

CH. MULLON

Essais de reconstitution d'un tableau ternaire dont on connaît deux des marges binaires : exemple des reports de votes aux scrutins présidentiels

Les cahiers de l'analyse des données, tome 9, n° 4 (1984), p. 473-487

http://www.numdam.org/item?id=CAD_1984__9_4_473_0

© Les cahiers de l'analyse des données, Dunod, 1984, tous droits réservés.

L'accès aux archives de la revue « Les cahiers de l'analyse des données » implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques

<http://www.numdam.org/>

ESSAIS DE RECONSTITUTION D'UN TABLEAU TERNAIRE
DONT ON CONNAIT DEUX DES MARGES BINAIRES :
EXEMPLE DES REPORTS DE VOTES AUX
SCRUTINS PRÉSIDENTIELS

[REC. TERN.]

par Ch. Mullon

1 Un exemple : l'estimation des reports de votes

Un exemple servira à poser le problème en introduisant des notations. .

En France, lors des scrutins nationaux d'élection présidentielle, tels ceux de mai 1981, les électeurs sont appelés à exprimer leurs voix à deux semaines d'intervalle ; au premier tour de scrutin, l'ensemble I des attitudes possibles comprend outre l'abstention (et le vote nul) le vote en faveur d'assez nombreux candidats ; au deuxième tour, on a un ensemble J plus restreint, car seuls subsistent les deux candidats les plus favorisés au premier tour. Une question se pose donc : comment de l'attitude i adoptée au premier tour, l'électeur passe-t-il à l'attitude j adoptée au deuxième tour ? Un modèle probabiliste simple se présente naturellement : à chaque attitude i, correspond une loi de probabilité f_J^i sur J ; f_j^i étant la probabilité qu'un électeur qui a voté i au premier tour, vote j au 2-ème, l'ensemble de ces lois f_J^i formant une transition probabiliste de I vers J.

$$f_J^I = \{f_J^i \mid i \in I\} = \{f_j^i \mid i \in I ; j \in J\} ;$$

$$\forall i, j : f_j^i \geq 0 ; \quad \forall i : \sum \{f_j^i \mid j \in J\} = 1$$

Dans le langage politique on dit que f_j^i est la probabilité qu'un électeur ayant voté i, reporte sa voix sur j ; en terme probabiliste, on appellerait $f_j^i = f(j/i)$ la probabilité conditionnelle de j quand i. Reste à déterminer les f_j^i ...

Si les attitudes de votes de l'ensemble S des électeurs (sujets) étaient connues, il suffirait de calculer f_J^I , d'après le tableau de correspondance entre attitudes aux premier et deuxième tours :

$k(i, j)$ = nombre des sujets ayant voté i, puis j ;

$f_{ij} = k(i, j)/k$; où $k = \text{Card } S$.

$f_j^i = f_{ij}/f_i$; où $f_i = k(i)/k$ suivant les notations usuelles...

Cependant le vote étant secret, un tel calcul n'est pas possible : tout au plus peut-on tenter un sondage sur un sous-ensemble de quelques milliers d'électeurs.

(1) Statisticien à l'O.R.S.T.O.M..

Une autre voie s'ouvre donc : estimer les corrélations entre i et j d'après la répartition géographique des attitudes de vote. Soit en effet M une partition de l'ensemble S en parties m , au niveau de laquelle le détail des votes soit dénombré ; (par exemple, au niveau le plus fin M sera l'ensemble des bureaux de vote ; où, à un niveau accessible d'après la presse quotidienne, l'ensemble des cantons, ou des départements ; c'est le terme département que nous retiendrons ; le lien entre les $k(m,i)$ (nombre de votes i dénombrés dans m au 1-er tour) et les $k(m,j)$ (nombre de votes j dénombrés dans m au 2-ème tour) donne une idée des reports de voix.

De façon précise l'ensemble des votes définit une correspondance ternaire $I \times J \times M$ avec

$k(i,j,m)$ = nombre des électeurs du département m ayant adopté successivement les attitudes i et j ; le tableau ternaire est inconnu ; mais on a deux de ses marges publiées comme résultats des votes aux 1-er et 2-ème tours :

$k(i,m)$ = nombre des électeurs de m ayant voté i , au 1-er tour ;
 $k(j,m)$ = nombre des électeurs de m ayant voté j , au 2-ème tour ;

avec, suivant les notations usuelles, des lois associées, des transitions etc. :

$$f_{IJM} = \{f_{ijm} \mid i \in I ; j \in J ; m \in M\} ; f_{ijm} = k(i,j,m)/k ;$$

$$f_{IM} = \{f_{im} \mid i \in I ; m \in M\} ; f_{im} = k(i,m)/k ;$$

$$f_M^I = \{f_m^i \mid i \in I ; m \in M\} ; f_m^i = f_{im}/f_i ; \text{ etc.}$$

Seuls sont connus f_{IM} et f_{JM} , liés par la condition de compatibilité d'avoir même loi marginale f_M sur M ; le reste, en particulier f_{IJ} (ou la transition f_J^I) associé au report des voix est caché : notre problème est de l'estimer par le calcul moyennant diverses hypothèses.

