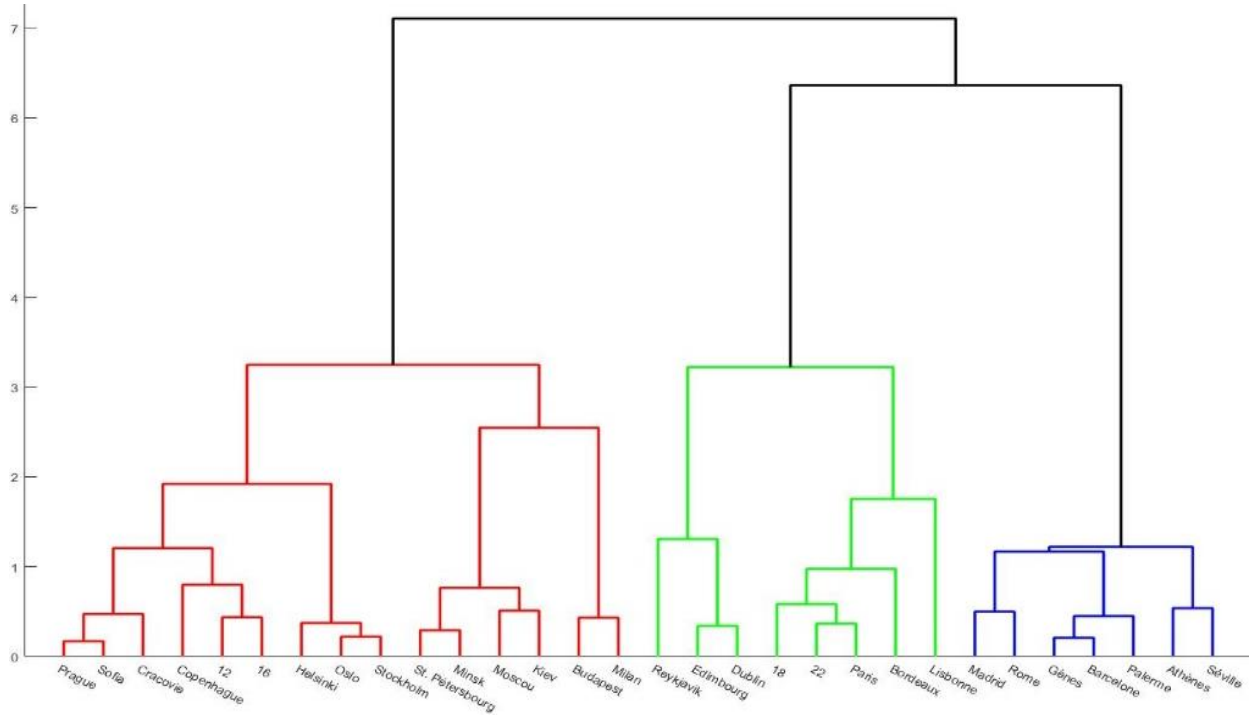
Q7. **(1 pt)** Comment interprétez-vous maintenant le cercle de corrélation des variables ? Qu'est-ce que ces informations ont apporté ?

Q8. **(1 pt)** De cette interprétation, quels sont les individus dont la contribution à l'axe factoriel F1 est supérieure à la moyenne ?

Q9. **(1 pt)** Quels sont les individus dont la contribution à l'axe factoriel F2 (2ème composante principale) est de forte amplitude ?

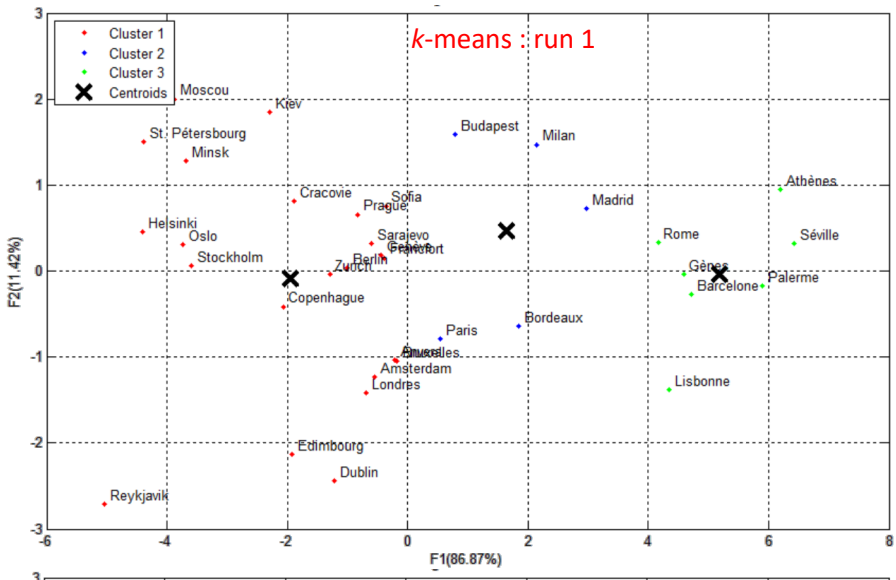
Classification : CAH, k-means et Silhouettes.

