ANALYSE BIVARIÉE DE VARIABLES QUALITATIVES
LE TEST DU *Chi2*

Dominique LAFFLY

Maître de Conférences, Université de Pau
Laboratoire Société Environnement Territoire
UMR 5603 du CNRS et Université de Pau
Domaine Universitaire, IRSAM, 64000 PAU
Tél : 05 59 92 31 23 Fax : 05 59 80 83 39
Mail : dominique.laffly@univ-pau.fr

Le test du *Chi2* consiste à mesurer l'écart entre une situation observée et une situation théorique et d'en déduire l'existence et l'intensité d'une liaison mathématique. Par exemple, en théorie il y a autant de chance d'obtenir « pile » que « face » au lancer d'une pièce de monnaie, en pratique il n'en est rien. Le *Chi2* mesure alors l'écart entre la distribution théorique (une chance sur 2) est celle observée à la suite des lancements successifs.

En sciences sociales – notamment en géographie – on utilise le test du *Chi2* dans la même logique que celle appliquée au calcul du coefficient de corrélation linéaire pour des variables quantitatives : existe-t-il une liaison entre deux variables, si oui quelle est son intensité ?

Avec des données qualitatives (tranche d'âge, mode de déplacement, CSP…) il est nécessaire de reformuler les hypothèses initiales. D'un point de vue mathématique, il existe une situation théorique d'indépendance de deux variables qualitatives (notons dès à présent qu'ici on démontrera l'indépendance pour démontrer *a contrario* la dépendance éventuelle). On confronte une situation observée et une situation théorique d'indépendance mathématique. La première représente les effectifs observés lorsque l'on croise les différentes modalités des deux variables initiales, la seconde les effectifs théoriques. Les tests qui suivront seront fondés sur les écarts – distances – entre ces deux cas.

D'un point de vue mathématique on dit que la variable X est indépendante de la variable Y si la proportion des unités qui sont dans X_i et Y_j parmi toutes celles qui sont dans Y_j est la même que la proportion de celles qui sont dans X_i, dans la population totale, ceci étant vrai pour toutes valeurs de i et j, ce qui s'écrit :

$$\frac{n_{i,j}}{n_j} = \frac{n_i}{n} \text{ pour } i = 1, 2, \ldots, h \text{ et } j = 1, 2, \ldots, k$$

Ou encore

$$n_{i,j} = \frac{\left(n_i * n_j\right)}{n_{\bullet\bullet}}$$

En pratique, afin de tenir compte des fluctuations d'échantillonnage, on calcule des effectifs théoriques n'_{ij} en tenant compte des distributions conditionnelles notées $n_{i\bullet}$ pour somme des lignes, $n_{\bullet j}$ pour la somme des colonnes et $n_{\bullet\bullet}$ pour la somme de toutes les cellules. Soit :

$$n'_{i,j} = \frac{\left(n_{i\bullet} * n_{\bullet j}\right)}{n_{\bullet\bullet}}$$

	HOM	FEM	<25ans	25-35a	35-45a	45-55a	55-65a	>65ans	SA	Agri	Artisa	CadSup	ProfInt	Empl	Ouv	Etud	Retrai
HOM	69		19	10	8	17	5	10	7	5	5	11	4	7	6	15	9
FEM		68	20	9	10	17	5	7	13	3	5	4	4	10	1	18	10
<25ans			39						4			1		3	1	30	
25-35a				19					5		2	1	2	4	3	2	
35-45a					18				4	1	4	3	1	4	1		
45-55a						34			4	6	3	10	4	5	1		1
55-65a							10		2	1	1			1	1		4
>65ans								17	1				1			1	14
SA									20								
Agri										8							
Artisa											10						
CadSup												15					
ProfInt													8				
Empl														17			
Ouv															7		
Etud																33	
Retrai																	19

Le tableau ci-dessus présente un extrait d'une matrice de *Burt* – de contingences multiples – issue d'une enquête auprès d'une population de 137 individus. Pour réaliser l'analyse bivariée on sélectionne dans cette matrice les cellules correspondant aux modalités des deux variables retenues. Par exemple, les CSP (SA, Agri, CadSup, PorfInt, Empl, Ouv, Etud et Retrai) et les classes d'âge (moins de 25 ans, de 25 à 35, de 35 à 45, de 45 à 55, de 55 à 65 et plus de 65 ans). Soit la matrice observée suivante :

Tableau observé

	SA	Agri	Artisa	CadSup	ProfInt	Empl	Ouv	Etud	Retrai	$n_{i.}$
<25ans				1		3	1	30		39
25-35a	5		2	1	2	4	3	2		19
35-45a	4	1	4	3	1	4	1			18
45-55a	4	6	3	10	4	5	1		1	34
55-65a	2	1	1			1	1		4	10
>65ans	1				1			1	14	17
$n._j$	20	8	10	15	8	17	7	33	19	137

On calcule alors la matrice théorique. Par exemple, effectif théorique pour la modalité *<25ans* et celle *SA* :

$$n'_{1,1} = \frac{(n_{1\bullet} * n_{\bullet 1})}{n_{\bullet\bullet}}$$

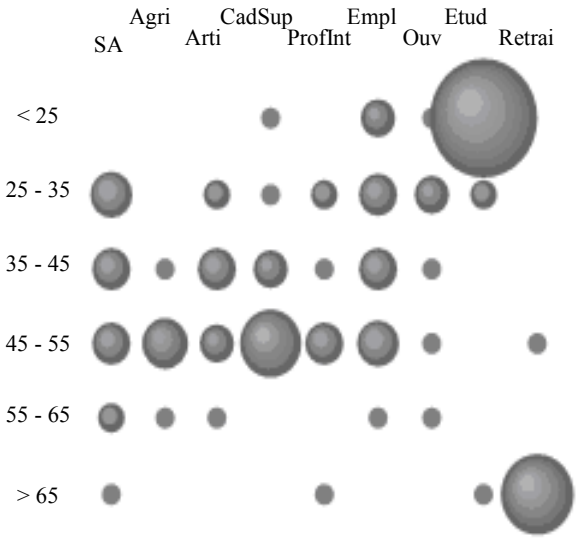
$$n'_{1,1} = \frac{(39 * 20)}{137} = 5.69$$

