La Statistique au service des Données

Quelques macros Excel
pour faire de l'analyse exploratoire des données

Jacques Vaillé

Février 2010

Table des macros disponibles

Avant-propos

La Statistique propose de nombreux outils et de nombreuses méthodes pour étudier un ensemble de données. Les recherches en ce domaine conduisent à des méthodes de plus en plus complexes pour étudier des problèmes spécifiques. D'un autre côté, le développement de l'informatique met ces outils à la portée d'un très large public. Un simple clic permet de mettre en œuvre les procédures les plus sophistiquées.

On ne peut que se réjouir de cette évolution, mais elle n'est pas sans risque.

- Puisqu'il est si facile d'obtenir une moyenne et un écart-type, pourquoi ne pas résumer ainsi sa série statistique, même si pour elle ces paramètres ne sont pas pertinents ?
- Pourquoi se priver de faire une analyse des correspondances sur un tableau de mesures hétérogènes ou une analyse en composantes principales sur un codage numérique de variables purement nominales, puisque les programmes fonctionnent très bien ?
- Pourquoi se préoccuper de la façon dont l'outil utilisé traite les données manquantes dans un tableau puisque de toute façon il donne un résultat ?

Et l'auteur en tire de belles conclusions cautionnées par la méthode statistique utilisée ***mais qui malheureusement n'ont rien à voir avec les données quand on les regarde d'un peu plus près***.
On pourrait multiplier les exemples tous pris dans de bonnes revues !

Il me semble qu'ils traduisent tous une primauté accordée à l'outil sur les données alors que la première fonction de l'outil devrait être de donner une vue plus claire des données surtout quand le tableau est un peu conséquent.

Les possibilités de publication actuelles permettent de joindre les tableaux de données utilisés. On peut souhaiter que ce soit toujours le cas ; cela permet au lecteur d'avoir une approche plus critique et de chercher les réponses qu'il peut légitimement se poser.

C'est dans le but de permettre à l'utilisateur de mieux appréhender ses données que j'ai conçu le fichier de macros complémentaires pour Excel : Explore.xla.
Quand on le charge, il ajoute un menu « Exploration » (sous Excel 2007, on le trouve dans les « Compléments »). Il regroupe diverse macros déjà proposées dans les cahiers Excel'Ense de Modulad et il permet d'utiliser directement certaines procédures contenues dans ces macros. Elles ont été corrigées de certains petits défauts et revues pour fonctionner aussi bien avec Excel 2003 qu'avec Excel 2007.

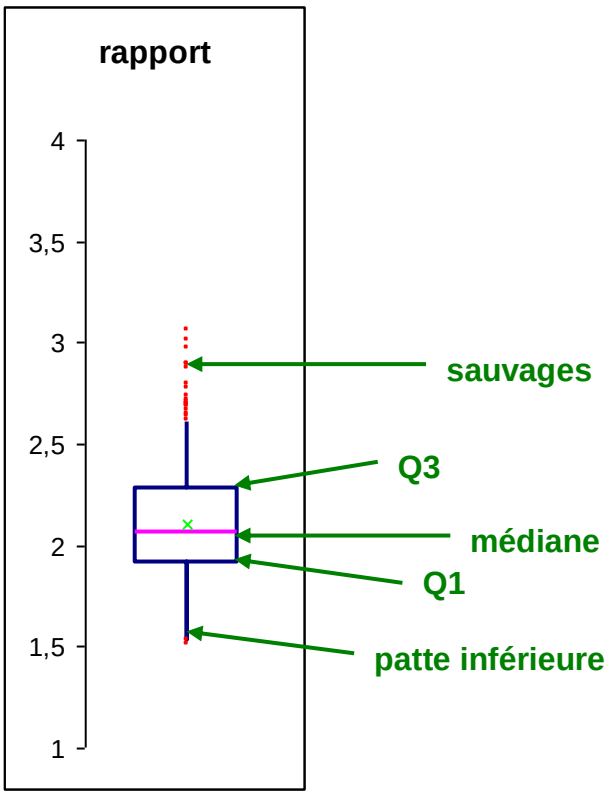
Ce document explique comment mettre en œuvre les méthodes proposées et ce qu'elles permettent de faire.

Je serai toujours heureux de connaître les utilisations que vous ferez de ces macros et de recevoir vos remarques et suggestions pour leur amélioration à l'adresse : jacques.vaille@free.fr

Boites de distribution

Le graphique

Une boite de distribution ou boite à pattes ou boite à moustaches est un graphique résumant la distribution d'une variable numérique.
Le corps de la boite est limité par le premier et le troisième quartile, la barre transversale correspondant à la médiane. La croix signale la position de la moyenne.
La « patte » inférieure a pour extrémité la première valeur de la série qui est inférieure ou égale à ***Q1 – 1.5 * (Q3 – Q1)*.**
La « patte » supérieure a pour extrémité la première valeur de la série qui est supérieure ou égale à ***Q3 + 1.5 * (Q3 – Q1).***
Les valeurs qui se trouvent au-delà des pattes sont appelées « sauvages », elles correspondent aux points rouges.
C'est un résumé graphique de la série statistique qui permet d'avoir une idée plus précise de la distribution. Par exemple, la série représentée ici a une distribution dissymétrique, étalée du côté des plus grandes valeurs



Réalisation

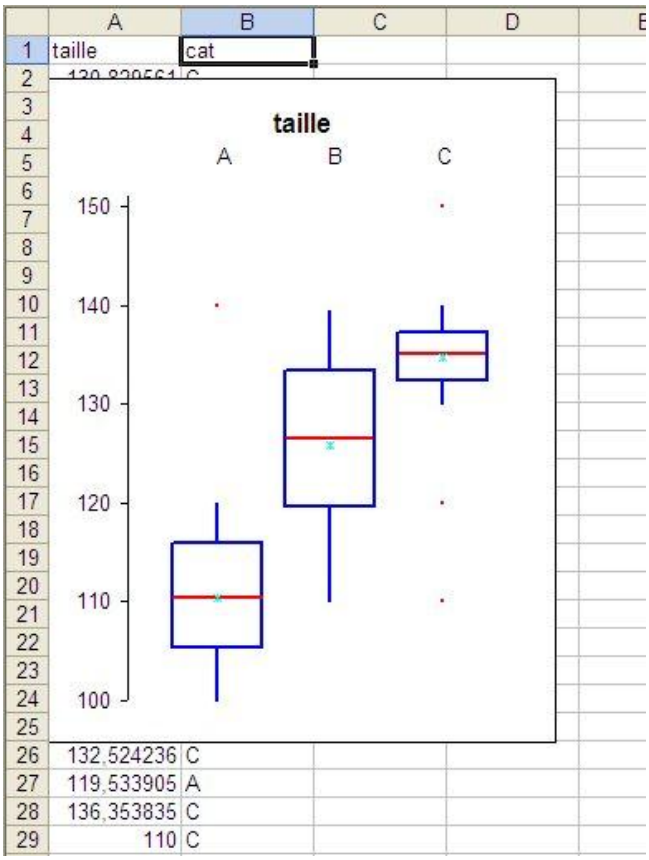
Sélectionner les séries de données pour lesquelles on veut obtenir les boites de distribution, en incluant la ligne contenant le nom de la série (*rapport* dans l'exemple).
Lancer le graphique en allant dans le menu **Exploration**>>**Boites de distribution**.
La représentation reflète les modifications des données ; si une valeur est modifiée et **sort de l'intervalle des valeurs originales**, il faut modifier le maximum ou le minimum sur l'axe vertical pour les faire apparaître sur le graphique qui a été optimisé pour les données originales (faire un double-clic sur cet axe).

