A.5 Un petit exemple d'ACP commenté

Le commentaire de l'ACP du petit tableau présenté ci-après permet d'illustrer les règles et la démarche d'interprétation d'une ACP (voir aussi chapitre 9).

A.5.1.Description des données

Pour 15 villes de France, on dispose des moyennes des températures mensuelles calculées sur 30 ans (entre 1931 et 1960). Ces données sont extraites du Quid 1986, page 507 (édition Robert Laffont). Elles sont rassemblées dans le tableau 1 qui croise les 15 villes (en lignes) et les 12 mois de l'année (en colonnes). On a ajouté quatre colonnes supplémentaires commentées par la suite. Les deux dernières lignes, la moyenne et l'écart-type des colonnes, ne sont là que pour information, elles ne sont pas introduites dans l'analyse.

Tableau 1. Moyennes des temperatures mensuelles de 15 villes de France

	janv	févr	mars	avri	mai	juin	juil	aoû	sept	octo	nove	déce	lati	longi	moy	ampli
Bordeaux	5.6	6.6	10.3	12.8	15.8	19.3	20.9	21.0	18.6	13.8	9.1	6.2	44.5	-0.34	13.3	15.4
Brest	6.1	5.8	7.8	9.2	11.6	14.4	15.6	16.0	14.7	12.0	9.0	7.0	48.2	-4.29	10.8	10.2
Clermont	2.6	3.7	7.5	10.3	13.8	17.3	19.4	19.1	16.2	11.2	6.6	3.6	45.5	3.05	10.9	16.8
Grenoble	1.5	3.2	7.7	10.6	14.5	17.8	20.1	19.5	16.7	11.4	6.5	2.3	45.1	5.43	11.0	18.6
Lille	2.4	2.9	6.0	8.9	12.4	15.3	17.1	17.1	14.7	10.4	6.1	3.5	50.4	3.04	9.7	14.7
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	45.5	4.51	11.4	18.6
Marseille	5.5	6.6	10.0	13.0	16.8	20.8	23.3	22.8	19.9	15.0	10.2	6.9	43.2	5.24	14.2	17.8
Montpellier	5.6	6.7	9.9	12.8	16.2	20.1	22.7	22.3	19.3	14.6	10.0	6.5	43.4	3.53	13.9	17.1
Nantes	5.0	5.3	8.4	10.8	13.9	17.2	18.8	18.6	16.4	12.2	8.2	5.5	47.1	-1.33	11.7	13.8
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	43.4	7.15	14.8	15.2
Paris	3.4	4.1	7.6	10.7	14.3	17.5	19.1	18.7	16.0	11.4	7.1	4.3	48.5	2.20	11.2	15.7
Rennes	4.8	5.3	7.9	10.1	13.1	16.2	17.9	17.8	15.7	11.6	7.8	5.4	48.1	-1.41	11.1	13.1
Strasbourg	0.4	1.5	5.6	9.8	14.0	17.2	19.0	18.3	15.1	9.5	4.9	1.3	48.4	7.45	9.7	18.6
Toulouse	4.7	5.6	9.2	11.6	14.9	18.7	20.9	20.9	18.3	13.3	8.6	5.5	43.4	1.26	12.7	16.2
Vichy	2.4	3.4	7.1	9.9	13.6	17.1	19.3	18.8	16.0	11.0	6.6	3.4	46.1	3.26	10.7	16.9
Moyenne	4.0	4.8	8.2	11.0	14.4	17.8	19.8	19.6	17.0	12.3	7.9	4.9	46.0	2.58	11.8	15.9
Ecart-type	1.94	1.81	1.48	1.37	1.45	1.73	2.06	1.94	1.79	1.77	1.74	1.89	2.22	3.21	1.55	2.25

A.5.2.Problématique

Le but général de l'étude est bien évidemment de comparer les températures mensuelles des différentes villes. Précisons quelques questions auxquelles les résultats de l'ACP permettent

de répondre en abordant le tableau successivement à travers ses lignes et à travers ses colonnes.

