Les Cahiers du CEDIMES

2022, Volume 17, Numéro 4, https://doi.org/10.69611/cahiers17-4-08

ISSN 2110-6045

QUANTIFIER DES CRITERES DE SELECTION BOOLÉENS

Philippe RYCKELYNCK
Maître de Conférences, Agrégé et Docteur,
Université du Littoral-Côte d'Opale

France

philippe.ryckelynck@univ-littoral.fr

Résumé :
Les décisions que prennent les membres des conseils d'administration ou des conseils stratégiques ou bien encore les décideurs lors des campagnes de recrutement s'appuient bien souvent sur de nombreux critères binaires (ou encore booléens), à la fois dépendants et souvent contradictoires les uns avec les autres. En outre, de nombreux décideurs réagissent suivant des préordres incompatibles qu'il s'agit toutefois d'agréger. La question de fournir des mesures quantitatives pour obvier à toutes ces complications est délicate en général, et cependant fondamentale. On propose ici un cadre méthodologique qui repose sur l'hypothèse que les booléens employés sont rassemblés en quelques groupements homogènes, grâce auxquels sont formés des statistiques quantitatives facilement interprétables, comme un potentiel de connaissances ou un potentiel de ressources humaines, ou un potentiel financier, par exemple. Le nuage représentatif des alternatives donne lieu à des mesures explicites et numériques de ces potentiels en recherchant l'inertie de ces potentiels. Ce faisant, en travaillant dans un espace de faible dimension, correspondant à tous les potentiels, la sélection d'alternatives se termine par la méthode multicritère Electre. On applique ce cadre conceptuel à une enquête statistique faite sur des éco-entrepreneurs du territoire de Dunkerque.

Mots-clés *: booléens, statistiques, méthode Electre, optimisation multicritère, recrutement.*

Abstract:
The decisions taken by boards of directors or strategic councils or even decision-makers during recruitment campaigns are often based on numerous binary (a.k.a. boolean) criteria, both dependent and often contradictory with each other. In addition, many decision-makers react according to incompatible pre-orders which, however, need to be aggregated. The question of providing quantitative measures to obviate all these complications is delicate in general, and yet fundamental. We propose here a methodological framework based on the assumption that the Booleans used are gathered into a few homogeneous groupings, thanks to which are formed easily interpretable quantitative statistics, such as a knowledge potential or a human resources potential, or a financial potential, for example. The representative cloud of alternatives gives rise to explicit and numerical measurements of these potentials by looking for the inertia of these potentials. In doing so, by working in a low-dimensional space, corresponding to all the potentials, the selection of alternatives ends with the Electre multi-criteria method. We apply this conceptual framework to a statistical survey of eco-entrepreneurs in the Dunkirk area.

Keywords: *Booleans, statistics, Electre, multiobjective optimization.*

Classification JEL *: C44, C69, C61, C65, C44.*

1. Introduction et position du problème

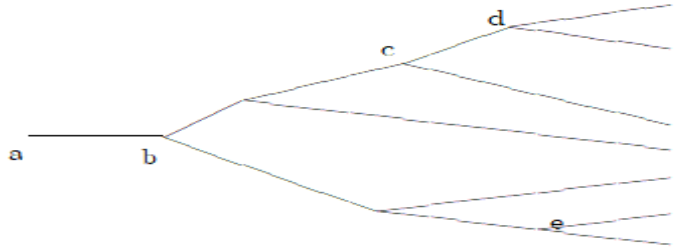
Lorsqu'un recrutement doit être fait, différents candidats $C_1, C_2, \dots, C_K$ se présentent devant un jury formé de juges $A_1, A_2, \dots, A_J$ et ceux-ci, après qu'ils aient auditionné les aspirants, leur attribuent, au regard de critères de sélections $S_1, S_2, \dots, S_I$ des notes, des valeurs, ou des impressions qualitatives traduites par des valeurs binaires (0/1, oui/non, apte/inapte, etc.) permettant des jugements personnels, relativement à chaque arbitre, sur l'ensemble de ces auditions ou entretiens. Ainsi formulée, la prise en compte sincère des résultats d'un recrutement impose de gérer des données comprises dans un tableau à trois dimensions de tailles respectives *(I, J, K)*. Un « tableau » de cette sorte est appelé un *tenseur* en mathématiques[1]. Le nombre total des entrées d'un tenseur peut dépasser l'entendement ; c'est ainsi que, pour recruter sur un poste pour lequel seront évaluées une vingtaine de compétences de chaque candidat dans une liste de cinquante personnes, un jury d'une dizaine d'arbitres aura à manipuler un tenseur comportant dix mille entrées. Il ne saurait être question de s'en remettre exclusivement à un vote, faute de quoi le scénario d'un recrutement honnête paraîtrait assurément biaisé, et le facteur humain aurait dès lors une importance excessive. C'est à cette problématique que s'attaque cet article. Avant de poursuivre, on notera dès maintenant que le contexte du recrutement n'est en rien le seul auquel ce qui suit doit s'appliquer ; on pourra avoir en vue les délibérations d'un conseil d'administration sur les stratégies d'une entreprise, les travaux des chefs d'un projet industriel pour choisir les orientations à adopter afin de mener à bien les solutions qui y mènent. Bien entendu dans ces deux nouveaux contextes, la mise en œuvre sera, au vocabulaire près, mutatis-mutandis identique, et peut-être même plus probante.

2. Décisions dans des ensembles partiellement ordonnés

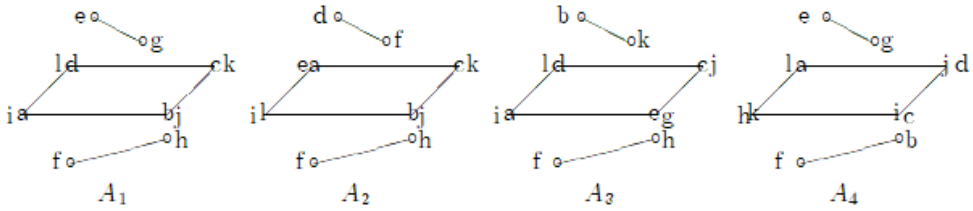
Un arbitre A_j peut ainsi se représenter les candidats par niveaux comme dans la figure ci-après. Cette représentation fait naturellement apparaître que chaque juge A_j aura structuré l'ensemble $\{C_1, C_2, \dots, C_K\}$ en un ensemble partiellement ordonné (*poset*, en franglais). Un poset n'est autre chose qu'un ensemble avec une relation binaire de comparaison R, traduisant la préférence, xRy signifiant que y est préféré à x, de façon à satisfaire aux deux axiomes habituels *réflexivité* et *transitivité*[2]. On insiste sur le fait que la rationalité des choix individuels est nécessaire au développement qui va suivre. C'est ainsi que les préférences manifestées par les électeurs lors d'élections uninominales n'est nullement par essence transitive. Si l'ensemble ainsi formé, $\{C_1, C_2, \dots, C_K\}$ muni de la relation binaire R_j associée au jury A_j, est un arbre et possède une racine, la sélection du meilleur candidat ou de la meilleure candidate est immédiate pour A_j. Mais des niveaux peuvent surgir et faire en sorte qu'on ne puisse discriminer entre les candidats.

[1] Ce mot est méconnu en dehors de la géométrie différentielle et de la Relativité, car la quasi-totalité des quantités employées par les scientifiques forment soit des scalaires, soit des vecteurs, soit des matrices et ne requièrent ainsi qu'au plus deux indices.

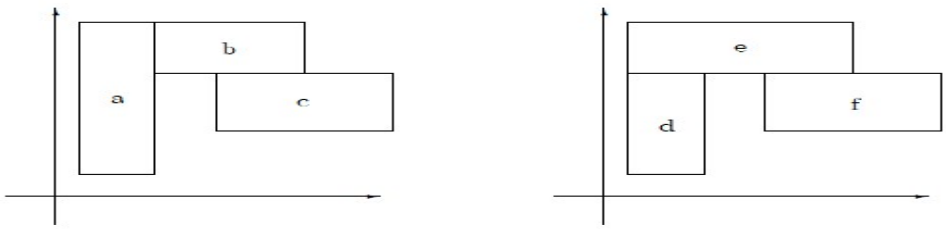
[2] La *réflexivité :* xRx pour tout x, et la *transitivité :* $xRy\ et\ yRz$ entraine xRz pour tous x, y et z, ce qui traduit la cohérence ou si, l'on veut, la rationalité des choix. On n'impose pas le caractère *antisymétrique* : xRy et yRx entraine que $x = y$ pour tous x et y.



