Les tendances récentes en matière de techniques d'analyse des données : La modélisation est-elle toujours le cadre dominant des analyses de données ?

Dr. Abderrazek ELKHALDI

Université de Sousse

Abstract

Ce papier propose des nouveaux outils d'analyse et d'exploitation des sondages politiques. Il met l'accent sur l'importance de procéder à une lecture en deuxième niveau des données issues des sondages politiques. Il se base sur une simulation inspirée des élections présidentielles française pour mettre en valeur l'étendue des conclusions tirées à travers une analyse en profondeur des données du sondage politique.

Mots clé : Sondage politique, analyse factorielle, carte perceptuelle.

Remerciement : *L'auteur tient à remercier l'équipe dirigeante de l'université Européenne de Tunis pour tout le soutien dont il a bénéficié pour mettre à terme le présent travail. Un remerciement spécial au comité d'organisation du SYMPOSIUM portant sur : Les Nouvelles Tendances et Méthodes de Sondage Politique, tenu le Mercredi 20 Décembre 2017 au campus de l'Université Européenne de Tunis*

1. Introduction

L'histoire de sondage politique remonte au début du XIX^{ème} siècle dans le cadre des « straw votes » ou votes de paille. En effet, aux Etats-Unis, des simulations furent entreprises par les journaux auprès de leur clientèle pour statuer sur leurs intentions de vote. Les modalités de ses sondages furent différentes, mais la plus répondue prenait la forme d'un coupon découpé du journal et récupéré une fois rempli ou tout simplement des journalistes qui interrogeaient des passants dans la rue. Toutefois, ces techniques manquaient de beaucoup de rigueur et souffraient d'un énorme problème de représentativité en dépit des efforts des journaux pour augmenter le nombre des participants.

Au cours de la première moitié du XX^{ème} siècle, Roper publie une enquête basée sur 3000 Américains à propos de leur attitude à l'égard de plusieurs sujets d'actualité de l'époque. Cette innovation a conduit Gallup, un agent publiciste à fonder l'American Institute of Public Opinion. Les élections présidentielles de 1936 étaient une bonne occasion pour s'imposer sur ce nouveau domaine d'activité. Ainsi, s'appuyant sur un échantillon de 4000 citoyens a prévu la victoire de Roosevelt sur London alors que plusieurs autres nommant la revue Literary Digest avec un vote de paille portant sur 10 millions de personnes, prévoyaient la victoire de London. A la surprise de tout le monde, Roosevelt l'emportait sur London avec 62% (Gallup avait prévu une victoire à 56%). Ce premier exercice, marqua la date de naissance de la pratique de sondage politique contemporain. En 1948, la pratique de sondage a connu une crise après une victoire inattendue de Truman sur Dewey. Cet incident a conduit à l'abandon de la méthode d'échantillonnage stratifié en faveur d'autres techniques notamment l'échantillonnage aléatoire.

En France, la pratique de sondage politique n'a pas tardé avec la fondation en 1938 de l'Institut Français d'Opinion Publique par Jean Stoetzel. Plusieurs exercices ont été menés par l'IFOP plus particulièrement un sondage sur les accords de Munich concordant la décision de la chambre des députés. Malgré ces succès, plusieurs critiques ont été adressées à la pratique du sondage dans le domaine politique conduisant plusieurs instances nationales et internationales à réglementer sa pratique et à renforcer le contrôle de sa mise en œuvre. D'un coté académique, le choix des méthodes d'échantillonnage a souvent été sujet à controverse entre les chercheurs.

L'objet de ce papier est de décrire la pratique des sondages dans le domaine politique et d'avancer la possibilité de combiner les techniques de modélisation et les techniques géométriques comme l'analyse des correspondances, les cartes perceptuelles ou l'analyse discriminante. Dans ces conditions, la seconde section sera dédiée à l'étude de la genèse de la pratique de sondage dans le domaine politique, la troisième section avancera les conditions de réussite d'un sondage politique. La quatrième section rapportera quelques bugs de sondage politique. La sixième section proposera des nouveaux outils applicables pour mieux exploiter les sondages politiques. Enfin la septième section fournira une conclusion.