2 Hypothèses et modèles

Outre le modèle de transition (§ 2.2) qui nous intéresse le plus, on considérera d'autres modèles, qui offrent des termes de comparaison utile.

2.0 Lois ternaires compatibles avec deux marges données

Avant de formuler des hypothèses, il importe de considérer quel est l'ensemble des lois ternaires f_{IJM} admettant pour lois marginales binaires deux lois données f_{IM} et f_{JM} (ayant même marge f_M). Pour chaque m de M , on a un étage f_{IJm} du tableau f_{IJM} ; et f_{IJm} est une mesure positive sur $I \times J$, ayant pour masse totale f_m ; et dont les deux marges f_{Im} et f_{Jm} sont connues ; l'ensemble des f_{IJm} possibles (pour m donné) est donc un convexe de R_{IJ} . De même l'ensemble des f_{IJM} possibles est un convexe de R_{IJM} ; et par projection, l'ensemble des marges f_{IJ} possibles pour une loi f_{IJM} dont on connaît les marges f_{IM} et f_{JM} est également un convexe. Certes, étant définis par de

nombreuses équations, ces convexes ne sont pas faciles à décrire : mais d'une part ils définissent en toute rigueur l'ensemble des hypothèses permises par les données, indépendamment de tout modèle ; et d'autre part on pourra (cf. § 3.2) s'inspirer de la définition de ces convexes pour achever la construction d'un modèle.

2.1 Hypothèses d'indépendance

2.1.1. Absence d'interaction entre I et J : Sous l'hypothèse d'absence d'interaction ternaire (cf. A. Bener, [INTER. CORR. MULT.] in C.A.D. Vol VII n° 1 ; 1982) un tableau ternaire s'exprime comme suit en fonction de ses marges binaires (et simples) :

$$f_{ijm} = f_{ij} f_m + f_{jm} f_i + f_{mi} f_j - 2f_i f_j f_m$$

Dans le cas présent où f_{ij} est inconnu, on peut postuler

$$f_{ij} = f_i f_j ; \text{ d'où la formule :}$$

$$f_{ijm} = f_{jm} f_i + f_{mi} f_j - f_i f_j f_m.$$

Il importe de noter que cette formule peut introduire des valeurs négatives pour des f_{ijm} . L'indépendance entre i et j n'est pas présentement une hypothèse réaliste ; en revanche, l'absence d'interaction ternaire est une simplification qu'il convenait de rappeler.

2.1.2 Absence d'interaction au niveau de chaque département : En termes probabilistes usuels on écrit :

$$f(i, j/m) = (f_i/m) f(j/m) ;$$

ce qui permet de reconstituer comme suit le tableau ternaire :

$$f_{ijm} = f_{im} f_{jm} / f_m ;$$

(ici n'apparaît jamais aucune valeur négative) ; d'où pour les f_{ij} :

$$f_{ij}^{\text{loc}} = \Sigma \{ f_{im} f_{jm} / f_m \mid m \in M \} ;$$

(où l'indice "loc" rappelle l'hypothèse d'indépendance locale).

Il est clair que le tableau des f_{ijm} ainsi reconstitué a bien les marges imposées sur $I \times M$ et $J \times M$; de plus, quoique l'indépendance entre 1-er et 2-ème tour au niveau du département soit inacceptable, la loi conjointe obtenue sur $I \times J$ par cumul des départements n'est pas (en général) la loi produit $f_I \times f_J$; les différences politiques entre départements suffisent à produire des corrélations globales même sous l'hypothèse d'indépendance locale.

Il est d'ailleurs intéressant de construire de manière analogue des lois sur $I \times I$ ou $J \times J$; ces lois correspondent au modèle fictif suivant. Chacun des électeurs vote deux fois sur les candidats du 1-er tour : les deux votes successifs étant indépendants entre-eux, et régis tous deux par une même loi qui est la loi f_I^m propre au département de résidence de l'électeur ; d'où pour les deux votes fictifs la loi conjointe :

$$f_{ii'}^{loc} = \Sigma \{ f_{im} f_{i',m} / f_m \mid m \in M \}.$$

Et de même pour le 2-ème tour, on aurait avec deux votes fictifs:

$$f_{jj'}^{loc} = \Sigma \{ f_{jm} f_{j',m} / f_m \mid m \in M \}.$$

Remarque : Si on applique à la reconstitution des f_{ijm} , la méthode RAS généralisée (cf. CAD Vol VIII n° 3 pp 359-370 ; 1983) on aboutit à $f_{im} \cdot f_{jm} / f_m$.

2.2 Le modèle de transition

2.2.1 Hypothèses du modèle et processus-générateur : Les données sont deux lois f_{IM} et f_{JM} ayant même marge f_M , (i.e., dans l'exemple des élections présidentielles, les deux tableaux recensent sur l'ensemble M des départements, les votes au 1-er et 2-ème tour) ; l'inconnue est le tableau ternaire f_{IJM} ; ou (on se bornant à un objectif plus modeste), le tableau de marge manquant f_{IJ} ; (ou ce qui est équivalent la transition f_J^I qui lui est associée).