Tableau théorique

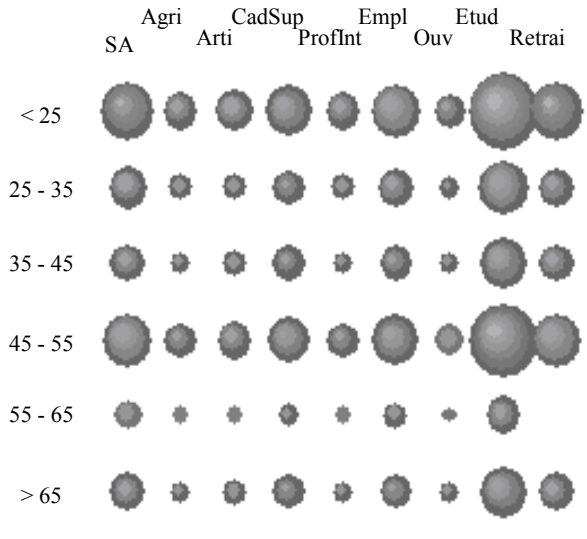
	SA	Agri	Artisa	CadSup	ProfInt	Empl	Ouv	Etud	Retrai	$n_{i.}$
<25ans	5.69	2.28	2.85	4.27	2.28	4.84	1.99	9.39	5.41	**39**
25-35a	2.77	1.11	1.39	2.08	1.11	2.36	0.97	4.58	2.64	**19**
35-45a	2.63	1.05	1.31	1.97	1.05	2.23	0.92	4.34	2.50	**18**
45-55a	4.96	1.99	2.48	3.72	1.99	4.22	1.74	8.19	4.72	**34**
55-65a	1.46	0.58	0.73	1.09	0.58	1.24	0.51	2.41	1.39	**10**
>65ans	2.48	0.99	1.24	1.86	0.99	2.11	0.87	4.09	2.36	**17**
$n_{.j}$	**20**	**8**	**10**	**15**	**8**	**17**	**7**	**33**	**19**	**137**

Rq. Les distributions conditionnelles des deux matrices sont identiques, ce qui permet de réaliser un rapide test pendant les calculs avec un tableur.

Il est possible de réaliser des cartogrammes pour visualiser les différences d'effectif. Comme pour une carte, les surfaces des cercles sont proportionnelles au valeurs. Afin de rendre comparables les graphes il faut retenir la valeur maximale de référence au sein des deux matrices (ce type de graphique est facilement réalisable avec un tableur).



Cartogramme des effectifs observés



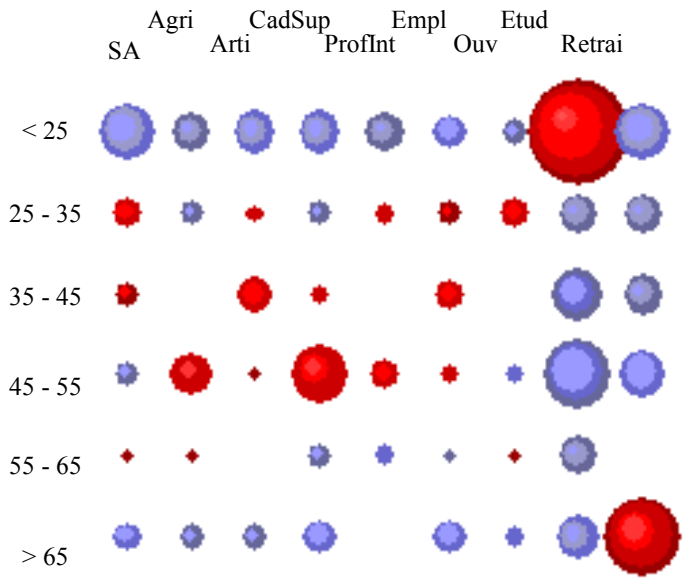
Cartogramme des effectifs théoriques

L'étape suivante consiste à dresser une matrice des différences entre situation observée et situation théorique. Une forte différence positive représente une surévaluation de la réalité par rapport au cas théorique et *vice versa*.

Tableau des différences

	SA	Agri	Artisa	CadSup	ProfInt	Empl	Ouv	Etud	Retrai
<25ans	-5.69	-2.28	-2.85	-3.27	-2.28	-1.84	-0.99	20.61	-5.41
25-35a	2.23	-1.11	0.61	-1.08	0.89	1.64	2.03	-2.58	-2.64
35-45a	1.37	-0.05	2.69	1.03	-0.05	1.77	0.08	-4.34	-2.50
45-55a	-0.96	4.01	0.52	6.28	2.01	0.78	-0.74	-8.19	-3.72
55-65a	0.54	0.42	0.27	-1.09	-0.58	-0.24	0.49	-2.41	2.61
>65ans	-1.48	-0.99	-1.24	-1.86	0.01	-2.11	-0.87	-3.09	11.64

Un cartogramme peur facilement être réalisé à nouveau, on joue sur la teinte pour distinguer les différences positives et négatives.



Cartogramme des différences (bleu, négatives)

Les résultats sont à manipuler avec précaution, il s'agit de dénombrements et les chiffres peuvent induire en erreur. Par exemple, une différence de 10 individus ne représente pas la même signification pour une population initiale de 100 individus ou de 10 000 individus.

On préfère alors une autre estimation des écarts fondés sur une pondération des masses, il s'agit de la métrique du *Chi2 : Chi2* = $\frac{(n_{i,j} - n'_{i,j})^2}{n'_{i,j}}$

Tableau du *Chi2*

	SA	Agri	Artisa	CadSup	ProfInt	Empl	Ouv	Etud	Retrai	$n_{i.}$
<25ans	5.69	2.28	2.85	2.50	2.28	0.70	0.49	45.20	5.41	**67.40**
25-35a	1.79	1.11	0.27	0.56	0.71	1.14	4.24	1.45	2.64	**13.91**
35-45a	0.72	0.00	5.49	0.54	0.00	1.40	0.01	4.34	2.50	**14.99**
45-55a	0.19	8.12	0.11	10.59	2.04	0.14	0.31	8.19	2.93	**32.62**
55-65a	0.20	0.30	0.10	1.09	0.58	0.05	0.47	2.41	4.92	**10.12**
>65ans	0.88	0.99	1.24	1.86	0.00	2.11	0.87	2.34	57.49	**67.79**
$n_{.j}$	**9.47**	**12.80**	**10.06**	**17.14**	**5.62**	**5.54**	**6.39**	**63.92**	**75.88**	**206.83**

où $n\bullet\bullet = 206.33$ est le *Chi2* total (somme des cellules)

$n_i\bullet$ est le Chi2 de chaque ligne de la matrice

$n\bullet_j$ est le Chi2 de chaque colonne de la matrice

$\chi^2_{i,j}$ est le χ^2 de chaque cellule

Si l'hypothèse d'indépendance mathématique est vérifiée, les valeurs du *Chi2* total sont distribuées selon une loi de *Pearson* dont la table qui suit donne les valeurs pour un risque d'erreur α choisi (colonnes, en pourcentage) et un nombre *v* de degré de liberté (en lignes, v = *(h-1)*(k-1)* avec *h* et *k* le nombre de modalités des variables *1* et *2).*