Point de vue des lignes (ou individus : les villes)

Les villes sont caractérisées par leurs 12 températures moyennes mensuelles. Quelles sont, de ce point de vue, les villes qui se ressemblent ? Quelles sont celles qui diffèrent ? Plus généralement peut-on faire une typologie des villes mettant en évidence l'ensemble des ressemblances ainsi définies ? En ACP la dissemblance entre les individus est mesurée par une distance (cf. **section 1.1**). Ici le carré de la distance entre deux villes est la somme des carrés des différences entre les moyennes des températures des 12 mois. Ceci traduit bien la notion de proximité souhaitée.

Cette typologie faite, on peut se demander si ces ressemblances (ou dissemblances) correspondent à des proximités (ou des éloignements) géographiques.

Point de vue des colonnes (ou variables : les mois)

Les mois sont caractérisés par les températures moyennes mensuelles des 15 villes. Le problème n'est pas de séparer les mois chauds des mois froids pour l'ensemble des 15 villes (ce qui arriverait si nous les considérions comme des individus) mais de comparer la répartition des 15 villes (des plus chaudes aux plus froides) pour deux mois différents *sans tenir compte du fait que d'un mois à l'autre les températures sont globalement plus ou moins élevées* (l'élimination de cet effet de moyenne est assurée par le centrage). La notion de ressemblance entre mois est traduite par celle de liaison, plus précisément de corrélation, entre variables numériques. Deux mois sont d'autant plus liés que pour chacun on observe la même répartition des 15 villes selon leur température moyenne. A l'inverse, ils sont peu liés si ce ne sont pas dans les mêmes villes que l'on trouve les températures les plus élevées (ou les plus basses).
Ceci posé, les questions sont les suivantes : Quels sont les mois qui sont liés entre eux ? Quels sont ceux qui le sont peu ? Plus généralement peut-on faire un bilan des liaisons entre les 12 mois ? Les températures mensuelles sont-elles liées à leur position géographique ? D'autre part, si les mois sont liés, l'information donnée par les 12 colonnes est redondante. Peut-on la résumer en remplaçant les 12 mois par un plus petit nombre de variables synthétiques ?

Ajout de variables supplémentaires (ou explicatives, ou illustratives)

Il apparaît dans la problématique que les températures doivent être analysées en ayant à l'esprit la position géographique des villes. On peut formaliser cette position par leur latitude et leur longitude, données introduites dans l'analyse en tant que variables supplémentaires.

Deux autres variables supplémentaires ont été ajoutées pour des raisons qui apparaîtront au cours de l'interprétation.

Faut-il réduire les données ?

Lorsque les unités de mesure diffèrent d'une variable à l'autre, le recours à la réduction des variables est systématique. Ce n'est pas le cas ici et la question mérite d'être posée.

Ne pas réduire revient ici à considérer que un écart de 1 degré entre deux villes a la même importance quel que soit le mois où il est observé, que ce soit un mois où les écarts entre les températures des 15 villes sont plutôt faibles ou au contraire importants. Dans les distances entre les villes, un mois aura alors d'autant plus d'influence que l'on y observe de grandes différences de températures d'une ville à l'autre. On montre facilement que ne pas réduire les variables revient à accorder aux variables réduites un poids égal à leur écart-type. A l'inverse, en réduisant, on accorde à chaque mois de l'année la même importance a priori dans l'analyse.

Sur ce jeu de données les deux points de vue sont également défendables. Pour cet exemple didactique, nous choisissons de réduire les données. L'ACP est alors dite *normée*, terme omis la plupart du temps car correspondant au cas le plus fréquent. Comme les écarts-types varient peu d'un mois à l'autre (minimum : 1.37 et maximum : 2.06) les deux analyses, normée et non normée, conduisent certainement à des résultats très proches. Ceci a été vérifié : pour les quatre premiers facteurs, les coefficients de corrélation entre les facteurs de même rang des deux analyses sont tous supérieurs à 0.99.

