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

François ROBIN¹

Banque de France – DGS DESS SEEC

Francois.ROBIN@banque-france.fr

Mots clés : Google Trends, Nowcasting, Conjoncture, E-commerce, Bayesian averaging, Sélection de variables

*Document provisoire, ne pas diffuser ou
citer sans l'accord de l'auteur*

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 auprès des particuliers, depuis 2012. La Banque de France procède à des estimations en attendant les livraisons. La croissance du e-commerce renforce l'enjeu des estimations ; jusqu'ici effectuées *via* un modèle autorégressif (SARIMA(12)). Moyennement satisfaisant en termes de résultats, il peut désormais être complété par d'autres modèles statistiques s'appuyant sur des données exogènes grâce à un historique plus long de données. Les requêtes Google notamment, mesurées par Google Trends, permettent d'améliorer la capacité prédictive du modèle final, obtenu en combinant les modèles simples.

Cet article détaille les différents choix opérés conduisant à la prévision finale : transformation des données, modèles à sélection de variables et stratégie pour la prévision.

Abstract

This paper investigates how to improve a nowcasting estimation, by adding external data and in particular Googler trends data. More precisely, FEVAD (Federation of e-commerce and mail-order selling) delivers monthly data – aggregated turnover of their members – to Banque de France. Nevertheless, these data come too late to be integrated to the first release of the monthly survey in retail trade. Estimation has to be done. It is done so far with a SARIMA model. Moreover, contrary to FEVAD data, Google Trends and retail stores data from the survey are available for the estimating period. An adaptive lasso is used to choose the most relevant information, especially in Google Trends potentially numerous researches. As only one estimation is needed, a choice has to be done between pooling the information sets and combining simple models. Adaptive lasso is also used for the encompassing model whereas Bayesian averaging enables to combine the three individual models (SARIMA and the two adaptive lasso based on Google Trends and retail stores data from the survey).

JEL Codes : C51 (Model Construction and Estimation), C53 (Forecasting and Prediction Methods), C11 (Bayesian Analysis : General), E17 (Forecasting and Simulation : Models and Applications)

¹ François Robin travaille au Service des Enquêtes Économiques de Conjoncture (Direction Générale des Statistiques). Il remercie tout particulièrement les participants à un atelier interne de recherche de la Banque de France, Patrick Kwok d'avoir continuellement suivi, contribué et soutenu ces travaux. Un très grand merci aussi à Martial Ranvier et Valérie Chauvin pour leurs remarques pertinentes et leurs conseils méthodologiques précieux. Enfin, merci à François Guinouard ; et à François Brunet, à l'origine du sujet.

1. Introduction

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-commerce² B2C (Business to Consumer). Cependant, ces livraisons sont trop tardives pour être intégrées dans la première publication de l'enquête mensuelle de conjoncture (EMC) du Commerce de Détail, où elle est remplacée par des estimations.

Jusqu'à présent, la faible profondeur historique des données restreignait le champ des possibles techniques de prévision : le modèle autorégressif utilisé jusqu'ici, peut aujourd'hui être complété par des modèles utilisant des données exogènes : les données Google Trends et les indices quantitatifs conjoncturels du commerce traditionnel (issus de l'EMC Commerce de Détail). L'intérêt des données exogènes est la disponibilité. Concrètement, en période d'enquête (au début du mois $M+1$) et plus précisément lors de l'estimation des données FEVAD du mois M , les indices quantitatifs conjoncturels du mois M sont en construction ; alors que les données Google Trends du mois M sont définitives. Ces estimations rentrent pleinement dans le cas d'un exercice de *nowcasting*. Cette évolution rencontre deux problématiques.

D'une part, Google Trends fournit de nombreuses variables explicatives, parmi lesquelles les meilleures doivent être sélectionnées. Une méthode de machine learning répond à la contrainte duale du sujet, à savoir combiner la faible profondeur de l'historique des données FEVAD (les livraisons ont démarré en 2012) avec l'immense champ des requêtes Google possibles : l'adaptive lasso, développé par Zou (2006).

D'autre part, disposant de plusieurs modèles avec des variables exogènes d'origines différentes, il est intéressant de vérifier si la combinaison de modèles permet d'obtenir de meilleurs résultats. Ce sujet est largement débattu, comme l'expliquent Bec et Mogliani (2015).

Après avoir situé l'article dans la littérature, la deuxième partie est consacrée aux données, avec la présentation de l'EMC Commerce de Détail (CD), des données FEVAD et des données Google Trends. Les particularités et l'opacité de la méthodologie de construction de ces dernières nécessitent la mise en place de tests de robustesse et l'automatisation de certaines corrections. La troisième partie concerne les choix de modélisation. Y sont traités successivement le traitement de la stationnarité des séries, puis le processus complet de test des modèles. La quatrième partie est dédiée aux résultats et à leur interprétation. La dernière partie conclut.

² D'après la FEVAD, la part du e-commerce dans le commerce de détail (hors alimentaire, conformément au champ de l'enquête de conjoncture Commerce de Détail de la Banque de France) était de 7 % en 2013, 8 % en 2014 et 9 % en 2015.

2. Revue de littérature

2.1. Données Google Trends

Un enjeu majeur du sujet est l'intégration d'une nouvelle source de données : Google Trends. Disponibles en temps quasi-réel, ces indices (associés aux requêtes effectuées par les utilisateurs sur le moteur de recherche Google) constituent un flux d'information et une source de données massives. Encore peu utilisée par les institutions pour des travaux récurrents, elle a été l'objet de publications. Les travaux d'Ettredge, Gerdes et Karuga (2005) ou ceux d'Askistas et Zimmerman (2009), consacrés à la prédiction du taux de chômage en utilisant des mots-clés recherchés dans Google, connaissent des résultats encourageants. Choi et Varian (2009 et 2011) se montrent plus mitigés quant à l'apport des Google Trends. Par ailleurs, leur revue de littérature recense de nombreuses publications utilisant les recherches Google, principalement dans le domaine de l'épidémiologie ; l'outil alors utilisé et développé par Google (Google Flu) a été supprimé, suite à des défaillances déjà mentionnées par Bortoli et Combes (2015).