Remarques

Le graphique n'est pertinent que s'il y a suffisamment de valeurs dans la série.
La macro ajoute une feuille cachée pour chaque boite dessinée et des noms avec le préfixe BP_ ; il ajoute aussi un nom pour chaque colonne utilisée qui est fourni par la première ligne du tableau.
Les résultats obtenus ne sont pas liés à Explore.xla. Vous pouvez donc les expédier sans problème à un correspondant.

Boites de distribution conditionnelle

Le graphique



	A	B
1	taille	cat
26	132,524236	C
27	119,533905	A
28	136,353835	C
29	110	C

Le principe de représentation est le même que pour les boîtes de distribution. Mais ici, on veut comparer les distributions d'une variable numérique dans les sous-populations repérées par les modalités d'une variable nominale (ces modalités peuvent être alphanumériques : homme, femme par exemple ou un simple codage par des numéros).
On obtient sur le même graphique, une boite pour chacune des modalités de la variable nominale.
Dans le graphique ci-contre, on a représenté la variable taille avec une boite pour chacune des modalités de la variable « cat », qui sont ici les trois lettres A, B et C.

Réalisation

Sélectionner les séries de données : la colonne de la variable nominale, puis les colonnes des valeurs numériques (si la **variable nominale n'est pas juste à gauche de la variable numérique**, faire une sélection multiple en utilisant la touche CTRL).
Lancer le graphique en allant dans le menu **Exploration**>>**Boites de distribution conditionnelle**.
On obtient un graphique par variable numérique avec une boite pour chaque modalité de la condition.
La représentation reflète les modifications des données ; si une valeur est modifiée et **sort de l'intervalle des valeurs originales**, il faut modifier le maximum ou le minimum sur l'axe vertical pour les faire apparaître sur le graphique qui a été optimisé pour les données originales (faire un double-clic sur cet axe). De même on peut changer la catégorie d'une ligne ***à condition de rester dans les catégories originales***. Par exemple, si pour le point de la catégorie C qui a pour taille 150, on remplace C par B, le point rouge va se déplacer sous le B ou la patte supérieure de cette boite va s'allonger jusqu'à ce point (s'il ne dépasse pas Q3+1.5*(Q3-Q1) dans la catégorie B).
Par contre si on change la codification de la variable nominale (1 remplacé par homme par exemple) la boite correspondante disparaît car les modalités sont extraites par la macro et ne restent pas liées à la colonne de départ.

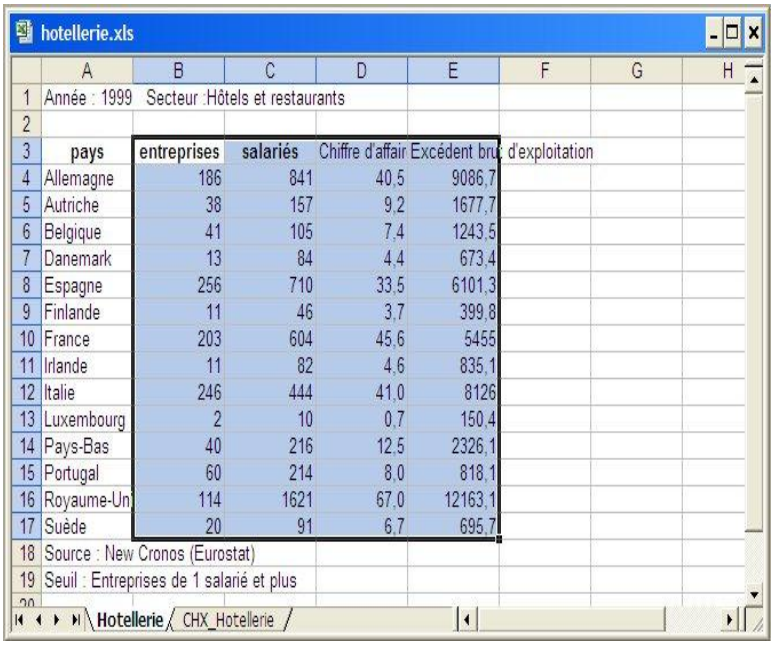
Remarques

Le graphique n'est pertinent que s'il y a suffisamment de valeurs dans la série.
La macro ajoute une feuille cachée pour chaque boite dessinée et des noms avec le préfixe BP_ ; il ajoute aussi un nom pour chaque colonne utilisée qui est fourni par la première ligne du tableau.
Les résultats obtenus ne sont pas liés à Explore.xla. Vous pouvez donc les expédier sans problème à un correspondant.

Nuages à partir d'un tableau

Le but

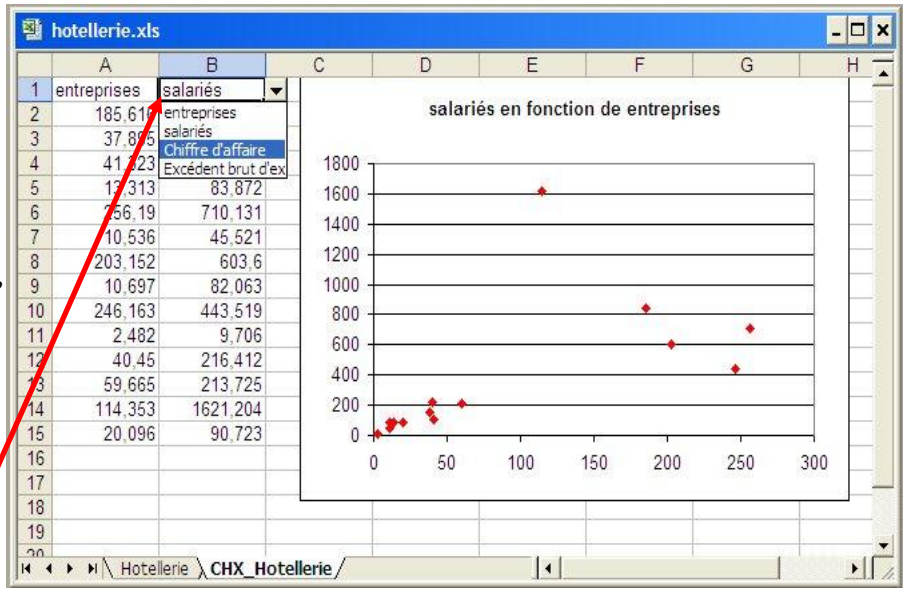
On a un tableau regroupant plusieurs variables mesurées sur les mêmes individus. Pour explorer les liaisons entre ces diverses variables, on veut examiner les nuages de points obtenus à partir de deux quelconques d'entre elles.



Réalisation

Sélectionner la partie du tableau contenant les données numériques (la ligne du haut doit contenir les noms des variables associées à chacune des colonnes).