2. Genèse de la pratique des sondages politiques

Avant d'aborder la méthode des sondages, intéressons-nous un peu à leur genèse. Les sondages politiques naissent vraiment en 1936, à l'occasion de l'élection présidentielle américaine. Roosevelt, président depuis 1932, se présente pour un second mandat contre London.

La presse américaine affirme que London va gagner, après avoir eu recours, comme à l'habitude depuis le début du 19ème siècle, à ce que l'on appelle les «straw votes» (votes de paille). Cette technique consiste, pour les journaux, à demander à leurs lecteurs de renvoyer un coupon mentionnant leur choix. Ainsi, le Literary Digest reçoit près de 2 millions de réponses qui pronostiquent la victoire de London. De son côté, Georges Gallup, qui vient de créer en 1935 l'un des premiers instituts de sondages, pronostique la victoire de Roosevelt, en interrogeant un échantillon représentatif de 4.000 personnes seulement. L'élection lui donne raison.

Dès 1938, Jean Stoetzel, créateur de l'IFOP (Institut Français de l'Opinion Publique) importe cette technique en France et la désigne par le mot «sondage», à consonance scientifique. Mais c'est en 1965 seulement que les instituts de sondage réalisent pour la première fois une estimation de vote le soir de l'élection présidentielle. Les Français découvrent à 20 heures, que le Général de Gaulle est mis en ballottage par François Mitterrand.

Cette date marque l'entrée en force des sondages dans le paysage politique français. La nouveauté cette année est que ces sondages vont pouvoir être publiés, y compris dans la semaine précédant les élections (jusqu'au vendredi minuit). En effet, la loi de 1977 qui interdisait toute publication pendant la semaine précédant les élections a été mise à mal par les nouvelles technologies de la communication. On se rappelle tous du sondage sur le second tour des élections présidentielles de 1995 publié dans la semaine de l'élection sur le site internet de «La tribune de Genève». De même, les sites de la presse française ont placé, lors des législatives de 1997, des liens vers des sites étrangers présentant les derniers sondages pré-électoraux. Dans un arrêt de septembre 2001, la Cour de Cassation a jugé que la loi de 1977 était contraire à la Convention européenne des droits de l'Homme, au nom du droit à l'information. Le Parlement en a tiré les conclusions, en légalisant, le 7 février dernier la publication des sondages jusqu'au vendredi minuit précédant l'élection.

3. Les conditions pour réussir un sondage politique

Malgré les critiques adressées aux cabinets de sondage politique, il demeure néanmoins une part de vérité dans les exercices qu'ils entreprennent. En effet, le respect d'un certain nombre des conditions fondamentales de mise en œuvre d'un sondage politique en bon et due forme ne semble pas être une illusion. Des conditions telles que le choix de la méthode d'échantillonnage, les modalités d'administration de l'enquête ou de collecte des données, la maitrise de la marge d'erreur, l'ajustement des résultats et les anticipations auto-réalisatrices, sont autant des facteurs qui conditionnent la réussite d'un sondage politique.

S'agissant de l'échantillon, l'expérience a montré que la meilleure technique d'échantillonnage est sans doute l'échantillonnage probabiliste, dans le sens où chaque individu a la même chance de faire partie de l'échantillon qu'il soit jeune ou vieux, instruit ou illettré, aisé ou pauvre. Ceci s'explique par le fait que le jour du scrutin, personne ne peut prévoir ceux qui vont participer de ceux qui ne manifesteront pas d'intérêt pour voter. Cette suprématie de la méthode probabiliste a permis de rompre avec la méthode d'échantillonnage stratifié qui, pendant plusieurs années était considérée comme la plus scientifique. Cependant, la mise en œuvre de la technique probabiliste exige un certain nombre de critères restrictifs dont l'expérience a montré que leur respect est quelque peu difficile. Il s'agit tout d'abord de s'assurer que tous les individus composants l'échantillon seront susceptibles de répondre. De même, leurs réponses doivent correspondre à leur opinion. Par ailleurs, il faudrait que le sondeur dispose des coordonnées de tous les électeurs pour maximiser la chance de joindre tous les individus composant l'échantillon. Il est évident que dans le cadre d'un sondage national, le respect de ces critères se heurte à plusieurs contraintes liées aux coûts, à la disponibilité de l'information et à la qualité des membres de m'échantillon.