Ces outils sont uniquement à la main de Google : la méthodologie de construction est opaque, comportant de nouveaux risques pour les utilisateurs. Du reste, la vigilance est aussi requise en amont du choix des requêtes : les changements de requêtes Google significatives, qui peuvent venir des comportements des utilisateurs ou de la mouvance du paysage du phénomène étudié – bien illustré par McLaren et Shanbhogue (2011) dans leur application au chômage avec la baisse de popularité de la requête « ANPE » au profit de « Pôle Emploi » -, demandent une attention particulière.

2.2. Sélection de variables

Les méthodes de machine learning apportent une solution à la sélection de variables, et plus particulièrement l'adaptive lasso développé par Zou (2006). Pour rappel, l'équation du lasso classique introduit par Tibshirani (1996) est :

$$\hat{\beta}_{lasso} = \underset{\beta}{\operatorname{argmin}} (Y - X' \beta)^2 + \lambda (\beta), \lambda \geq 0$$

$$\hat{\beta}_{lasso} = \underset{\beta}{\operatorname{argmin}} \left(Y - \sum_{j=1}^p x_j \beta_j \right)^2 + \lambda \sum_{j=1}^p (\beta_j), \lambda \geq 0$$

Dans la régression lasso, la même pénalité λ est appliquée à toutes les variables. Zou (2006) propose d'adapter la pénalité en fonction des variables dans l'adaptive lasso (*adalasso*) :

$$\hat{\beta}_{adalasso} = \underset{\beta}{\operatorname{argmin}} \left(Y - \sum_{j=1}^p x_j \beta_j \right)^2 + \lambda \sum_{j=1}^p w_j (\beta_j), \begin{pmatrix} \lambda \geq 0 \\ w_j \geq 0 \end{pmatrix}$$

L'adaptive lasso est un lasso pondéré. Zou (2006) démontre ses propriétés d'Oracle (consistance dans la sélection de variables et dans l'estimation paramétrique : convergence asymptotique en loi normale). Si Zou (2006) définit les pénalités individuelles $\hat{w} = 1/(\hat{\beta})^\gamma$, avec $\hat{\beta}$ l'estimateur des moindres carrés ordinaires et $\gamma > 0$ (en pratique, $\gamma \in (0,5 ; 1 ; 2)$), une alternative consiste à utiliser l'estimateur issu de la régression ridge, introduite par Hoerl et Kennard (1970), pour définir le vecteur de pénalités individuelles. Son utilisation permet notamment d'éviter une mauvaise estimation des pénalités due à la présence de multi-colinéarité parmi les régresseurs. L'estimateur ridge s'écrit comme suit³ :

$$\hat{\beta}_{ridge} = \text{arg min}_{\beta} (Y - X' \beta)^2 + \kappa (\beta)_2^2, \quad \kappa \geq 0$$

L'optimisation de l'adaptive lasso se fait donc en deux étapes. Premièrement, les pénalités individuelles sont déduites d'une régression ridge ; la valeur de la pénalité κ est alors obtenue par validation croisée « *leave one out* »⁴ (Hyndman et Athanasopoulos 2018). Ensuite, $\hat{w} = \hat{\beta}_{ridge}$ mène à l'équation du lasso (dont la pénalité λ est également optimisée par validation croisée « *leave one out* ») :

$$\hat{\beta}_{adlasso} = \text{arg min}_{\beta} \left(Y - \sum_{j=1}^p x_j \beta_j \right)^2 + \lambda \frac{\sum_{j=1}^p (\beta_j)}{(\hat{\beta}_{ridge, j})}, \quad \lambda \geq 0$$

L'avantage du lasso est de pouvoir fonctionner en grande dimension (nombre de variables supérieur au nombre d'observations, *i.e.* à la taille de la fenêtre temporelle dans notre cas). Le lasso est aussi réputé pour sa parcimonie. Ces deux propriétés en font la solution idoine au problème ; à savoir le grand nombre de requêtes Google possibles couplé aux courts historiques des livraisons FEVAD.

2.3. Combinaison de modèles ou modèles global ?

Trois modèles individuels ont été mis en œuvre : le modèle Google Trends, le modèle CD⁵ issu de l'enquête commerce de détail, et le modèle SARIMA qui était utilisé jusqu'à présent.

Bec et Mogliani (2015) recensent les méthodes les plus fréquentes en termes de combinaison d'information. Selon eux, Bates et Granger (1969) sont les premiers à se positionner, en faveur de l'agrégation de prévisions issues de modèles distincts. Plus tard, Diebold (1989) favorise l'utilisation d'un seul modèle, combinant différentes sources d'information. Plus récemment, Huang et Lee (2010) affirment qu'un modèle global est meilleur s'il est bien spécifié. Par ailleurs, Clements et Galvão (2006, 2008) ou Kuzin, Marcellino et Schumacher (2013) tranchent en faveur de l'agrégation lors d'applications empiriques. Dans l'exercice de prévision de l'indice de consommation, Bec et Mogliani (2015) ont de meilleurs résultats avec l'agrégation. Cet article vise à enrichir ce débat d'un nouveau cas d'application en comparant les résultats du modèle agrégé avec ceux d'un modèle global dont la spécification est la même que pour les modèles individuels, soit l'adaptive lasso appliqué à tous les régresseurs (Google Trends, indices CD et SARIMA) simultanément. De Gooijer et Hyndman (2006) mettent en avant les bénéfices de l'agrégation, notamment sa lisibilité lorsque les

³ La régression ridge et la régression lasso sont des régressions pénalisées de normes respectivement L2 et L1.

⁴ Concrètement, l'échantillon de validation est constitué d'une observation ; celui d'apprentissage, des $n-1$ autres observations (pour un échantillon de taille n). Les n valeurs de κ , obtenues sur chaque échantillon d'apprentissage (chacune minimisant la RMSE) font l'objet d'une moyenne pour obtenir la valeur finale de κ .