On utilise le menu **Exploration>>Nuages à partir d'un tableau**.
On obtient un nuage de point associé aux deux premières colonnes dans une feuille nommée *CHX_NomDeLaFeuille*, ou *NomDeLaFeuille* est le nom de la feuille où se trouvent les données d'origine.
Les deux colonnes de gauche contiennent une copie des deux colonnes utilisées. En cliquant sur l'en-tête de la colonne, on obtient une liste déroulante permettant de choisir la variable à placer sur l'axe correspondant. On peut ainsi explorer tous les nuages possibles.



Si l'on veut ajouter des étiquettes aux points, voir plus loin : **« Étiquettes d'un nuage »**.

Remarques

La macro ajoute des noms avec le préfixe CHX_.
Les résultats obtenus ne sont pas liés à Explore.xla. Vous pouvez donc les expédier sans problème à un correspondant.

Nuage avec étiquettes

Le graphique

Il s'agit de tracer un nuage de points à partir de deux séries numériques en mettant comme étiquette un texte correspondant à chaque ligne de la série.

Réalisation

Les deux séries sélectionnées comme abscisses et ordonnées doivent être des colonnes d'un tableau dont la première ligne contient les noms des variables et la première colonne, les étiquettes de chaque ligne. Sélectionnez, ***dans cet ordre***, la colonne des étiquettes, la colonne des abscisses puis la colonne des ordonnées, en incluant chaque fois la ligne des noms de variables. Si les colonnes ne sont pas consécutives dans le tableau, ajouter chaque colonne en maintenant la touche CTRL enfoncée pendant que vous sélectionnez avec la souris.
Vous pouvez maintenant réaliser le graphique par **Exploration**>>**Nuage avec étiquettes**. On peut déplacer les étiquettes avec la souris pour rendre le graphique plus lisible.

C3 : salariés

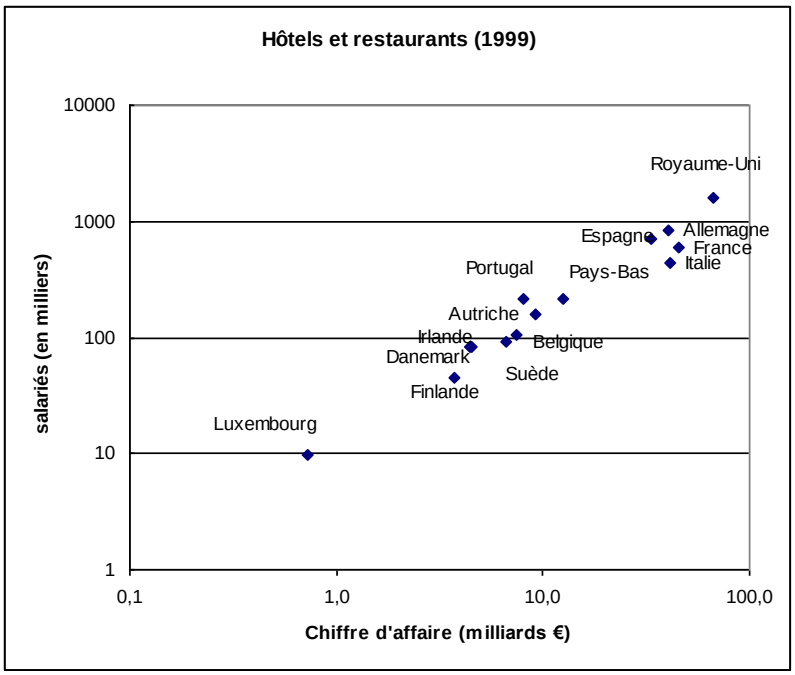
	A	B	C	D	E	F
1	Année : 1999	Secteur :Hôtels et restaurants				
2						
3	**pays**	**entreprises**	**salariés**	Chiffre d'affair	Excédent brut d'exploitation	
4	Allemagne	186	841	40,5	9086,7	
5	Autriche	38	157	9,2	1677,7	
6	Belgique	41	105	7,4	1243,5	
7	Danemark	13	84	4,4	673,4	
8	Espagne	256	710	33,5	6101,3	
9	Finlande	11	46	3,7	399,8	
10	France	203	604	45,6	5455	
11	Irlande	11	82	4,6	835,1	
12	Italie	246	444	41,0	8126	
13	Luxembourg	2	10	0,7	150,4	
14	Pays-Bas	40	216	12,5	2326,1	
15	Portugal	60	214	8,0	818,1	
16	Royaume-Uni	114	1621	67,0	12163,1	
17	Suède	20	91	6,7	695,7	
18	Source : New Cronos (Eurostat)					
19	Seuil : Entreprises de 1 salarié et plus					
20						

Les colonnes ont été sélectionnées dans l'ordre :
pays, chiffre d'affaire, salariés.

Pour rendre le graphique plus lisible, on a choisi des échelles logarithmiques pour les chiffres d'affaire et pour les effectifs.

Remarque

Les étiquettes restent même si Explore.xla n'est pas chargé, par exemple si vous avez envoyé le classeur à un correspondant.

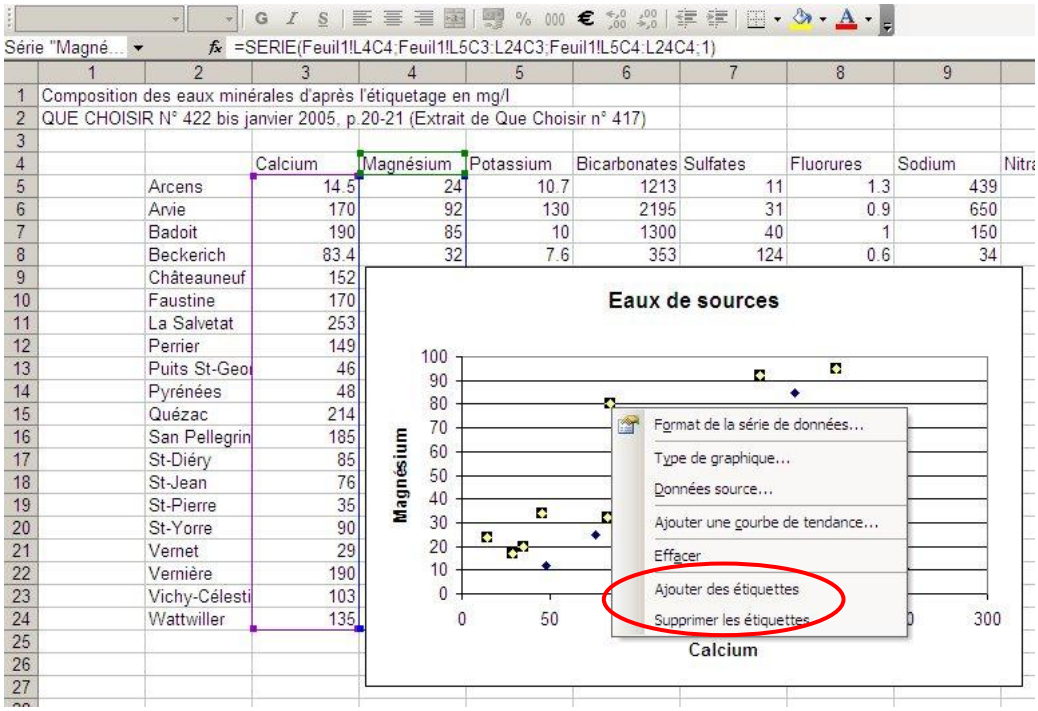


Étiquettes d'un nuage

Cela complète les possibilités du menu précédent.
Explore.xla ajoute deux lignes au menu contextuel des séries graphiques :

- Ajouter des étiquettes
- Supprimer les étiquettes

Ce menu s'obtient en faisant un click droit sur un des points de la série pour laquelle on veut travailler sur les étiquettes ***si on travaille avec Excel 2003***.