Vient alors la question de l'agrégation des préordres de préférence R_j. La figure suivante, où les quatre préordres de quatre décideurs et douze candidats sont orientés de la gauche vers la droite et du bas vers le haut, concerne deux critères seulement. Chaque décideur a rendu quatre fois équivalents deux candidats, mais les paires de candidats ainsi mis sur de mêmes paliers différent entre deux jurys. La situation est loin d'être dépourvue d'épines.



Mais la décision doit se baser sur des réalités bien plus complexes que celles de l'agrégation d'arborescences. Un premier temps consiste alors à classer plus librement les candidats. Chaque juge peut classer, à la façon de celle employée dans la logique floue, les candidats de façon à ce que des hésitations sur les préférences soient tenues en compte. Dans l'exemple de la figure ci-dessous, on peut classer six candidats a,b, …f, suivant deux compétences X et Y, représentées en abscisse et ordonnée, en les ordonnant mutuellement, mais sans être catégoriques en cela.



Dans cet exemple, il est clair que la pente de la direction transverse influera pour décider si b<c ou b>c, voire a>c, et de même si e>f ou f>e.

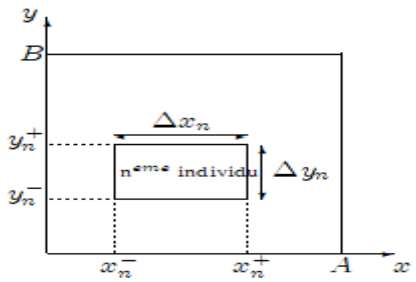
Si l'intégralité des juges et des critères permettent d'établir une fonction discriminante au regard du poste à pourvoir, c'est que les données ont été assemblées en des fonctions d'utilité[1] et agrégées en une mesure unique, qu'on peut assimiler à une note, qui rassemble les décisions individuelles, malgré les divergences que la variété des juges et la pluralité des critères

[1] La question de la continuité d'une fonction d'utilité a été clarifiée par Gérard Debreu sous certaines hypothèses de convexité ; voir l'ouvrage de Jacques Attali pour une démonstration topologique.

rendent inéluctables. La figure précédente montre deux situations possibles où le seul accord possible est un optimum de Pareto[1], dans lesquelles les candidats ne sont estimés qu'au regard de deux critères seulement. Telle est généralement la conclusion d'une démarche dans laquelle on n'aura pas trouvé de mesures agrégatives correctes, dans le sens où elles respectent le plus possible les optima de Pareto des préordres des divers juges. Et c'est ainsi que la problématique de l'existence de ces fonctions numériques compatibles avec les préordres se pose, puisqu'elle vise incidemment à détecter de vraies décisions conformes, dans un sens précis qui apparaîtra infra, aux choix des juges.

3. Représentations par niveaux de critères et orientation des polyèdres

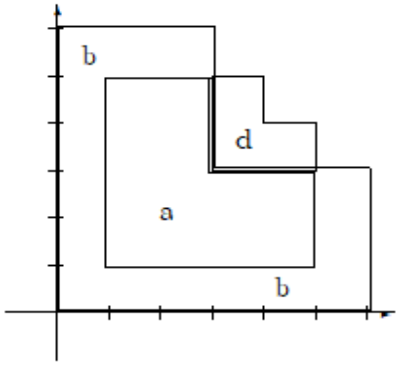
Une technique pour quantifier ces situations est la suivante, que nous détaillons, pour simplifier, lorsque deux critères seulement président aux choix de décisions ; la généralisation à toute dimension ne pose pas de difficulté conceptuelle. Tout d'abord, si les données recueillies contiennent des valeurs non booléennes mais numériques et sujettes à discussion, on range celles-ci par intervalles, ou classes, de façon à obtenir une évaluation binaire en recueillant autant de booléens que de niveaux de compétence. C'est ainsi que procèdent certains organismes de certification, notamment dans l'apprentissage de la langue anglaise.



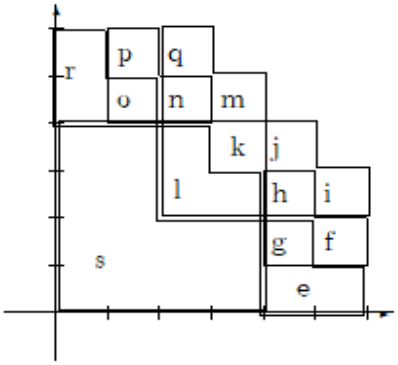
On isole dès lors les candidats dans des polygones compacts connexes mais non pas forcément convexes[2] et, disjoints deux à deux, par exemple des losanges, ou des rectangles, de barycentres $\overrightarrow{C_k} = (x_k, y_k)$.

[1] Une décision est dite *optimale au sens de Pareto* lorsqu'elle n'est pas dominée par toute autre décision relativement à tous les critères. En d'autres termes, aucune autre décision n'est meilleure pour chacun des critères de décision pris en compte, ou bien quel que soit le juge.

[2] Si le candidat a atteint un niveau de compétence X (resp. Y) dans la connaissance 1 (resp. 2), alors tous les points (x,0) et (0,y) avec x (resp.y) moindre que X (resp. Y) font partie du domaine de compétences du candidat. Mais l'exigence de convexité impose dès lors que tous les points (x,y) aussi sont dans le domaine définissant le candidat. Un tel domaine est dit *étoilé*. Si l'on admet la simultanéité des compétences mises en œuvre ou des éléments d'ordre impondérable dans l'exercice de celles-ci, la convexité peut être mise en défaut.

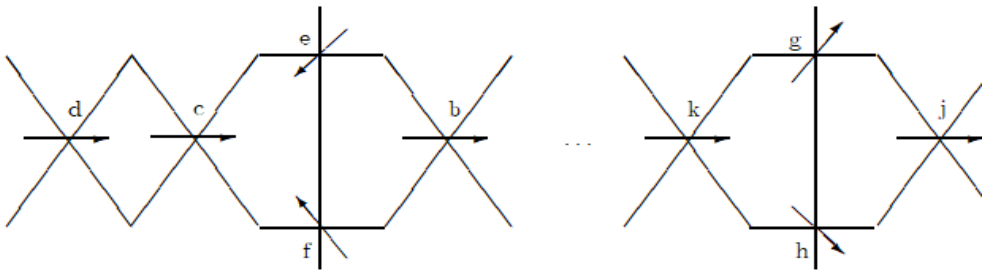


Quatre candidats



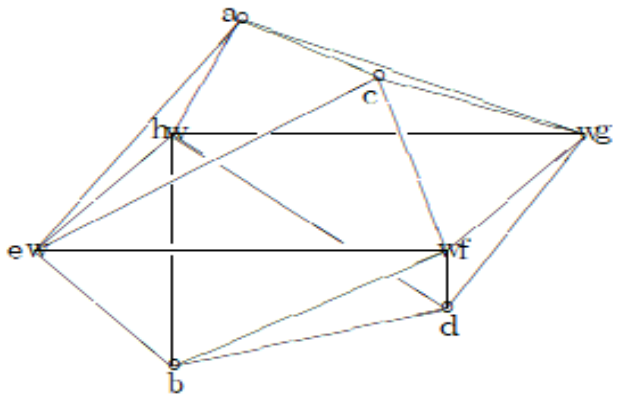
Quinze candidats

Ensuite, on cherche une direction dans le plan $\vec{d} = (d_1, d_2)$ qui soit telle que les projections orthogonales des candidats $\overrightarrow{C_k}$ sur cette droite soient deux à deux distinctes. Un raisonnement simple de topologie montre que de telles directions existent toujours et sont en quantité infinie[1], de sorte que la question de la sensibilité de l'ordre au choix des directions $\vec{d}$ ne se pose pas. Dès lors, le meilleur choix est la personne $\overrightarrow{C_k}$ qui rende maximale les quantités $s_k = d_1 x_k + d_2 y_k$. Une telle fonction pondère ainsi, et quantifie à la fois, les compétences. Une telle technique s'applique quelle que soit la dimension, mais la visualisation de la démarche n'est guère possible. On peut dès lors procéder en agrégeant progressivement les critères au regard desquels on juge les futures recrues. La figure montre la complexité croissante d'une telle approche, ne serait-ce qu'en dimension deux, lorsque deux critères qualitatifs seulement sont employés. L'exemple complexe proposé dans la figure ci-dessous montre qu'avec une certaine stabilité dans le critère $d_1 x + d_2 y$ employé comme transverse au graphe, l'ensemble des candidats admet une sous-suite croissante formée des candidats (d,c,b,k,j); parmi ceux-ci j est la meilleure recrue, et il surclassera le candidat h, donc aussi le candidat f.