L'hypothèse fondamentale relative au tableau ternaire, (ou plus précisément aux transitions associées à ses marges binaires) est :

$$f_J^M = f_J^I \circ f_I^M ; \text{ i. e. :}$$

$$\forall m, j : f_j^m = \Sigma \{ f_j^i f_i^m \mid i \in I \}.$$

Cette formule résulte nécessairement du modèle de processus-générateur suivant : un sujet résidant dans le département m vote d'abord au premier tour suivant des probabilités dont la loi f_I^m est propre à ce département ; puis il vote au 2-ème tour suivant une loi f_J^i conditionnée seulement par son vote au 1-er tour ; mais non par le département m où il réside. Ce modèle produit le tableau ternaire suivant :

$$\forall i, j, m : f_{ijm} = f_j^i f_i^m f_m$$

Dans les notations du calcul des transitions, (cf. TII B n° 1 [NOTE LIM.]) ce modèle s'écrit :

$$f_{IJM} = ((\delta_I^I \times f_J^I) \otimes \delta_M^M) \circ (f_I^M \times \delta_M^M) \circ f_M$$

2.2.2 Résolution algébrique du modèle : Répétons-le : les données sont f_{IM} et f_{JM} , l'inconnue est f_J^I (la loi de report des voix) ; la solution formelle apparaît en composant à droite avec f_M^I les deux membres de l'hypothèse fondamentale :

$$f_J^M = f_J^I \circ f_I^M ; \text{ d'où :}$$

$$f_J^M \circ f_I^M = f_J^I \circ f_I^M \circ f_I^M ;$$

dans cette dernière équation, apparaissent deux transitions composées g_I^I et q_J^I .

$$g_I^I = f_I^M \circ f_M^I ; q_J^I = f_J^M \circ f_M^I.$$

Il est curieux de noter que ces transitions ne sont autres que celles associées à des modèles considérés au § 2.1.2 : g_I^I correspond aux deux votes indépendants sur le 1-er tour ; et q_J^I à deux votes successifs indépendants sur deux tours :

$$g_i^i = f_{ii}^{loc} / f_i ; q_j^i = f_{ij}^{loc} / f_i.$$

Quoiqu'il en soit, on peut inverser la transition g_I^I : cette inversion est possible *algébriquement* si seulement le tableau f_{IM} est de rang Card I (ce qui est quasi certain ; puisque Card M, nombre des départements, dépasse de beaucoup Card I, nombre d'attitudes de vote). L'inverse obtenu peut être appelé une transition conservant la masse de I vers I ; en ce sens qu'elle transforme une mesure de masse totale 1 sur I en une mesure de masse totale 1 sur J (en effet, en bref cette propriété est vraie pour $(g^{-1})_I^I$ parce qu'elle l'est pour g_I^I) ; et de plus g^{-1} laisse invariant f_I (comme le fait g). Mais en revanche il est assuré que g^{-1} comporte des composantes négatives. (En bref, une transition probabiliste ne peut jamais avoir pour inverse une transition probabiliste : car la transition se fait avec perte d'information : le seul cas où il n'y a pas de telle perte, et où donc l'inversion est possible, est celui d'une transition déterministe, associée à une fonction ; cf. [NOTE LIM.] § 5). L'équation composée peut être réécrite :

$$q_J^I = f_J^I \circ g_I^I ; \text{ et sa solution est :}$$

$$f_J^I = q_J^I \circ (g^{-1})_I^I ; \text{ ou explicitement :}$$

$$f_J^I = f_J^M \circ f_M^I \circ (f_I^M \circ f_M^I)^{-1} ; \text{ ou encore :}$$

$$f_j^i = \sum \{ f_j^m \cdot f_m^{i'} (g^{-1})_i^{i'} \mid m \in M ; i' \in I \}.$$

Il importe de noter que

- 1°) Si f_{IM} est de rang Card I, la solution formelle est unique ;
- 2°) Cette solution formelle n'est acceptable que si les f_j^i obtenus sont tous positifs ou nuls ; (condition bien plus faible que celle impossible, qui imposerait aux $(g^{-1})_i^{i'}$, d'être tous positifs).
- 3°) Acceptable ou non quant à la positivité, il reste à vérifier que la solution formelle trouvée pour f_J^I , donne la valeur requise pour f_J^M ; i.e. que :

$$f_J^I \circ f_I^M \stackrel{?}{=} f_J^M ; \text{ ou :}$$

$$f_J^M \circ f_M^I \circ (f_I^M \circ f_M^I)^{-1} \circ f_I^M \stackrel{?}{=} f_J^M .$$

cette condition n'est pas en général vérifiée : comme on le verra clairement dans l'interprétation géométrique du § 2.2.3, elle équivaut à imposer que les lignes (indicées par j) du tableau f_{JM} soient toutes combinaisons linéaires des lignes (indicées par i) du tableau f_{IM} ; (la condition pouvant être équivalamment formulée pour f_J^M et f_I^M ; ou pour f_M^I et f_M^J ; cf. § 2.2.3). En revanche, il est clair que la transition f_J^I , calculée formellement satisfait bien à :

$$f_J^I \circ f_I = f_J ;$$

car f_I est invariant pour g^{-1} ; et $f_J = f_J^M \circ f_M$; $f_M = f_M^I \circ f_I$; et de même il est clair que f_J^I est une transition conservant la masse (i.e. conserve la masse totale : $\forall \mu_I \in R_I$, $\text{masse}(f_J^I \circ \mu_I) = \text{masse}(\mu_I)$, comme composé de transitions conservant la masse (f_J^M et f_M^I sont des transitions probabilistes ; et g^{-1} est l'inverse d'une telle transition).