	1%	2.50%	5%	10%		1%	2.50%	5%	10%
1	6.63	5.02	3.84	2.71	16	32	28.84	26.3	23.54
2	9.21	7.38	5.99	4.61	17	33.41	30.19	27.59	24.77
3	11.34	9.35	7.81	6.25	18	34.8	31.53	28.87	25.99
4	13.28	11.14	9.49	7.78	19	36.19	32.85	30.14	27.2
5	15.09	12.83	11.07	9.24	20	37.57	34.17	31.41	28.41
6	16.81	14.45	12.59	10.64	21	38.93	35.48	32.67	29.61
7	18.47	16.01	14.07	12.02	22	40.29	36.78	33.92	30.81
8	20.09	17.53	15.51	13.36	23	41.64	38.08	35.17	32.01
9	21.67	19.02	16.92	14.68	24	42.98	39.37	36.41	33.2
10	23.21	20.48	18.31	15.99	25	44.31	40.65	37.65	34.38
11	24.72	21.92	19.67	17.27	26	45.64	41.92	38.88	35.56
12	26.22	23.34	21.03	18.55	27	46.96	43.19	40.11	36.74
13	27.69	24.74	22.36	19.81	28	48.28	44.46	41.34	37.92
14	29.14	26.12	23.68	21.06	29	49.59	45.72	42.56	39.09
15	30.58	27.49	25	22.31	30	50.89	46.98	43.77	40.26

Table des valeurs du Chi2

Lorsque *v* est supérieur à 30, la valeur du *Chi2* s'obtient par la formule suivante :

$$\chi^2 = \frac{\left(u + \sqrt{(2v-1)}\right)^2}{2}$$

où $u = 1.2816$ pour $\alpha = 10\ \%$;

$u = 1.6449$ pour $\alpha = 5\ \%$;

$u = 1.96$ pour $\alpha = 2.5\ \%$;

$u = 2.3263$ pour $\alpha = 1\ \%$.

Avec notre exemple, $v = (6-1)*(9-1) = 40$.
La valeur du *Chi2* théorique calculée avec la formule précédente est égal à 55.47 pour un risque d'erreur $\alpha = 5$ %.

Lorsque la valeur du *Chi2* issue du tableau des observations est inférieure à celle issue de la table théorique, le test d'indépendance mathématique est vérifiée, il n'y a alors pas de lien entre les deux variables. Inversement, lorsque le *Chi2* « observé » est supérieur au *Chi2* « théorique », le test d'indépendance mathématique n'est pas vérifié, les variables sont donc dépendantes (corrélées dirait-on avec des variables quantitatives).

Dans notre exemple, $Chi2_{observé} = 206.83$ supérieur à $Chi2_{théorique} = 55.47$, donc la variable « tranche d'âge » est celle « catégorie socioprofessionnelle » sont liées dans la population enquêtée.

À ce niveau, la liaison entre les variables étant démontrées, il est possible de la quantifier par un coefficient variant de 0 à 1. Nous retenons celui de *Tschuprow* qui mesure, à la racine carrée près, le rapport entre le $Chi2_{théorique}$ et la $Chi2_{maximum}$ si les variables étaient indépendantes. On peut traduire ce coefficient comme un pourcentage d'information expliquée par la liaison (équivalent au coefficient de détermination avec des variables quantitatives). Il s'obtient par la formule :

$$T = \sqrt{\frac{\chi^2_{observé}}{(N.\sqrt{v})}}$$

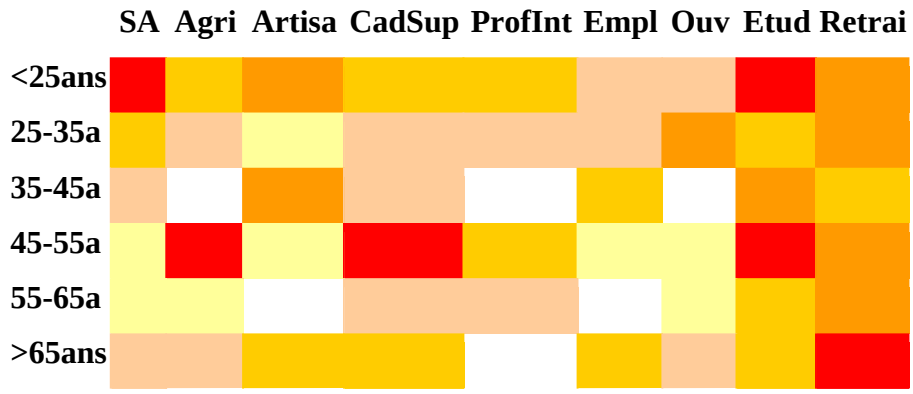
Dans notre exemple, T = 0.41 soit 41 % d'information expliquée.

Une dernière étape consiste à déterminer la contribution de chaque cas au *Chi2*. La contribution d'une cellule correspond à sa part relative dans la valeur du *Chi2*. D'où la matrice suivante :

Table des contributions au CHI2

	SA	Agri	Artisa	CadSup	ProfInt	Empl	Ouv	Etud	Retrai	$n_{i.}$
<25ans	2.75	1.10	1.38	1.21	1.10	0.34	0.24	21.85	2.62	32.59
25-35a	0.87	0.54	0.13	0.27	0.34	0.55	2.05	0.70	1.28	6.73
35-45a	0.35	0.00	2.65	0.26	0.00	0.68	0.00	2.10	1.21	7.25
45-55a	0.09	3.93	0.05	5.12	0.99	0.07	0.15	3.96	1.42	15.77
55-65a	0.10	0.15	0.05	0.53	0.28	0.02	0.23	1.17	2.38	4.89
>65ans	0.43	0.48	0.60	0.90	0.00	1.02	0.42	1.13	27.80	32.77
$n_{.j}$	4.58	6.19	4.86	8.29	2.71	2.68	3.09	30.91	36.69	100.00

La contribution est une variable quantitative, en y appliquant les règles de cartographie statistique on obtient un cartogramme synthétique. Notons que si les individus de la matrice initiale représentaient des entités géographiques, on pourrait dresser une carte des valeurs de contribution.



Le test du *Chi2* est souvent utilisé pour l'analyse des résultats d'une enquête, le but recherché étant d'identifier des ensembles de variables dépendantes ou indépendantes de manière à progresser dans la compréhension de l'analyse globale. En aucun cas on ne doit réduire l'analyse des données à celle du *Chi2*, il faut poursuivre au contraire vers la voie de l'analyse multivariée exploratoire seule capable de dégager de véritables structures dans l'organisation des données.