⁵ Les indices quantitatifs du commerce de détail font également l'objet d'une modélisation lasso.

modèles agrégés sont facilement interprétables. Ici, l'agrégation concerne les trois modèles individuels, chacun portant des effets qui lui sont propres :

- Le modèle SARIMA reproduit le schéma passé de la série
- Le modèle CD exploite les données issues du commerce de détail traditionnel
- L'information d'internet est extraite du modèle basé sur les indices Google Trends

La question est de pondérer chaque prévision :

$$\hat{Y}_{t+1} = \gamma \hat{Y}_{t+1}^{SARIMA} + \mu \hat{Y}_{t+1}^{gTrends} + \vartheta \hat{Y}_{t+1}^{CD}$$

Plusieurs stratégies d'agrégation sont possibles : des plus simples, telles que la pondération par la moyenne ($\gamma = \mu = \vartheta = \frac{1}{3}$) ou par l'inverse des erreurs – *In Sample* ou *Out Of Sample* (cf. Aiofli et

Timmerman (2006)) -, aux plus élaborées. Par exemple, l'inférence bayésienne, dont le fondement est le théorème de Bayes⁶ (cf. Marin et Robert (2010)), déduit la probabilité d'un évènement à partir d'autres évènements déjà évalués. La statistique bayésienne, particulièrement utilisée lorsque les échantillons sont petits, débouche sur des méthodes de classification ; ou d'agrégation ici. Hoeting, Madigan, Raftery et Volinsky (1999) mettent en avant les bonnes performances de l'agrégation bayésienne. Zeugner (2011) a développé un package R sur le sujet. L'idée est de tester les modèles d'une classe M donnée ; et, de les pondérer selon leurs probabilités d'être le bon modèle. La classe M est celle des modèles linéaires. Habituellement, le grand nombre de modèles rend difficile l'agrégation bayésienne (cf. Hoeting, Madigan, Raftery et Volinsky (1999)). Ce n'est pas le cas ici : avec trois régresseurs (correspondant aux modèles Google Trends, CD et SARIMA), huit modèles linéaires sont possibles. En notant D les données et M_j ($1 \leq j \leq 8$) un modèle donné, le théorème de Bayes donne :

$$P(M_j, D) = \frac{P(D, M_j) P(M_j)}{\sum_{1 \leq i \leq 8} P(D, M_i) P(M_i)}$$

Il s'agit de préciser les deux termes du numérateur pour évaluer la probabilité *a posteriori*⁷ :

- $P(M_j)$ correspond à la probabilité *a priori*⁸ que le modèle M_j soit le bon

⁶ L'écriture commune du théorème est : $P(A, B) = \frac{P(B, A) P(A)}{P(B, A) P(A) + P(B, -A) P(-A)}$ avec P la mesure de probabilité, A et B deux évènements

⁷ La probabilité *a posteriori* est déterminée à l'aide des données, *In Sample*.

⁸ Il y a différentes manières de déterminer les probabilités *a priori*, comme le montre Zeugner (2011). Dans notre cas, plusieurs tests ont été faits (prior binomial, uniforme, déterministe...) sans affecter véritablement les résultats.

- $P(D, M_j) = \int pr(D, \beta_j, M_j) pr(\beta_j, M_j) d\beta_j$ avec β_j les paramètres du modèle :
 $\beta_j = (\gamma_j, \mu_j, \vartheta_j)$ estimé sur le modèle M_j . Ici, les paramètres sont la quantité d'intérêt.

Concrètement, les valeurs des coefficients obtenus dans chaque modèle M_j de la classe M sont pondérées par la probabilité que chaque modèle M_j soit le bon : $\gamma = \sum_i \gamma_i P(M_i | D)$ avec $\gamma_i = E(\gamma | D, M_i)$ la valeur du coefficient dans le modèle M_i . Il en est de même pour μ et ϑ .

1. Données

1.1. L'Enquête Mensuelle de Conjoncture – Commerce de Détail

Parmi les enquêtes de conjoncture réalisées par la Banque de France, l'une est consacrée au Commerce de Détail⁹ (CD). Publiée mensuellement, elle suit l'évolution des chiffres d'affaires (CA) TTC de ses déclarants. Chacune des 6 800 entités de l'échantillon – réparties selon plus de 4 000 entreprises – fournit son CA total et la part des principaux produits (si elle n'est pas « mono produit »). Chaque mois, le taux de réponse est d'environ 90 %. Les indicateurs quantitatifs conjoncturels sont établis en regroupant ces données individuelles selon des caractéristiques communes aux entreprises : par mode de distribution (petit commerce traditionnel, grande surface spécialiste et succursaliste, hyper et supermarchés, grand magasin et magasin populaire), puis par produit.

1.1.1. Les indices quantitatifs conjoncturels

Selon la méthodologie de l'enquête, les chiffres d'affaires sont cylindrés. Chaque indice de chiffre d'affaires γ issu de l'enquête est ainsi construit (avec X le montant de chiffre d'affaires associé à l'indice) :

$$Y_M = Y_{M-12} \frac{X_M}{X_{M-12}}$$

Le périmètre utilisé pour X_M est le même que pour X_{M-12} . Ce sont les mêmes entreprises qui sont considérées dans ces deux (sommées de) chiffres d'affaires. Le cylindrage évite les variations extrêmes non représentatives de l'échantillon (points aberrants). La fermeture (ou l'ouverture) d'un magasin est en effet le phénomène extrême le plus fréquent auquel l'enquête est confrontée. La baisse (respectivement hausse) de chiffre d'affaires qui en découle est compensée par des évolutions opposées dans l'ensemble de ses concurrents, ensemble qui ne sera pas pris en compte dans sa totalité dans l'échantillon. Par ailleurs, il est plus probable de capter une fermeture qu'une ouverture de magasins (magasin ou nouvelle enseigne non encore entrés dans l'échantillon), si bien que la mesure serait biaisée en l'absence de cylindrage.

