Filteris, Enigma… Face aux instituts de sondage, la défaite des prévisions « alternatives »

Plusieurs entreprises et instituts annonçaient ces derniers mois des résultats très différents des sondages. Tous se sont largement trompés.

LE MONDE | 24.04.2017 à 14h50 • Mis à jour le 24.04.2017 à 17h22

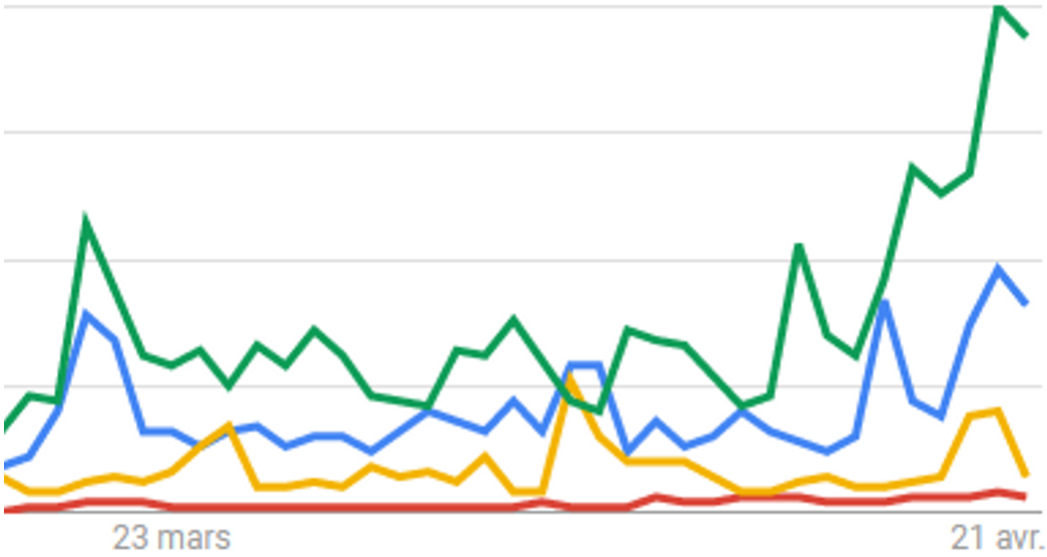
Elles étaient particulièrement prisées des militants : depuis des mois, des études se basant non pas sur des sondages mais sur le « big data », l'analyse des messages publiés sur les réseaux sociaux ou encore du profil sociodémographique des Français, prétendaient mesurer plus efficacement que les instituts de sondage les intentions de vote au premier tour de la présidentielle.

Ces études donnaient pour la plupart des résultats très différents des sondages classiques. L'entreprise canadienne Filteris affirmait ainsi que le second tour opposerait François Fillon et Marine Le Pen. Vigiglobe, une start-up française, ne donnait pas de classement prédictif, mais affirmait que François Fillon serait probablement au second tour, et jugeait Marine Le Pen en perte de vitesse. Plus affirmatif, Predict my president (http://www.lepoint.fr/presidentielle/presidentielle-les-deux-qualifies-pour-le-second-tour-sont-18-04-2017-2120431_3121.php?utm_campaign=Echobox&utm_medium=Social&utm_source=Facebook&link_time=1492499204#xtor=CS1-31-%5BEchobox%5D) , un programme d'étudiants de l'école Telecom Paris Tech, donnait des scores très précis pour les quatre principaux candidats : Marine Le Pen (24,13 %), François Fillon (21,77 %), Emmanuel Macron (20,32 %), et Jean-Luc Mélenchon (18,66 %).

Elles avaient prévu la victoire de Trump

Toutes ces prédictions se sont donc révélées fausses, tout comme celles d'entreprises parues dans la presse étrangère. Le fonds d'investissement Leonie Hill Capital (http://www.cnbc.com/2017/02/15/marine-le-pen-is-on-course-to-be-frances-next-president-leonie-hill-capitals-arun-kant-says.html) avançait ainsi que son programme d'intelligence artificielle voyait Marine Le Pen emporter 28 % des voix au premier tour, devançant Emmanuel Macron à 19 % ou 20 % et François Fillon à 16,4 %. Un peu plus prudente dans ses affirmations, l'agence suisse Enigma envisageait (https://enigma.swiss/fr/2017/04/12/election-presidentielle-francaise-et-big-data/) , elle, *« l'hypothèse d'un duel Fillon-Mélenchon »*, qui *« paraît tout à fait cohérente avec d'autres analyses, qualitatives cette fois »*.