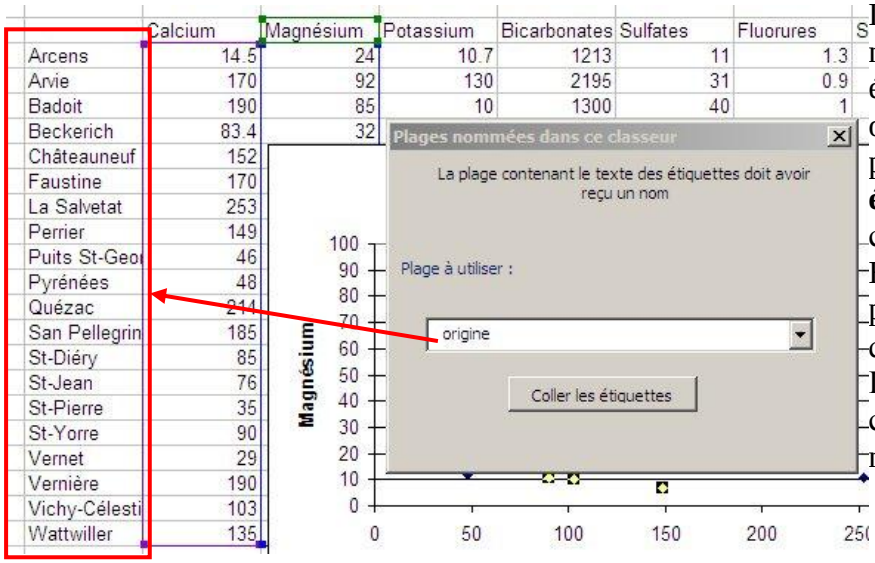


Avec Excel 2007, on trouve cela dans la deuxième partie du menu contenant les autres macros. Ce menu n'est plus accessible quand un graphique est sélectionné ! La macro ne fonctionne que si :

- il n'y a qu'un graphique dans la feuille
- il n'y a qu'une série représentée sur ce graphique. (Quel progrès cet Excel 2007 !)

Pour l'appeler, il faut sélectionner une case de la feuille de calcul et non le graphique.

Ajouter des étiquettes



Le nuage de points a été fait de la manière habituelle. Pour ajouter des étiquettes aux points de la série : ouvrez le menu contextuel sur un point de la série et >> **Ajouter des étiquettes**. La boite de dialogue ci-contre s'ouvre.
Elle permet de choisir parmi les plages nommées du classeur, celle qu'on va utiliser pour les étiquettes. Ici c'est la colonne de gauche contenant l'origine de l'eau minérale.

La taille de la plage est vérifiée, mais **l'utilisateur reste responsable de la correspondance entre les étiquettes et les données !**

Supprimer les étiquettes :

Menu contextuel >>**Supprimer les étiquettes**. Les étiquettes attachées aux points sont supprimées. Cela fonctionne aussi avec les graphiques faits dans la section précédente.

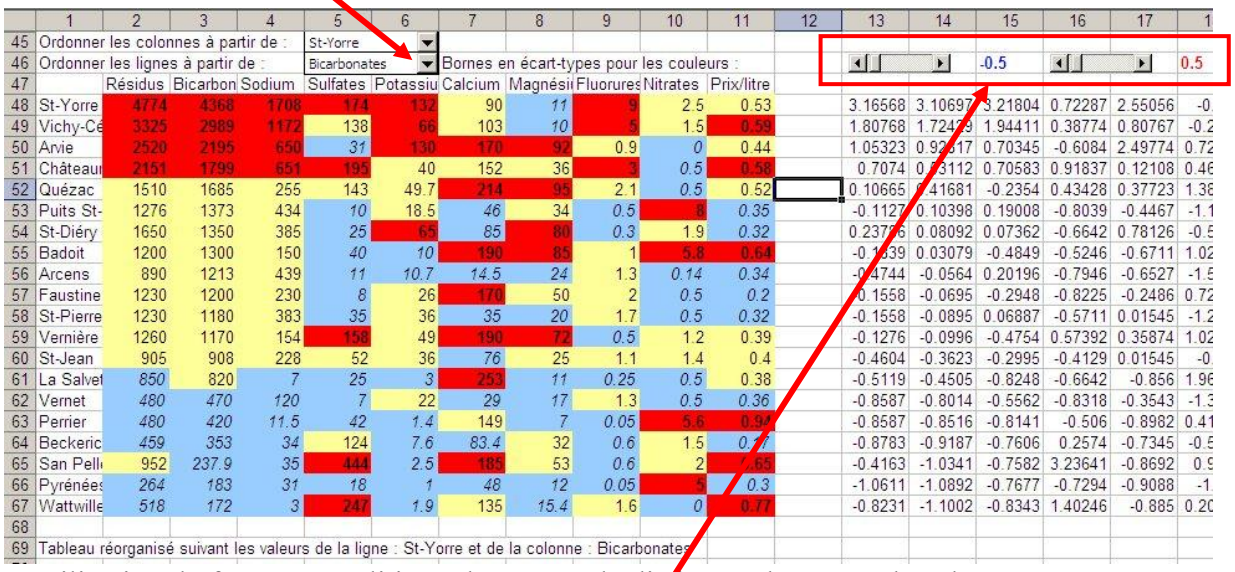
Remarque

Les étiquettes restent même si Explore.xla n'est pas chargé, par exemple si vous avez envoyé le classeur à un correspondant.

Graphique de Bertin sur tableau de mesures

Principe

Dans *La graphique et le traitement graphique de l'information*, Jacques Bertin proposait en 1977 de mettre en évidence certaines propriétés d'un tableau de données en réordonnant ses lignes et ses colonnes. C'est ce que cette macro permet de mettre en œuvre.
Comme il s'agit de **mesures qui peuvent être d'unités et d'ordre de grandeur différents**, on utilise ici les données centrées réduites pour ordonner les colonnes à partir des valeurs d'une ligne exprimées en nombre d'écart-type. Elles sont calculées dans la partie de droite de la feuille. Elles servent à réordonner le tableau d'origine dans la partie de gauche.
Deux **listes déroulantes** permettent de choisir la colonne et la ligne par rapport auxquelles le tableau est réordonné.