⁹ Les derniers résultats sont disponibles à cette adresse : <https://www.banque-france.fr/statistiques/chiffres-cles-france-et-etranger/enquetes-de-conjoncture/conjoncture-commerce-de-detail>

1.1.2. La FEVAD

Conformément à la méthodologie de l'enquête, la FEVAD nous livre chaque mois les chiffres d'affaires du mois M et leur révision à échantillon¹⁰ constant du mois $M-12$; en effet, la liste des adhérents de la FEVAD évolue. Tirée par le e-commerce, la vente à distance (VAD) a pris de l'ampleur. La FEVAD regroupe de nombreux acteurs du e-commerce en France et est le partenaire idoine pour permettre à la Banque de France de bien traiter les évolutions de chiffre d'affaires dans le e-commerce. Depuis janvier 2012, la fédération nous livre les chiffres d'affaires mensuels agrégés de ses adhérents. Cependant, l'enquête du mois M est publiée aux alentours du milieu du mois $M+1$, alors que les livraisons de la FEVAD arrivent à la fin du mois $M+1$. Ces livraisons concernent le chiffre d'affaires total et ceux de cinq produits : textile (habillement et textile de maison – dénommé habillement dans la suite), chaussures (maroquinerie incluse), électronique grand public (EGP), électroménager et mobilier (meubles essentiellement). Renseignant le champ des produits industriels hors automobile, le chiffre d'affaires total ne correspond donc pas à la somme de ceux des produits.

1.1.3. Rapprochement des données FEVAD avec celles de l'EMC

Voici les indices quantitatifs conjoncturels des autres modes d'exploitation utilisés¹¹ pour chaque estimation :

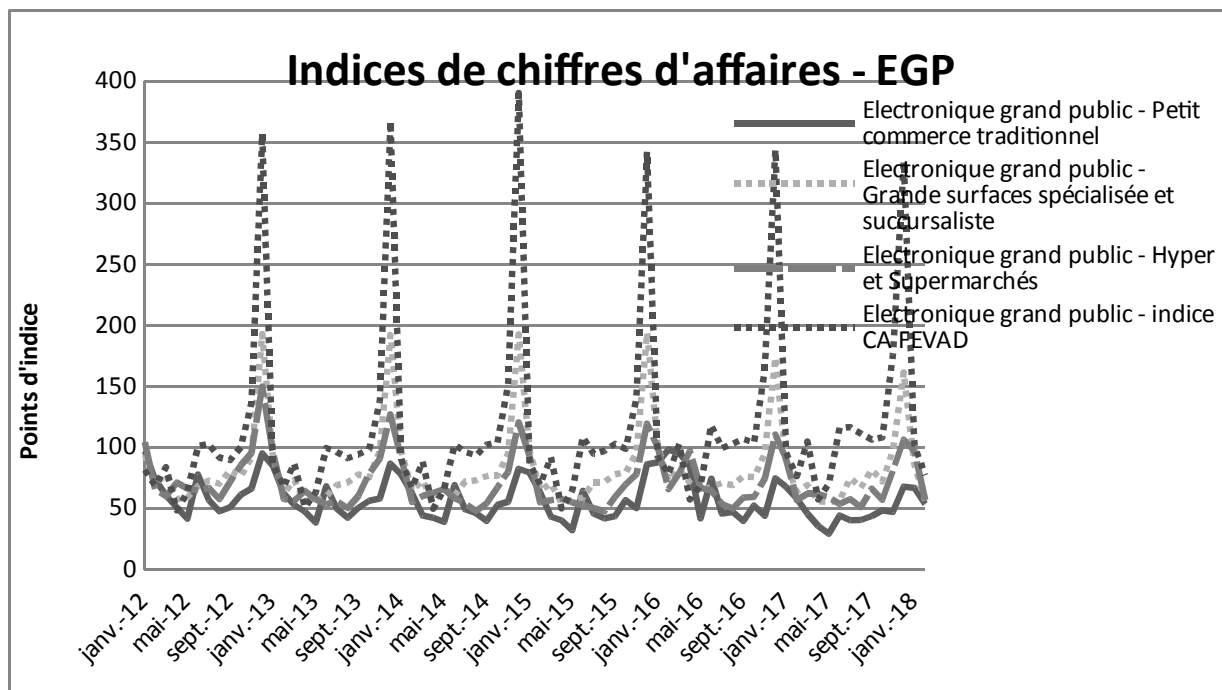
	Petit commerce traditionnel	Grande surface spécialiste et succursaliste	Hyper et supermarchés	Grand magasin et magasin populaire	Ensemble des ventes physiques
Total Produits industriels hors automobiles	O	O	O	O	X
Chaussures	X	X	O	O	O
EGP	X	X	X	O	O
Électroménager	X	X	X	O	O
Meubles	X	X	X	O	O
Habillement	X	X	X	X	O

Tableau 1 : Existence (X) ou non (O) des indices conjoncturels des modes d'exploitation, par produit. Source : Banque de France DGS SEEC

Ces indices existent depuis le début des années 1990. Ci-dessous, les indices de CA dans l'EGP (issus de l'enquête pour les modes d'exploitation concernant les magasins ; et obtenu par construction pour la FEVAD) :

¹⁰ L'échantillon est constitué de 70 plus gros déclarants

¹¹ Certains croisements ne sont pas calculés faute d'un échantillon suffisant et pour des raisons de confidentialité des données



Graphique 1 : indices de l'électronique grand public - Sources : Banque de France DGS SEEC – FEVAD

Dans l'EGP, le pic des ventes en décembre est commun aux différents modes d'exploitation. Les corrélations¹² entre l'indice de CA de la FEVAD et les indices de CA des ventes physiques pour l'EGP (en %) complètent l'information graphique :

EGP - Petit commerce traditionnel	44%
EGP - Grandes surfaces spécialisée et succursaliste	96%
EGP - Hyper et Supermarchés	48%

Tableau 2 : Corrélations de l'indice de CA FEVAD avec les indices du commerce traditionnel issus de l'enquête dans l'EGP. Sources : Banque de France DGS SEEC – FEVAD

La corrélation entre l'indice de CA VAD (ventes à distance) et celui des grandes surfaces spécialisée et succursaliste incite à utiliser les données des ventes physiques pour estimer les données FEVAD. D'une manière générale, rapprocher ces données permet l'observation de mécanismes économiques simples. Par exemple, un effet de substituabilité peut être traduit, à long terme, par une baisse des ventes dans les points de ventes physique ; le corollaire est une hausse des ventes à distance. En revanche, à court terme, une hausse (ou baisse) des ventes physiques peut annoncer une hausse (baisse, respectivement) des ventes à distance : ces évolutions communes traduisent celle de la consommation des ménages.

