

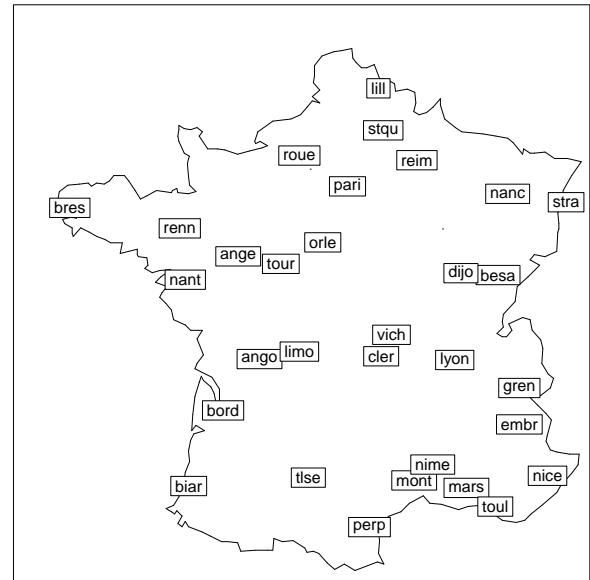
IS d'analyse de données

mardi 14 mars 2023 — durée : 1 heure 30 minutes — documents **non** autorisés

1 La température en France (15 points)

On dispose de relevés mensuels moyens de température (en degrés Celsius) dans 32 stations météorologiques en France que l'on reproduit ci-dessous, suivis de la représentation des stations sur la carte de France et du tableau de corrélation entre les mois. Chaque colonne correspond à un mois de l'année (de janvier à décembre) et chaque station est représentée par le nom de la ville la plus proche. On notera que ajac correspond à la ville d' Ajaccio en Corse, et que cette île est trop au sud dans la méditerranée pour figurer sur la carte.