2.2.3 Discussion géométrique du modèle : Plaçons-nous dans l'espace R_M des mesures sur M (en fait on considérera exclusivement des lois de probabilité sur M ; i.e. des points du simplexe P_M) ; la donnée des lois f_{IM} et f_{JM} équivaut à celle des deux nuages dans R_M .

$$N(I) = \{(f_M^i, f_i) \mid i \in I\} ; \quad N(J) = \{(f_M^j, f_j) \mid j \in J\}.$$

D'autre part on peut récrire l'équation du modèle de telle sorte qu'apparaissent ces deux nuages (ou les coordonnées de leurs points). On a :

$$\Sigma \{(f_{ij}/f_i) (f_{im}/f_m) \mid i \in I\} = (f_{jm}/f_m) ;$$

d'où après multiplication par (f_m/f_j) :

$$\Sigma \{(f_{ij}/f_j) (f_{im}/f_i) \mid i \in I\} = (f_{jm}/f_j) ; \text{ i.e.}$$

$$f_M^I \circ f_I^J = f_M^J ; \text{ ou encore :}$$

$$\forall j \in J : f_M^j = \Sigma \{f_i^j f_M^i \mid i \in I\} ;$$

en d'autres termes, les f_i^j sont les coordonnées barycentriques des f_M^j relativement au nuage des f_M^i . On notera qu'apparaissent ici les f_i^j et non les coefficients f_j^i lesquels s'interprètent directement comme des probabilités de report des votes : mais peu importe : les deux étant reliés par les f_i et f_j qui sont connus et tous non-nuls : $f_j^i f_i = f_i^j f_j = f_{ij}$.

Comme au § 2.2.2 on peut donner au système barycentrique une forme algébrique résoluble ; ici on composera à gauche avec f_I^M ; il vient :

$$f_I^M \circ f_M^I \circ f_I^J = f_I^M \circ f_M^J ; \text{ soit :}$$

$$g_I^I \circ f_I^J = q_I^J ; \text{ d'où : } f_I^J = (g_I^I)^{-1} \circ q_I^J$$

(où g_I^I est le même qu'au § 2.2.2 ; et $q_I^J = f_{ij}^{\text{loc}}/f_j$) . Le système ainsi obtenu est le même qu'au § 2.2.2 ; comme on le voit sous la forme :

$$f_{ij} = \Sigma (g^{-1})_i^{i'} q_{i'j}.$$

Reste à discuter l'existence et l'unicité des f_{ij} définis par cette voie.

1°) Pour que les f_M^j puissent s'exprimer comme des barycentres des f_M^i , il faut que les f_M^j soient dans le support du nuage $N(I)$ des f_M^i : c'est la condition nécessaire rencontrée à la fin du § 2.2.2. Condition qui dans l'exemple des votes n'est *a priori* pas vérifiée vue la grande dimension de R_M . Toutefois, il suffit de substituer aux f_M^j leurs projections orthogonales sur le support de $N(I)$, (au sens de la métrique du χ^2 de centre f_M).

2°) L'unicité des coordonnées barycentriques n'est assurée, que si les f_M^i sont les sommets d'un simplexe ; autrement dit si la dimension affine du support de $N(I)$ est $\text{Card } I - 1$.

3°) La positivité des f_i^j (ou des f_{ij}) requiert que chaque point f_M^j (ou sa projection sur le support de $N(I)$) tombe à l'intérieur de l'enveloppe convexe du nuage $N(I)$ (i.e., sous la condition de 2°, à l'intérieur du simplexe ayant pour sommets les f_M^i).

4°) Il importe de vérifier que le f_{IJ} construit a bien pour lois marginales f_I et f_J . Les f_i^j étant des coordonnées barycentriques, on a bien :

$$\forall j \in J : \Sigma \{f_i^j | i \in I\} = 1 ; \Sigma \{f_{ij} | i \in I\} = \Sigma \{f_i^j f_j\} = f_j$$