A.5.3. Résultats de l'ACP

A.5.3.1 Inertie des facteurs

Dans une ACP normée l'inertie totale de chacun des nuages (celui des villes et celui des mois) est égale au nombre de variables actives (ici 12). Avec une inertie de 9.58, qui représente 80% de l'inertie des nuages dans l'espace tout entier, le premier facteur est largement prépondérant. L'inertie du deuxième facteur vaut 2.28 et celle du troisième 0.07. Les deux premiers facteurs totalisent donc 98.8% de l'inertie totale. Les deux nuages de points (individus et variables) sont donc pratiquement bidimensionnels : leur projection sur le premier plan factoriel en donne une représentation quasiment parfaite. On se limite dans l'interprétation à l'étude de ces deux premiers facteurs et du plan qu'ils engendrent.

Contribution des individus (tableau 2)

Le premier facteur est dû essentiellement à 5 villes (Lille, Marseille, Montpellier, Nice et Strasbourg) qui totalisent 77.4% de son inertie. Compte tenu du faible nombre de villes étudiées, cette situation est banale et n'attire pas d'observation particulière.

Le deuxième facteur est dû pour moitié (49.1%) à la ville de Brest, qui est donc assez particulière du point de vue climatique. Remarquons toutefois que la différence d'inertie entre le deuxième et le troisième facteur (2.28 - 0.07 = 2.20) est beaucoup plus grande que l'inertie de Brest le long de ce deuxième axe (2.28 X 0.49 = 1.12). Même sans la ville de Brest, ce deuxième facteur serait donc apparu. Il semble que le cas de Brest est, certes, particulier mais s'inscrit dans une tendance générale, ce qui sera confirmé lors de l'interprétation.

Tableau 2. Coordonnées, qualité de représentation et contributions des 15 villes pour chacun des 2 premiers facteurs

	coordonnées		contributions		qualités de rep.	
	1	2	1	2	1	2
Bordeaux	3.12	-.11	6.8	0.0	.95	.00
Brest	-2.27	-4.09	3.6	49.1	.23	.76
Clermont	-1.73	.59	2.1	1.0	.88	.10
Grenoble	-1.53	1.69	1.6	8.3	.43	.52
Lille	-4.22	-.60	12.4	1.0	.97	.02
Lyon	-0.83	1.79	0.5	9.4	.18	.82
Marseille	4.83	0.83	16.2	2.0	.96	.03
Montpellier	4.15	.44	12.0	0.6	.99	.01
Nantes	-0.28	-1.11	0.1	3.6	.06	.89
Nice	6.01	-0.79	25.1	1.8	.98	.02
Paris	-1.24	0.16	1.1	0.1	.89	.01
Rennes	-1.44	-1.67	1.4	8.2	.42	.57
Strasbourg	-4.11	2.17	11.7	13.8	.78	.22
Toulouse	1.74	0.14	2.1	0.1	.95	.01
Vichy	-2.20	0.58	3.4	1.0	.92	.06

A.5.3.2. Interprétation du premier facteur

Coordonnées des variables actives (tableau 3 et figure 4)

Tableau 3. Coordonnées des variables actives et supplémentaires pour chacun des 2 premiers facteurs

	janv	févr	mars	avri	mai	juin	juil	aoû	sept	octo	nove	déce	lati	longi	moy	ampli
facteur 1	.76	.88	.97	.97	.87	.86	.84	.90	.97	.98	.90	.77	-.84	.17	1.00	.10
facteur 2	-.64	-.47	-.16	.20	.47	.50	.53	.43	.21	-.17	-.41	-.62	-.31	.79	-.02	.99

Les 12 variables sont corrélées fortement et positivement au premier facteur. Etant ainsi liées à une même variable, elle sont liées entre elles ; ceci peut être constaté sur la matrice des corrélations (tableau 4) dont toutes les valeurs sont positives.