En ce qui concerne les modalités de collecte des données, elles reposaient auparavant sur les entretiens face à face avec les personnes ciblées. Mais cette méthode a été abandonnée en raison de ses coûts suffisamment élevés laissant la place aux interviews téléphoniques moins chers et mettant plus à l'aise les répondants. Ainsi, les premières expériences ont révélé un taux élevé des réponses qui peut être expliqué par une auto-réalisation du répondant (le téléphone des années 60 était un luxe et faire partie de l'échantillon était un privilège voire une distinction). A cette époque, on parlait d'un taux de réponse de l'ordre de 80%. Toutefois, on ne se retrouvait pas dans un cadre d'échantillonnage probabiliste car on excluait les personnes ne disposant pas chez elles de téléphone, ce qui constituait un véritable biais. De même, le nombre des répondants a considérablement chuté on parle aujourd'hui d'un taux de réponse de l'ordre de 20%. Avec l'évolution de la technologie, on a développé une nouvelle technique de sondage encore moins coûteuse s'appuyant sur les interviews par boite vocale interactive où le répondant était appelée à appuyer sur les boutons de son appareil pour exprimer un choix. Les premières expériences furent prometteuses, mais progressivement, cette méthode a eu le même sort que l'entretien téléphonique (latitude des répondants et impossibilité d'aboutissement final de l'entretien). Ces problèmes ont incité les sondeurs à réfléchir sur des méthodes plus souples et moins coûteuses administrées via internet. Ainsi, les cabinets de sondage ont aujourd'hui tendance à cibler des individus d'âge et de catégorie socioprofessionnelle différentes qu'ils invitent périodiquement à répondre à des sondages moyennant un système de gratification (cadeaux, bon d'achat, tickets de match…). Cette nouvelle technique a donné un nouvel élan à l'activité de sondage. Cependant, elle souffre tout de même de plusieurs insuffisances

notamment l'homogénéité de l'échantillon, la subjectivité des réponses (on répond pour bénéficier d'une gratification), impossibilité de vérifier si le répondant est réellement propriétaire du compte (demander à un ami de le répondre à sa place alors qu'il est indisponible). Par ailleurs, rien ne garantit que la personne ait dit la vérité quant à ses données personnelles (Age, genre, profession, fumeur/non-fumeur…) de même, avec cette méthode on n'est plus dans une logique d'échantillonnage probabiliste (les gens n'ont pas tous des ordinateurs ou des connexions internet).

Ces différents biais incontournables sont à l'origine d'un troisième souci auquel devraient penser les cabinets de sondage. Il s'agit de la marge d'erreur ; cet élément fondamental qui témoigne de la crédibilité et qui conditionne la pérennité du sondeur. La marge d'erreur est le verdict qui tombe le jour du scrutin et qui donne de l'élan pour le cabinet « vainqueur » sur ses concurrents. Ainsi, les raisons de coûts poussent plusieurs sondeurs à limiter la taille de leur échantillon ou à joindre les personnes les plus faciles à contacter ou même à privilégier une technique d'administration sur une autre. Ceci conduit inéluctablement à éloigner les prévisions des résultats effectifs. La plupart des chercheurs expliquent l'occurrence des erreurs :

- ✓ à la composition de l'échantillon
- ✓ à la volatilité électorale
- ✓ au timing du sondage
- ✓ à la mobilité des électeurs
- ✓ à l'ampleur des indécis
- ✓ à la submergence du vote utile
- ✓ à la quasi-disparition du clivage gauche-droite (notamment en France)

Une dernière critique adressée au métier du sondage est relative à un phénomène purement comportemental en l'occurrence les anticipations auto-réalisatrices. Concept couramment utilisé en finance pour expliquer le comportement des investisseurs face à un dysfonctionnement probable du marché. Ainsi, une rumeur d'une éventuelle faillite d'une entreprise inciterait les investisseurs à abandonner l'action en question et pourrait conduire à sa faillite même si la réalité fait que l'entreprise est en bonne santé financière. Parallèlement, les résultats d'un sondage favorisant un candidat sur un autre pourrait conduire à la victoire du premier. Ainsi, le sondage est accusé parfois de faire le marketing pour un parti contre un autre et généralement pour le bon payant. Mais les cabinets de sondage continuent toujours à nier cette pratique et à prouver la scientificité de leur pratique.