[1] L'argument est le suivant. Si on choisit par exemple une droite d'équation $g_1 x + g_2 y = g_0$ assez proche de l'axe d'inertie du nuage mais ne rencontrant en aucun point le nuage, alors, avec le théorème de séparation des convexes, on peut trouver un voisinage ouvert du triplet (g_1, g_2, g_0) formé de coefficients d'équations de droites qui répondent toutes à la question. Ce raisonnement n'est nullement spécifique à la dimension 2.

Dans la dimension trois, les polygones deviennent des polyèdres, et l'imagination seule ne permet pas de décider quel sera le choix retenu. L'exemple donné par la figure est celui d'un dodécaèdre, ayant pour interprétation que huit candidats sont jugés à l'aune de trois compétences, et l'orientation choisie est de la gauche vers la droite, du bas vers le haut, de l'arrière vers l'avant. La frontière de Pareto sera dès lors formée des points wg, a,c. Là encore, sans quantification, via une direction transverse au polyèdre, nul moyen ne se présente d'éviter les compromis parétiens.



La dimension du préordre d'un juge n'est autre que le nombre de critères de jugement ou de compétences détectées dans les candidats qui lui sont présentés ; dans l'étude numérique ci-après cette dimension vaut 92. Lorsque cette dimension est grande, le choix d'une direction $\vec{d} = (d_1, .., d_n)$ transverse au polyèdre est encore possible si et seulement si le préordre est transitif, et en ce cas ce choix équivaut à celui d'une fonction linéaire $d_1 x_{k,1} + ... + d_n s_{k,n}$ qui soit croissante par rapport à l'ordre, et à nouveau ces directions forment un ensemble ouvert infini qu'hélas nulle représentation graphique ne permet de détecter. Tel est le sens de la méthode explicitée ci-après.

4. Enquêtes générant des statistiques booléennes (ou binaires)

Lorsqu'un sondage ou une enquête sont menés dans une population, il peut se faire que certaines questions qualitatives posées à la population $P = \{I_1, I_2, ..., I_N\}$ ne soient pas d'une clarté absolue, et les individus $I_1, ..., I_N$, quand ils répondent, sont dans trois situations : répondre oui ou non, ou bien NSP (ne se prononçent pas), ou alors ne fournissent aucune réponse (ce qu'on appelle des données censurées), ou même encore, se trompent, et auraient assurément préférer répondre avec une échelle de valeur, une graduation entre 0 et 10 par exemple. La troisième situation mène à ce qu'on appelle des caractères quantitatifs de logique floue, et la première fournit des caractères booléens. On récolte ainsi, pour chaque question Q_j un vecteur booléen $\Phi_j = (\phi_j(I_1), \phi_j(I_2), ..., \phi_j(I_N)$, dont les composantes sont 0 ou 1, et nous supposerons dans ce qui suit que ces vecteurs sont complets, autrement dit ne contiennent aucune donnée censurée[1].

[1] Lorsque la matrice d'un questionnaire ne contient pas de donnée censurée, si l'objet de l'étude n'est ni la sélection ni l'optimisation, mais par exemple la détection de certains profils, alors la méthode à

On peut rencontrer des booléens discriminants, dans le sens sociétal contemporain de ce terme. Sont de cette nature les six exemples suivants : $\phi_1(I_k)$ est le genre de I_k, $\phi_2(I_k)$ est la pratique religieuse de I_k, $\phi_3(I_k)$ est l'adhésion de I_k à un parti politique fixé, et $\phi_4(I_k)$ est la catégorie d'âge (junior/senior) de I_k, $\phi_5(I_k)$ est la nationalité (français/étranger) de I_k, $\phi_6(I_k)$ est la situation de I_k vis à vis de l'emploi, etc. D'autres booléens peuvent permettre de discriminer « valablement » la population, et la procédure pour ce faire peut être ainsi décrite. Lorsque l'enquête contient un questionnaire assez riche, et une base d'exploitation assez large, il convient de filtrer les données au regard de booléens qui ont été recueillis, mais non exploités dans l'analyse statistique. La démarche procède ainsi. Ayant mesuré des booléens $\phi_j: \{I_1, I_2, \dots, I_N\} \to \{0,1\}$ où $1 \leq j \leq M$, on a donc formé un vecteur statistique X constitué de Q mesures, indépendant des M booléens précédents (au sens où aucune composante de X ne contient les ϕ_j). On peut déduire un vecteur statistique Y sur la sous-population P' formée des individus I_i appartenant au groupe discriminé, c'est-à-dire tels que $\phi_j(I_i) = Vrai = 1$ pour toute discrimination ϕ_j examinée.

Les booléens peuvent être eux-mêmes rangés en différentes classes homogènes, en ce sens que les questions auxquelles répondent ces booléens sont apparentées. Par exemple, on formera ainsi plusieurs listes de questions qui porteront respectivement sur les motivations et les convictions (M), sur la scolarité et les diplômes et les connaissances tacites (S), sur les réseaux humains formels ou informels mais noués par les individus (R), sur les questions financières ou entrepreneuriales (F), etc. Il est important que cette préparation des booléens soit bien pensée pour équilibrer les booléens et garantir aussi une certaine indépendance stochastique entre les différents groupements.

On renvoie aux traités et ouvrages récents de synthèse de L.A. Goodman (2000), de J. Han et M. Kamber (2006), ainsi que C. J. Lloyd, (1999), qui explorent l'analyse des données, tout particulièrement discrètes et/ou qualitatives, de façon moderne et complète. L'auteur étant français, qu'il lui soit permis aussi de recommander le remarquable traité de G. Saporta, plutôt unique dans son style, mais fécond.

5. Un exemple effectif d'enquête relative à des éco-entrepreneurs dunkerquois

Donnons quelques exemples de questions qualitatives qu'on posera volontiers sur le sens civique et collectif qu'auront des entrepreneurs ou des salariés vis à vis des questions écologiques et politiques de notre époque. Le questionnaire a été effectué au printemps 2015 auprès de 60 éco-entrepreneurs de la région de Dunkerque, grâce notamment à un staff d'étudiants. Les interviews in situ contenaient un peu plus d'une centaine de questions essentielles mais, somme toute, pas moins de 92 booléens ne contenaient aucune censure. De même, il a fallu exclure une soixantaine de répondants n'ayant pas parachevé correctement l'enquête. Une première analyse faite avec Sphinx a permis de voir quels booléens étaient réellement discriminants et d'étudier la liaison desdits booléens. Le détail de cette enquête a

employer est bien connue et déjà classique. En effet, on traduira la contrainte qui donne lieu au sous-ensemble recherché par une expression booléenne traduisant globalement les conditions de recrutement. A l'aide d'un calcul booléen ou d'un *tableau de Karnaugh*, on peut aisément retrouver de façon exhaustive le groupe voulu. Dans Maple ou Excel, les calculs correspondants se font aisément (aux références absolues près).

été donné dans l'article de l'auteur et de Sophie Boutillier en 2017, auquel on renvoie. Les indices donnés ci-après correspondent à ceux employés dans cet article. Quatre premiers booléens ϕ_j « personnels » forment le début de l'enquête et n'ont pas été retenus comme variables explicatives (elles étaient assurément discriminantes). On retrouve alors quatre grands groupes de booléens :

a) Vis à vis du territoire, les sondés I_n, où $1 \leq n \leq N$, disposent de ressources du territoire géographique ϕ_5 à ϕ_{13} , ainsi que de ressources du territoire organisationnelles ϕ_{14} à ϕ_{19} et enfin de ressources cognitives ϕ_{20} à ϕ_{24}.
b) Vis à vis des motivations, les sondés I_n vont évoquer leurs motivations personnelles ϕ_{25} à ϕ_{28}, économiques ϕ_{29} à ϕ_{34}, les éventuels indicateurs liés à l'innovation ϕ_{36} à ϕ_{39}, ou bien à la réglementation ϕ_{40}, voire des motivations très actuelles liées à l'intérêt pour la protection de l'environnement ϕ_{41}, ϕ_{45}.
c) Les sondés I_n ont des ressources sociales rassemblées en un réseau informel que décrivent les booléens ϕ_{46} à ϕ_{49}.
d) Les sondés I_n ont des ressources sociales rassemblées en un réseau formel et qui forment d'autres booléens ϕ_{50} à ϕ_{57}.