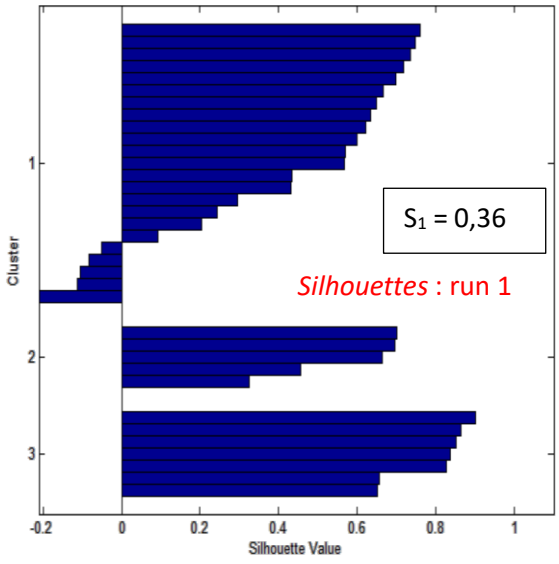
A l'issue de l'ACP, nous souhaitons réaliser une classification en utilisant les coordonnées des individus sur les deux premiers axes factoriels. L'application de la méthode de classification ascendante hiérarchique CAH a donné le dendrogramme suivant :

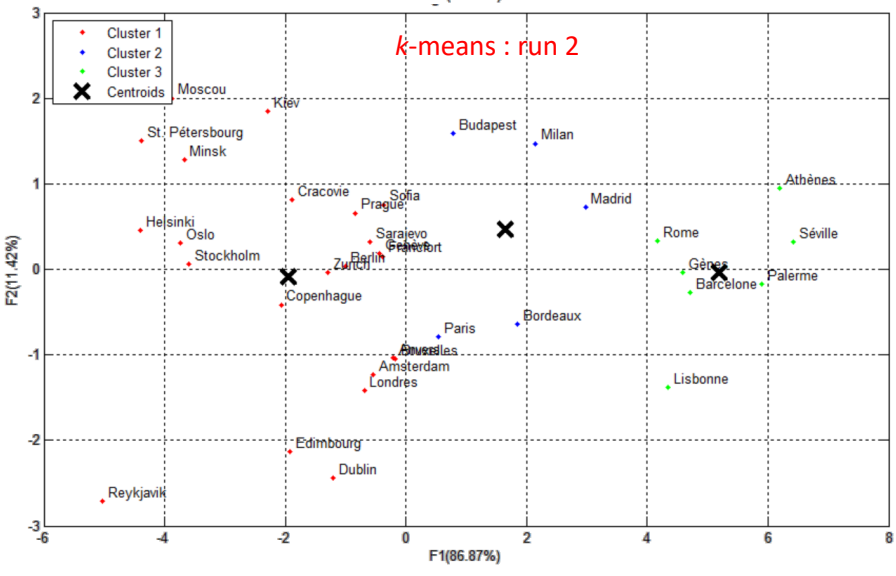


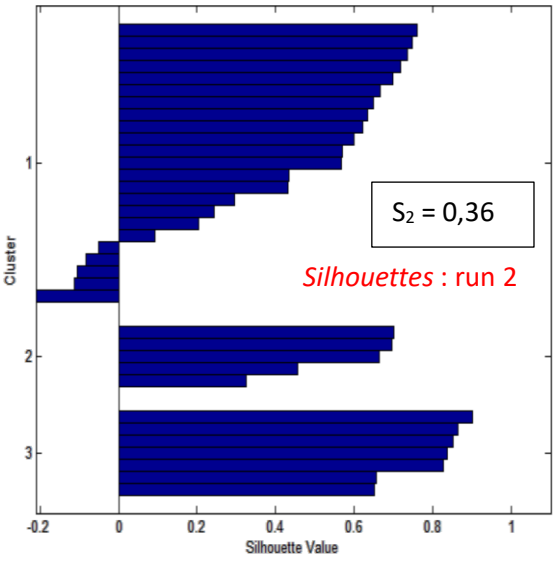
Q10. **(0,5 pt)** L'examen de ce graphique suggère de réaliser une partition des individus en combien de classes ? Justifier.