Ce type de facteur est classique et est appelé "effet taille". Il exprime que certains individus ont de grandes valeurs pour l'ensemble des variables et d'autres de petites valeurs pour l'ensemble des variables. Dans notre exemple cela indique que certaines villes sont plus chaudes que d'autres quel que soit le mois de l'année.

Tableau 4. Matrice des corrélations entre toutes les variables

	janv	févr	mars	avri	mai	juin	juil	aoû	sept	octo	nove	déce	lati	longi	moy	ampli
janvier	1.00															
février	.97	1.00														
mars	.84	.93	1.00													
avril	.61	.76	.92	1.00												
mai	.36	.55	.77	.95	1.00											
juin	.34	.52	.76	.94	.99	1.00										
juillet	.30	.49	.72	.91	.98	.99	1.00									
août	.41	.59	.80	.95	.98	.99	.99	1.00								
septembre	.60	.76	.91	.98	.94	.94	.93	.97	1.00							
octobre	.85	.94	.97	.91	.77	.76	.74	.81	.93	1.00						
novembre	.95	.99	.93	.78	.59	.57	.55	.64	.80	.96	1.00					
décembre	.99	.97	.83	.62	.38	.36	.32	.43	.62	.87	.96	1.00				
latitude	-.42	-.60	-.81	-.85	-.84	-.87	-.88	-.90	-.90	-.78	-.64	-.44	1.00			
longitude	-.39	-.22	-.04	.29	.54	.53	.59	.50	.35	.07	-.13	-.35	-.31	1.00		
moyenne an.	.77	.89	.97	.96	.86	.85	.83	.89	.97	.98	.91	.79	-.83	.16	1.00	
amplitude	-.57	-.38	-.06	.28	.55	.58	.62	.52	.31	-.06	-.30	-.54	.-42	.83	.08	1.00

Coordonnées des individus (tableau 2 et figure 5)

Compte tenu des relations entre les coordonnées des individus et celles des variables (cf. relations de transition, section 1.7) on s'attend à trouver, le long de l'axe 1, les villes chaudes

du côté des coordonnées positives et les villes froides du côté des coordonnées négatives. C'est bien ce qu'on observe, l'axe 1 opposant principalement Nice, Marseille et Montpellier (à droite) à Lille et Strasbourg (à gauche). Cette opposition se retrouve facilement dans les données. Ainsi, quel que soit le mois de l'année, les températures mesurées à Nice, Marseille et Montpellier se situent au dessus de la moyenne (calculée sur les 15 villes) tandis que celles mesurées à Lille et Strasbourg se situent au dessous de cette moyenne. Attention, la première formule de transition relie la coordonnée d'une ville à *l'ensemble des coordonnées* des variables. Ainsi, Lille a la plus faible coordonnée sur le premier axe, mais il serait faux d'en conclure qu'elle est, quel que soit le mois, la ville la plus froide. La fausseté de cette affirmation se constate immédiatement sur les données : bien que toujours plus froide que la moyenne, Lille n'est la ville la plus froide que 2 mois sur 12 (septembre et avril). La position extrême de Lille provient du fait que cette ville est la plus froide sur *l'ensemble de l'année*. Certains mois de l'année, une autre ville, ou même plusieurs, sont plus froides qu'elle mais elles sont sensiblement moins froides que Lille pendant beaucoup d'autres mois.

La position des villes proches de l'origine s'interprète dans le même esprit. La faible coordonnée sur le premier axe de Nantes, Lyon ou Paris indique que sur l'ensemble de l'année la température de cès villes est moyenne. Mais on ne peut en déduire que les températures y sont toujours moyennes car elles peuvent aussi être tantôt élevées et tantôt basses. Le deuxième facteur sera éclairant sur ce point.