	janv	fev	mars	avri	mai	juin	juil	aout	sept	oct	nov	dec
ajac	7.7	8.7	10.5	12.6	15.9	19.8	22.0	22.2	20.3	16.3	11.8	8.7
ange	4.2	4.9	7.9	10.4	13.6	17.0	18.7	18.4	16.1	11.7	7.6	4.9
ango	4.6	5.4	8.9	11.3	14.5	17.2	19.5	19.4	16.9	12.5	8.1	5.3
besa	1.1	2.2	6.4	9.7	13.6	16.9	18.7	18.3	15.5	10.4	5.7	2.0
biar	7.6	8.0	10.8	12.0	14.7	17.8	19.7	19.9	18.5	14.8	10.9	8.2
bord	5.6	6.6	10.3	12.8	15.8	19.3	20.9	21.0	18.6	13.8	9.1	6.2
bres	6.1	5.8	7.8	9.2	11.6	14.4	15.6	16.0	14.7	12.0	9.0	7.0
cler	2.6	3.7	7.5	10.3	13.8	17.3	19.4	19.1	16.2	11.2	6.6	3.6
dijo	1.3	2.6	6.9	10.4	14.3	17.7	19.6	19.0	15.9	10.5	5.7	2.1
embr	0.5	1.6	5.7	9.0	13.0	16.4	18.9	18.3	15.3	10.1	4.6	0.5
gren	1.5	3.2	7.7	10.6	14.5	17.8	20.1	19.5	16.7	11.4	6.5	2.3
lill	2.4	2.9	6.0	8.9	12.4	15.3	17.1	17.1	14.7	10.4	6.1	3.5
limo	3.1	3.9	7.4	9.9	13.3	16.8	18.4	17.8	15.3	10.7	6.7	3.8
lyon	2.1	3.3	7.7	10.9	14.9	18.5	20.7	20.1	16.9	11.4	6.7	3.1
mars	5.5	6.6	10.0	13.0	16.8	20.8	23.3	22.8	19.9	15.0	10.2	6.9
mont	5.6	6.7	9.9	12.8	16.2	20.1	22.7	22.3	19.3	14.6	10.0	6.5
nanc	0.8	1.6	5.5	9.2	13.3	16.5	18.3	17.7	14.7	9.4	5.2	1.8
nant	5.0	5.3	8.4	10.8	13.9	17.2	18.8	18.6	16.4	12.2	8.2	5.5
nice	7.5	8.5	10.8	13.3	16.7	20.1	22.7	22.5	20.3	16.0	11.5	8.2
nime	5.7	6.8	10.1	13.0	16.6	20.8	23.6	22.9	19.7	14.6	9.8	6.5
orle	2.7	3.6	6.9	9.8	13.4	16.6	18.4	18.2	15.6	10.9	6.6	3.6
pari	3.4	4.1	7.6	10.7	14.3	17.5	19.1	18.7	16.0	11.4	7.1	4.3
perp	7.5	8.4	11.3	13.9	17.1	21.1	23.8	23.3	20.5	15.9	11.5	8.6
reim	1.9	2.8	6.2	9.4	13.3	16.4	18.3	17.9	15.1	10.3	6.1	3.0
renn	4.8	5.3	7.9	10.1	13.1	16.2	17.9	17.8	15.7	11.6	7.8	5.4
roue	3.4	3.9	6.8	9.5	12.9	15.7	17.6	17.2	15.0	11.0	6.8	4.3
stqu	2.0	2.9	6.3	9.2	12.7	15.6	17.4	17.4	15.0	10.5	6.1	3.1
stra	0.4	1.5	5.6	9.8	14.0	17.2	19.0	18.3	15.1	9.5	4.9	1.3
toul	8.6	9.1	11.2	13.4	16.6	20.2	22.6	22.4	20.5	16.5	12.6	9.7
tlse	4.7	5.6	9.2	11.6	14.9	18.7	20.9	20.9	18.3	13.3	8.6	5.5
tour	3.5	4.4	7.7	10.6	13.9	17.4	19.1	18.7	16.2	11.7	7.2	4.3
vich	2.4	3.4	7.1	9.9	13.6	17.1	19.3	18.8	16.0	11.0	6.6	3.4