4. Les bugs de sondage

L'histoire de la pratique de sondage nous révèle plusieurs échecs connus par les cabinets du métier. A titre d'illustration, en 2001, lors des élections italiennes, les prévisions montraient la dominance de Giuliano Amato sur son rival Silvio Berlusconi. Ces derniers, l'emportait contrant toutes les pronostiques. En 2002, lors des élections présidentielles françaises, on a assisté à deux résultats inattendus avec l'élimination de Lionel Jospin et la montée en puissance de Jean-Marie Le Pen. Plus récemment, les sondeurs britanniques ont raté leur gloire lorsqu'ils prévoyaient un « Non » pour le référendum du Brexit et encore plus récent, la victoire inattendue de Trump sur Clinton vient confirmer les problèmes dont souffrent les sondeurs à travers le monde.

L'une des explications apportées à ce problème est celle avancée par Gault qui affirme que le centre de gravité de la société n'est plus autour de groupes d'opinions, mais plutôt centrée vers l'individu. Or ce dernier a tendance à changer rapidement de camp au gré des évènements de la dernière minute. De ce fait, il devient de plus en plus difficile de regrouper les gens et de se trouver des valeurs partagées, notamment en matière de politique.

5. Faut-il changer de techniques de sondage ?

Les changements technologiques et sociologiques imposent une révision des méthodes jusqu'à présent adoptées pour mettre en œuvre un sondage politique. Ainsi, poser des questions directes pour mesurer les chances d'un candidat sur un autre n'est plus approprié. Il faudrait donc réfléchir sur des nouveaux outils favorisant l'agrégation des données individuelles, la prise en compte des variables médiatrices, modératrices et d'attributs individuels. Les politiciens américains l'ont compris beaucoup plus tôt et ont souvent mis l'accent sur l'importance des éléments relatifs à leur vie privée et familiale et à l'image que devrait porter un citoyen à son candidat préféré.

Dans cette perspective, et considérant les évolutions spectaculaires en matière de data mining et progiciels d'analyse des données, des nouveaux outils peuvent être employés pour mieux faire parler les chiffres et les données collectés par les études et les sondages politiques. des outils tels que l'analyse factorielle avec ses différentes variétés (analyse de correspondance, analyse en composante principale…), carte perceptuelle, analyse discriminante etc…

5.1 Utilisation de l'analyse factorielle

Malgré quelle remonte au début du XX^{ème} siècle, l'analyse factorielle n'a pas été suffisamment employée pour analyse les résultats des sondages d'opinion. Ceci pourrait être expliqué par l'intérêt que portent les politiciens et les donneurs

d'ordre des sondages à la lecture directe des chiffres. Or, cette lecture n'est valable que sur le très court terme. Alternativement, l'utilisation des l'analyse factorielle permettrait de tirer des conclusions qui pourraient être valables sur le long terme. Spearman et Benzécri furent parmi les premier à utiliser la technique de l'Analyse factorielle. L'utilisation fut élémentaire, mais grâce à l'introduction de l'outil informatique, ses conditions et ses modalités d'exploitation furent élargies. Les principaux objectifs de l'analyse factorielle visent la restructuration des données et leur représentation sur un plan factoriel en vue de faciliter la lecture et l'interprétations des données plus compliquées. Par ailleurs, et D'un point de vue théorique, l'analyse rejoint un aspect psychométrique de mesure de concepts non observables. En effet, on essaie de passer des mesures directes des résultats de l'enquête vers la visualisation des variables initialement latentes.