Surfant sur la méfiance envers les sondages, après les victoires de Donald Trump aux Etats-Unis et le Brexit britannique, les entreprises spécialisées dans le « big data » ont suscité l'engouement. La plupart en utilisant un argument massue : elles avaient, contrairement aux sondages, prévu la victoire de Donald Trump. C'était notamment l'un des points sur lesquels s'appuyait Filteris, dont les prédictions avaient la faveur de certains supporteurs de François Fillon, que l'entreprise donnait en tête à l'issue du premier tour. L'entreprise mettait aussi volontiers en avant ses prédictions passées en France, et notamment le fait qu'elle avait prévu le bon classement à la présidentielle de 2012, ou encore la victoire de François Fillon à la primaire de la droite et du centre.



Détail d'une recherche Google Trends sur des noms des candidats de l'élection présidentielle.

Mais l'entreprise oubliait aussi de préciser que son modèle prédictif avait connu des ratés majeurs – elle imaginait que François Fillon affronterait Nicolas Sarkozy au second tour de la primaire, et avait largement sous-estimé le score de Marine Le Pen en 2012 – sans oublier qu'elle s'était totalement trompée dans sa prédiction pour la primaire de la gauche gagnée par Benoît Hamon, voyant une victoire de Manuel Valls face à Arnaud Montebourg au deuxième tour.

Aucune de ces entreprises ne dévoile précisément la méthodologie utilisée pour établir leurs prédictions. Mais toutes s'appuient sur des principes similaires : la mesure d'opinions exprimées sur Internet, parfois pondérées par des données démographiques, électorales ou sociologiques. Certaines mesurent directement le « buzz » sur les réseaux sociaux (Vigiglobe), d'autres effectuent des sondages qui ne mesurent pas les intentions de vote mais des positionnements sur des sujets (Filteris) ou analysent l'évolution des recherches sur Google (Enigma). Le tout passé à la moulinette d'un algorithme dont le fonctionnement est tenu secret, pour pondérer les résultats bruts en fonction de critères définis par ces entreprises.

Des méthodologies secrètes (et douteuses)

C'est d'ailleurs l'un des paradoxes de ces études, dont les défenseurs raillent volontiers les sondeurs et leur habitude de *« redresser »* les sondages de manière peu transparente. Aucun de ces instituts ne communique en effet la méthodologie précise utilisée, se bornant à donner les principes généraux appliqués.

Mais plus encore que le secret qui entoure ces outils de prédiction, c'est, dans la plupart des cas, la méthodologie générale utilisée qui pose question. Les prédictions basées sur les publications sur les réseaux sociaux souffrent de plusieurs défauts structurels. La sociologie des utilisateurs de Facebook et surtout de Twitter n'est, par exemple, pas représentative du corps électoral ; la mesure sur les réseaux donne également une *« prime »* aux candidats dont les militants sont les plus actifs sur ces réseaux, sans préjuger du vote des électeurs. Enfin, certains signaux captés en ligne sont beaucoup plus difficiles à interpréter que la réponse à une question du type *« pour qui comptez-vous voter ? »*.

Les courbes de Google Trends, qui mesurent les requêtes sur le moteur de recherche, fournissent ainsi des indications particulièrement difficiles à interpréter : elles montraient ces dernières semaines un très fort volume de recherches pour Emmanuel Macron et Jean-Luc Mélenchon. Mais ces recherches ne se traduisent pas automatiquement par un regain de votes, tant s'en faut, puisqu'un internaute peut effectuer des recherches sur un candidat pour des raisons très différentes. Et le changement de variables aussi simples que la présence (https://trends.google.com/trends/explore?date=today%203-m&q=emmanuel%20macron,jean-luc%20m%C3%A9lenchon,fran%C3%A7ois%20fillon,marine%20le%20pen) ou l'absence (https://trends.google.com/trends/explore?date=today%203-m&q=m%C3%A9lenchon,macron,fillon,le%20pen) du prénom d'un candidat peut conduire à des résultats très différents : dans le premier cas, Marine Le Pen arrive derrière Emmanuel Macron, dans le second elle le devance. Avant même l'interprétation des

données par des algorithmes plus ou moins évolués, la collecte des données elles-mêmes reste un défi majeur pour tous les outils prédictifs qui voudraient remiser les sondages au placard.