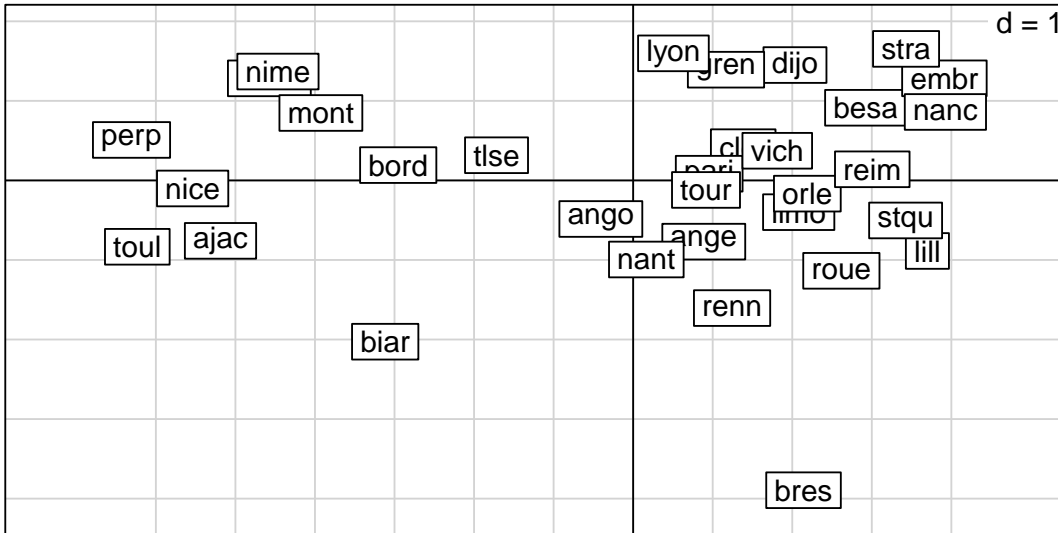
	janv	fev	mars	avri	mai	juin	juil	aout	sept	oct	nov	dec
janv	1.00	0.99	0.93	0.80	0.63	0.59	0.56	0.65	0.80	0.93	0.98	0.99
fev	0.99	1.00	0.97	0.87	0.72	0.69	0.66	0.74	0.87	0.97	0.99	0.99
mars	0.93	0.97	1.00	0.95	0.84	0.81	0.78	0.85	0.94	0.97	0.96	0.92
avri	0.80	0.87	0.95	1.00	0.96	0.94	0.92	0.96	0.98	0.94	0.88	0.80
mai	0.63	0.72	0.84	0.96	1.00	0.99	0.98	0.98	0.95	0.84	0.75	0.64
juin	0.59	0.69	0.81	0.94	0.99	1.00	0.99	0.99	0.94	0.82	0.72	0.60
juil	0.56	0.66	0.78	0.92	0.98	0.99	1.00	0.99	0.93	0.80	0.69	0.57
aout	0.65	0.74	0.85	0.96	0.98	0.99	0.99	1.00	0.97	0.87	0.77	0.66
sept	0.80	0.87	0.94	0.98	0.95	0.94	0.93	0.97	1.00	0.96	0.89	0.81
oct	0.93	0.97	0.97	0.94	0.84	0.82	0.80	0.87	0.96	1.00	0.98	0.93
nov	0.98	0.99	0.96	0.88	0.75	0.72	0.69	0.77	0.89	0.98	1.00	0.98
dec	0.99	0.99	0.92	0.80	0.64	0.60	0.57	0.66	0.81	0.93	0.98	1.00

1.1 Analyse des corrélations (2 point)

Question 1 Expliquez pourquoi les valeurs des corrélations placées juste au dessus de la diagonale sont grandes. D'autre part, montrer qu'il y a deux blocs de mois consécutifs et très corrélés entre eux ($r \geq 0,98$). Comment est la corrélation entre ces blocs ?

1.2 Analyse en composante principales (9 points)

On effectue une analyse en composantes principales sur les données centrées réduites. On donne ci-dessous la projection des individus sur le premier plan principal, puis, pour les 3 premiers axes principaux, les valeurs propres, les corrélations entre les variables et les axes, les coordonnées des individus sur les axes et enfin le tableau des qualités de représentation des individus par les sous espaces principaux (en %). On notera que ces deux derniers tableaux sont coupés en deux.



Valeurs propres

Axis1 Axis2 Axis3
10.47 1.41 0.06

Corrélations

	Comp1	Comp2	Comp3
janv	-0.88	-0.47	-0.01
fev	-0.93	-0.35	-0.01
mars	-0.97	-0.15	-0.13
avri	-0.98	0.14	-0.12
mai	-0.92	0.38	-0.06
juin	-0.90	0.43	-0.01
juil	-0.88	0.46	0.05
aout	-0.93	0.36	0.06
sept	-0.99	0.13	0.07
oct	-0.98	-0.15	0.08
nov	-0.95	-0.31	0.06
dec	-0.88	-0.46	0.02

	Axis1	Axis2	Axis3
ajac	-5.19	-0.76	0.58
ange	0.89	-0.77	-0.12
ango	-0.41	-0.48	-0.37
besa	2.92	0.91	0.05
biar	-3.12	-2.03	-0.22
bord	-2.95	0.19	-0.67
bres	2.14	-3.89	0.12
cler	1.38	0.42	-0.01
dijo	2.03	1.45	-0.22
embr	3.91	1.28	0.41
gren	1.17	1.39	-0.10
lill	3.70	-0.89	0.32
limo	2.08	-0.40	-0.23
lyon	0.51	1.60	-0.08
mars	-4.57	1.28	0.19
mont	-3.93	0.85	0.08

