Franck Gintrand est directeur général de Global conseil corporate après avoir été directeur conseil chez Euro-Rscg C&O et consultant chez Bernard Krief.

Décryptage

Aux limites du Big data prédictif

Cela aurait pu être une révolution aux dernières présidentielles françaises. Ce ne fut qu'une confirmation : l'analyse du big data ne permet pas de disposer d'une photographie plus claire de l'opinion et encore moins de prédire ce que cette opinion sera demain.

A chaque élection, c'est la même fébrilité dans les bureaux des instituts de sondage. Les résultats viendront-ils confirmer ou infirmer les chiffres publiés à la veille du scrutin ? Les sondologues pourront-ils commenter les résultats avec cette grave assurance qui caractérisent les experts sûrs de leur fait ou devront-ils commencer par se justifier et se défendre face aux politiques qui, dans ces cas-là, se font un malin plaisir de rappeler combien les sondages doivent être pris avec des pincettes ? La question hante tous les esprits depuis 2002. Pourtant, cette année-là, les présidentielles françaises semblaient jouées d'avance. L'unique incertitude portait sur l'ordre d'arrivée de Lionel Jospin et de Jacques Chirac au premier tour, le candidat socialiste étant largement donné gagnant au second tour. C'est dire la stupeur des Français le soir du 21 avril lorsque les écrans affichèrent le visage du président du Front national aux côtés de celui de Jacques Chirac. Lionel Jospin ne s'en est jamais relevé. Les experts des sondages, eux, y ont perdu beaucoup de leur superbe. D'autant que le 21 avril n'a fait qu'ouvrir une longue série de déconvenues. De ce point de vue, 2015 restera une « annus horribilis ». Législatives israéliennes, législatives turques, élections générales britanniques, référendum grec... Rien ne se sera passé comme le prévoyaient les sondages. Et 2016 n'aura pas été plus faste. Là encore, les instituts se seront vus reprochés de n'avoir ni vu venir le Brexit, ni anticipé la victoire de Donald Trump quand, dans le même temps, des entreprises revendiquant d'autres méthodes parvenaient, elles, à pronostiquer le bon résultat.

BIG DATA CONTRE MÉTHODE DES QUOTAS

En France, la société Filteris envisageait ainsi dès octobre 2015 la victoire de François Fillon tandis que les instituts le plaçaient deux jours avant le scrutin derrière Alain Juppé et Nicolas Sarkozy. De quoi donner des ailes à Filteris qui dès la proclamation de la victoire de François Fillon s'empressait de proclamer la victoire de « l'analyse Big data » sur la méthode des quotas avant de se lancer dans le suivi des présidentielles. Dans son sillage, Vigiglobe, une startup française dirigée par un ancien directeur de TNS et Predict The President, un programme initié et animé par cinq étudiants de l'école Télécom Paris Tech pour l'hebdomadaire le Point. Alors que les instituts laissaient entrevoir la probable sélection d'Emmanuel Macron et de Marine Le Pen au second tour, ces trois acteurs du big data pariaient sur un duel de François Fillon / Marine Le Pen. A la veille du premier tour, Filteris reprenait même sur Twitter l'interrogation formulée par le quotidien La Tribune : "le big data donne Fillon au second tour, pas les sondages. Qui aura raison ?" On connait depuis la réponse. François Fillon éliminé, le big data a été officiellement mis K.O. par la méthode des quotas.

OPINION ASSISTÉE CONTRE OPINION VOLONTAIRE

Le séisme n'aura donc pas eu lieu. Les présidents d'instituts auront tremblé, retenu leur souffle avant de respirer. Après avoir justifié leur échec par un revirement de dernière minute de l'opinion, les partisans du Big data se font aujourd'hui plus discrets. Mais le débat n'est pas clos. Critiqués pour avoir longtemps cultivé l'ambiguïté sur le caractère prédictif de leur approche, les instituts font désormais preuve d'une modestie à toute épreuve. Par prudence, tous mettent désormais un point d'honneur à rappeler combien les sondages ne sont qu'une « photographie de l'opinion » et, en aucun cas, une prévision des résultats définitifs. Pour les acteurs du big data, ce point de vue n'est pas acceptable. Dans la mesure où le sondage ne reflète qu'une opinion sollicitée et volontaire, cette photographie serait selon eux profondément biaisée. Par ailleurs, elle serait dans l'incapacité de rendre compte de la dynamique de l'opinion et de son point ultime de cristallisation. Par rapport à cette double limite, le big data présenterait l'avantage de porter sur le traitement de données non sollicitées et sur l'attention portée aux signaux faibles. Quand les sondages se contenteraient de prendre une série de clichés, le big data permettrait d'anticiper la fin du film électoral.