1.1.4. Utilisation des données FEVAD des produits pour le total VAD

Pour l'EMC, l'estimation du CA VAD total est la plus importante. Pour l'estimer, outre l'indice de CA de l'ensemble des ventes physiques (cf. Tableau 1), les indices de CA VAD des cinq produits sont utilisés. En effet, par construction, les CA des cinq produits contribuent au CA total ; représentant en moyenne 68% sur l'historique des livraisons. Pour prolonger chaque indice de CA des cinq produits, une modélisation SARIMA est utilisée.

¹² Ces corrélations sont obtenues sur les valeurs différenciées (cf. plus loin, conformément aux données utilisées lors de la modélisation)

1.2. Google Trends

Google Trends fournit les indices mensuels des termes recherchés sur le moteur de recherche Google par les utilisateurs (cf. annexe). Élaborés par Google selon une méthodologie non publique, ces indices sont constitués selon les champs – précisés dans la recherche – géographique (la France ici), temporel (l'historique maximum remonte à 2004), fréquentiel (mensuel ici) et d'appartenance à une catégorie (e.g. « Shopping »). Disponibles si le volume des recherches est « suffisant » (au sens de Google), ces indices sont constitués de valeurs entières comprises entre 0 et 100 et réalisés sur des échantillons des recherches totales effectuées. Outre le fait que la méthodologie de construction des indices Google Trends soit opaque, certains des points précédents appellent à effectuer des tests de robustesse.

1.2.1. L'échantillonnage de Google

Construit sur un échantillon aléatoire des recherches, un indice Google Trends diffère entre deux tirages. Pour vérifier la robustesse de l'outil, nous avons comparé les séries d'un même terme, recherchées plusieurs fois. À titre illustratif, voici la table des corrélations obtenues pour deux tirages distincts réalisés à quelques jours d'intervalle (i.e. à méthodologies Google et Google Trends constantes, a priori) :

Corrélations entre deux tirages (en %)

Amazon	Cdiscount	Fnac	eBay	Carrefour
99,96	99,50	99,95	99,98	99,21

Note : il s'agit de de séries mensuelles de janvier 2004 à février 2018, soit 170 points

Tableau 3 : Source Google Trends - Banque de France, Banque de France DGS SEEC

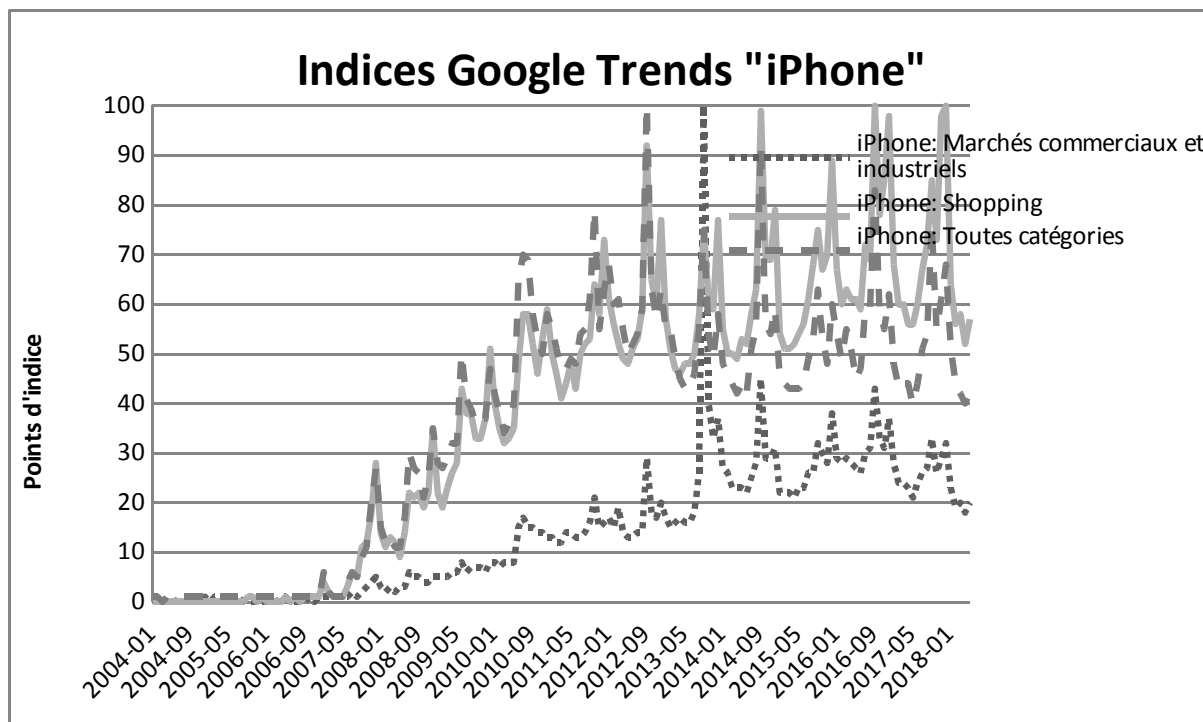
Cette expérience a été répétée de nombreuses fois, sans obtenir de corrélation inférieure à 98 %. Dans ces conditions, la méthode d'échantillonnage paraît suffisamment fiable pour requêter régulièrement les indices Google Trends.

1.2.2. Indices de valeurs entières

Ensuite, il aurait été pratique d'extraire les indices Google Trends simultanément. Mais, lors d'une extraction commune de deux indices (ou plus : jusqu'à cinq maximum *via* l'outil), la valeur 100 est attribuée à la plus grande valeur sur l'historique de l'ensemble des indices ; les autres valeurs des indices sont au *pro rata*. Si les volumes de recherches diffèrent franchement, un indice écrase l'autre et alors l'indice écrasé prend un nombre restreint de valeurs, qui reflètent mal ses variations. Or, dans un modèle statistique, le nombre de décimales des variables, et plus généralement la précision des variables, peut avoir une influence sur l'estimation finale selon Kozicki et Hoffman (2004). Afin d'avoir les valeurs les plus précises possibles chaque série Google Trends est donc extraite individuellement.