Par ailleurs, le questionnaire peut mettre en lumière des caractères booléens qui participent au potentiel de ressources de différentes espèces, par exemple certaines des ressources financières ϕ_{58} à ϕ_{62}, ou bien certaines ressources en connaissance codifiées ϕ_{63} à ϕ_{69} (il s'agit là essentiellement de niveaux de diplômes), voire des ressources en connaissance tacites ϕ_{68} à ϕ_{70}.

Ce sont donc là les 92 booléens à la base de l'enquête et sur lesquels s'appuie l'étude qui suit.

6. La matrice booléenne de l'enquête et ses interprétations combinatoires

Supposons que l'enquête contienne donc M booléens $\phi_j: \{I_1, I_2, \dots, I_N\} \to \{0,1\}$ mesurés sur N individus. On forme une matrice booléenne A de format $M \times N$ qui contient les réponses des individus aux questions. Nous donnons un exemple réel de mesure. Il est essentiel que la matrice A soit de rang égal à N. Les sommes des colonnes sont évidemment les cardinaux des différents sous-ensembles discriminés par les booléens, autrement dit par les fonctions caractéristique ϕ_k. Les groupes dominants sont donc ceux pour lesquels $|\phi_k| > N/2$, si on note $|E|$ le cardinal d'un ensemble. Le produit matriciel $A \cdot A^T$ contient bien sûr tous les cardinaux d'intersections $|\phi_{j_1} \cap \phi_{j_2}|$. On note qu'il y a $(M+1) \times M/2$ choix de deux caractères booléens. Les couples de questions antinomiques sont détectés par $|\phi_{j_1} \cap \phi_{j_2}| = 0$. Les couples majoritaires sont ceux qui recueillent des intersections de cardinaux autour de $N/2$, plus précisément tels que $|\phi_{j_1} \cap \phi_{j_2}| \in [3N/8{,}5N/8]$.

Par exemple, pour l'enquête dunkerquoise, on note qu'il y a $91 \times 90/2 = 4095$ choix de deux caractères booléens. De plus les booléens pour lesquels $|\phi_k| > 25$ sont ϕ_{41}, ϕ_{51}, ϕ_{76}, ϕ_{91}, ϕ_{131}, ϕ_{132}, ϕ_{161}, ϕ_{138}, ϕ_{23}, ϕ_{25}, ϕ_{14}, ϕ_{17}.

On introduit ensuite les trois *tenseurs* suivants à M^3 composantes, contenant

$$U_{j_1,j_2,j_3} = |\phi_{j_1} \cap \phi_{j_2}| + |\phi_{j_3}| - 2|\phi_{j_1} \cap \phi_{j_2} \cap \phi_{j_3}|,$$
$$V_{j_1,j_2,j_3} = |\phi_{j_1}| + |\phi_{j_2}| + |\phi_{j_3}| -$$
$$-2|\phi_{j_1} \cap \phi_{j_2}| - 2|\phi_{j_1} \cap \phi_{j_3}| - 2|\phi_{j_2} \cap \phi_{j_3}| + 4|\phi_{j_1} \cap \phi_{j_2} \cap \phi_{j_3}|$$

et enfin

$$W_{j_1,j_2,j_3} = \sum_{i=1}^{N} \phi_{j_1}(i)\phi_{j_2}(i)\phi_{j_3}(i) = |\phi_{j_1} \cap \phi_{j_2} \cap \phi_{j_3}|$$

Les triplets d'indices (j_1, j_2, j_3) tels que $U_{j_1,j_2,j_3} = 0$ et les triplets d'indices (j_1, j_2, j_3) tels que $V_{j_1,j_2,j_3} = 0$ sont d'un intérêt particulier, et caractérisent d'une part les minorités ϕ_{j_3} incluses dans l'intersection de deux autres groupes ϕ_{j_1}, ϕ_{j_2} et d'autre part celles qui sont contenues dans la réunion des deux autres, quelle que soit le choix d'une des minorités parmi les trois. Comme il convient de ne pas choisir les populations sondées sans aucune restriction, c'est parce que certaines homogénéités sont présentes que ces calculs permettent de mieux entrevoir ou de révéler des sortes de coïncidence statistiques[1] qui sont « le fait du hasard » mais révélatrices. On pourra se reporter à l'article de S. Boutillier et de l'auteur donnant sans davantage d'explications quelques-unes de celles-ci.

7. Tests d'indépendance des caractères booléens

Les populations auxquelles une batterie de questions booléennes et quantitatives sont posées sont peu souvent hétérogènes ou prises au hasard. Elles résultent d'un groupe humain précis, comme par exemple les employés d'une entreprise, les membres d'un comité d'entreprise, les candidats à un recrutement, etc. Aussi, il n'est pas surprenant que des corrélations puissent avoir lieu, dans des questions ou des mesures qui n'auraient, dans une population bien plus grande et dont celle que l'on scrute fait partie, aucune marque de dépendance nette.

Ainsi, si la population de référence a quelques traits caractéristiques, conviendra-t 'il de poser différentes questions binaires et de faire quelques mesures pour affiner ce qui peut en quelque sorte diviser une masse homogène en sous-groupes ayant d'autres particularités. Cela étant, on peut tester le caractère indépendant de deux fonctions booléennes ϕ_{j_1} et ϕ_{j_2} avec le test du khi-deux puis avec le test de Goodman-Kruskal. Que les variables soient booléennes apporte une simplification importante dans le calcul de la matrice du khi-deux. Les couples de fonctions très corrélés sont ceux qui vérifient $\chi^2(j_1, j_2) > 3N/4$. Les couples absolument indépendants sont ceux ayant un khi-deux nul $\chi^2(j_1, j_2) = 0$. Nous ne détaillerons pas davantage le test de Goodman-Kruskal $[j_1, j_2]$ employé et concluant, et renvoyons derechef à l'article de S. Boutillier et l'auteur, dans lequel les techniques qui suivent sont absentes.

[1] Tous ces calculs sont aisés mais longs avec le logiciel Maple (ou le logiciel libre Scilab), mais impossible avec Excel (sauf si on a très peu de juges ou très peu de critères) et impossibles dans Sphinx.

8. Les caractères déterminants $\Phi_g(i)$ associés aux sous-groupes de critères

Dans ce paragraphe, on aborde la question de l'agrégation de booléens. Supposons que les critères aient été répartis en plusieurs sous-groupes, portant un sens mutuellement, et bien distingués les uns des autres. Pour fixer les notations et rendre claire la démarche, supposons qu'il y ait quatre groupes de booléens ayant pour significations respectives de rassembler les informations binaires et variées des individus relativement (i) à leurs connaissances, (ii) leurs ressources financières, (iii) leurs ressources sociales en intégrant à la fois les réseaux formels et informels, et enfin (iv) les motivations pour agir. Pour former les quatre caractères statistiques quantitatifs $\Phi_c(i)$, $\Phi_f(i)$, $\Phi_s(i)$ $\Phi_m(i)$ essentiels, sur lesquels l'inertie du nuage de points est maximale, on sépare les différentes fonctions booléennes en quatre groupes d'intérêt et on agrège celles-ci en polynômes formels ayant de nombreux coefficients réels à déterminer ultérieurement.

Pour former les trois caractères statistiques quantitatifs $\Phi_c(i)$, $\Phi_f(i)$, $\Phi_s(i)$ essentiels, sur lesquels l'inertie du nuage de points est maximale, on sépare les différentes fonctions booléennes en trois groupes d'intérêt qu'on agrége en polynôme formel et pour cela on introduit 69 coefficients réels à déterminer ultérieurement.