		Résidus	Bicarbon	Sodium	Sulfates	Potassiu	Calcium	Magnésit	Fluorures	Nitrates	Prix/litre
45	Ordonner les colonnes à partir de :				St-Yorre						
46	Ordonner les lignes à partir de :				Bicarbonates		Bornes en écart-types pour les couleurs :				
48	St-Yorre	4774	4368	1708	174	132	90	11	9	2.5	0.53
49	Vichy-Cé	3325	2989	1172	138	66	103	10	5	1.5	0.59
50	Arvie	2520	2195	650	31	130	170	92	0.9	0	0.44
51	Châteaur	2151	1799	651	195	40	152	36	3	0.5	0.58
52	Quézac	1510	1685	255	143	49.7	214	95	2.1	0.5	0.52
53	Puits St-	1276	1373	434	10	18.5	46	34	0.5	8	0.35
54	St-Diéry	1650	1350	385	25	65	85	80	0.3	1.9	0.32
55	Badoit	1200	1300	150	40	10	190	85	1	5.8	0.64
56	Arcens	890	1213	439	11	10.7	14.5	24	1.3	0.14	0.34
57	Faustine	1230	1200	230	8	26	170	50	2	0.5	0.2
58	St-Pierre	1230	1180	383	35	36	35	20	1.7	0.5	0.32
59	Vernière	1260	1170	154	158	49	190	72	0.5	1.2	0.39
60	St-Jean	905	908	228	52	36	76	25	1.1	1.4	0.4
61	La Salvet	850	820	7	25	3	253	11	0.25	0.5	0.38
62	Vernet	480	470	120	7	22	29	17	1.3	0.5	0.36
63	Perrier	480	420	11.5	42	1.4	149	7	0.05	5.6	0.94
64	Beckeric	459	353	34	124	7.6	83.4	32	0.6	1.5	0.77
65	San Pell	952	237.9	35	444	2.5	185	53	0.6	2	1.65
66	Pyrénées	264	183	31	18	1	48	12	0.05	5	0.3
67	Wattwille	518	172	3	247	1.9	135	15.4	1.6	0	0.77
68											
69	Tableau réorganisé suivant les valeurs de la ligne : St-Yorre et de la colonne : Bicarbonates										

45																		
46													-0.5				0.5	
47																		
48	3.16568	3.10697	3.21804	0.72287	2.55056	-0												
49	1.80768	1.72429	1.94411	0.38774	0.80767	-0.2												
50	1.05323	0.92617	0.70345	-0.6084	2.49774	0.72												
51	0.7074	0.53112	0.70583	0.91837	0.12108	0.46												
52	0.10665	-0.41681	-0.2354	0.43428	0.37723	1.38												
53	-0.1127	0.10398	0.19008	-0.8039	-0.4467	-1.1												
54	0.23776	0.08092	0.07362	-0.6642	0.78126	-0.5												
55	-0.1039	0.03079	-0.4849	-0.5246	-0.6711	1.02												
56	-0.4744	-0.0564	0.20196	-0.7946	-0.6527	-1.5												
57	0.1558	-0.0695	-0.2948	-0.8225	-0.2486	0.72												
58	-0.1558	-0.0895	0.06887	-0.5711	0.01545	-1.2												
59	-0.1276	-0.0996	-0.4754	0.57392	0.35874	1.02												
60	-0.4604	-0.3623	-0.2995	-0.4129	0.01545	-0												
61	-0.5119	-0.4505	-0.8248	-0.6642	-0.856	1.96												
62	-0.8587	-0.8014	-0.5562	-0.8318	-0.3543	-1.3												
63	-0.8587	-0.8516	-0.8141	-0.506	-0.8982	0.41												
64	-0.8783	-0.9187	-0.7606	0.2574	-0.7345	-0.5												
65	-0.4163	-1.0341	-0.7582	3.23641	-0.8692	0.9												
66	-1.0611	-1.0892	-0.7677	-0.7294	-0.9088	-1												
67	-0.8231	-1.1002	-0.8343	1.40246	-0.885	0.20												
68																		
69																		

L'utilisation de formats conditionnels permet de distinguer les cases dont le contenu est supérieur de plus de α écart-type de la moyenne (rouge) et celles dont la valeur est inférieure de plus de β écart-type de la moyenne (bleue).
Ces deux seuils sont fixés par les **deux ascenseurs** situés à droite.
Les choix faits ci-dessus font apparaître une liaison importante entre les résidus, les bicarbonates et le sodium. Par contre le calcium n'a aucun rapport avec les colonnes précédentes.

Réalisation

Il faut sélectionner le tableau des **valeurs numériques avec les identificateurs dans la colonne de gauche et la ligne du haut**. Ce tableau doit comporter sur sa gauche, une colonne contenant le **nom des lignes** et au-dessus, une ligne contenant le **nom des colonnes** ou variables mesurées.
Lancer le graphique par le menu **Exploration**>>**Bertin sur tableau de mesures**.
Le résultat ci-dessus est placé dans une nouvelle feuille nommée BRM_*NomDeLaFeuille*,
Où *NomDeLaFeuille* est le nom de la feuille contenant le tableau des données.

Remarques

La macro rajoute un certain nombre de noms dans le classeur qui permettent au graphique de fonctionner. Ils commencent tous par le préfixe BRM_.
Le graphique ne peut fonctionner que si la feuille Explore.xla est chargée. Si vous envoyez le classeur à un correspondant, il faut donc lui envoyer aussi Explore.xla.

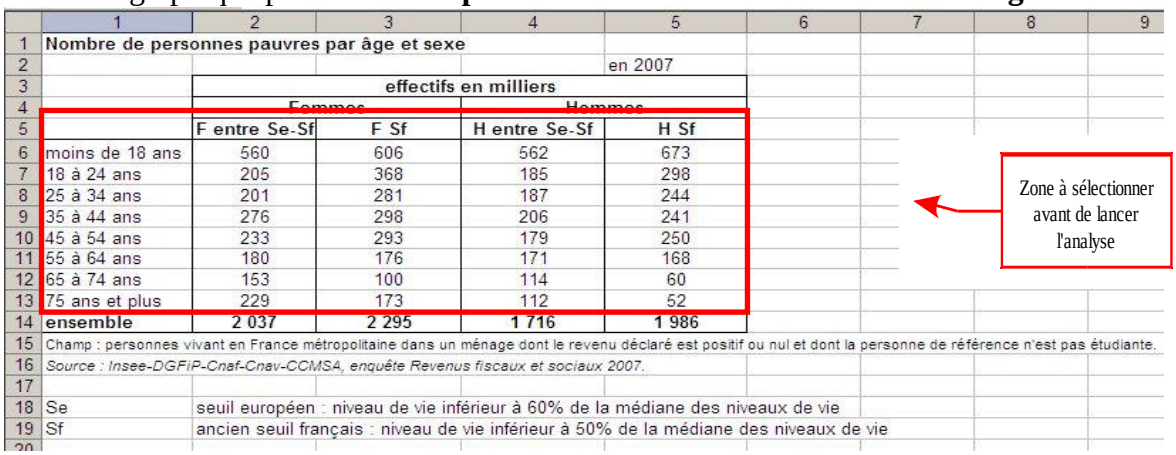
Graphique de Bertin sur tableau de contingence

Principe

Le principe est le même que pour un graphique de Bertin sur un tableau de mesures. Ici toutes **les données sont des effectifs**, on peut utiliser les répartitions marginales.
Les interprétations les plus pertinentes se font en comparant des profils lignes ou des profils colonnes. Les publications donnent souvent des tableaux de pourcentages, il faut autant que possible utiliser les données de départ : les effectifs. *Si on a les pourcentages de chaque colonne par exemple, les profils des lignes n'ont plus de sens.*