Quant à la marge f_I on peut écrire :

$$f_M = f_M^I \circ f_I^J \circ f_J = f_M^J \circ f_J$$

(relation qui ne sera toutefois qu'approchée si f_M^J a été modifié par projection des f_M^j sur le support de $N(I)$) ; ou encore :

$$f_M = \Sigma \{ \Sigma \{ f_i^j f_j | j \in J \} f_M^i | i \in I \}.$$

Si les f_M^i sont les sommets d'un véritable simplexe (i.e. la dimension du support de $N(I)$ est $\text{Card } I - 1$), le centre de gravité f_M s'exprime de manière unique en combinaison des f_M^i , avec pour masses les f_i , et on a donc bien :

$$\forall i \in I : \sum_j \{f_{ij}^j\} = \sum_j \{f_{ij}\} = f_i$$

5°) A la vérité il est quasi certain, vu la dimension de M , que $N(I)$ est un simplexe. Mais comme le montre l'analyse factorielle, ce simplexe est aplati sur le sous-espace de son support engendré par les premiers axes d'inertie ; c'est véritablement dans ce sous-espace qu'il convient de se placer ; alors les coordonnées barycentriques f_i^j des f_M^j relativement aux f_M^i sont indéterminées ; et il convient d'imposer la condition supplémentaire

$$\forall i \in I : \sum_{j \in J} \{f_i^j\} = f_i.$$

6°) Finalement on aboutit à une solution en trois étapes : (cf. § 3.2).

a) réduire en dimension par analyse factorielle le support de $N(I)$ dans R_M .

b) projeter $N(J)$ sur ce support réduit

c) chercher pour les f_M^j projetés, relativement aux f_M^i (projetés sur le support réduit) des coordonnées barycentriques assujetties d'une part à être positives et d'autre part à satisfaire à la condition de marge de 5°. L'approximation se faisant au sens des moindres carrés pour la métrique du χ^2 de centre f_M .

3 Utilisation de l'analyse des correspondances : Nous suivrons le plan en trois étapes (a,b,c) proposé au § 2.2.3, 6°.

3.1 Représentation approchée du nuage $N(I)$ dans R_M : L'analyse du tableau de correspondance $k(I,M)$ (résultats du premier tour) introduit dans R_M un système d'axes orthonormés de centre f_M avec pour vecteur de base les $u_{\alpha M}$. Il importe, en vue des calculs ultérieurs, de rappeler l'expression des vecteurs unitaires de base en fonction des quantités ordinairement calculées. On a :

$$u_{\alpha M} = \{u_{\alpha m} | m \in M\} ; u_{\alpha m} = \varphi_{\alpha}(m) f_m = \lambda_{\alpha}^{-1/2} G_{\alpha}(m) f_m ;$$

il résulte de l'équation même de définition des axes factoriels, qu'on a :

$$u_{\alpha M} = \lambda_{\alpha}^{-1/2} \sum \{F_{\alpha}(i) f_i \cdot f_M^i | i \in I\} ;$$

(cette formule peut d'ailleurs être vérifiée, en remplaçant les f_M^i par leurs valeurs données par la formule de reconstitution).

La formule :

$$f_M^i = f_M + \sum_{\alpha} F_{\alpha}(i) u_{\alpha M}$$

est une formule exacte si la somme est étendue à tous les facteurs non triviaux ; elle est approchée si on se borne aux premiers facteurs.

3.2 Projection du nuage $N(J)$ sur les axes ajustés à $N(I)$: Le support de $N(I)$ est rapporté aux axes $u_{\alpha M}$; projeter $N(J)$ sur ces axes, c'est adjoindre les j en éléments supplémentaires à l'analyse du tableau $I \times M$. Nous noterons désormais $F_{\alpha}(j)$ les valeurs de facteurs issus de l'analyse de $I \times M$ pour les j considérés comme éléments supplémentaires. La formule :

$$f_M^j \approx f_M + \sum_{\alpha} F_{\alpha}(j) u_{\alpha M}$$

n'est pas en général une formule exacte, même si l'on prend tous les facteurs non triviaux issus de l'analyse de $I \times M$: en effet puisque $\text{Card } I < \text{Card } M$, les axes factoriels ne constituent dans R_M qu'une base incomplète : plus précisément, ils constituent une base pour le support de $N(I)$. Donc en étendant la somme à tous les facteurs non triviaux, on aura seulement la projection orthogonale de f_M^j sur ce support, et si on ne retient que quelques axes, on aura la projection sur le sous-espace engendré par ces axes.

Il importe de noter que la formule proposée, peut en vertu de l'égalité rappelée au § 3.1, s'exprimer en fonction des profils f_M^i : on a :

$$\begin{aligned} f_M^j &\approx f_M + \sum \{ \lambda_{\alpha}^{-1} F_{\alpha}(i) F_{\alpha}(j) f_i^1 | i \in I, \alpha \in A \} \\ &\approx f_M + \sum \{ \lambda_{\alpha}^{-1} F_{\alpha}(i) F_{\alpha}(j) f_i^1 | \alpha \in A \} f_M^i | i \in I \} \end{aligned}$$

comme f_M n'est autre que $\sum_i f_i^1 f_M^i$, le profil f_M^j apparaît ainsi comme une combinaison linéaire des profils f_M^i . Avec les notations du § 2.2.3, on peut encore écrire :

$$f_i^j \approx f_i^1 (1 + \sum_{\alpha} \lambda_{\alpha}^{-1} F_{\alpha}(i) F_{\alpha}(j)) ;$$

cette formule a avec la formule classique de reconstitution une similitude frappante : la seule différence étant que λ_{α}^{-1} remplace $\lambda_{\alpha}^{-1/2}$. Il faut toutefois noter que les facteurs $F(i)$, $F(j)$ ne sont pas issus de l'analyse d'un tableau $I \times J$, mais d'un tableau $I \times M$ avec $J \times M$ en supplémentaire.