Ainsi, les ressources en connaissance codifiées et les ressources en connaissance tacites ϕ_{c1} à ϕ_{c9} seront agrégées dans la fonction numérique telle que

$$\begin{gathered}\Phi_c(i) = \alpha_0 + \alpha_1\phi_{c1}(i) + \alpha_2\phi_{c2}(i) + \alpha_3\phi_{cc}(i) + \alpha_4\phi_{c4}(i) + \\ +\alpha_5\phi_{c5}(i) + \alpha_6\phi_{c1}(i)\phi_{c2}(i) + \alpha_7\phi_{c1}(i)\phi_{c2}(i) + \alpha_8\phi_{c1}(i)\phi_{c3}(i) + \ldots + \\ +\alpha_{21}\phi_{c1}(i)\phi_{c2}(i)\phi_{c3}(i) + \alpha_{22}\phi_{c1}(i)\phi_{c2}(i)\phi_{c4}(i)\alpha_{23}\phi_{c2}(i)\phi_{c3}(i)\phi_{c4}(i)\end{gathered}$$

Les ressources financières ϕ_{f1} à ϕ_{f5} sont agrégées dans la fonction numérique telle que

$$\begin{gathered}\Phi_f(i) = \beta_0 + \beta_1\phi_{f1}(i) + \beta_2\phi_{f2}(i) + \beta_3\phi_{f3}(i) + \beta_4\phi_{f1}(i)\phi_{f2}(i) \\ +\beta_5\phi_{f1}(i)\phi_{f3}(i) + \beta_6\phi_{f1}(i)\phi_{f4}(i) + \beta_7\phi_{f1}(i)\phi_{f2}(i)\phi_{f3}(i) \\ +\beta_8\phi_{f2}(i)\phi_{f3}(i)\phi_{f4}(i) + \beta_9\phi_{f2}(i)\phi_{f4}(i)\phi_{f25}(i)\end{gathered}$$

Les ressources sociales liées au réseau informel ou au réseau formel ϕ_{s1} à ϕ_{s8} s'agrégent dans la troisième fonction qui est similaire à

$$\begin{gathered}\Phi_s(i) = \gamma_0 + \gamma_1\phi_{r1}(i) + \gamma_2\phi_{r2}(i) + \gamma_3\phi_{r3}(i) + \gamma_4\phi_{r4}(i) + \\ \gamma_5\phi_{115}(i) + \gamma_6\phi_{116}(i) + \gamma_7\phi_{117}(i) + \gamma_8\phi_{101}(i) + \gamma_9\phi_{102}(i) + \\ \gamma_{s1}\phi_{111}(i)\phi_{s0}(i) + \gamma_{s2}\phi_{112}(i)\phi_{s0}(i) + \gamma_{s3}\phi_{113}(i)\phi_{s0}(i) \\ +\gamma_{s4}\phi_{114}(i)\phi_{s0}(i) + \gamma_{s5}\phi_{115}(i)\phi_{s0}(i) + \gamma_{s6}\phi_{116}(i)\phi_{s0}(i) + \\ \gamma_{s7}\phi_{117}(i)\phi_{s0}(i) + \gamma_{s8}\phi_{101}(i)\phi_{s0}(i) + \gamma_{s9}\phi_{102}(i)\phi_{s0}(i) + \\ \gamma_{31}\phi_{111}(i)\phi_{s1}(i) + \gamma_{32}\phi_{112}(i)\phi_{s1}(i) + \gamma_{33}\phi_{113}(i)\phi_{s1}(i) + \\ \gamma_{34}\phi_{114}(i)\phi_{s1}(i) + \gamma_{35}\phi_{115}(i)\phi_{s1}(i) + \gamma_{36}\phi_{116}(i)\phi_{s1}(i) \\ +\gamma_{37}\phi_{117}(i)\phi_{s1}(i) + \gamma_{38}\phi_{101}(i)\phi_{s1}(i) + \gamma_{39}\phi_{102}(i)\phi_{s1}(i)\end{gathered}$$

Enfin, les caractères analysant les motivations et agents de la décision donneront lieu à la fonction

$$\Phi_m(i) = \delta_0 + \delta_1\phi_{m1}(i) + \delta_2\phi_{m2}(i) + \delta_3\phi_{m3}(i) + \delta_4\phi_{m1}(i)\phi_{m2}(i) + \delta_5\phi_{m1}(i)\phi_{m3}(i) + \delta_6\phi_{m1}(i)\phi_{m4}(i) + \delta_7\phi_{m1}(i)\phi_{m2}\phi_{m3}(i)(i) + \delta_8\phi_{m2}(i)\phi_{m3}(i)\phi_{m4}(i) + \delta_9\phi_{m2}(i)\phi_{m4}(i)\phi_{m25}(i).$$

On notera que, comme les fonctions indicatrices $\phi_k(i)$ sont booléennes, aucune autre puissance que la première n'est requise, et on voit aisément que le nombre de polynômes possibles pour modéliser un potentiel de ressources d'un type donné est moindre que 2^p où p est le nombre de booléens s'agrégeant dans le potentiel de ressources concerné.

Pour déterminer les modèles quantitatifs les plus adéquats, il convient d'examiner *l'inertie du nuage* quadridimensionnel[1] formé par les points $\Psi(i) = \left(\Phi_c(i), \Phi_f(i), \Phi_r(i), \Phi_m(i)\right)$. Les techniques pour opérer ainsi sont bien connues et n'ont nullement à être expliquées ici. Cependant, donnons quelques étapes du travail à fournir, sur ordinateur. Toutes les étapes peuvent être accomplies avec n'importe quelle version d'un quelconque tableur. On introduit quatre formes quadratiques à I variables au total représentant les parts d'inertie associées à ces caractères

$$I_c = \sum_{i=1}^{N} \Phi_c\,(i)^2,\ I_f = \sum_{i=1}^{N} \Phi_f\,(i)^2,\ I_m = \sum_{i=1}^{N} \Phi_m\,(i)^2,\ I_s = \sum_{i=1}^{N} \Phi_s\,(i)^2.$$

Les valeurs extrêmes sur les sphères unités respectives sont obtenues par des résolutions de systèmes linéaires, et sont des nombres qu'il faut déterminer. On obtient ainsi les nombreux *points critiques* pour chacune des fonctions d'inertie I_g, sur la sphère de l'espace des paramètres. Pour chacune des trois fonctions I_c, I_f, I_s on trouve respectivement 30, 20 et 46 points critiques sur la sphère multi-dimensionnelle dans l'espace de phase associés à chacune d'elles. Pour chacune d'elles encore, deux points critiques seulement sont opposés et fournissent le maximum global lié de ladite fonction sur la sphère. C'est de cette manière qu'on calcule les coefficients (les lettres grecques) dans les formules ci-dessus[2]. On obtient les valeurs suivantes (on n'a affiché que trois décimales) :

$\alpha_0 = .798$	$\alpha_3 = .370$	$\alpha_{33} = .095$	$\alpha_{34} = .015$	$\alpha_{35} = .016$
$\alpha_{36} = .028$	$\alpha_{38} = .012$	$\alpha_4 = .080$	$\alpha_5 = .113$	$\alpha_6 = .028$
$\alpha_7 = .062$	$\alpha_8 = .120$	$\alpha_{23} = .339$	$\alpha_{24} = .080$	$\alpha_{25} = .113$
$\alpha_{26} = .028$	$\alpha_{27} = .062$	$\alpha_{28} = .096$	$\alpha_{40} = .168$	$\beta_0 = .520$
$\beta_1 = .350$	$\beta_2 = .446$	$\beta_3 = .335$	$\beta_4 = .047$	$\beta_5 = .270$
$\beta_6 = .312$	$\beta_7 = .257$	$\beta_8 = .033$	$\beta_9 = .240$	$\gamma_0 = .731$
$\gamma_1 = .138$	$\gamma_3 = .330$	$\gamma_4 = .071$	$\gamma_5 = .031$	$\gamma_6 = .171$
$\gamma_7 = .097$	$\gamma_8 = .150$	$\gamma_9 = .430$	$\gamma_{21} = .075$	$\gamma_{23} = .075$

[1] Cela étant, on voit qu'on a remplacé l'analyse discriminante du nuage de 50 points dans un espace à 92 dimensions par celle d'un nuage de 50 points en dimension 4.

[2] La méthode est laborieuse eu égard à la taille des expressions et au nombre de paramètres mais n'offre aucune difficulté théorique ni pratique, car il s'agit d'optimisation quadratique sous contrainte. Le logiciel employé pour traiter cette question est Maple.