1.2.3. Catégorie

L'outil Google Trends répertorie les requêtes Google par catégorie. L'exemple de la requête « iPhone » appelle à la vigilance lors des extractions :



Graphique 2 : Requêtes Google Trends « iPhone ». Source Google Trends

Si la catégorie « Marchés commerciaux et industriels » n'est pas utile pour étudier la VAD, la série tracée illustre l'importance du choix de la catégorie : son maximum, atteint en septembre 2013, n'est pas synonyme d'une explosion des ventes. En l'absence de plus d'information sur les catégories, toutes les requêtes utilisées dans la suite de l'article appartiennent à la catégorie « Shopping », correspondant *a priori* le mieux au e-commerce.

1.2.4. Ruptures de séries

Comme dit précédemment, la méthodologie d'élaboration des indices Google Trends n'est pas publique. Cependant, sur la page d'extraction, deux remarques figurent :

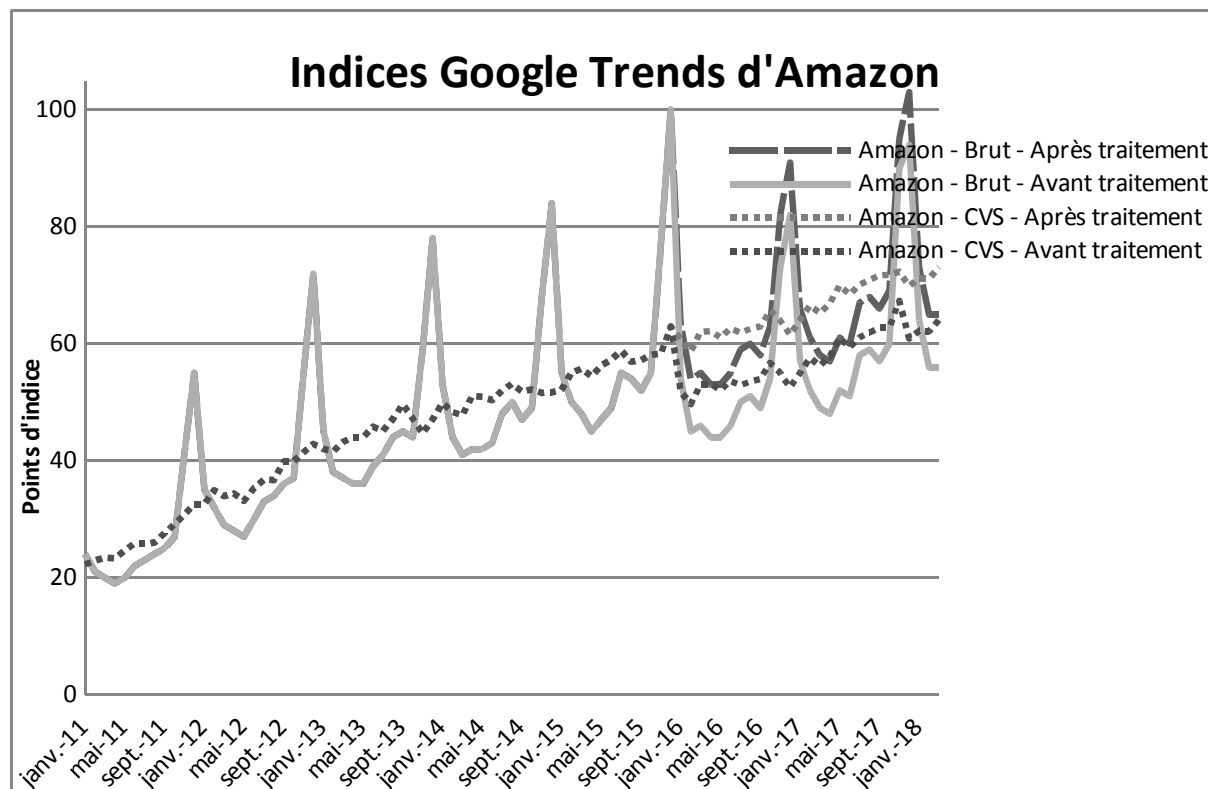
- « La fonctionnalité de détermination de la position géographique a été améliorée. Cette mise à jour a été appliquée à partir du 1^{er} janvier 2011. »
- « Notre système de collecte de données a été amélioré. Cette mise à jour a été appliquée à partir du 01/01/2016. »

Il semble que Google préviennent les utilisateurs de Google Trends lors des principales modifications opérées sur l'outil. Les indices de CA FEVAD démarrant en janvier 2012, la seconde remarque mérite une attention particulière¹³. L'analyse des données Google Trends à l'aide de X13 a permis de détecter un grand nombre de valeurs aberrantes sur les indices Google Trends, notamment en janvier 2016. Le changement méthodologique opéré par Google impacte la qualité des séries Google Trends. N'ayant pas accès à la méthodologie de construction de ces indices, il a été décidé de traiter toutes les valeurs aberrantes, à savoir les changements de niveau (*Level shift*), les variations passagères (*Transitory change*) et les valeurs singulièrement aberrantes (*Additive outlier*) ; excepté

¹³ Afin d'améliorer la robustesse, pour les besoins de cette opération, les indices Google Trends sont extraits depuis janvier 2011. S'il est globalement convenu qu'une dessaisonnalisation ne peut se faire avec un historique inférieur à 3 ans, ajouter un an d'historique permet de stabiliser les séries CVS et, donc, d'améliorer la détection des valeurs aberrantes.

celles traduisant l'apparition d'une nouvelle requête (d'un nouvel acteur, d'une nouvelle marque... ; cf. exemple de la chaussure infra).