Test du *Chi2* et cartographie

	TZAU							
NIVEQUIP	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**Total**
0		6	3	23	1	1	27	**61**
1	2	6	3	29	2	6	28	**76**
2	4	13	2	21		2	19	**61**
3		9	1	12	1	1	33	**57**
4	1	9	3	12		1	15	**41**
5	3	8	4	5		2	16	**38**
6	2	7		12		2	11	**34**
7	4	5		6		2	6	**23**
8		3		6				**9**
9		6		5	2	2	6	**21**
10		1		1	1		3	**6**
11	1	1					2	**4**
12	1	1		4			4	**10**
13	4	1			1	1		**7**
14	1	1		2			1	**5**
15	2	2		3	2		1	**10**
16	1	1		1	1			**4**
17	1	2	1	2			3	**9**
18	1	2		1				**4**
19	1	1		1			1	**4**
20	1	1	1	1				**4**
21	1			1				**2**
22	1		1					**2**
23	1	2		1			1	**5**
24	1			1				**2**
26		1		1			1	**3**
27	4	1						**5**
28	1	1		1				**3**
29	1	1					1	**3**
30				1		1	1	**3**
31	1						2	**3**
32	1	1		1		2		**5**
33				3			1	**4**
34	1			1	2			**4**
35	1				1		3	**5**
36	6							**6**
Total	**50**	**93**	**19**	**158**	**14**	**23**	**186**	**543**

1 :pôle urbain, unité urbaine qui offre au moins 5 000 emplois sur son territoire ; 2 : commune péri-urbaine, au moins 40% des habitants actifs travaillent dans un même pôle urbain ; 3 :commune multipolarisée, commune envoyant au moins 40% de ces actifs vers plusieurs pôles urbains ; 4 : commune sous faible influence urbaine ; 5 : pôle rural ; 6 : commune sous influence du pôle rural ; 7 : rural isolé.

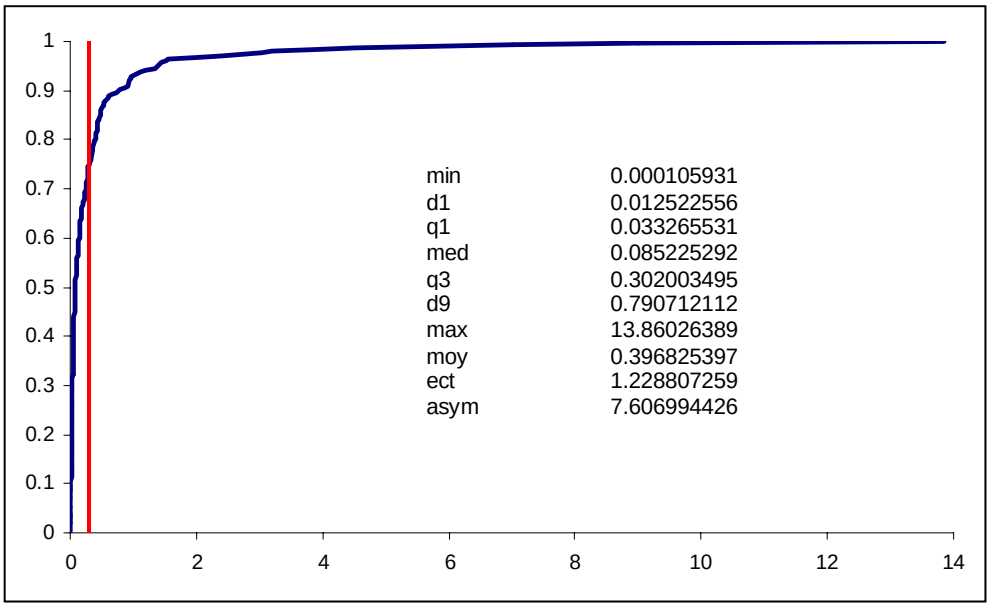
Dans un dernier exemple nous présentons une analyse bivariée de deux variables qualitatives – le zonage INSEE des communes (7 modalités) et leur niveau d'équipement (37 modalités) – pouvant se traduire par une représentation cartographique. Notons que le niveau d'équipement peut être envisagée soit sous l'angle d'une variable qualitative soit quantitative discrète en tant que dénombrement d'équipements présents sur la commune. Le tableau initial présenté plus haut donne à voir comment se ventilent les modalités des deux variables les unes par rapport aux autres.