Sur le plan pratique, il s'agit de partir d'un ensemble d'items sensés représenter le même phénomène, tels que les attributs personnels d'un homme politique. Il s'agit d'une batterie d'items déduits sur la base d'une synthèse théorique ou d'une analyse qualitatives menées auprès d'une population électorale. Les données collectées à cette occasion peuvent être synthétisées dans un ou plusieurs axes factoriels. Le même exercice pourrait être mené sur un nombre d'items reflétant les attributs professionnels d'un homme politique. L'objectif étant de statuer sur les principaux items résumant le phénomène étudié. Ceci est d'une utilité cruciale pour le donneur d'ordre car elle permet d'orienter ses actions futures et de fixer les points essentiels de son programme électoral. Le résultat d'une telle analyse pourrait être visualisé comme suit (Exemple fictif inspiré des élections présidentielles françaises en 2017) :
Pour les items distinguant les attributs personnels d'un homme politique on distingue :

Tableau 1 : Qualité de représentation des variables personnelles

	Initial	Extraction
VIE PRIVEE	1,000	,962
LOOK	1,000	,974
MEDIA IMPACT	1,000	,971
SUPPORT FAMILIAL	1,000	,967

Méthode d'extraction : Analyse en composantes principales.

Les résultats du tableau 1 montrent que tous les items sont très bien représentés dans le modèle (Communalité supérieure à 0,5). Reste à vérifier si tous les items restituent ou pas le même phénomène en l'occurrence les attributs personnels d'un candidat. La réponse à cette question est fournie par le tableau 2.

Tableau 2 : Variance totale expliquée par les variables personnelles

Composante	Valeurs propres initiales			Extraction Sommes des carrés des facteurs retenus		
	Total	% de la variance	% cumulés	Total	% de la variance	% cumulés
1	3,875	96,876	96,876	3,875	96,876	96,876
2	,062	1,547	98,423			
3	,034	,849	99,272			
4	,029	,728	100,000			

Méthode d'extraction : Analyse en composantes principales.

D'après le tableau 2, les quatre items restituent un même axe et permettent d'expliquer 96% de la variance totale expliquée.

Le même travail devrait être entrepris pour les items relatifs aux attributs professionnels. les tableau 3 et 4 confirment l'unidimensionnalité de la variable attributs professionnels avec 92% de variance restituées par les items considérés.

Tableau 3 : Qualité de représentation des variables professionnelles

	Initial	Extraction
CONSEILLERS	1,000	,942
PROGRAMME ELECTORAL	1,000	,904
ANTECEDENTS	1,000	,921
LEADERSHIP	1,000	,937
CONFIANCE	1,000	,898

Méthode d'extraction : Analyse en composantes principales.

Tableau 4 : Variance totale expliquée par les variables professionnelles

Composante	Valeurs propres initiales			Extraction Sommes des carrés des facteurs retenus		
	Total	% de la variance	% cumulés	Total	% de la variance	% cumulés
1	4,602	92,046	92,046	4,602	92,046	92,046
2	,168	3,363	95,410			
3	,088	1,766	97,176			
4	,074	1,477	98,653			
5	,067	1,347	100,000			

Méthode d'extraction : Analyse en composantes principales.

5.2 Utilisation de la carte perceptuelle

La carte perceptuelle a été initialement utilisée dans le domaine du marketing pour étudier le positionnement perceptuel des différentes marques sur un produit quelconque. Le mérite de cette méthode est de prendre en considération la perception du client sur un produit selon l'origine et les conditions de consommation. L'application dans le domaine de la politique serait d'un apport incontestable car le processus de vote s'appuie inéluctablement sur l'appréciation que porte un électeur sur les candidats potentiels.

La mise en œuvre de cette technique exige au préalable la réalisation d'une étude qualitative visant à statuer sur les déterminants les plus importants qui amènent un électeur à voter pour un candidat plutôt qu'un autre.

Nous procédons dans le cadre de ce travail à une simulation sur les élections présidentielles françaises de 2017.

L'analyse qualitative a permis de distinguer deux familles des critères :

- Les attributs personnels du candidat ;
- Les attributs Professionnels du candidat.