La méthode X13 a permis de mieux détecter les valeurs aberrantes des indices Google Trends¹⁴, pas nécessairement explicables ; à l'instar des nombreuses ruptures de séries suite au changement méthodologique opéré par Google en janvier 2016. Dans ces conditions, une correction automatique des valeurs aberrantes a été mise en place, en déterminant le type de valeur aberrante et en analysant son ampleur. Par exemple, pour la série d'Amazon, en janvier 2016, un saut de niveau est détecté ; après évaluation, la série Google Trends d'Amazon peut être corrigée selon le graphique ci-dessous :



Graphique 3 : Source Google Trends - Banque de France DGS SEEC

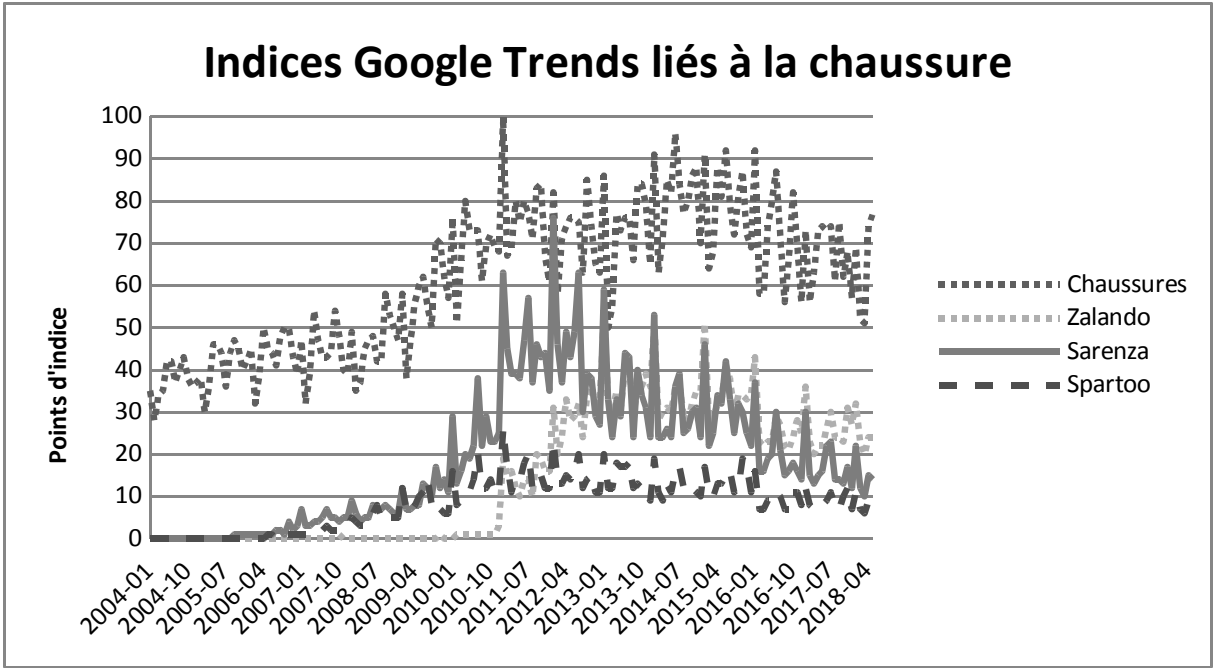
Le saut de niveau (*Level Shift*) intervient lors du changement méthodologique opéré par Google. Le traitement démarre par la dessaisonalisation de l'indice, les valeurs aberrantes étant détectées sur les deux séries : CVS et brute. Le traitement est différencié selon le type de valeur aberrante ; dans le cas d'Amazon, l'estimation du saut de niveau correspond à l'écart entre le point de la série CVS¹⁵ de janvier 2016 et la prévision de la même série tronquée en décembre 2015. Cette correction est alors appliquée au reste de la série, contrairement aux valeurs aberrantes singulières (*Additive Outlier*), dont le traitement est ponctuel.

Comme de nombreux autres indices, celui de Cdiscount, un autre très gros acteur du e-commerce en France, connaît le même problème (cf. annexe). Au vu du nombre d'indices Google Trends (plus de 140), susceptibles d'augmenter avec la mouvance du paysage du e-commerce, l'automatisation du

¹⁴ En effet, si une valeur aberrante peut être détectée sur les indices issus de l'enquête, le SEEC est en mesure de les expliquer (cf. exemple en annexe : juin 2016 pour l'électronique grand public) ; contrairement à celles détectées sur les indices Google Trends

¹⁵ Dans cet exemple, l'estimation du saut de niveau est faite à l'aide des CVS car celle fournie par les données brutes semblait moins cohérente. Néanmoins, la détection s'opère sur les deux indices (brut et CVS), ce qui permet de capter un maximum de valeurs aberrantes.

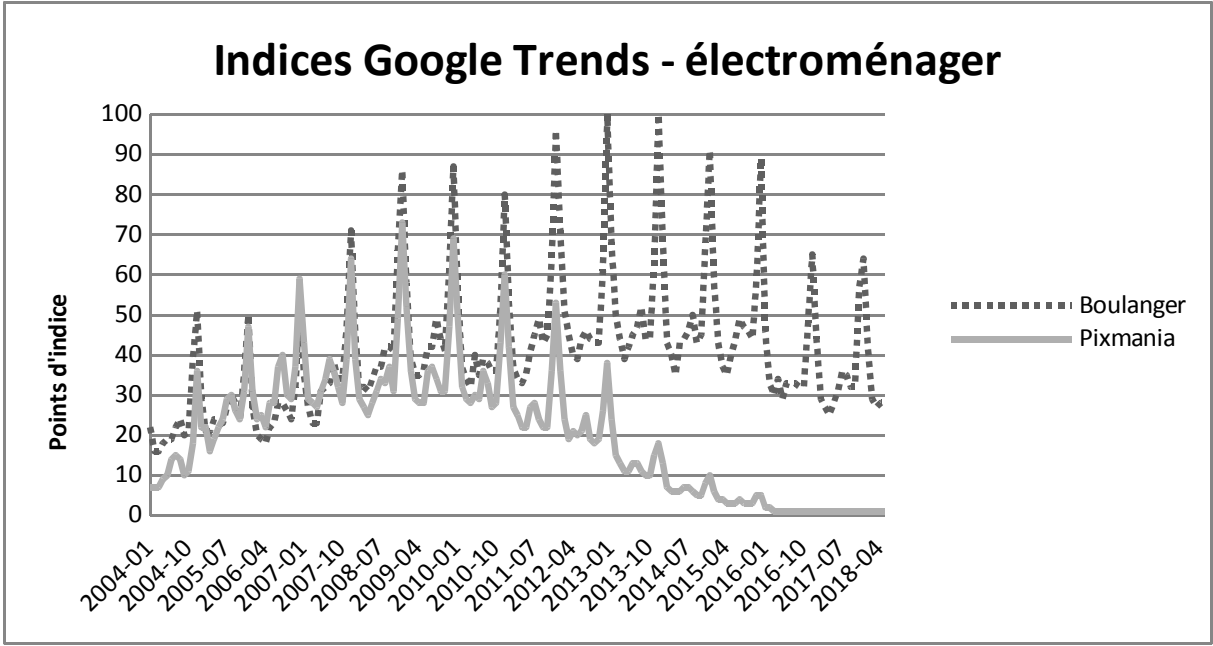
traitement s'est imposée. Seules les valeurs aberrantes traduisant l'apparition de requêtes Google Trends ne sont pas traitées. Effectivement, l'émergence du e-commerce s'accompagne de nouveaux acteurs. Pour les chaussures par exemple, les trois Pure Players (vendant uniquement sur internet) leaders du e-commerce sur le marché français sont relativement récents :