NB TZAU	TZAU							
NIVEQUIP	1	2	3	4	5	6	7	**Total**
0	1.449426688	0.488559993	0.090575073	0.400778049	0.053821937	0.250516102	0.460278512	**3.193956355**
1	0.921154556	0.976000016	0.011263482	0.553269418	0.000216172	0.619881793	0.038345321	**3.12013076**
2	0.120111781	0.160920025	0.002185037	0.153602722	0.405839473	0.03403747	0.044349056	**0.921045563**
3	1.354382315	0.015365238	0.127954474	0.327162235	0.038723436	0.213804166	2.399798207	**4.477190071**
4	0.526464459	0.143760036	0.440754554	0.000105931	0.272777351	0.080631494	0.016785502	**1.481279326**
5	0.018368826	0.088226674	1.383869766	0.856216273	0.252818032	0.024437496	0.176453348	**2.800390415**
6	0.105386319	0.061367194	0.306993325	0.115774379	0.226205608	0.056161258	0.009258034	**0.881146117**
7	0.431618813	0.07371071	0.207671955	0.018487883	0.153021441	0.27870909	0.115572891	**1.278792782**
8	0.213849839	0.356140464	0.081262939	1.12652847	0.059877955	0.098370926	0.795521402	**2.731551995**
9	0.498982958	0.414395656	0.189613524	0.052078043	1.013910095	0.357753514	0.051087445	**2.577821236**
10	0.14256656	0.000191622	0.054175293	0.08222391	1.191907354	0.065580617	0.112064214	**1.648709569**
11	0.279546886	0.037354738	0.036116862	0.300340219	0.026612424	0.043720412	0.074709476	**0.798401017**
12	0.00175737	0.076530649	0.090292154	0.105410225	0.066531061	0.109301029	0.024870966	**0.474693455**
13	4.507389777	0.008514574	0.063204508	0.525595383	0.960263366	0.430722208	0.618738868	**7.114428683**
14	0.163189288	0.006217742	0.045146077	0.052705113	0.033265531	0.054650514	0.076530649	**0.431704913**
15	0.326378575	0.012435483	0.090292154	0.000722157	3.037740361	0.109301029	0.443154353	**4.020024113**
16	0.279546886	0.037354738	0.036116862	0.005956064	2.012640973	0.043720412	0.353565068	**2.768901002**
17	0.009133764	0.035202187	0.384579547	0.037729007	0.059877955	0.098370926	0.000574865	**0.625468252**
18	0.279546886	0.651253242	0.036116862	0.005956064	0.026612424	0.043720412	0.353565068	**1.396770957**
19	0.279546886	0.037354738	0.036116862	0.005956064	0.026612424	0.043720412	0.025805697	**0.455113083**
20	0.279546886	0.037354738	1.363692912	0.005956064	0.026612424	0.043720412	0.353565068	**2.110448504**
21	0.932618158	0.088391267	0.018058431	0.077492745	0.013306212	0.021860206	0.176782534	**1.328509553**
22	0.932618158	0.088391267	3.189301477	0.150170109	0.013306212	0.021860206	0.176782534	**4.572429963**
23	0.163189288	0.394118355	0.045146077	0.03669976	0.033265531	0.054650514	0.076530649	**0.803600175**
24	0.932618158	0.088391267	0.018058431	0.077492745	0.013306212	0.021860206	0.176782534	**1.328509553**
26	0.07128328	0.118713488	0.027087646	0.004773273	0.019959318	0.032790309	0.000191622	**0.274798935**
27	**7.022037952**	0.006217742	0.045146077	0.375425273	0.033265531	0.054650514	0.441956334	**7.978699424**
28	0.489316946	0.118713488	0.027087646	0.004773273	0.019959318	0.032790309	0.265173801	**0.95781478**
29	0.489316946	0.118713488	0.027087646	0.225255164	0.019959318	0.032790309	0.000191622	**0.913314492**
30	0.07128328	0.1325869	0.027087646	0.004773273	0.019959318	1.54740504	0.000191622	**1.803287079**
31	0.489316946	0.1325869	0.027087646	0.225255164	0.019959318	0.032790309	0.237426976	**1.164423259**
32	0.163189288	0.006217742	0.045146077	0.03669976	0.033265531	3.896162247	0.441956334	**4.622636979**
33	0.095044373	0.176782534	0.036116862	0.747428498	0.026612424	0.043720412	0.025805697	**1.1515108**
34	0.279546886	0.176782534	0.036116862	0.005956064	**9.002908511**	0.043720412	0.353565068	**9.898596336**
35	0.163189288	0.220978167	0.045146077	0.375425273	1.518870181	0.054650514	0.249670837	**2.627930338**
36	**13.86026389**	0.265173801	0.054175293	0.450510328	0.039918637	0.065580617	0.530347601	**15.26597017**
Total	**38.34272915**	5.850969396	8.745844118	7.530684375	**20.77370937**	9.058113813	9.697949776	**100**

Le *Chi2* calculée (387.52) est largement supérieur à celui donné par la loi de probabilité (260.93), l'hypothèse d'indépendance statistique est donc rejetée. Le coefficient de

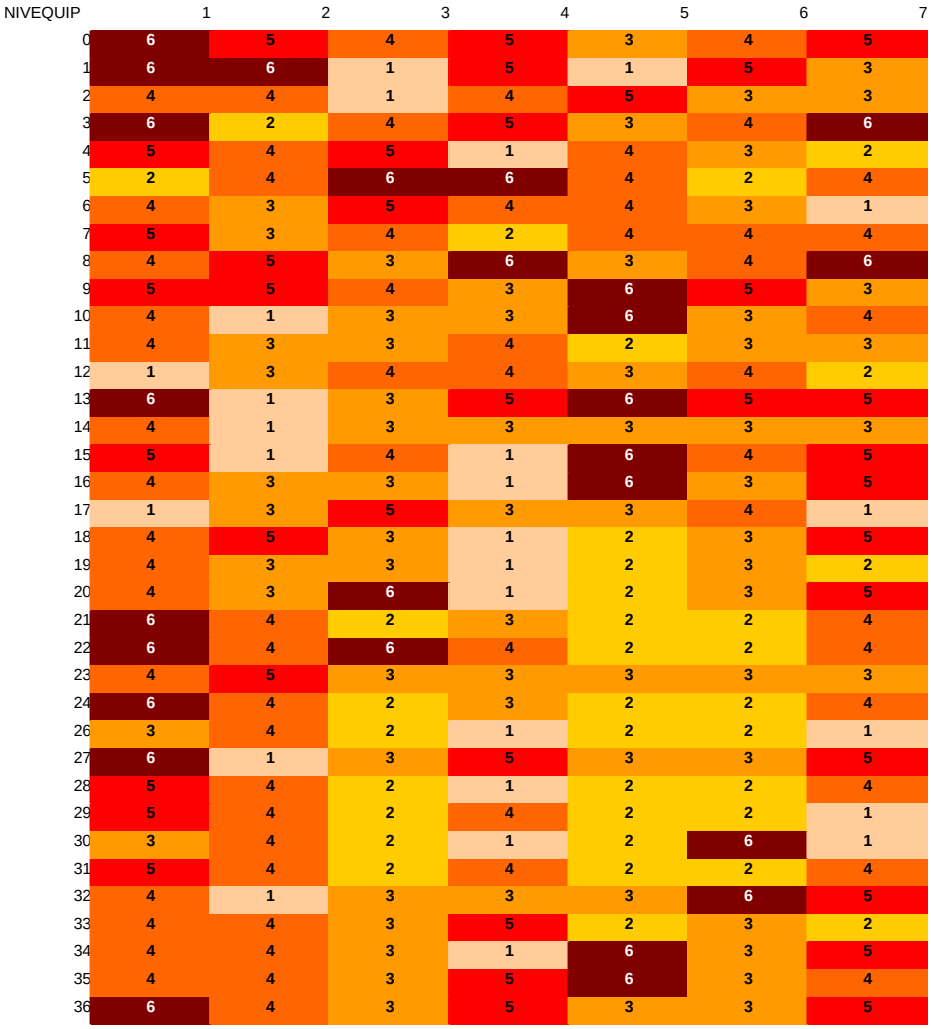
Tschuprow est de 0.22, il exprime une relation de l'ordre de 22% entre les modalités des deux variables, c'est peu mais l'analyse de la contribution au *Chi2* permet d'observer comment se ventilent de manière différentielle les liens entre les modalités.

La figure sui suit présente la forme de la distribution des contributions au *Chi2* et une synthèse statistique. La forte asymétrie est nette, seules quelques rares cas contribuent efficacement au *Chi2* (les cellules en rouge dans le tableau de contribution).