COMPRÉHENSION DU PRÉSENT CONTRE ANTICIPATION DU FUTUR

Le débat entre big data et méthode des quotas ne concernerait que le suivi des élections, sa portée n'aurait qu'un intérêt mineur. Mais les divergences de méthode dépassent largement la sphère politique. Au-delà des sondages, c'est le marché autrement plus rémunérateur des entreprises qui est visé. Les outils du marketing traditionnel que sont les panels ont l'inconvénient trop souvent de ne refléter que les attentes présentes des consommateurs. Henry Ford s'amusait à dire que « si (il) avais(t) demandé au consommateur ce qu'il voulait lorsque (il a) conçu la Ford T, il ne fait aucun doute qu'il (lui) aurait répondu « un cheval plus rapide » ». Soit. Mais la banalité de la réponse ne tiendrait-elle pas surtout à celle de la question ? L'analyse prédictive du big data pourrait-elle se révéler plus performante en relevant des corrélations dont l'opinion n'est pas (encore) consciente ? L'outil permettrait ainsi de lever les réticences qui pèsent à l'égard de la stratégie Océan bleu dans la mesure où, la plupart du temps, l'absence de choix stratégique clair – dénoncée par les concepteurs de cette stratégie – constitue la meilleure, sinon l'unique façon de limiter les risques.

La question centrale de l'évaluation du risque

Le débat sur les élections ne fait en réalité qu'essayer de répondre à cette question : le big data prédictif peut-il aider les entreprises à surmonter et dépasser ces stratégies de non-choix en limitant au maximum les risques associés aux changements majeurs ou aux décisions radicales ? Que des sociétés réussissent à prédire le chiffre d'affaire généré par la sortie d'un film pendant une année, juste en analysant les réactions des in-

ternautes au moment de la sortie en salle, est une chose. Qu'il puisse y avoir une méthode permettant de parvenir au même résultat avant la sortie en salle, avant même le tournage, en est évidemment une autre. Filteris ou Vigiglobe, fortes de pronostics qui se sont vérifiés dans les faits sur le Brexit, la victoire de Trump ou encore le résultat des primaires de gauche et de droite, sont convaincues que cet objectif n'a rien d'illusoire. Pour se livrer à ses prévisions dans le dernier tournant de l'élection, Filteris déclare avoir principalement utilisé le « social media monitoring » qui consiste à recueillir les opinions exprimées à grande échelle sur les réseaux sociaux et à les rediriger sur des algorithmes chargés de distribuer des scores en termes de « popularité » et de « sentimentalité ». De son côté, Vigiglobe dit utiliser des algorithmes de « machine learning » capable d'analyser quantitativement mais surtout qualitativement les messages postés sur les réseaux sociaux. Mais quel que soit la méthode revendiquée et les ingrédients du big data, la recette consiste pour l'essentiel en une veille des réseaux sociaux – en l'occurrence Twitter et Facebook –, à savoir un recueil et un traitement d'opinions à grande échelle.

LES LIMITES DU BIG DATA PRÉDICTIF

Les instituts ont bien entendu critiqués l'absence de représentativité de ces réseaux. Seulement 10% de la population serait inscrite sur Twitter principalement des jeunes urbains surdiplômés et Facebook ne concernerait quant à lui qu'un Français sur deux, ce que les animateurs de Predict The President ont admis. Ce biais s'accentuerait même durant les élections en fonction du cours de la campagne. On sait que pour ces présidentielles, les sympathisants de François Fillon ont fait preuve d'un activisme exceptionnel pour défendre leur champion face aux accusations dont il faisait l'objet. «Fillon est évidemment surcoté chez nous», reconnaissait le fondateur de l'application GOV à deux mois du scrutin. « Le décalage donne quand même une information ; Fillon dispose de partisans déterminés, convaincus, mobilisés, qui veulent le montrer. » Une indication, donc. Mais en aucun cas une prévision. Plus largement, les instituts ont dénoncé le manque de rigueur et de clarté des méthodes utilisées : indicateurs retenus, modélisation des données, nature des croisements, etc. Les algorithmes issus du machine learning sont-ils réellement en mesure d'identifier les aspects positifs et négatifs des commentaires mais aussi les nuances attachées au langage et à sa diversité. Comment distinguer le sarcasme, la démonstration par l'absurde, l'ironie, bref le second du premier degré ? Est-ce même possible ? En réalité si ces critiques sont recevables aujourd'hui, il n'est pas non plus absurde d'imaginer que des progrès technologiques puissent en venir à bout. Pourquoi ne pas imaginer des méthodes de redressement pour corriger les distorsions liées aux spécificités des internautes sur les réseaux sociaux. Après tout, les instituts y recourent depuis des années pour corriger la sous-déclaration de certains votes (d'extrême-droite, notamment) ou la sous-représentativité de certains groupes sociaux et de certaines classes d'âge avec un certain succès. De la même façon on ne voit pas pourquoi les progrès en matière d'intelligence artificielle ne parviendraient pas à maîtriser les nuances du langage au-delà d'un simple décompte des occurrences. On admettra sans crainte de beaucoup se tromper que le secteur du big data n'en est qu'à ses débuts et que l'expérience aidant les difficultés que l'on connait aujourd'hui pour parvenir à une appréciation juste des opinions sur les réseaux sociaux seront de mieux en mieux maitrisées au fil du temps.