Coordonnées des variables supplémentaires (tableau 3)

Ce facteur semble correspondre à la température moyenne annuelle. Pour s'en assurer, on peut faire la moyenne des 12 températures mensuelles pour chacune des 15 villes et calculer le coefficient de corrélation entre cette nouvelle variable et le premier facteur (défini sur les villes). Pratiquement, il suffit de relancer la même analyse, la température moyenne annuelle étant introduite en variable supplémentaire. Ce coefficient de corrélation vaut 1.00 (aux erreurs d'arrondi près), ce qui achève de justifier l'assimilation du premier facteur à la température moyenne annuelle. Remarquons que, bien que le coefficient de corrélation soit très proche de 1, ce premier facteur n'est pas exactement la moyenne annuelle. Comme toute composante principale, ce facteur est une combinaison linéaire des variables actives dont les coefficients sont proportionnels aux coordonnées des variables. Si ce facteur coïncidait exactement avec la moyenne, les 12 coefficients de la combinaison linéaire seraient égaux. Or cette combinaison est égale ici à : 0.76 janvier + 0.88 février + ... + 0.77 décembre

Considérer ce premier facteur comme une moyenne annuelle est une interprétation interne aux données traitées. On franchit un nouveau pas dans l'interprétation en le reliant à des données externes comme la position géographique des villes. Le nombre de villes étant faible,

on peut constater directement que parmi les 15 villes, les plus chaudes sont aussi les plus méridionales. La latitude et la longitude ayant été introduites dans l'analyse en tant que variables supplémentaires, on dispose aussi de leur coefficient de corrélation avec le premier facteur. Celui de la latitude vaut 0.84, ce qui exprime que l'ordre des coordonnées des 15 villes sur le premier axe correspond à peu près à leur latitude (à peu près seulement, des villes comme Vichy, Clermont, Grenoble et Lyon sont plus froides que ne le laisse attendre leur latitude). La longitude, elle, est très peu liée au premier facteur (corrélation 0.17).

A.5.3.3 Interprétation du deuxième facteur

Coordonnées des variables actives (tableau 3 et figure 4)

Les mois d'automne et d'hiver sont opposés aux mois de printemps et d'été. Les mois qui encadrent les solstices d'hiver et d'été sont les plus liés à ce facteur.

Cette opposition montre qu'à température moyenne annuelle égale (premier facteur fixé) certaines villes sont plutôt chaudes en été et plutôt froides en hiver alors que d'autres, à l'inverse, sont plutôt froides en été et plutôt chaudes en hiver. L'amplitude thermique plus importante dans les premières que dans les secondes semble correspondre à ce facteur.

Coordonnées des individus (tableau 2 et figure 5)

Compte tenu des relations de transition, on sait que les coordonnées des villes ayant une forte amplitude thermique sont positives tandis que celles des villes à faible amplitude sont négatives. Ainsi, Brest, dont la coordonnée sur ce facteur est la plus élevée, subit des températures au dessus de la moyenne depuis novembre jusqu'à février et très au dessous de la moyenne depuis avril jusqu'à septembre. Cette tendance se retrouve de façon atténuée pour la belle ville de Rennes. A l'opposé, Grenoble subit des températures très en dessous de la moyenne depuis novembre jusqu'à février et presque égales à la moyenne depuis mai jusqu'à août. Brest apparaît donc comme la situation la plus extrême d'une tendance générale.

Coordonnées des variables supplémentaires (tableau 3)

L'interprétation générale du deuxième facteur est confirmée par sa corrélation avec la variable supplémentaire *amplitude thermique* (température mensuelle maximum -température mensuelle minimum) égale à 0.99. Avec un coefficient de corrélation de 0.79 ce facteur est aussi lié à la longitude (qui, grossièrement, exprime la proximité avec l'océan atlantique, et, encore plus grossièrement la continentalité). Les villes sont à peu près placées par longitude croissante ; la seule exception notable est Nice qui, très à l'est, a pourtant une amplitude thermique annuelle légèrement inférieure à la moyenne.

A.5.3.4. Premier plan factoriel

Il est toujours intéressant d'étudier globalement un plan factoriel, même si, comme ici, chaque facteur est clairement interprétable.