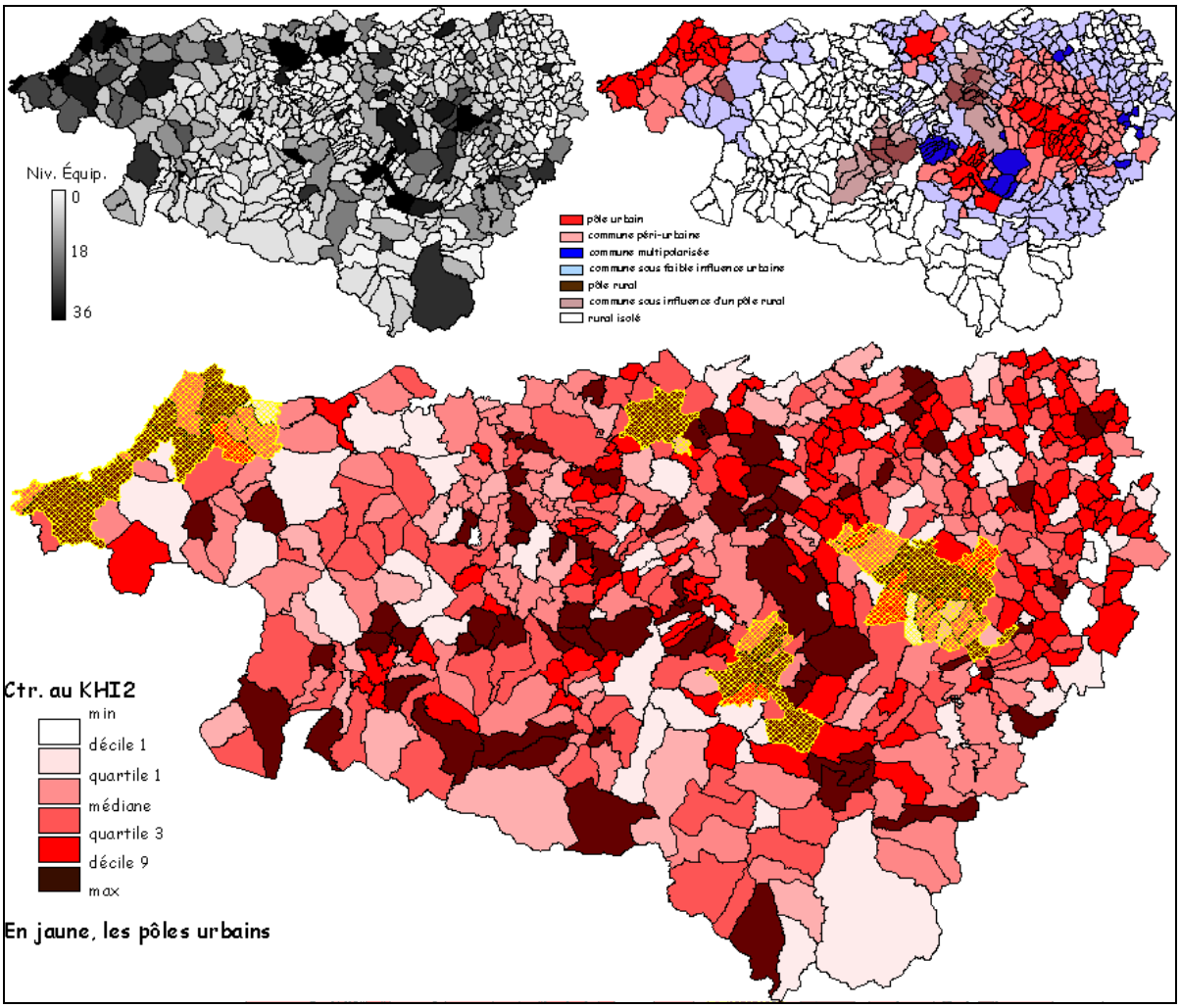
Nous utilisons les paramètres statistiques présentés dans la figure afin de réaliser une discrétisation de la variable « contribution au *Chi2*», soit en tout 6 classes. On obtient alors un cartogramme (cf. plus bas).

Cartogramme des contributions au *Chi2*

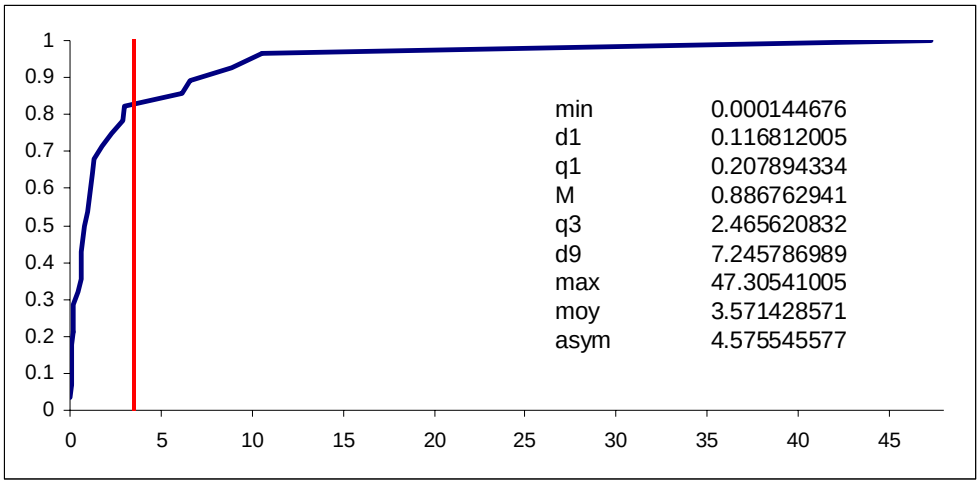


Il est ensuite possible de réaliser une carte statistique, chaque individu – ici des communes – est identifié par une modalité de niveau d'équipement et une modalité du zonage INSEE, soit en tout : 37 * 7 = 259 possibilités. On constate sur la matrice de contingence initiale que 115 cas théoriques ne sont pas observés et que d'autres ne sont que très peu présentes.