$\gamma_{24} = .072$	$\gamma_{25} = .020$	$\gamma_{26} = .063$	$\gamma_{27} = .047$	$\gamma_{28} = .131$
$\gamma_{29} = .161$	$\gamma_{31} = .040$	$\gamma_{33} = .051$	$\gamma_{34} = .055$	$\gamma_{36} = .025$
$\gamma_{37} = .036$	$\gamma_{38} = .070$	$\gamma_{39} = .077$		

et également tous les coefficients nuls suivants :

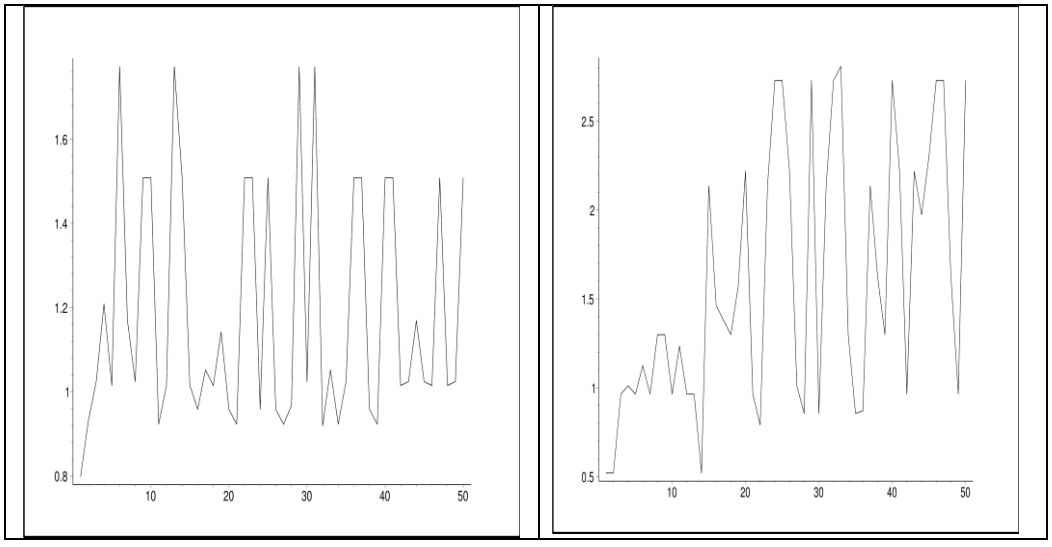
$$\alpha_1 = \alpha_2 = \alpha_9 = \alpha_{21} = \alpha_{22} = \alpha_{29} = \alpha_{31} = \alpha_{32} = \alpha_{37} = \alpha_{39} = 0$$
$$\gamma_2 = \gamma_{20} = \gamma_{22} = \gamma_{30} = \gamma_{32} = \gamma_{35} = 0$$

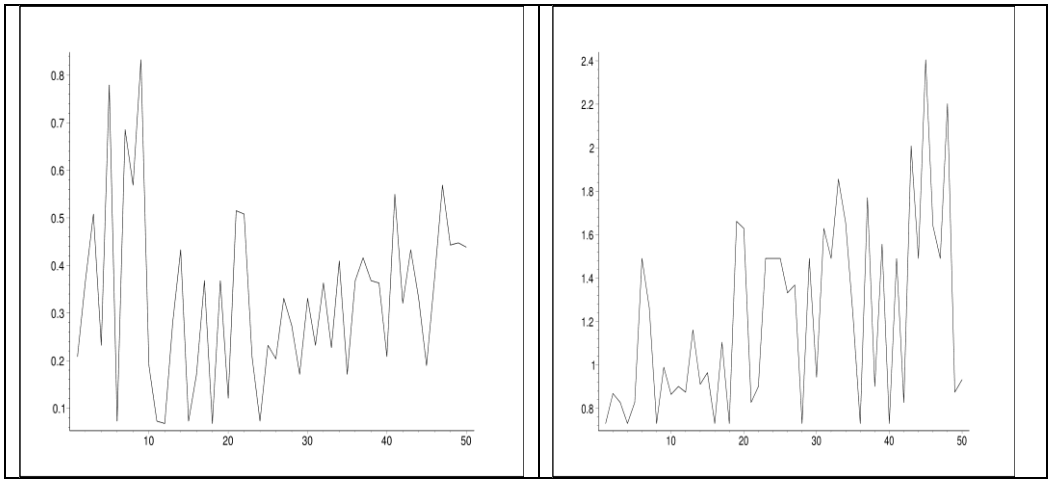
Les valeurs extrêmes sur les sphères unités respectives sont les suivantes

$$\max I_c = 74.24,\ \max I_f = 152.41,\ \max I_s = 83.43,$$

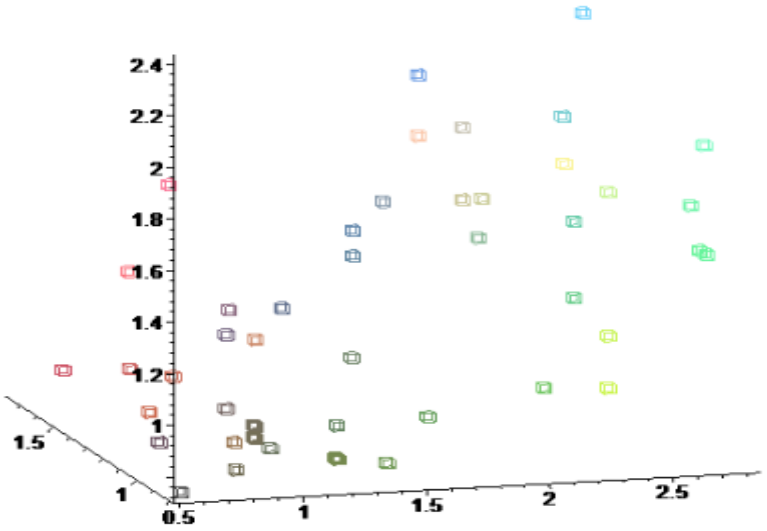
et indiquent quelles quantités « grecques » prendre afin d'absorber le plus possible de l'inertie totale du nuage. Dès lors on peut recalculer numériquement les points $\Psi(i) = \left(\Phi_c(i), \Phi_f(i), \Phi_m(i), \Phi_m(i)\right)$. Si la matrice des covariances des quatre caractères statistiques quantitatifs $\left(\Phi_c, \Phi_f, \Phi_m, \Phi_s\right)$ donne lieu à une analyse Anova pertinente, il n'y a nul lieu de changer le modèle, sinon on ajoute ou supprime certains des facteurs et on recommence sans compliquer à l'excès le modèle.

Les quatre courbes ci-dessous représentent respectivement les quatre fonctions $\left(\Phi_c(i), \Phi_f(i), \Phi_m(i), \Phi_m(i)\right)$ dans l'ordre indiqué, de haut en bas et de gauche à droite, le numéro de l'individu i étant en abscisse. Ces histogrammes rendent plus simple en principe la sélection des individus maximisant chacun des quatre potentiels de ressources, mais montrent aussi une grande volatilité des mesures de ces potentiels au sein de la population d'éco-entrepreneurs sur le territoire concerné.





Le nuage tridimensionnel des points $\left(\Phi_c(i), \Phi_f(i), \Phi_s(i)\right)$ est représenté ci-dessous. Les potentiels $\left(\Phi_c(i), \Phi_f(i), \Phi_s(i)\right)$ varient de 0 à respectivement 2.5, 1.5 et 2.4. Si l'on accorde un poids très important aux deux premières formes de ressources, connaissance et finance, l'image montre qu'un individu se dégage, l'éco-entrepreneur no. 36; on le retrouvera plus loin après l'application de la méthode d'Electre III.



9. La méthode Electre pour la classification des candidats

L'ouvrage de Y. Collette et P. Siarry parait être, pour un lecteur français, l'ouvrage le plus didactique pour maitriser les méthodes multicritères non agrégatives, telles que celles employées dans ce paragraphe.

Le principe est de chercher des candidats maximisant tout à la fois les quatre caractères statistiques $\Phi_c(i)$, $\Phi_f(i)$, $\Phi_m(i)$, $\Phi_s(i)$. Ainsi nous disposons de N=50 actions, chaque action n'étant autre que le choix d'un candidat, et de 4 critères, à savoir les quatre caractères $\Phi_c(i)$, $\Phi_f(i)$, $\Phi_m(i)$, $\Phi_s(i)$.

La méthode Electre I donnera des conclusions mais il advient parfois que celles-ci sont peu utilisables. Ce sera le cas si la matrice d'adjacence qu'on déduit de cette méthode est très chargée, même avec des seuils de concordance et de discordance très fins comme le sont les valeurs suivantes : seuilconc=0.99 et seuildisc=0.01. Il faut là aussi adapter la méthode et faire des essais.

Il peut être opportun d'employer la méthode Electre III qui donnera alors des conclusions plus appropriées. On doit choisir des poids pour les quatre statistiques $\Phi_c(i)$, $\Phi_f(i)$, $\Phi_m(i)$, $\Phi_s(i)$. On introduit et on calcule les matrices de concordance et discordance relatives aux différents critères.