	Axis1	Axis2	Axis3
nanc	3.92	0.86	0.11
nant	0.17	-0.99	-0.23
nice	-5.54	-0.10	0.12
nime	-4.47	1.37	0.07
orle	2.19	-0.17	0.07
pari	0.96	0.08	-0.29
perp	-6.31	0.51	-0.08
reim	3.01	0.14	0.17
renn	1.24	-1.60	-0.13
roue	2.62	-1.14	0.05
stqu	3.42	-0.50	0.19
stra	3.43	1.65	-0.08
toul	-6.23	-0.83	0.21
tlse	-1.72	0.32	0.07
tour	0.92	-0.12	-0.14
vich	1.81	0.37	0.18

	Axis1	Axis1:2	Axis1:3
ajac	96.6	98.6	99.8
ange	55.4	97.1	98.1
ango	25.9	62.8	84.2
besa	90.7	99.6	99.6
biar	69.4	98.7	99.1
bord	94.0	94.4	99.3
bres	23.2	99.8	99.9
cler	89.4	97.6	97.6
dijo	65.7	99.1	99.9
embr	88.1	97.6	98.6
gren	38.8	94.0	94.3
lill	93.7	99.1	99.8
limo	93.7	97.1	98.3
lyon	9.0	99.1	99.4
mars	92.4	99.6	99.8
mont	95.4	99.8	99.9

	Axis1	Axis1:2	Axis1:3
nanc	94.8	99.4	99.4
nant	2.5	92.9	98.0
nice	99.7	99.7	99.8
nime	91.1	99.6	99.6
orle	99.2	99.8	99.9
pari	83.9	84.6	92.5
perp	99.2	99.8	99.9
reim	99.0	99.2	99.5
renn	37.0	98.7	99.1
roue	83.7	99.5	99.5
stqu	97.2	99.3	99.6
stra	80.4	99.1	99.1
toul	98.0	99.7	99.8
tlse	93.0	96.2	96.3
tour	93.5	95.1	97.2
vich	94.0	98.0	98.9

Question 2 Combien de d'axes principaux retient-on a priori ? Que peut-on dire de la qualité globale de la représentation ainsi obtenue ?

Question 3 Quelles sont les variables qui déterminent les deux premières composantes principales (préciser les critères utilisés) ? Que peut-on dire de la première composante principale ?

Question 4 Comment peut-on interpréter les deux premiers axes principaux grâce aux corrélations avec les variables ? Quel rapport y a-t-il avec les variables décrites en question 1 ?

Question 5 Quels sont les individus qui déterminent les deux premiers axes principaux ? (préciser les critères utilisés)

Question 6 Quelle nouvelle interprétation des axes principaux est suggérée par la comparaison entre projection des individus sur le premier plan principal et la carte des stations météorologiques ? Y a-t-il un individu qui n'est pas vraiment où on l'attend ?

Question 7 Que peut-on dire de la qualité de la représentation des individus par le premier plan principal ? On expliquera notamment la signification de cette qualité de représentation.

Question 8 Calculez la contribution de bres aux axes 1 et 2. Commentez les valeurs.

1.3 Un individu supplémentaire (2 points)

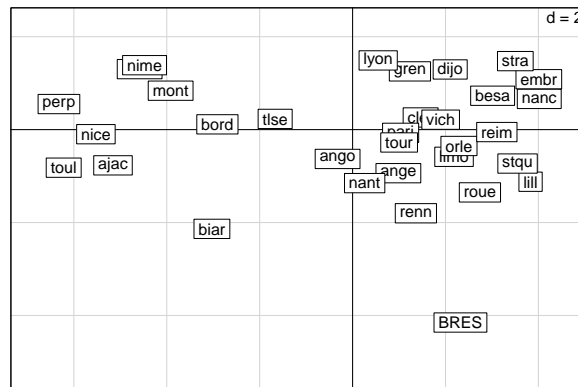
On refait l'ACP sur données centrées réduites en plaçant la station **bres** (Brest) en individu supplémentaire. La nouvelle répartition des valeurs propres, le cercle des corrélations des variables avec les 3 premières composantes principales et la projection des individus sur le premier plan principal sont donnés ci-dessous.