Le tableau 5 rapporte les critères obtenus sur la base d'une analyse qualitative auprès de 12 électeurs français (seuil de saturation).

Tableau 5 : Attributs Personnels et professionnels du candidat

Attributs Personnels	Attributs Professionnels
VIE PRIVEE	CONSEILLERS
LOOK	PROGRAMME ELECTORAL
MEDIA IMPACT	ANTECEDENTS
SUPPORT FAMILIAL	LEADERSHIP

Ces attributs ont fait l'objet d'une analyse en composante principale en vue de les regrouper en deux axes factoriels. Chemin faisant, il convient de procéder à la technique d'agrégation des données qui vise à ramener le nombre des observations de 4000 à seulement sept observations pour chacun des sept candidats et pour les deux catégories d'attributs. Le mérité de l'agrégation des variables réside dans la possibilité de réduire le nombre d'observations par candidat, sans pour autant réduire le contenu informationnel restitué par les scores calculés. Les résultats sont rapportés par le tableau 6.

Tableau 6 : Agrégation des variables

Candidat	Attributs personnels	Attributs professionnels
MELENCHON	-0,08	-0,12
MACRON	0,68	0,69
LASSALLE	-0,63	-0,64
FILLON	0,21	0,34
MARINE LEPEN	0,57	0,51
DUPONT-AIGNAN	-0,53	-0,52
HAMON	-0,21	-0,25

Les scores calculés seront finalement rapportés sur un plan factoriel visant à positionner les sept candidats potentiels sur une carte reflétant la perception des électeurs français sur ces derniers. Ce positionnement pourrait être représenté comme suit :