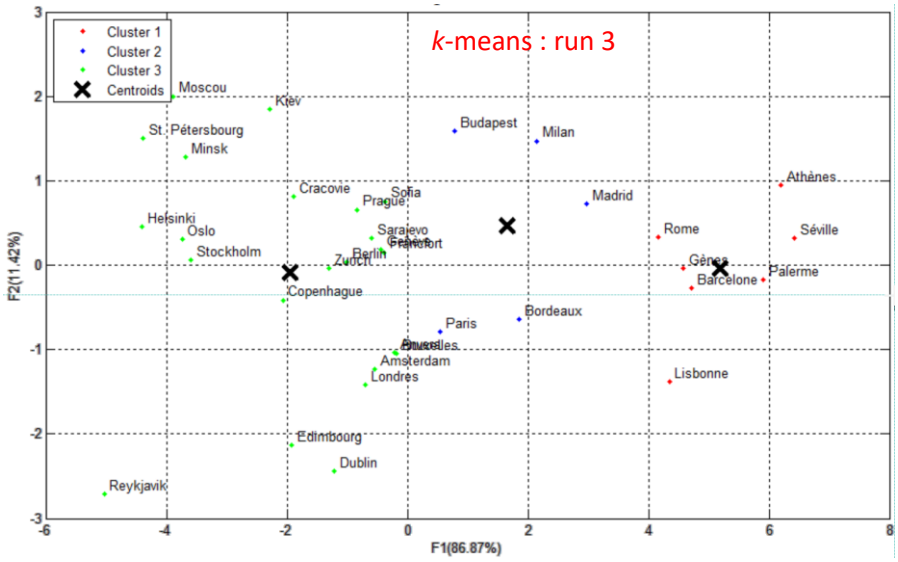
Nous allons réaliser une classification par la méthode *k*-means. Nous répétons l'opération 5 fois tout en évaluons chacune des exécutions par la méthode des silhouettes. Les figures suivantes montrent les résultats de classification de chaque exécution et son évaluation.