Remarques préalables sur la représentation des variables (figure 4)

La projection sur le premier plan factoriel conservant 98.8% de l'inertie du nuage des mois (construit dans un espace de dimension 15) la déformation des longueurs et des angles des vecteurs représentant ces 12 variables est presque négligeable. Les extrémités des flèches associés aux 12 mois n'atteignent pas le cercle de rayon 1 (appelé cercle de corrélation) mais il s'en faut de très peu. On peut vérifier sur ce plan la représentation géométrique du coefficient de corrélation par le cosinus de l'angle entre les vecteurs représentant les variables. Par exemple, la corrélation entre janvier et juillet vaut 0.30, ce qui correspond à un angle de 72 degrés, angle que l'on peut mesurer sur le plan.

Insistons sur le fait que cette propriété, toujours vraie dans l'espace complet, ne se vérifie sur les plans factoriels que pour les variables parfaitement bien représentées. Ainsi, l'angle observé dans le plan entre juillet et la variable supplémentaire longitude vaut 45 degrés, angle dont le cosinus vaut 0.70. Mais la longitude n'est pas bien représentée sur ce plan, comme sa distance au cercle de corrélation permet de le constater. Il n'est donc pas étonnant que la corrélation entre juillet et longitude (0.59) diffère de 0.70.

Bilan des liaisons entre variables (figure 4)

Tous les angles entre les vecteurs représentant les variables étant inférieurs à un angle droit, les 12 températures mensuelles sont corrélées positivement entre elles. En plus, il apparaît une structure qui correspond au cycle annuel avec deux périodes. De janvier à juin, et de juillet (très proche de juin) à décembre (très proche de janvier), les mois se répartissent dans l'ordre du calendrier : deux mois proches dans le calendrier sont fortement liés entre eux (la corrélation entre deux mois consécutifs n'est jamais inférieure à 0.92) et dans chacune des deux périodes, cette liaison décroît régulièrement avec l'éloignement. Ceci correspond vraisemblablement à une inertie thermique. D'autre part, les mois des deux périodes se superposent quasiment deux à deux : il semble que deux mois sont d'autant plus liés qu'ils correspondent à la même durée du jour.

Variables synthétiques

Il est clairement apparu que l'évolution thermique annuelle de l'ensemble des 15 villes peut