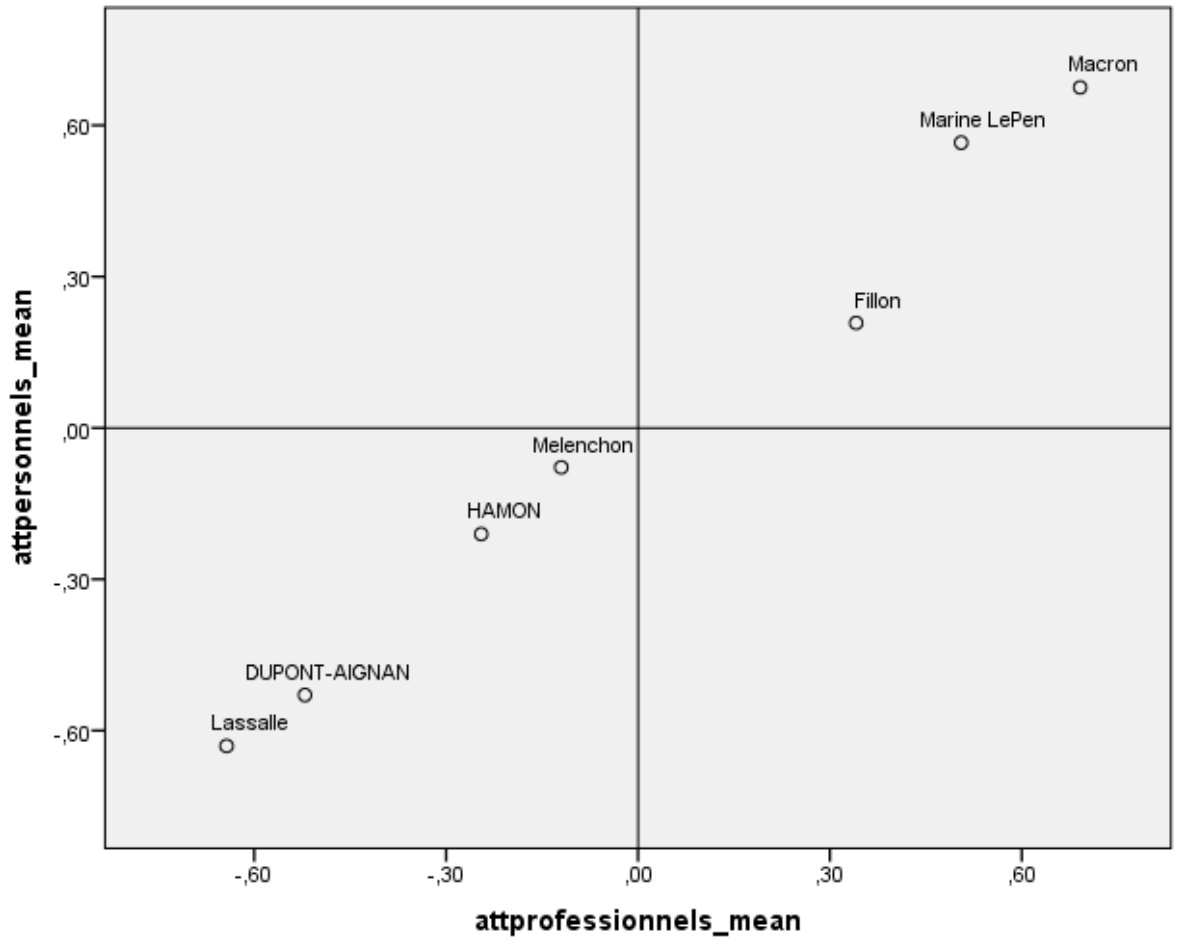

Figure 1 : Positionnement des sept candidats en fonction des attributs personnels et professionnel.

Les résultats de la carte perceptuelle obtenue pourraient être interprétés comme suit :

Le candidat le mieux perçu par les électeurs français est Macron suivi par LePen et un peu plus loin par Fillon. Ces trois candidats se situent dans le cadrant positif relatif aux attributs professionnels. Cependant les quatre autres candidats sont positionnés sur le cadrant négatif pour les deux attributs professionnels et personnels. Ce résultat pourrait être interprété comme étant un échec de communication de la part des quatre candidats sur leurs qualités professionnelles et personnelles. Ça pourrait être une forme de recommandation pour les candidats afin de soigner leurs politiques de communication pour des campagnes électorales futures.

6. Conclusion

L'objet de ce papier était de dresser un état de lieu sur la pratique des sondages dans le domaine de la politique. L'observation a permis de constater que les cabinets spécialisés se contentent d'interpréter les chiffres bruts et tirent leurs conclusions à partir d'une lecture superficielle des résultats obtenus. Nous soulignons par ailleurs, l'absence d'une vision prospective et de tout effort d'apprentissage des expériences passées. Ainsi, l'obsession majeure desdits cabinets est de réussir le pari des pronostiques annoncés. Dans le cadre de ce travail, nous avons montré que les résultats d'un sondage peuvent ouvrir les horizons sur des nouvelles actions à mettre en œuvre par les candidats afin de maintenir une position favorable ou corriger une insuffisance.

A cet effet, une lecture en second degré s'impose à travers les techniques d'agrégation des variables individuelles à l'instar de l'analyse factorielle ou de la carte perceptuelle.

Considérant un exemple fictif tiré des élections présidentielles françaises de 2017, nous avons pu montrer que l'utilisation des techniques précitée serait d'un apport incontestable et constituerait un point de départ pour mieux expliquer le déroulement du sondage politique.

Références bibliographiques

Bartholomew D.J. (2011), Spearman and the origin and development of factor analysis, British Journal of Mathematical and Statistical Psychology, Volume 48, Issue 2, pp 211-221.

Boudon, R., Bourricaud, F., Girard, A. (1981), « Science et théorie de l'opinion publique, hommage à Jean Stoetzel », Retz, *Actualités des sciences humaines*, Paris, 1981.

Blondiaux (1998), La Fabrique de l'opinion. Une histoire sociale des sondages, Le Seuil, 1998.
Bryce, J. (1911) The American Commonwealth, McMillan, 1911.

Gaïti, B. (2004)« L'opinion publique dans l'histoire politique : impasses et bifurcations », Le Mouvement Social, n° 221, 2004.

Blumer, H. (1948), « Public Opinion and Public Opinion Polling », American Sociological Review, vol. 13, n°5, 1948, p. 542-554.

Fishkin, J. (1997), The Voice of the People. Public Opinion and Democracy, Yale University Press, 1997.