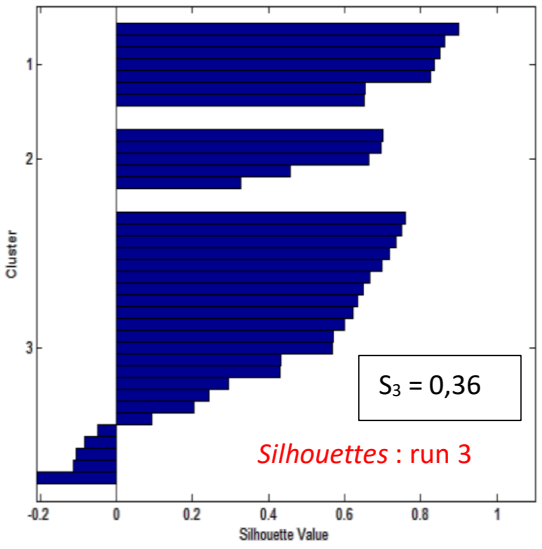


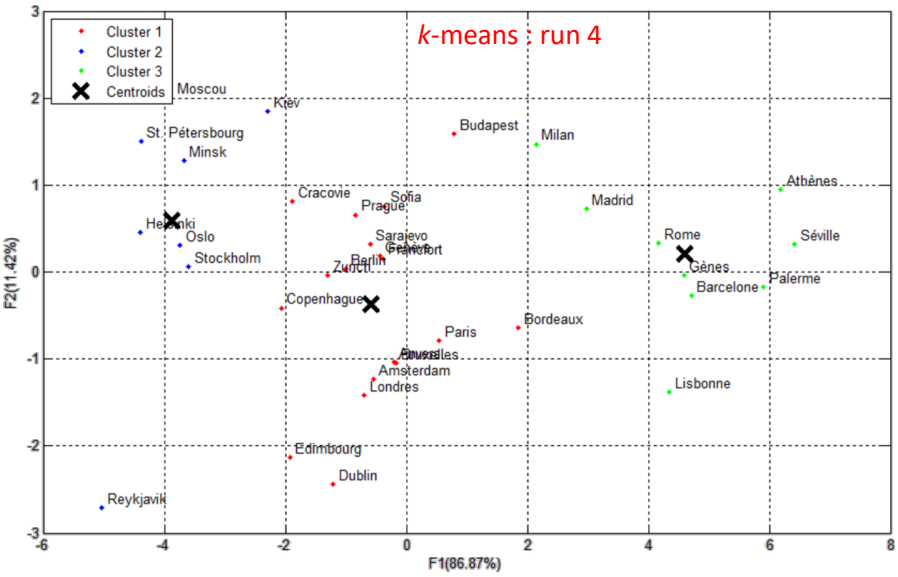


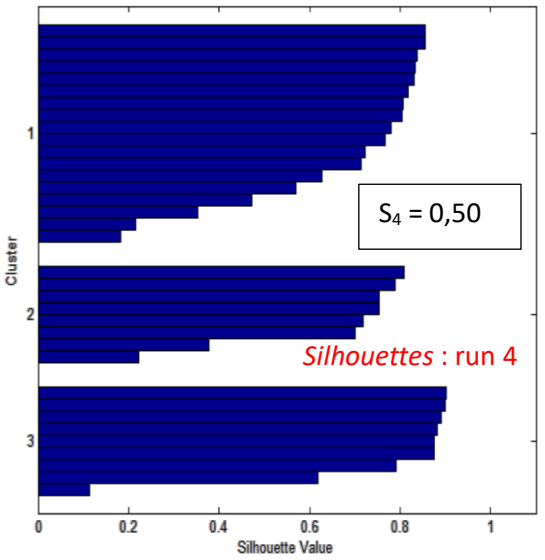


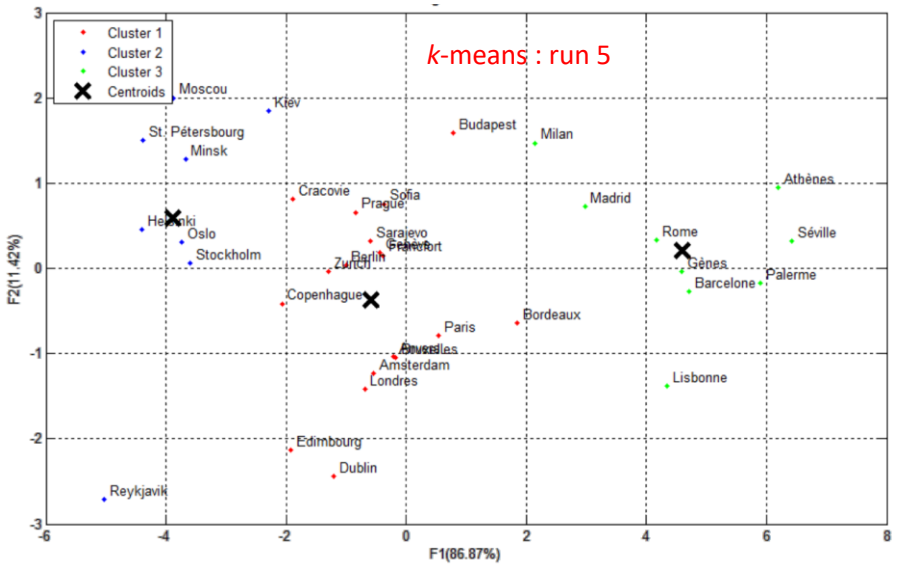


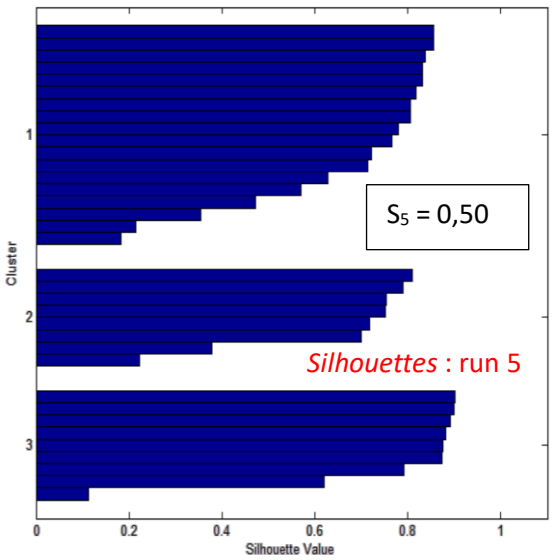












Q11. **(0,5 pt)** Pourquoi devrons nous réaliser plusieurs exécutions de la méthode *k*-means ?

Q12. **(0,5 pt)** Selon l'évaluation faite par la méthode des silhouettes, quelle serait la meilleure classification à retenir ? Justifier.

Q13. **(1 pt)** Expliquer la répartition des individus sur celle-ci.

Q14. **(1 pt)** Quelle sont les points de similarité entre les individus de la même classe et les points de dissimilarité entre les individus de classes différentes ?