Si i, k sont des indices de 1 à N et ℓ une des quatre lettres c, f, m, s alors

$$C_{ik}(\ell) = \max\Big(\min\Big(\big(S_\ell^q - S_\ell^p\big)^{-1} \cdot \big(\Phi_\ell(k) - \Phi_\ell(i) - S_\ell^p\big), 1\Big), 0\Big)$$

et

$$D_{ik}(\ell) = \max\Big(\min\Big(\big(S_\ell^v - S_\ell^p\big)^{-1}\big(\Phi_\ell(k) - \Phi_\ell(i) - S_\ell^p\big), 1\Big), 0\Big).$$

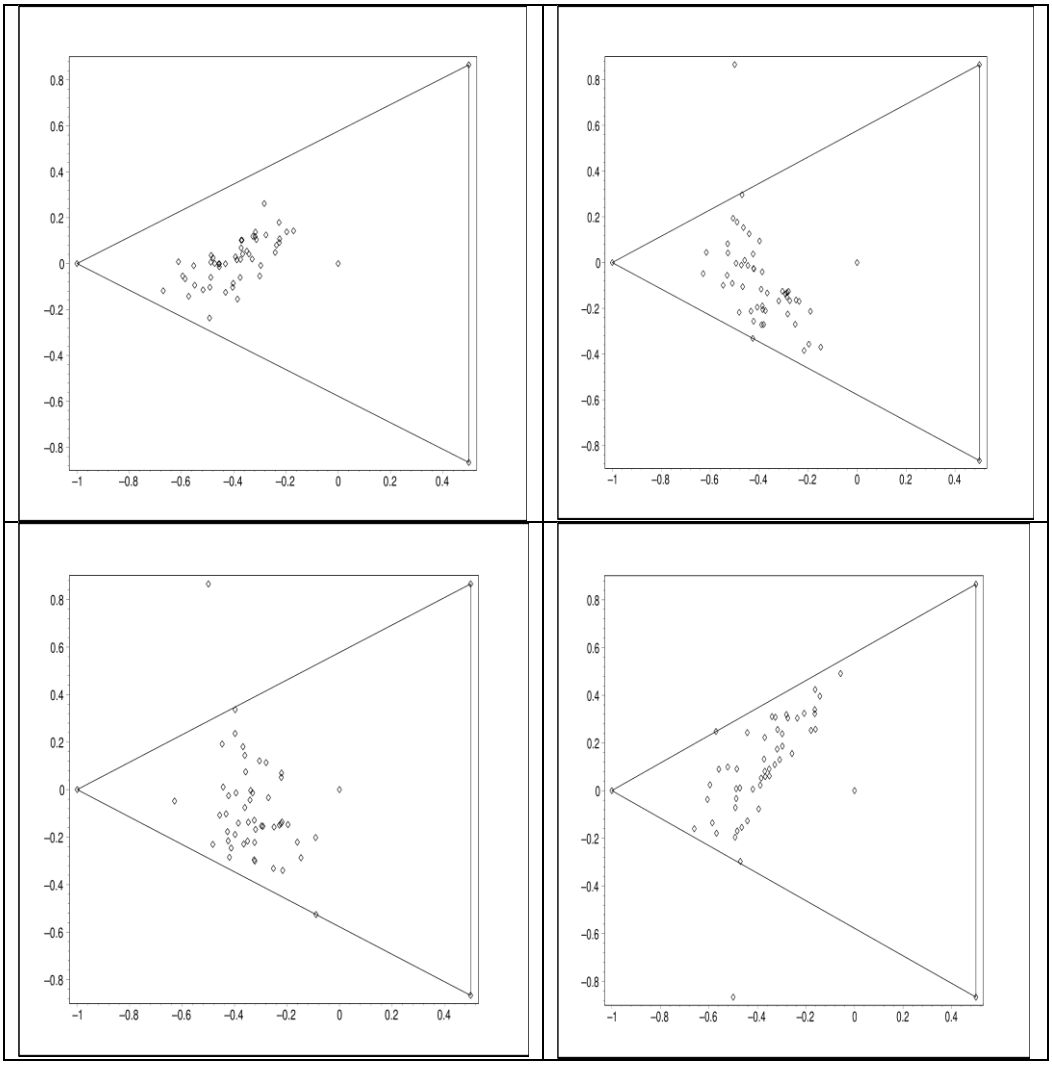
On cherche ensuite, avec certains seuils bien choisis, tous les triplets d'indices (i, k, ℓ) vérifiant $0 < C_{ik}(\ell) < 1$ et tous les triplets d'indices (i, k, ℓ) vérifiant $0 < D_{ik}(\ell) < 1$, ce qui assure que les comparaisons entre candidats ne sont pas triviales. On donne ensuite un seuil de crédibilité égal à 0.15 et on forme les matrices de crédibilité $\big(\gamma_{ik}(\ell)\big)$ par la formule suivante :

$$\gamma_{ik}(\ell) = \big(1 - D_{ik}(\ell)\big)/\big(1 - C_{ik}(\ell)\big) \cdot C_{ik}(\ell) \text{ si } D_{ik}(\ell) > C_{ik}(\ell).$$

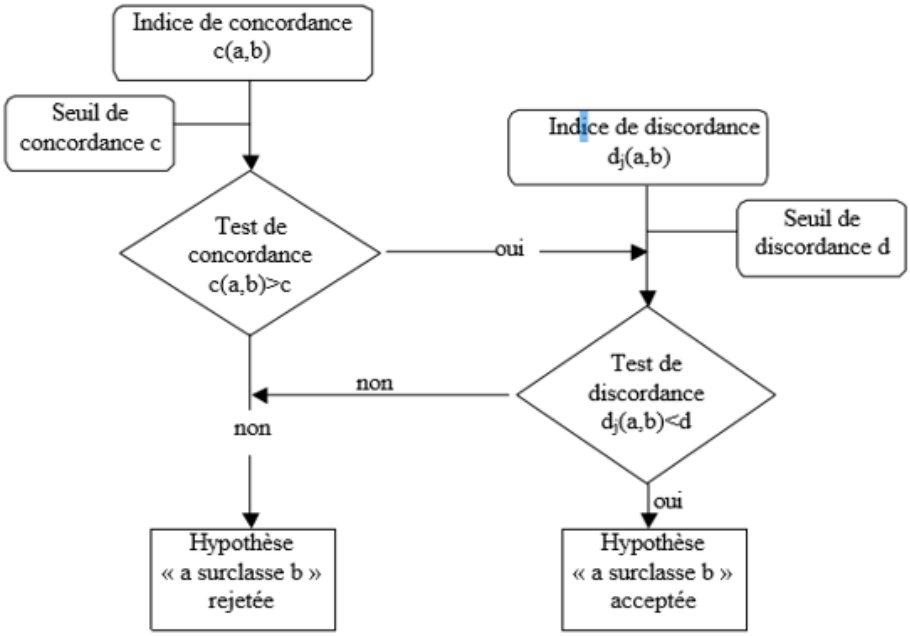
Les entrées de ces matrices sont des réels entre 0 et 2 et on calcule par tâtonnement un seuil de puissance pour qu'à ce seuil très peu de candidats, un ou deux, aient une puissance supérieure au seuil. En ce cas, la décision ne souffre plus d'arbitraire et le caractère discriminant peut évidemment être rejeté.

Lorsque ceci est fini, on peut donner des graphiques triangulaires[1] des caractères statistiques créés. Les individus recherchés sont les plus *centraux* dans les triangles en cela même qu'ils sont plus doués dans certaines combinaisons linéaires des différents potentiels et dans toute perturbation peu importante de celles-ci.

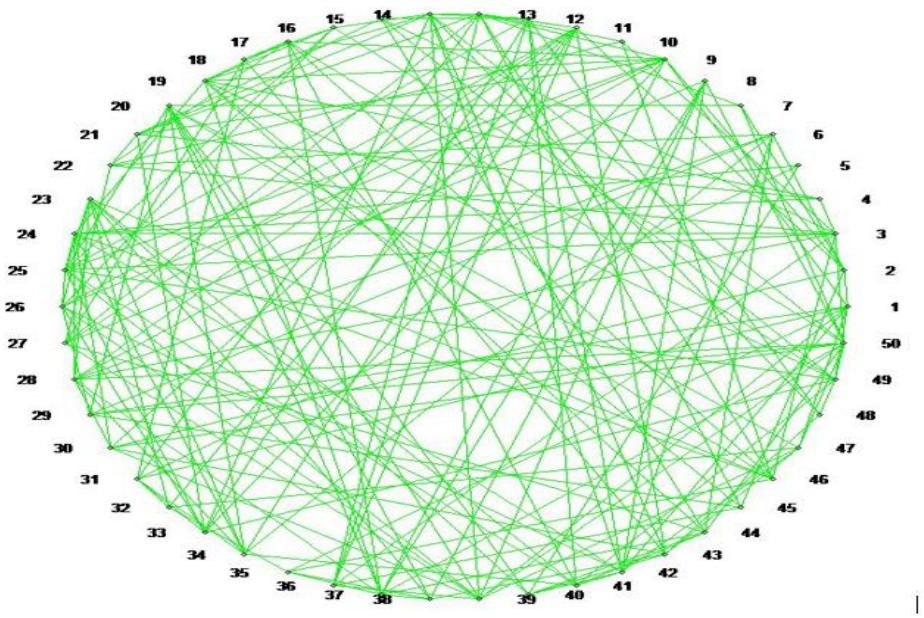
[1] Qu'on permette à l'auteur de faire observer la double analogie plaisante entre ces graphes triangulaires, d'une part, avec ceux employés dans la caractérologie de Marcel Berger, qui repose sur trois « sommets » ou axes principaux de caractère, et d'autre part la représentation des domaines physiques en mécanique quantique, imaginée par Tom Kibbles en 1960.