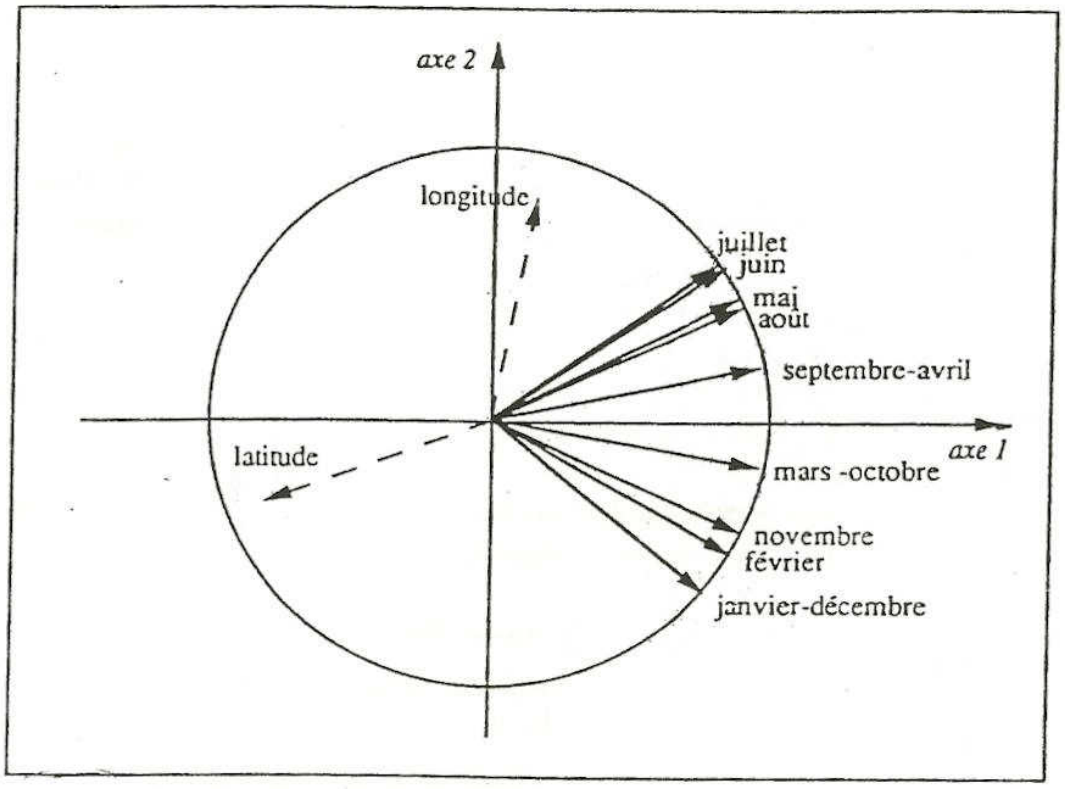

Figure 4 . *Projection des 12 variables actives et de 2 variables supplémentaires sur le plan des deux premiers facteurs*