3.3 Conditions complémentaires : Telles quelles les valeurs f_i^j suggérées au § 3.3 ont peu de chance d'être satisfaisantes : on sait que projeter sur un sous-espace, introduit généralement des valeurs négatives, comme on le voit en appliquant la formule de reconstitution. Intuitivement on peut dire que les termes en $F_{\alpha}(j)$ sont des termes de contraste par lesquels les f_i^j diffèrent quand j varie, et que nous désirons introduire un contraste aussi fort que possible dans les limites permises par les conditions imposées.

En toute rigueur, le problème pourrait être posé directement : trouver des f_i^j qui reconstruisent des f_M^j aussi proches que possible des originaux. Toutefois nous préférons partir de la formule du § 3.2, en la modifiant le moins possible : ainsi les calculs seront à la fois plus simples et d'interprétation plus sûre.

On pose donc simplement :

$$f_i^j = f_i (1 + \sum_{\alpha} N_{\alpha} \lambda_{\alpha}^{-1} F_{\alpha}(i) F_{\alpha}(j))$$

les coefficients N_{α} étant choisis aussi proche de 1 que possible. Nous devons d'une part vérifier la condition de marge du § 2.2.3, 5°; d'autre part fixer le critère d'ajustement optimum. La condition de marge se vérifie immédiatement : en effet l'égalité :

$$\forall i \in I : \sum_i \{f_i^j \mid j \in J\} = f_i$$

résulte de ce que sur l'ensemble J les facteurs $F_{\alpha}(j)$ ont moyenne nulle (parce que $N(J)$ et $N(I)$ ont même centre de gravité f_M).

Reste le critère d'ajustement et la condition de positivité. Le critère sera la distance quadratique, au sens de la métrique du χ^2 de centre $f_J \otimes f_M$, entre le tableau $J \times M$ donné (résultats du 2-ème tour) et le tableau reconstitué. Les deux tableaux ayant même marge, sur J et M , cette distance est donnée par la formule :

$$d^2 = \sum \{f_j \|f_M^j - f_M^j\|^2 \mid j \in J\}$$

où f_M^j désigne le profil reconstitué, et les distances se calculent dans la métrique du χ^2 de centre f_M . En fait le calcul des distances se fait très simplement dans le système orthonormé des axes factoriels :

$$\|f_M^j - f_M^j\|^2 = \sum_{\alpha} ((N_{\alpha} F_{\alpha}(j)) - F_{\alpha}(j))^2 ;$$

dans cette formule, manquent les termes correspondant aux directions perpendiculaires au support du nuage $N(I)$ (ou au sous-espace engendré par les premiers axes factoriels seuls retenus) : mais ces termes ne dépendant pas du choix des N_{α} , il est inutile de les faire entrer dans le critère. On a donc pour quantité critère :

$$\begin{aligned} \|f_{JM} - f_{JM}\|^2 &= \sum \{f_j \|f_M^j - f_M^j\|^2 \mid j \in J\} \\ &= \sum_{\alpha} (N_{\alpha} - 1)^2 \sum \{f_j F_{\alpha}(j)^2 \mid j \in J\} \\ &= \sum_{\alpha} (N_{\alpha} - 1)^2 \mu_{\alpha} \end{aligned}$$

ou on a noté μ_{α} l'inertie du nuage $N(J)$ en projection sur l'axe α (issu de l'analyse du nuage $N(I)$).

C'est cette quantité critère très simple qu'on doit minimiser sous les conditions complémentaires que les f_i^j soient tous positifs; i.e. :

$$\forall i, j : 1 + \sum_{\alpha} N_{\alpha} \lambda_{\alpha}^{-1} F_{\alpha}(i) F_{\alpha}(j) \geq 0.$$

On a abouti à un problème classique d'optimisation quadratique sous contrainte linéaire ; avec pour inconnues les N_{α} .

3.4 Formule de reconstitution approchée : Finalement on a pour le tableau ternaire et sa marge binaire sur I x J les formules de reconstitution approchées :

$$f'_{ijm} = f_j f_{im} (1 + \sum_{\alpha} N_{\alpha} F_{\alpha}(i) F_{\alpha}(j))$$

$$f'_{ij} = f_i f_j (1 + \sum_{\alpha} N_{\alpha} F_{\alpha}(i) F_{\alpha}(j))$$

Nous parlons de "reconstitution approchée" parce que la marge sur J x M du tableau f'_{ijm} reconstitué n'est pas exactement la marge f_{JM} donnée (résultats du 2-ème tour). Pour le tableau des k'(i,j,m) on aura de même : k'(i,j,m) = k f'_{ijm} ; ou pour les tableaux de reports proprement dits : k'(i,j) = k f'_{ij} ; (où k'(i,j) est le nombre estimé des électeurs ayant eu respectivement aux 1-er et 2-ème tours (les attitudes i et j).