Réalisation

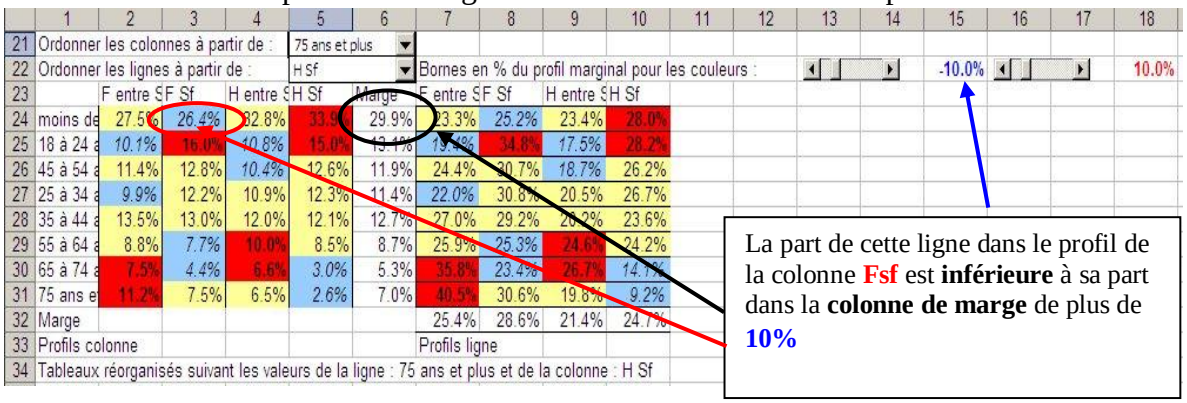
Il faut sélectionner le tableau des **effectifs**. Ce tableau doit comporter à gauche, une colonne contenant le **nom des lignes** et en haut, une ligne contenant le **nom des colonnes**.
Lancer le graphique par le menu **Exploration>>Bertin sur tableau de contingence**.



	1	2	3	4	5
1	Nombre de personnes pauvres par âge et sexe				
2					en 2007
3		effectifs en milliers			
4		Femmes		Hommes	
5		F entre Se-Sf	F Sf	H entre Se-Sf	H Sf
6	moins de 18 ans	560	606	562	673
7	18 à 24 ans	205	368	185	298
8	25 à 34 ans	201	281	187	244
9	35 à 44 ans	276	298	206	241
10	45 à 54 ans	233	293	179	250
11	55 à 64 ans	180	176	171	168
12	65 à 74 ans	153	100	114	60
13	75 ans et plus	229	173	112	52
14	ensemble	2 037	2 295	1 716	1 986
15	Champ : personnes vivant en France métropolitaine dans un ménage dont le revenu déclaré est positif ou nul et dont la personne de référence n'est pas étudiante.				
16	*Source : Insee-DGFiP-Cnaf-Cnav-CCMSA, enquête Revenus fiscaux et sociaux 2007.*				
17					
18	Se	seuil européen : niveau de vie inférieur à 60% de la médiane des niveaux de vie			
19	Sf	ancien seuil français : niveau de vie inférieur à 50% de la médiane des niveaux de vie			

Zone à sélectionner avant de lancer l'analyse

Le résultat ci-dessous est placé dans une nouvelle feuille nommée BRC_*NomDeLaFeuille*, Où *NomDeLaFeuille* est le nom de la feuille contenant le tableau des données.
On retrouve les listes pour choisir ligne ou colonne et les ascenseurs pour choisir les seuils.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
21	Ordonner les colonnes à partir de :				75 ans et plus													
22	Ordonner les lignes à partir de :				H Sf		Bornes en % du profil marginal pour les couleurs :								-10.0%			10.0%
23		F entre Sf	F Sf	H entre Sf	H Sf	Marge	F entre Sf	F Sf	H entre Sf	H Sf								
24	moins de	27.5%	26.4%	32.8%	33.5%	29.9%	23.3%	25.2%	23.4%	28.0%								
25	18 à 24 a	10.1%	16.0%	10.8%	15.0%	13.1%	19.4%	34.8%	17.5%	28.2%								
26	45 à 54 a	11.4%	12.8%	10.4%	12.6%	11.9%	24.4%	30.7%	18.7%	26.2%								
27	25 à 34 a	9.9%	12.2%	10.9%	12.3%	11.4%	22.0%	30.8%	20.5%	26.7%								
28	35 à 44 a	13.5%	13.0%	12.0%	12.1%	12.7%	27.0%	29.2%	26.2%	23.6%								
29	55 à 64 a	8.8%	7.7%	10.0%	8.5%	8.7%	25.9%	25.3%	24.6%	24.2%								
30	65 à 74 a	7.5%	4.4%	6.6%	3.0%	5.3%	35.8%	23.4%	26.7%	14.1%								
31	75 ans et	11.2%	7.5%	6.5%	2.6%	7.0%	40.5%	30.6%	19.8%	9.2%								
32	Marge						25.4%	28.6%	21.4%	24.7%								
33	Profils colonne						Profils ligne											
34	Tableaux réorganisés suivant les valeurs de la ligne : 75 ans et plus et de la colonne : H Sf																	

La part de cette ligne dans le profil de la colonne **Fsf** est **inférieure** à sa part dans la **colonne de marge** de plus de **10%**

Attention : les interprétations se font en comparant des profils ! On a 26% dans la case entourée en rouge, elle ressort sur fond bleu car dans le profil marginal (ensemble de la population étudiée) on a 29,9% ; on est plus de 10% ***en-dessous***. Cette catégorie (moins de 18 ans) est donc **sous-représentée** dans cette colonne ! (même si c'est la plus nombreuse).

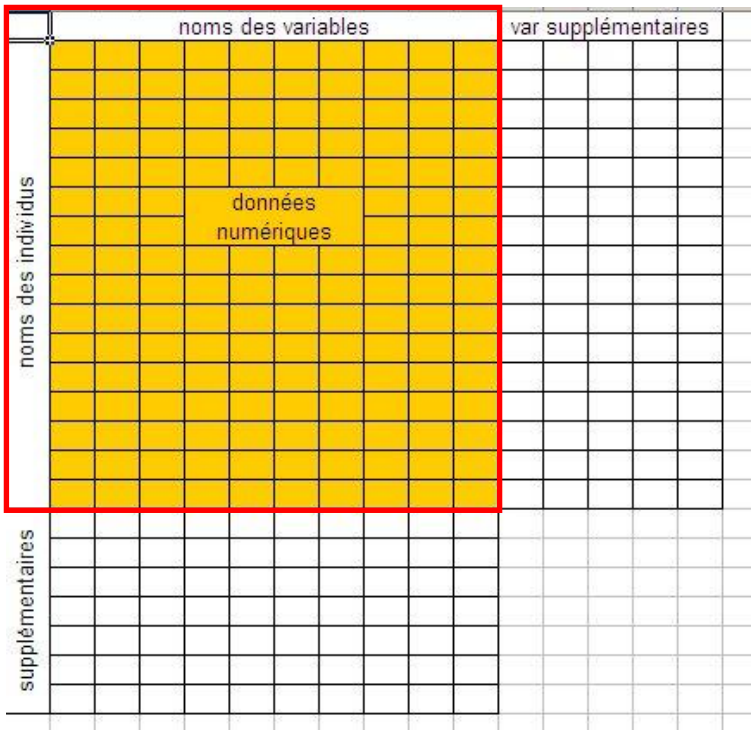
Remarques

La macro rajoute un certain nombre de noms dans le classeur qui permettent au graphique de fonctionner. Ils commencent tous par le préfixe BRC_.
Le graphique ne peut fonctionner que si la feuille Explore.xla est chargée. Si vous envoyez le classeur à un correspondant, il faut donc lui envoyer aussi Explore.xla.