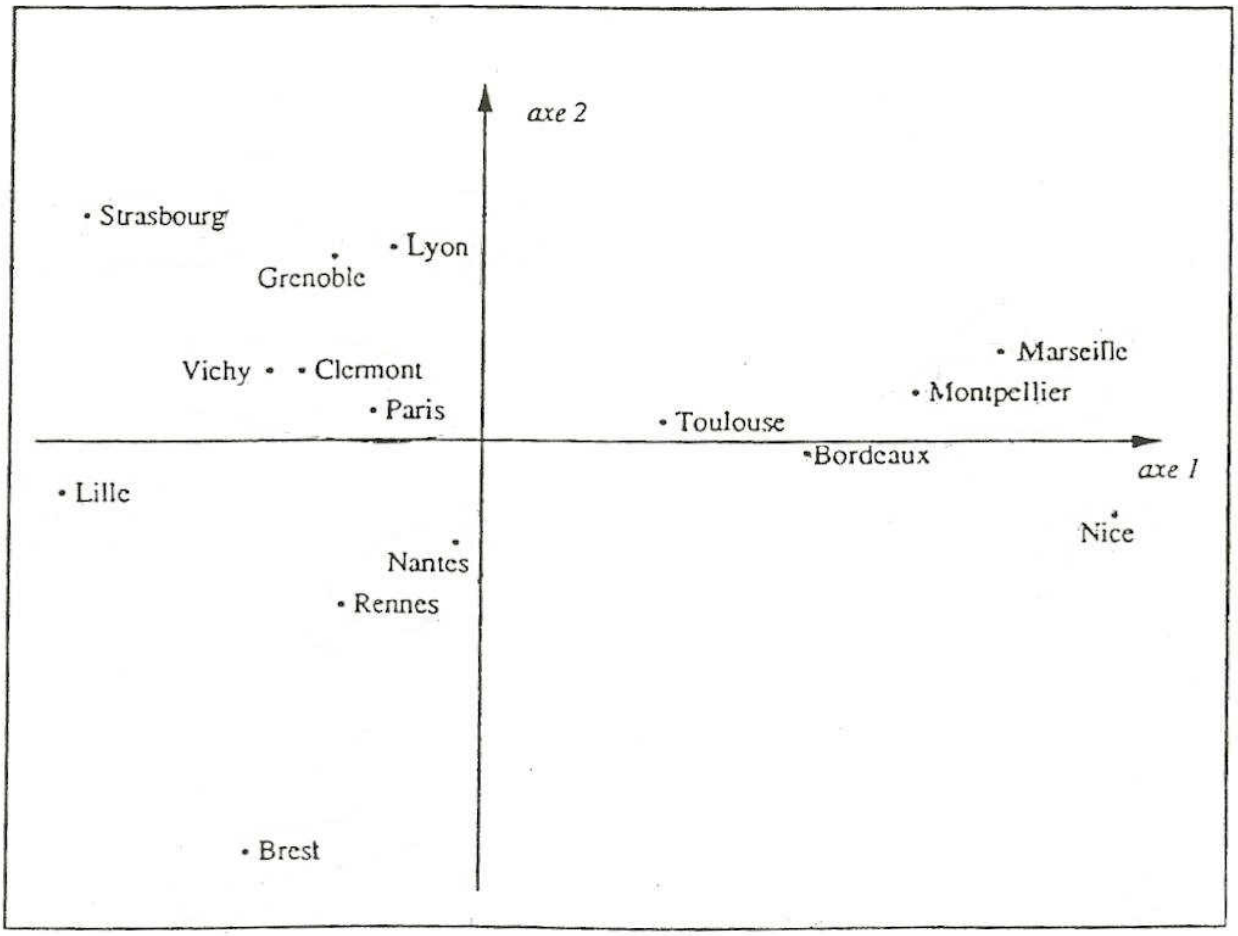

Figure 5 *Projection des 15 villes sur le premier plan factoriel*

être presque parfaitement synthétisée par deux variables : la température moyenne annuelle et l'amplitude thermique.

Typologies des villes (figure 5)

Sur ce plan, les deux axes correspondent aux deux variables synthétiques. Ainsi, plus une ville est froide, plus elle est située à gauche sur le plan et plus son amplitude thermique est grande, plus elle est située en haut.

Remarquons que les villes "chaudes", situées à droite, sont proches de l'axe horizontal : le deuxième facteur ne les différencie guère. Au contraire, pour les villes "froides" les différences d'amplitude thermique sont importantes.

La répartition sur le plan permet, un peu arbitrairement, de distinguer trois groupes de villes. L'interprétation des deux axes permet de caractériser ces groupes.

- Les villes à climat chaud : Marseille, Montpellier, Nice, Bordeaux et Toulouse
- Les villes à climat froid et continental (été chaud, hiver très froid) : Strasbourg, Lyon, Grenoble, Vichy, Clermont et Paris.
- Les villes à climat froid et océanique (été froid, hiver doux) : Brest, Rennes, Nantes et Lille.

Remarques sur la qualité de représentation des villes, (tableau 2)

La qualité de représentation d'un individu (par un axe, un plan ou un sous-espace) est une expression racourcie de "qualité de représentation de l'écart entre un individu et le point moyen" (par un axe, un plan ou un sous-espace). A la différence de celle des variables (dont la distance à l'origine est constante) la qualité de représentation des individus ne se lit pas directement sur le graphique. Il faut consulter le tableau 2 les indiquant.

Toutes les villes sont très bien représentées sur ce plan (ce qui n'est pas étonnant puisque la qualité globale est de 98.8). La moins bien représentée est Paris avec 0.89 + 0.01 =0.90. La différence entre les températures mensuelles de Paris et les températures mensuelles moyennes des 15 villes n'est pas totalement expliqué sur ce plan, pour l'examiner il faudrait consulter les facteurs suivants, le quatrième plus que le troisième puisque ses qualités de représentation sont respectivement de 0.03 et 0.07.

La coordonnée d'un individu est toujours interprétable, même si sa qualité de représentation par cet axe est mauvaise. Ainsi, bien que Paris soit mal représenté par le deuxième axe, sa

coordonnée presque nulle indique bien une amplitude thermique moyenne (vérifiable sur les données).

A.5.3.5. Conclusion

Ce cas est typique d'une ACP car il met un évidence un "effet taille" et une autre structure complémentaire qu'on peut appeler, en opposition à la première, "effet forme". En revanche il présente deux particularités. D'abord, le premier plan factoriel reconstitue presque parfaitement les données, ce qui d'autant plus rare que le nombre de variables est grand. Ensuite chacun des deux facteurs est facilement interprétable, ce qui est bien pratique pour un exemple à finalité pédagogique, mais l'utilisateur rencontre ordinairement des situations plus complexes.