4 Application à l'élection présidentielle de 1981

4.1 Résultat de l'analyse factorielle : On prend comme tableau principal le tableau M . I, où k(m,i) est le nombre des électeurs du département m ayant adopté au premier tour l'attitude i ; et on adjoit en supplémentaire le tableau M . J des attitudes de vote au 2-ème tour. De façon précise, on a pour le 1-er tour, cumulé en une seule attitude de vote notée "autres" les suffrages obtenus par les candidats ayant reçu moins de 5% des voix (M. Debré ; M.-F.Garaud ; A. Laguiller ; B. Lalonde ; etc.) ; et aux deux tours, les blancs et les nuls ont été comptés avec les abstentions (bien que les blancs représentent une attitude originale de refus actif ; tout autre que l'abstention). De plus le département du Bas-Rhin (très fortement associé à un candidat "autre" a été mis en supplémentaire.

VAL PROPRE	POURCENT	CUMUL	+	HISTOGRAMME DES VALEURS PROPRES DE LA MATRICE
0.01831	50.239	50.239	*	***** ***** ***** **
0.00671	18.410	68.649	*	***** ***** *****
0.00473	12.970	81.619	*	***** ***** *****
0.00379	10.410	92.030	*	***** ***** *****
0.00290	7.970	100.000	*	***** ***** *****

	J1	QLT	POID	INR	1#F	COR	CTR	2#F	COR	CTR	3#F	COR	CTR	4#F	COR	CTR	5#F	COR	CTR
1	GIS1	1000	221	164	103	388	127	-104	397	354	-65	157	199	35	45	71	19	13	28
2	CHI1	1000	146	164	89	196	64	158	612	544	-85	177	223	-25	15	24	0	0	0
3	MIT1	1000	210	82	-19	25	4	-43	129	57	25	45	29	-84	502	394	-65	299	307
4	MAR1	1000	126	400	-336	975	775	2	0	0	-39	13	40	17	2	9	34	10	49
5	AUT1	1000	102	98	70	138	27	10	3	2	109	337	255	-58	97	91	122	426	524
6	ABS1	1000	196	93	18	18	3	38	85	43	78	355	254	89	462	412	-37	79	92
			1000			1000			1000			1000			1000			1000	

Le tableau principal ayant 6 colonnes, on a cinq facteurs principaux : d'après l'histogramme des valeurs propres, et le tableau relatif aux attitudes de votes, on voit qu'aucun de ces facteurs n'est négligeable : en particulier le 5-ème facteur apporte à l'attitude "Mitterrand 1-er tour" une contribution relative de 30%. Quant aux attitudes du 2-ème tour, la qualité de représentation QLT est excellente pour les deux candidats (95% et 94% respectivement) ; et bonne pour les abstentions (87%). Nous nous bornerons à donner une interprétation détaillée avec nuage des département, pour le plan (1,2).

Sur l'axe 1, Marchais (à l'extrême gauche) s'oppose à Giscard et Chirac (celui-là étant un peu à droite de celui-ci) ; Mitterrand est presque au centre ; les "autres" tombent à droite (c'est-à-dire associés aux circonscriptions où prédominent les votes de droite). Sur l'axe 2 Chirac 1 s'oppose à Giscard 1 : on notera que quelques départements sont fortement associés à Chirac ($F_2 > 0$) : notamment la Corrèze, qui avec une ordonnée double de celle du point qui la suit (Corse n) sort nettement du cadre du graphique ; Mitterrand 1, sur l'axe 2, s'oppose à Chirac 1 : ce qui montre qu'au premier tour, Mitterrand a eu relativement plus de succès dans les départements votant pour Giscard que dans ceux votant pour Chirac.

Au deuxième tour, Mitterrand 2 et Giscard 2 s'opposent l'un à l'autre fort près de l'axe 1. On ne s'étonnera pas de voir Mitterrand passer du centre à la gauche par le report des voix de Marchais 1. Cependant Mitterrand 2 est dans le plan (1,2) au-dessus de la droite joignant Marchais 1 à Mitter. 1: ce qui ne peut s'expliquer que par le report sur Mitterrand de voix qui, au 1-er tour allaient à Chirac ; d'autant plus que le point Abstention s'élève, lui aussi sur l'axe 2, du 1-er au 2-ème tour. Ainsi, l'interprétation du plan (1,2) pose naturellement la question du report des voix ; question à laquelle on s'appliquera à répondre avec plus de précisions dans la suite, par la méthode proposée au § 3.3.

4.2 *Estimation des reports de voix* : On applique avec 5 facteurs la formule du § 3.2 ; sans nous prémunir contre les valeurs négatives, (autrement dit la formule § 3.4, où les N_α sont égaux à 1), soit :

$$k'(i,j)(k(i)k(j)/k) (1 + \sum \{F_\alpha(i)F_\alpha(j)/\lambda_\alpha \mid \alpha = 1, \dots, 5\}) .$$

Nous obtenons un tableau de report de voix assez satisfaisant.