Valeurs propres

Axis1	10.8927
Axis2	0.9821
Axis3	0.0579
Axis4	0.0330
Axis5	0.0149
Axis6	0.0059
Axis7	0.0053
Axis8	0.0042
Axis9	0.0015
Axis10	0.0014
Axis11	0.0008
Axis12	0.0003

Corrélations

	Comp1	Comp2	Comp3
janv	-0.92	-0.40	-0.01
fev	-0.95	-0.30	0.00
mars	-0.98	-0.14	-0.12
avri	-0.99	0.10	-0.13
mai	-0.94	0.31	-0.06
juin	-0.93	0.37	-0.02
juil	-0.91	0.40	0.06
aout	-0.95	0.30	0.06
sept	-0.99	0.10	0.07
oct	-0.99	-0.12	0.08
nov	-0.97	-0.24	0.06
dec	-0.92	-0.38	0.02



Question 9 Expliquez en quoi consiste la nouvelle analyse (comment l'individu supplémentaire est-il traité et comment est-il placé dans la projection des individus ?)

Question 10 Commentez les différences entre la nouvelle analyse et la première : répartition de l'inertie, corrélations et nouvelle projection des individus. Expliquez pourquoi d'après vous placer **bres** en individu supplémentaire ne change pas grand chose.

1.4 Une nouvelle variable (2 points)

On dispose en fait d'une donnée supplémentaire, l'altitude **alt** des stations météorologiques (mesurée en mètres). On la donne ci-dessous, ainsi que sa corrélation de **alt** avec les 3 premiers axes de l'analyse de la section précédente.

Altitude des stations

ajac	ange	ango	besa	biar	bord	bres	cler	dijo	embr	gren	lill	limo	lyon	mars	mont
4	50	132	307	69	47	94	331	219	871	220	47	402	198	5	3
nanc	nant	nice	nime	orle	pari	perp	reim	renn	roue	stqu	stra	toul	tlse	tour	vich
212	26	4	59	125	75	42	91	36	151	98	150	24	152	112	249

Corrélation de **alt** avec les axes

	Comp1	Comp2	Comp3
alt	0.51	0.32	0.15

Question 11 Y a-t-il une station très différente des autres ? Expliquez en quoi cette nouvelle information explique une anomalie constatée dans la question 6.

Question 12 Comment la corrélation de **alt** avec les axes a-t-elle été obtenue ? Interprétez ses valeurs.

2 Valeurs-tests pour des variables à deux modalités (5 points)

On s'intéresse dans le cadre d'une ACP à une variable supplémentaire qualitative à deux modalités (deux valeurs possibles), que l'on notera 1 et 2. On note $n^{(1)}$ et $n^{(2)}$ les effectifs de ces modalités (nombre d'individus de chaque sorte), et $c_k^{(1)}$ et $c_k^{(2)}$ leurs coordonnées sur l'axe k . Ces coordonnées de ces modalités peuvent s'écrire, pour $j = 1, 2$

$$c_k^{(j)} = \frac{1}{n^{(j)}} \sum_{i \text{ de mod. } j} c_{ki}$$

où la somme s'effectue sur tous les individus ayant la modalité j . Comme d'habitude, λ_k est la valeur propre associée à l'axe k et c_{ki} la coordonnée de l'individu i sur l'axe k . On rappelle que chaque vecteur \mathbf{c}_k est centré.

Question 13 Montrer que, pour chaque k ,

$$n^{(1)}c_k^{(1)} + n^{(2)}c_k^{(2)} = 0.$$

Question 14 Montrer que les valeurs tests associées à chacune des deux modalités sont égales au signe près.