ACP normée

Tableau des données

Le tableau que l'on veut analyser doit comporter en colonnes les variables mesurées pour des individus statistiques situés en lignes. Les données doivent être présentées sous la forme suivante :



la partie active du tableau (variables et individus qui vont servir à déterminer les axes factoriels et leurs identificateurs) **doit être sélectionnée** avant de lancer l'analyse ; c'est la partie **encadrée en rouge** du tableau ci-contre.

La ligne au-dessus doit contenir le nom de chaque variable et la colonne à droite de ce tableau, le nom de chaque individu.

S'il y a des données immédiatement à droite de la sélection, elles sont considérées comme des **variables supplémentaires**. Dès qu'une colonne contient autre chose que des valeurs numériques, cette colonne et les suivantes sont considérées comme des variables nominales (une modalité pour chaque valeur différente trouvée dans la colonne).

De même s'il y a des données numériques immédiatement en-dessous de la sélection, les lignes sont considérées comme des **individus supplémentaires ou illustratifs**.

Le tableau ne doit pas comporter de données manquantes.

Si c'est le cas il faut soit supprimer des lignes ou des colonnes, ou estimer ces valeurs manquantes; L'estimation la plus simple est de prendre la moyenne des autres lignes pour cette variable,

Réalisation de l'analyse

Les données **actives** étant sélectionnées,

on réalise l'analyse en choisissant **Exploration>>ACP normée.**

La feuille de départ est renommée « Données ».

Plusieurs feuilles ont été ajoutées au classeur :

- La feuille active se nomme « Résultats », on y trouve tous les éléments permettant de comprendre et d'interpréter les résultats de l'analyse.
- « ACP » : elle contient les calculs préliminaires à l'analyse : moyennes et écarts-types, matrice centrée réduite et matrice des corrélations.
- « Aux » : feuille cachée de calculs auxiliaires utilisés pour les représentations graphiques.
- « Permut » : feuille utilisée pour réordonner lignes et colonnes.
- « Diagonalisation » : qui contient la matrice de passage calculée à l'aide de la fonction =*matpassage*(matrice de corrélation), et les valeurs propres données par l'analyse.

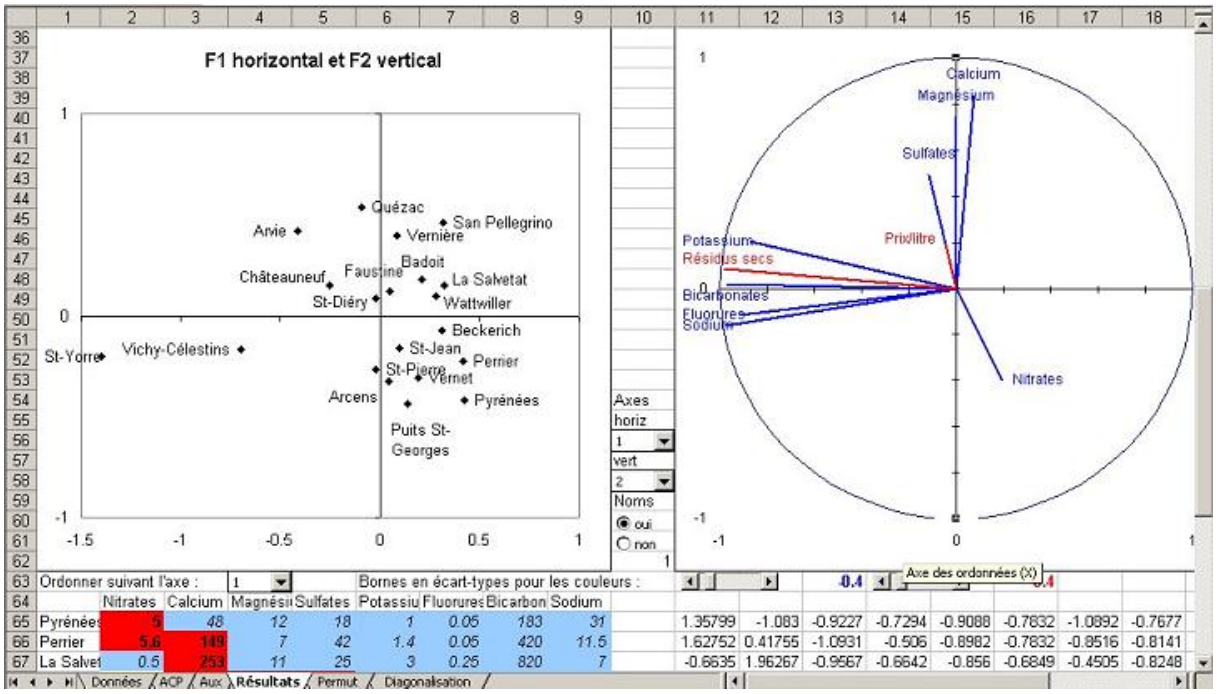
Les Résultats

Aides à l'interprétation :

Au-dessus des graphiques, on trouve les tableaux classiques des aides à l'interprétation d'une analyse factorielle : contributions aux axes et cosinus carrés.

Cartes factorielles :

On a en vis-à-vis, la carte des individus et le cercle des corrélations.



Remarque : certaines étiquettes ont été déplacées manuellement pour une meilleure lisibilité.

Au centre, deux listes déroulantes permettent de choisir les axes factoriels que l'on veut utiliser comme axes du graphique. Les colonnes ou variables actives sont en bleu, les illustratives en rouge.

Deux boutons radio permettent d'afficher ou non les noms des lignes ou individus.

S'il y a des variables nominales illustratives, une liste supplémentaire s'affiche au dessus des deux autres pour choisir la variable à faire apparaître sur le graphique : chaque modalité s'affiche en dessous dans la couleur utilisée pour cette catégorie dans le plan des individus.

Lecture des cartes factorielles :

- il s'agit de projections d'un espace multidimensionnel sur un plan : **attention aux effets de perspective** ! Deux projections peuvent paraître proches alors que les points sont situés de part et d'autre du plan de projection. Les proximités ne sont réelles que si la qualité de représentation sur le plan (somme des cosinus carrés sur chacun des deux axes) est bonne (proche de 1). Pour les variables, ce sont celles qui sont proches du cercle des corrélations.
- Le vecteur représentant une variable indique **dans quelle direction cette variable augmente**. Deux variables proches et bien représentées ont une corrélation importante.
- Pour les points, le **centre du graphique** correspond à un point qui a pour chaque variable **la moyenne** de cette variable. Si un point s'éloigne du centre dans la direction d'une variable, c'est qu'il a pour cette variable une valeur supérieure à la moyenne ; s'il s'éloigne à l'opposé, c'est qu'il a pour cette variable, une valeur inférieure à la moyenne.