L'IMPOSSIBLE ADDITION DES DONNÉES QUALITATIVES ET QUANTITATIVES

Même si le big data n'est pas en soi une discipline dédiée à la prédiction, la recherche et l'identification de corrélations ouvre la possibilité de construire des scénarios prédictifs. Gilles Badinet cite l'exemple de Recorded Future qui « travaille sur des prédictions aussi étonnantes que celles des attaques de sites Web, de banque… ». Et c'est loin d'être le seul. Cela dit, outre le fait que Badinet se montre prudent lorsqu'il dit Recorded Future qu'elle « travaille sur des prédictions » et non qu'elle « prédit » ou qu'elle « est en mesure de prédire » des attaques et que ces « prédictions » semblent limitées à l'univers du net, le Big data prédictif en matière d'opinion risque bien de se heurter à un autre obstacle, définitivement insurmontable : la différence de nature radicale, inconciliable, entre la donnée représentative et celle qui ne l'est pas. Le principe du big data est de traiter des données extrêmement hétérogènes. Mais donnée quantitative et donnée qualitative sont séparées par une frontière infranchissable : celle de la représentativité. Bien sûr les études qualitatives peuvent s'enrichir de l'étude de sondages. Elles sont également utilisées pour établir les questionnaires des sondages et éclairer les résultats. Mais il s'avère illusoire sinon absurde d'imaginer le processus inverse, à savoir pondérer les résultats d'une étude représentative par une étude qui ne l'est pas. Or c'est exactement ce que se proposaient de faire Filteris, Vigiglobe ou Predict The President : tirer des statistiques de l'étude sémantique des réseaux sociaux puis s'en servir pour corriger les chiffres des sondages. Une cuisine pour le moins curieuse. Prenons l'exemple d'un candidat progressant de 4% dans les intentions de vote selon les sondages et dont le nom connaitrait une augmentation de 8% d'occurrences positives durant la même période sur les réseaux sociaux ? Est-il possible d'imaginer procéder à une quelconque moyenne entre le premier et le second chiffre ? Evidemment non. Il ne s'agit pas des mêmes données.

Que l'analyse prédictive gagne en précision grâce au big data prédictif n'a rien d'un mythe. Les progrès des prévisions météo et leur utilisation en agriculture en constituent sans doute la preuve la plus évidente. Même constat dans le domaine de la santé où le développement d'une médecine préventive pourrait générer des économies considérables, de l'énergie pour le dimensionnement des unités de production et des réseaux ou des transports pour anticiper les flux principaux de déplacement et accompagner l'introductions de véhicules autonomes et connectés. Mais ces exemples, aussi intéressants soient-ils, reposent sur des données scientifiques et comportementales et non des données déclaratives. Quoi qu'en aient dit les instituts en leur temps et quoi qu'en disent aujourd'hui certains acteurs du big data, il est impossible de prévoir le score d'une élection, a fortiori en faisant un mixte de données qualitatives et quantitatives. Ne serait-ce que pour une raison : l'opinion peut évoluer jusqu'au jour du vote, dans le secret de l'isoloir. Chaque mesure de l'opinion n'est bel et bien qu'une photographie ou, alors dans le meilleur des cas, la confirmation ou l'infirmation d'une tendance passée. Ce qui n'est en soi déjà pas si mal. Encore faut-il qu'après les instituts, les partisans du big data se résignent à admettre cette limite. ■