UTILISATION DES DONNÉES GOOGLE TRENDS DANS L'ENQUÊTE DE CONJONCTURE COMMERCE DE DÉTAIL DE LA BANQUE DE FRANCE

François ROBIN

Banque de France – DGS DESS SEEC

Francois.ROBIN@banque-france.fr

Mots-clés : google trends, nowcasting, conjoncture, e-commerce (ou vente à distance), big data, bayesian averaging

Résumé

Dans le cadre du partenariat la liant à la Banque de France, la Fédération e-commerce et vente à distance (FEVAD) fournit mensuellement le chiffre d'affaires réalisé en e-commerce B2C. Cependant, ces livraisons sont trop tardives pour être intégrées à l'enquête mensuelle du Commerce de Détail lors de sa parution. Pour le pallier, il convient de procéder à des estimations. Outre le chiffre d'affaires général réalisé en e-commerce B2C, cette estimation concerne les 5 secteurs d'activité suivants : Chaussures, Habillement, Meubles, Électroménager Grand Public et Électronique. Les contraintes de la mise en production ont incité à retenir un seul processus, applicable à chaque secteur. Celui-ci découle de multiples essais effectués en amont.

Du modèle utilisant les Google Trends…

Le modèle mis en place a été construit selon certains choix méthodologiques, résultant d'une part de contraintes liées aux objectifs et données en présence, d'autre part de différents tests heuristiques et simulations. Ces choix sont basés sur la performance – dont l'indicateur est la RMSFE (erreur de prévision) observée sur la période de Backtest – et sur leur capacité à être mis en production, définissant ainsi le modèle. Les principaux points étudiés sont :

- Le choix d'un modèle log-linéaire
La transformation logarithmique résout, notamment, d'éventuels problèmes d'hétéroscédasticité.

- L'utilisation des données CVS
Appliquer le modèle aux séries désaisonnalisées permet de limiter l'impact des saisonnalités sur les coefficients du modèle linéaire.

- L'utilisation des données différenciées
La différenciation des données vise à éviter les régressions fallacieuses, fréquentes lors de régression entre séries temporelles.

- La période d'estimation
Le partenariat entre la FEVAD et la Banque de France est récent (2012). Il en découle une période de backtest relativement courte. Si ce point est problématique en apparence, le e-commerce a connu un développement conséquent ces dernières années, à l'image de sa forte croissance ; tel un phénomène émergent, son paysage est encore mouvant. En outre, les données récentes sont plus significatives.

- La sélection des termes recherchés

Suite aux résultats infructueux de l'outil Google Correlate (corrélations fortuites), plusieurs techniques de sélection de variables ont été testées : régression Stepwise, lasso... Aussi, une ACP a été réalisée sur différentes périodes et a permis d'utiliser le maximum de termes recherchés. Parallèlement, la question du type de termes recherchés a pris une place importante dans notre réflexion ; e.g. pour les Chaussures : faut-il privilégier des marques (Nike, Louboutin…), des catégories (bottes, tongs…), des Pure Players (Sarenza, Zalando…) ou les combiner indifféremment ? Pérenniser la production implique de travailler à méthodologie constante et donc de travailler avec le même pool de termes recherchés d'un mois sur l'autre. Cependant, le monde du e-commerce est mouvant et le modèle retenu doit être adaptable, i.e. capable d'intégrer de nouveaux termes de recherche.

… Vers un modèle agrégé

Si la littérature issue du monde de la recherche universitaire semble s'accorder sur l'apport des données Google Trends aux travaux de « nowcasting » (cf. Choi et Varian [2011]; Askitas Zimmermann [2009]), les résultats observés par les praticiens ne sont pas toujours à la hauteur des espérances (cf. Bortoli et Combes [2015], Insee; McLaren et Shanbhogue [2011], Banque d'Angleterre). Ici, l'application se prête particulièrement bien aux données Google Trends puisqu'il s'agit d'estimer la conjoncture du e-commerce. Néanmoins, une certaine vigilance vis-à-vis de l'outil Google Trends est nécessaire, du fait d'une méthodologie rendue globalement opaque par Google. Aussi, les modes de consultation d'internet peuvent évoluer. Dans ces conditions, utiliser plusieurs modèles permet de pallier ces problèmes. Jusqu'à présent, l'estimation était le fruit d'un modèle autorégressif (SARIMA(12)), qui a joué le rôle de benchmark dans cette étude. Peu satisfaisant en termes de résultats, il peut désormais être complété par d'autres modèles statistiques s'appuyant sur des données exogènes. Ces données exogènes, et en particulier Google Trends, permettent d'améliorer sensiblement la capacité prédictive du modèle final. La meilleure solution est l'agrégation de 3 modèles :

- Un modèle autorégressif SARIMA(12)

- Un modèle de régression (pondérée) dont les facteurs explicatifs sont des Google Trends et des variables muettes de saisonnalité

- Un modèle de régression dont les variables explicatives sont issues de chiffres d'affaires dans les autres formes de commerce (collectés par la Banque de France auprès des commerçants dès le début de M+1, donc plus précocement que les chiffres d'affaires fournis par la FEVAD), ainsi que des variables muettes saisonnières.

L'article détaille également la stratégie d'agrégation retenue, inspirée de la méthode d'agrégation Bayésienne. Simplement, l'agrégation est une pondération – dont les coefficients sont réévalués chaque mois – des 3 modèles ; ainsi, l'expert métier dispose de 4 estimations pour décider.

Par rapport au modèle précédemment utilisé, le modèle agrégé permet une réduction de la RMSFE (d'un facteur 4 pour le modèle global) sur la période de Backtest pour chaque secteur.