- Deux **points sont proches** si leurs valeurs sur l'ensemble des variables sont voisines. Ici encore il faut faire attention aux effets de perspective : la qualité de représentation dans le plan ne se traduit pas par une propriété géométrique.

Graphique de Bertin :

La partie basse de la feuille utilise un graphique de Bertin pour visualiser les propriétés mises en évidence par l'analyse **sur un axe factoriel.**

- Une liste déroulante permet de choisir **l'axe à partir duquel le tableau sera réordonné**. Les colonnes sont rangées de gauche à droite suivant leur abscisse sur l'axe factoriel choisi, tandis que les lignes sont rangées de bas en haut suivant le même principe.
- Les cases rouges correspondent à des valeurs élevées, les bleues à des valeurs faibles et les jaunes à des valeurs médianes.

Ordonner suivant l'axe :	1			Bornes en écart-types pour les couleurs			
	Nitrates	Calcium	Magnési	Sulfates	Potassiu	Fluorures	Bicarbon	Sodium
Pyrénées	5	48	12	18	1	0.05	183	31
Perrier	5.6	149	7	42	1.4	0.05	420	11.5
La Salvet	0.5	253	11	25	3	0.25	820	7
San Pell	2	185	53	444	2.5	0.6	237.9	35
Beckeric	1.5	83.4	32	124	7.6	0.6	353	34
Wattwille	0	135	15.4	247	1.9	1.6	172	3
Badoit	5.8	190	85	40	10	1	1300	150
Vernet	0.5	29	17	7	22	1.3	470	120
Puits St-	8	46	34	10	18.5	0.5	1373	434
St-Jean	1.4	76	25	52	36	1.1	908	228
Vernière	1.2	190	72	158	49	0.5	1170	154
Faustine	0.5	170	50	8	26	2	1200	230
Arcens	0.14	14.5	24	11	10.7	1.3	1213	439
St-Pierre	0.5	35	20	35	36	1.7	1180	383
St-Diéry	1.9	85	80	25	65	0.3	1350	385
Quézac	0.5	214	95	143	49.7	2.1	1685	255
Châteaur	0.5	152	36	195	40	3	1799	651
Arvie	0	170	92	31	130	0.9	2195	650
Vichy-Cé	1.5	103	10	138	66	5	2989	1172
St-Yorre	2.5	90	11	174	132	9	4368	1708
Tableau réorganisé suivant les positions sur l'axe F1								

On retrouve sur la partie gauche, les quatre variables corrélées négativement avec le premier axe. Cette corrélation est surtout due à l'opposition entre les eaux chargées en bicarbonates, sodium, potassium et fluorure (rouge dans ce tableau) et eau très peu minéralisées (en bleu sur ce tableau). La position de la variable illustrative « Résidus secs », vient conforter cette interprétation.

Deux barres de défilement permettent de choisir les bornes supérieures et inférieures utilisées pour les couleurs. Pour l'ACP, les bornes sont en écarts-type au-dessus ou au-dessous de la moyenne.

Remarques

La macro rajoute un certain nombre de noms dans le classeur qui permettent au graphique de fonctionner. Ils commencent tous par le préfixe ACP_.

Le graphique ne peut fonctionner que si la feuille Explore.xla est chargée. Si vous envoyez le classeur à un correspondant, il faut donc lui envoyer aussi Explore.xla.

AFC binaire

Tableau des données

Le tableau que l'on veut analyser doit être un **tableau de contingence**, c'est-à-dire que la valeur dans une case correspond à l'**effectif** des individus qui appartiennent aux catégories de la ligne et de la colonne considérée. Par extension on peut l'appliquer sur un tableau de **données positives et homogènes** : la condition étant que la somme de chaque ligne et de chaque colonne ait un sens par rapport aux données analysées.
Il vaut mieux que le tableau ait plus de lignes que de colonnes. On peut éventuellement le transposer avant l'analyse.
La présentation du tableau est la même que pour l'ACP normée, avec une ligne au-dessus et une colonne à gauche pour les identificateurs.

Réalisation de l'analyse

Les données **actives** étant sélectionnées,
on réalise l'analyse en choisissant **Exploration**>>**AFC binaire.**
On obtient les mêmes feuilles que dans le cas précédent.

Les Résultats

Aides à l'interprétation :

Au-dessus des graphiques, on trouve les tableaux classiques des aides à l'interprétation d'une analyse factorielle : contributions aux axes et cosinus carrés.

Cartes factorielles :

Dans un tel tableau, lignes et colonnes jouent le même rôle, on n'a pas à priori de variables et d'individus, mais des catégories en lignes et d'autres en colonnes.
On obtient donc une seule carte factorielle où sont représentées simultanément les lignes et les colonnes.

Lecture des cartes factorielles :

Comme pour toute analyse factorielle, il s'agit de projections d'un espace multidimensionnel sur un plan : **attention aux effets de perspective** !
Une difficulté supplémentaire par rapport à l'ACP, c'est qu'il n'y a pas de cercle des corrélations : il faut **additionner les cosinus carrés par rapport aux deux axes** pour avoir la qualité de représentation d'un point dans le plan.
Ce genre d'analyse repose sur des profils (pourcentage de la case par rapport à la somme de la ligne : **profil-ligne** ou par rapport à la somme de la colonne : **profil-colonne**) et non sur les effectifs bruts. **Toute lecture doit donc se faire en termes de profils.**
Le centre du graphique correspond aux profils marginaux (pourcentage de la ligne par rapport au total du tableau ou pourcentage de la colonne par rapport à ce même total). Une ligne (colonne) s'éloigne du centre si son profil est différent de celui de la ligne (colonne) de marge. Deux lignes (colonnes) sont proches si leurs profils se ressemblent (même si leurs effectifs sont très différents).
La **proximité** entre un point représentant une ligne et un point représentant une colonne **n'a pas de sens** !
Par contre, la position **d'une colonne** s'interprète en considérant les **lignes comme des variables** : si une colonne s'éloigne du centre dans la direction de certaines lignes, c'est que, dans son profil, ces lignes sont sur représentées par rapport à l'ensemble de la population (par exemple 20% dans la colonne et 10% dans l'ensemble). De même si une colonne s'éloigne du

centre à l'opposé de certaines lignes, c'est que, dans son profil, ces lignes sont sous-représentées (par exemple 60% au lieu de 80%).
La position d'une ligne s'interprète de même en considérant cette fois les colonnes comme des variables.

Graphiques de Bertin :

Il y en a deux ici : un sur les profils-ligne et l'autre sur les profils-colonne. Pour l'interprétation, voir ce qui est dit pour les graphiques de Bertin sur tableaux de contingence,

Remarques

La macro rajoute un certain nombre de noms dans le classeur qui permettent au graphique de fonctionner. Ils commencent tous par le préfixe AFC_.
Le graphique ne peut fonctionner que si la feuille Explore.xla est chargée. Si vous envoyez le classeur à un correspondant, il faut donc lui envoyer aussi Explore.xla.