L’organigramme d’utilisation des matrices de concordance et de discordance est donné ci-dessous et sert de canevas à un programme Maple aisé à écrire, ayant permis toutes les simulations numériques.



Deux alternatives sont déclarées équivalentes lorsqu'elles sont suffisamment concordantes et pas trop discordantes, ceci sous-entendant que des seuils aient été choisis. Il est indispensable de ``jouer" finement avec ces seuils pour obtenir le moins possible d'alternatives équivalentes. Par exemple pour les seuils 0.1 et 0. 9, le graphe indiquant les éco-entrepreneurs équivalents dans des mesures trop larges contient beaucoup trop d'arêtes, et de ce fait ressemble à une clique, voir le graphique ci-après.



On a ajouté 4 entrepreneurs fictifs représentants deux classes d'équivalence dans l'ensemble des alternatives A, afin d'améliorer la lisibilité du graphe. On notera qu'il est possible d'utiliser une autre définition de l'équivalence entre des alternatives (ici donc les éco-entrepreneurs) en employant la distance de Hammings.

On emploie la méthode Electre III qui donne des conclusions plus appropriées. On choisit les valeurs suivantes des poids et seuils

poids=[0.3,0.5,0.2], S^p=[0.9,0.5,0.8], S^q=[2.5,2.6,4.1], S^v =[1.6,2.2,4.1].

On trouve que le test de concordance est réalisé pour 984 paires et le test de discordance rejette 957 autres paires. On donne ensuite un seuil de crédibilité égal à 0.15 et on forme la matrice des crédibilités $\left(\gamma_{ik}(\ell)\right)$. Les entrées de ces matrices sont des réels entre 0 et 2 et on fixe un seuil de puissance égal à 1.5. Quelques comparaisons supplémentaires donnent deux individus ayant été retenus au terme de la méthode Electre III dans cette enquête. Cet ensemble est le *noyau* du graphe, formé des éco-entrepreneurs équivalents pour la définition donnée antérieurement avec les seuils précisés ci-dessus[1]. Il s'agit des éco-entrepreneurs numéros 28 et 36 dont la puissance est supérieure au seuil. Nous donnons ici pour chacun d'eux un vecteur se lisant

[i,puissance(i),faiblesse(i),$\Phi_c(k)$,$\Phi_f(k)$,$\Phi_s(k)$]
[28, 2, 47, .967, 0.855, 0.731] et [36, 2, 47, 1.508, 0.870, .731].

Conclusion et vue d'ensemble

Pour aborder la question de la quantification de choix binaires, incohérents en partie, non hiérarchisés, nous sommes partis de l'outil essentiel qu'est la matrice booléenne formée par chaque décideur, et qui traduit les degrés de compétences des recrues ou bien les caractéristiques binaires des éléments de décision associés à un projet. Cet outil matriciel est d'une grande souplesse d'emploi, car il donne lieu à de belles interprétations combinatoires sur l'ensemble des éventualités parmi lesquelles il faut arbitrer. Mais il permet également d'observer des liens de dépendance stochastique ou non, et donc des classements de booléens parmi des groupes de booléens apparentés et fort dépendants au sein des divers groupes, mais peu dépendants entre les groupes. Dès lors des vecteurs statistiques peuvent être créés en petite dimension, et ceci donne lieu à des nuages de candidats dans des espaces de taille modeste. Cela étant, les méthodes d'optimisation multicritères viennent naturellement à l'appui des questions de tri et de sélection dans les candidatures ou les alternatives présentes.

En ce qui concerne l'agrégation de préordres booléens, l'emploi de fonctions d'utilité doit être *évité a priori pour convertir un préordre en arbre* parce que ces fonctions sont en général de deux sortes. D'une part les fonctions purement topologiques (qui apparaissent dans la preuve de l'existence des fonctions d'utilité dans le théorème de G. Debreu) et d'autre part les combinaisons affines convexes des variables booléennes elles-mêmes. Il est préférable de « passer » aux formules numériques lorsque les différents critères évalués par les différents

[1] Un noyau dans un graphe orienté est un ensemble de sommets deux à deux non-adjacents et tel que de tout sommet émane un arc dont l'autre sommet appartient au noyau.

arbitres ont été soigneusement inspectés, de façon notamment à les assembler dans des groupes de booléens apparentés, ainsi qu'on l'a dit plus haut dans cette conclusion. C'est ainsi que les formules polynômiales qui ont été employées dans la mise en œuvre concrète de la méthode avec pour base l'enquête dunkerquoise, permettent de trouver des fonctions d'utilité bien plus significatives, car elles absorbent de surcroit l'inertie du nuage projeté dans le sous-espace porté par ce groupe de booléens homogènes.

Deux limites à cette méthode sont, d'une part, la bonne connaissance des modèles existants en économétrie et en recherche opérationnelle, mais aussi, d'autre part, la question des choix de maints paramètres apparaissant dans les modèles intermédiaires d'une telle étude. Par ailleurs, le fait que des logiciels comme Sphinx ou les tableurs soient particulièrement incommodes pour mener à bien l'étude peut être un frein dans la diffusion de cette méthodologie. On doit cependant noter que dans des logiciels de programmation courants (Maple, Matlab ou Scilab), la feuille de calcul constitue un fichier de taille modeste. On doit enfin regretter que, face à l'importance des données recueillies dans de tels sondages, et faute de mieux, pour que les décisions soient prises, on puisse être tenté de s'en remettre au vote (qui dénature tout le long travail d'évaluation circonstanciée des alternatives), ou à des méthodes plus frustres d'analyse des liaisons entre données qualitatives ou bien ordinales (qui mettent *à plat* les différents critères et arbitres).

L'auteur tient à la disposition des lecteurs et lectrices intéressés un fichier indispensable pour œuvrer, qui contient notamment la matrice booléenne de l'enquête (§5,§6), ainsi que les commandes et résultats des calculs des statistiques agrégées (§8) et de la méthode Electre (§9).

Vue d'ensemble en langue malgache, dûe à Mme V. Rabemanisa Andriamampandry.
Fijery faobe. *Ity lahatsoratra ity dia mikasika ny fomba ataon'ny orin'asa iray raha hampiditra mpiasa vaovao na haka fanapahan-kevitra amin'ny fampan-drosoana ny orin'asa. Mba ho tontosa soa aman-tsara izany dia hampiasainy ny mombamomban' ireo mpiasa ireo amin'ny alalan'ny toro-marika "marina sa diso".*

Bibliographie

- Attali, J.,(1981) Analyse économique de la vie politique, Quadrige, Puf.
- Boutillier, S., Ryckelynck, Ph. (2017), Sustainable entrepreneurs: quantifying opportunities and social networks. The case study on sustainable entrepreneurs in a heavy industrial area, *International Journal of Entrepreneurship and Small Business*, vol. 31, issue 1, 85-102.
- Collette,Y., Siarry, P., 2002, Optimisation multiobjectif, Eyrolles, Paris.
- Goodman, L.A., (2000), The analysis of cross-classified data: notes on a century of progress in contingency table analysis, and some comments on its prehistory and its future, *Statistics for the 21st Century*, editors: C.R. Rao, G. J. Szekely, 189-231, Marcel Dekker.
- Han, J., Kamber, M. (2006), Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers.
- Lloyd, C.J., (1999), Statistical analysis of categorical data, John Wiley & Sons.
- Saporta, G., (2006), Probabilités, analyse des données et statistique, Technip, Paris.