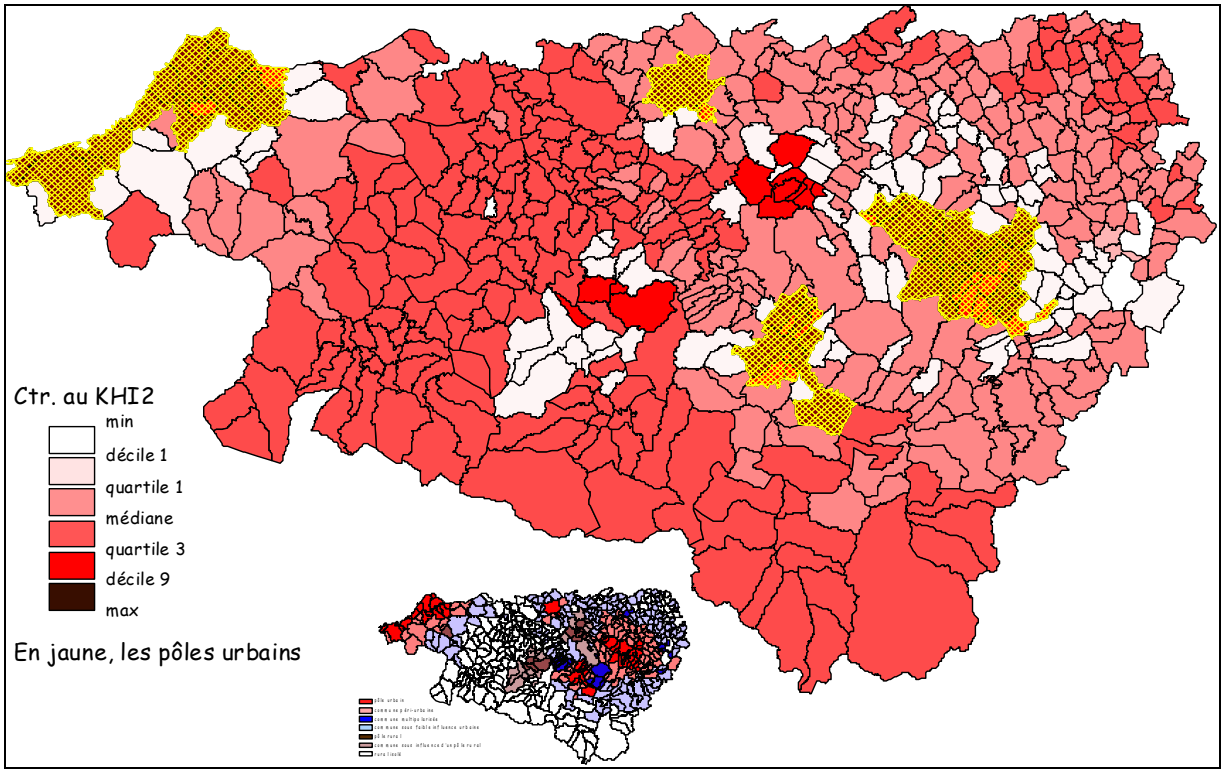
La figure qui suit présente le résultat cartographique. La classe du dernier décile s'étend de 0.79 à 14 % de contribution au *Chi2*. Cette forte hétérogénéité rend difficile l'interprétation, on observe cependant une forte différenciation spatiale qui fait ressortir sans doute des particularités locales telles qu'une commune « suréquipée » en rural isolée ou alors « sous équipée » en pôle urbain.



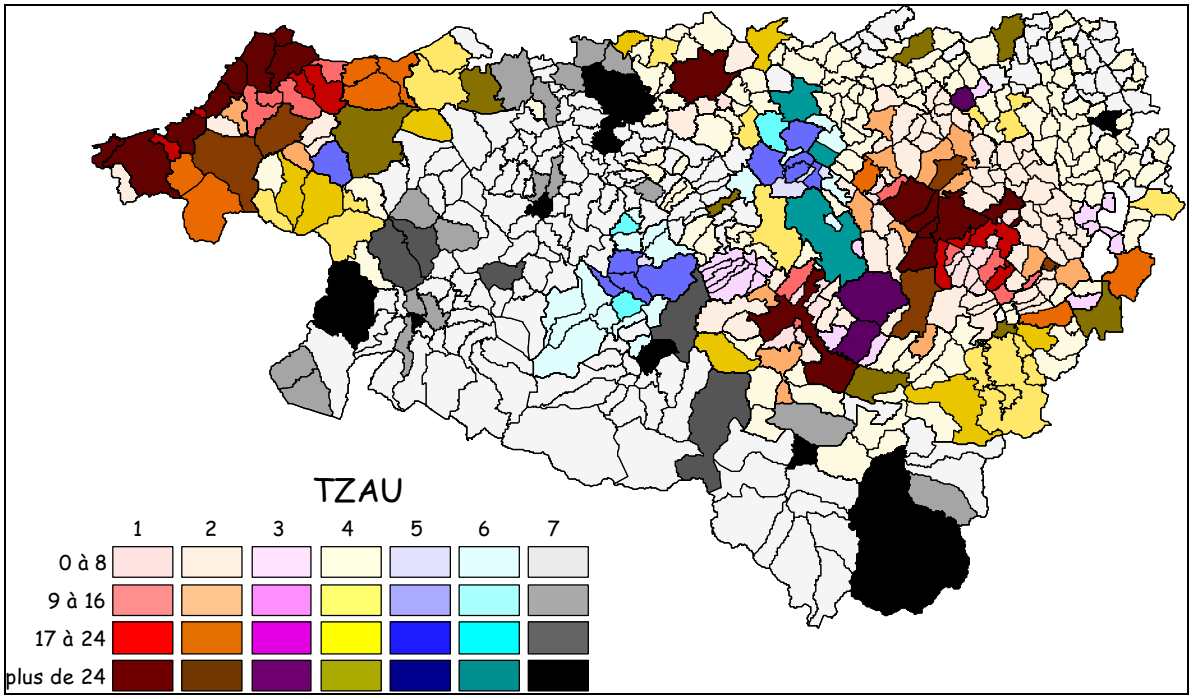
On peut légitimement se poser la question de la légitimité d'une telle analyse qui relève plus de l'exception. Nous proposons donc de regrouper les niveaux d'équipements en classe d'abondance d'équipement selon le même principe que celui appliqué à une discrétisation d'une variable continue (4 modalités réparties de 0 à 8, de 9 à 16, de 17 à 24 et plus de 24 équipements). La nouvelle distribution des contributions au *Chi2* reste fortement asymétrique à droite, le coefficient de *Tschuprow* garde les mêmes proprtions : T = 0.24 soit 24% d'information expliquée.



La carte obtenue est en revanche très différente. Elle est proche de celle de la répartition des zones INSEE. En regroupant les modalités de niveau d'équipement on fait ressortir le fait que ceux-ci se répartissent préférentiellement selon l'influence dominante dans la commune, d'où la similitude avec la carte du zonage INSEE.