	GIS2	MIT2	ABS2		
GIS1	80643.	-3193.	-429.	0.	77036.
CHI1	31746.	12661.	6539.	0.	50913.
MIT1	-3276.	86093.	-9711.	0.	73113.
MAR1	-1146.	39203.	5901.	0.	43953.
AUT1	18852.	12713.	3912.	0.	35482.
ABS1	12287.	6330.	49623.	0.	68243.
TOUT	139102.	153803.	55834.	0.	348729.

Ce tableau montre les principaux liens attendus entre les attitudes de vote aux premier et second tours ; mais il contient quelques nombres négatifs toutefois ceux-ci sont assez faibles relativement aux marges ; ce qui encourage à appliquer la méthode d'optimisation évoquée au § 3.3. On obtient pour les coefficients N_α les valeurs suivantes :

$$N_1 = 0,923$$

$$N_2 = 0,433$$

$$N_3 = 1,128$$

$$N_4 = 0,684$$

$$N_5 = 0,759$$

APRES AJUSTEMENT					
	GIS2	MIT2	ABS2		
GIS1	73243.	1259.	2535.	0.	77036.
CHI1	38413.	12520.	-0.	0.	50913.
MIT1	1672.	71399.	39.	0.	73113.
MAR1	62.	39130.	4766.	0.	43953.
AUT1	14825.	13228.	7422.	0.	35482.
ABS1	10891.	16271.	41073.	0.	68243.
TOUT	139102.	153803.	55834.	0.	348729.

Selon ce tableau (où l'unité est la centaine des suffrages), sur 5.091.300 voix obtenues par Chirac au 1-er tour ; 1.252.000 (soit environ le quart) se seraient reportées sur Mitterrand ; le reste allant à Giscard ; sans aucune abstention. Pour aider à la lecture on a calculé les pourcentages par ligne, en coefficient f_j^i de report : e.g. 60% des abstentionnistes du 1-er tour le seraient restés au 2-ème; 24% se décident à voter pour Mitterrand ; et 16% seulement pour Giscard;

	GIS2	MIT2	ABS2		
GIS1	951.	16.	33.	0.	77036.
CHI1	754.	246.	-0.	0.	50913.
MIT1	23.	977.	1.	0.	73113.
MAR1	1.	890.	108.	0.	43953.
AUT1	418.	373.	209.	0.	35482.
ABS1	160.	238.	602.	0.	68243.
TOUT	139102.	153803.	55834.	0.	348729.

11% des électeurs de Marchais se seraient abstenus au second tour ; le reste allant à Mitterrand ; sans aucun à Giscard ; etc. .

4.3 *Reconstitution des données au niveau des départements* : Ainsi qu'on l'a dit au § 3.4, les coefficients de report de voix donnent un tableau $k'(i,j,m)$, dont la marge $k'(j,m)$ ne coïncide pas avec les données réelles du 2-ème tour $k(j,m)$. Nous considérons d'abord cet écart globalement au niveau des 3 attitudes de vote du 2-ème tour ; puis nous signalerons les départements pour lesquels la formule de report proposée au § 4.2 s'écarte le plus des données réelles.

Le tableau ci-dessous donne les écarts entre profils des colonnes réelles et reconstituées. A la première ligne on a les distances

j	GIS2	MIT2	ABS2	MOY
$\ f_M^j - f_M\ ^2$	0,0120	0,0105	0,0176	0,0122
$\ f_M^j - f_M^j\ ^2$	0,0008	0,0009	0,0045	0,0015
poids	399	441	160	

des profils de colonnes réelles, du profil de marge, avec pour moyenne l'inertie de $N(J)$; la deuxième ligne donne les écarts entre profils réels et profils reconstitués ; avec pour moyenne la distance $\|f_{JM} - f_{JM}'\|^2$ calculée dans la métrique du χ^2 de centre $f_j \times f_M$. L'écart entre tableau réel et tableau reconstitué apparaît faible relativement à l'inertie du tableau réel. Reste à comparer case par case le tableau f_j^m des profils de vote par département (au 2-ème tour) au tableau des f_j^m reconstitués suivant la formule de report des voix. On donne pour chacune des trois attitudes de vote j la

liste des départements m où l'écart entre f_j^m et f_j^m dépasse 3% de f_j^m . Certains de ces écarts s'expliquent par les particularités politiques du département : en particulier en appliquant à la Corrèze la formule générale de report des voix, on surestime grandement les voix de Giscard et on sous-estime les voix de Mitterrand et les abstentions: ce qu'explique l'influence de Chirac, d'ailleurs notée sur l'axe 2.

5 Conclusion

Faut-il parler ici de validité statistique des résultats? Nous répondrons: non! D'une part les grands nombres en jeu (des millions de votes) feront conclure que tout écart par rapport au modèle est significatif. D'autre part, quel que soit l'intérêt du modèle, il apparaît *a priori* que la réalité est plus complexe : les reports de voix dépendant certainement des départements (notamment parce qu'un même programme politique n'est pas adopté ou rejeté pour les mêmes raisons à la campagne et à la ville ; dans les usines et dans les bureaux etc.). Mais les coefficients de report de voix que nous avons calculés donnent du comportement des électeurs une image schématique vraisemblable, digne d'être comparée à la réalité politique.