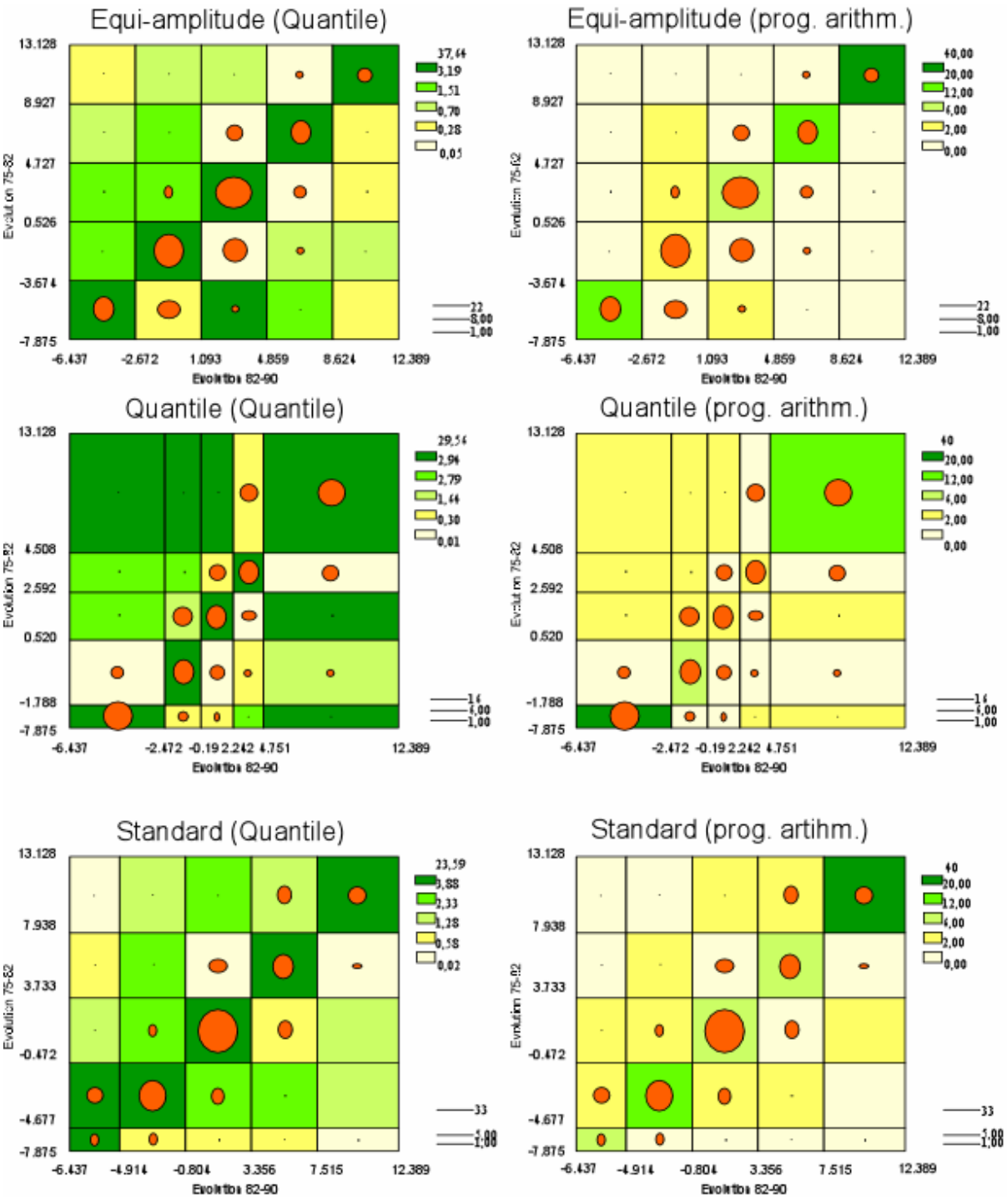
Cependant, on constate que le « rural isolé » apporte une contribution plus significative que celle des « communes mulitpolarisées », ce qui se comprend aisément puisqu'on peut supposer que l'on trouvera proportionnellement plus d'équipement dans le rural isolé – il y en a peu mais il en faut un minimum – que dans les secteurs multipolarisés où l'essentiel des services est concentré sur les pôles les plus proches. L'analyse de la couronne de l'agglomération paloise est riche de renseignements en ce sens.



Ces deux exemples illustrent les précautions à prendre au cours de l'analyse des données. Il est fondamental de rester très critique quant à la méthode et aux données elles-mêmes. On a pu le constater, un codage modifie considérablement les résultats. Prudence donc…

D'une variable quantitative à une variable qualitative

On est souvent emmené à coder une variable quantitative en une variable qualitative pour permettre l'analyse globale de la matrice. La figure qui suit présente les artéfacts de cette transformation.



On retient deux variables quantitatives très fortement corrélées à titre d'exemple : Évolution de la population des communes des Pyrénées-Atlantiques de 1975 à 1982 et de 1982 à 1990. On applique tout d'abord à chacune de ces deux variables une méthode de discrétisation couramment utilisée (cf. cours *La cartographie statistique*) : standardisation ; progression arithmétique ; quantile d'ordre *k*. Les limites des cases des cartogrammes correspondent aux bornes des classes, les cercles proportionnels sont relatifs à l'effectif des classes.

Pour chaque cas, un tableau de contingence est dressé pour réaliser un test du *Chi2* et calculer une matrice de contributions. Cette matrice subit une discrétisation selon des quantiles d'ordre k et une progression arithmétique (la méthode est indiquée entre parenthèses dans la figure). Les deux colonnes de la figure distinguent les deux cas, les cellules des différents cartogrammes sont associées à une teinte relative à leur valeur de contribution.

La lecture rapide des différents cas souligne de forts écarts directement induits par les méthodes de discrétisation utilisées pour coder les variables quantitatives initiales. On constate de plus que les valeurs de contribution varient également de manière significative selon des différents scénarios. L'exemple le plus pertinent, c'est-à-dire celui qui respecte les formes de distribution naturelle des différentes variables, est celui en bas à gauche de la figure. Les deux variables initiales étant très fortement corrélées, on retrouve logiquement les plus fortes valeurs de contribution au *Chi2* dans la diagonale du cartogramme. Celles-ci décroissent au fur et à mesure que l'on s'en éloigne, quelques cellules soulignent des écarts marqués par rapport à la tendance globale.

Cet exemple renforce l'invitation à la prudence évoquée plus haut. Néanmoins, une discrétisation fondée sur le respect de la forme des distributions donnent des résultats fiables. Il faut savoir perdre sur les détails pour gagner sur la global comme nous le verrons avec l'analyse exploratoire des données et l'exemple des Analyses Factorielles des Correspondances.