Graphique 4 : Indices Google Trends liés à la chaussure - Source : Google Trends

La croissance de l'indice « Chaussures » de 2004 à 2011 correspond à l'émergence des sites internet vendeurs de chaussures. Graphiquement, le lancement de Zalando en France, en décembre 2010, est très clair (l'indice passe de 1 à 19 en deux mois : 11/2010-01/2011).

À l'inverse, certains acteurs du e-commerce, présents au début, immergent. Dans le domaine de l'électroménager, l'indice Google Trends de Pixmania en témoigne :



Graphique 5 : Source Google Trends

Si la variable présentait un intérêt il y a quelques années, il semble que ce spécialiste du e-commerce pour l'électroménager et l'EGP soit devancé par la concurrence, par exemple Boulanger.

Le e-commerce connaît une évolution permanente, bien retranscrite par les indices Google Trends. Dans cet univers et dans le contexte de nos travaux, il est impérieux de régulièrement réviser les variables utilisées, en particulier celles concernant les acteurs du e-commerce pour les différents produits. L'aspect évolutif des termes recherchés ne doit pas être négligé.

La fusion de la Fnac avec Darty en est un autre exemple : désormais, l'indice Google Trends associé est « Groupe Fnac Darty ». Par ailleurs, la tendance de popularité d'un site internet n'est pas nécessairement la même que celle de l'indice Google Trends associé car tous les internautes ne passent pas par Google : les modes de consultation des sites internet évoluent, notamment avec l'émergence du m-commerce¹⁶ où les applications évitent l'utilisation du moteur de recherche.

1.1.1. Listes de variables

Voici les jeux de variables initiaux, *i.e.* avant sélection, cette étape faisant partie intégrante de la modélisation. Pour le total, cette pré-sélection correspond essentiellement aux acteurs majeurs du e-commerce en France. Les pré-sélections des cinq produits sont un mélange de pure players (ex : Sarenza, pour les chaussures), d'enseignes (La Halle), de termes génériques (ex : Chaussures femme) et de marques (Converse). Des travaux préliminaires tels qu'une recherche documentaire sur les produits concernés par l'estimation ou un sondage auprès d'utilisateurs sur les comportements avant un achat, ont été menés. Ils ont conduit à retenir des listes de variables hétérogènes :

Total	Habillement	EGP	Chaussures	Électroménager	Meubles
Amazon	Vertbaudet	iPhone	Chaussures	Clubic	But
eBay	Kiabi	Apple	Chaussure	Boulanger	Legallais
Vente privée	H&M	Cdiscount	Ceinture	Cdiscount	Cuisine
Cdiscount	C&A	PC Gaming	Maroquinerie	Four	Raja
Fnac	Jules	iPad	Botte	Réfrigérateur	Staples
Groupe Fnac Darty	Linge de maison	Groupe Samsung	Chaussures de sport	Machine à laver	Leroy Merlin
PriceMinister	Zara	FNAC	Vans	Darty	Castorama
Leroy Merlin	Costume	Télévision	Converse	Bosch	Conforama
UGAP	Culotte	Boulanger	Zalando	Electrolux	Vega
Castorama	3 Suisses	Sony	Spartoo	Conforama	Bureau
Boulanger	Devred	LDLC Pro	Sarenza	Amazon	Meuble
Carrefour	Robe	Amazon	Showroomprive	Cuisinière	Roche Bobois
Showroomprive	Etam	Phillips	Prada	Électro-dépôt	Armoire + étagère
E. Leclerc	La Redoute	LG Group	Escarpin	Brandt	Ikea
La Redoute	Jeans + Chino + Pantalon	Samsung Electronics	Adidas Stan Smith	Four à micro-ondes	Couteaux
Auchan	Manteau + Blouson	Téléphone + smartphone	Chaussures femme	Groupe Fnac Darty	Maisons du Monde
Raja	Veste	Darty	Ballerine	Aspirateur	Cinna
Rue du commerce	Vêtement femme	Tablette tactile	Chaussures homme	Whirlpool Corporation	Meuble en bois
3 Suisses	ASOS	Enceinte	Timberland	Mistergooddeal	Roset
Promos + soldes + blackfriday	Maisons du Monde	Appareil photographique reflex numérique	Chaussures Nike	GrosBill	Buffet + commode + vaisselier

¹⁶ Selon la FEVAD, 36,6 millions de Français achètent sur internet, dont 9,3 millions ont déjà effectué un achat à partir de leur mobile.

Alibaba	Lingerie	Ordinateur portable +PC	Chaussures enfant	Pulsat	Table + chaise + canapé
Groupon	Jennyfer	Bose	San Marina	Ubaldi	Fauteuil
PhotoBox	Vêtement	JBL	Eram	But	
Galeries Lafayette	Galeries Lafayette	Groupe Fnac Darty	Chaussures de ville		
Yves Rocher	Bonobo	Barre de son	J.M Weston		
Sephora	Brandalley	Appareil photo	Chaussea		
Decathlon	Camaïeu	Marshall	Bexley		
	Showroomprive		Gémo		
	Vente Privée		Sac à main		
	Rideau		La Halle		
	Blanche Porte		Chaussures de foot		
	Drap				
	Coussin				
	Homemaison				
	Sous-vêtement				
	La Halle				
	Decathlon				

Tableau 4 : Source Banque de France DGS SEEC

Note de lecture : l'utilisation du « + » permet de constituer un indice Google Trends correspondant au cumul des requêtes

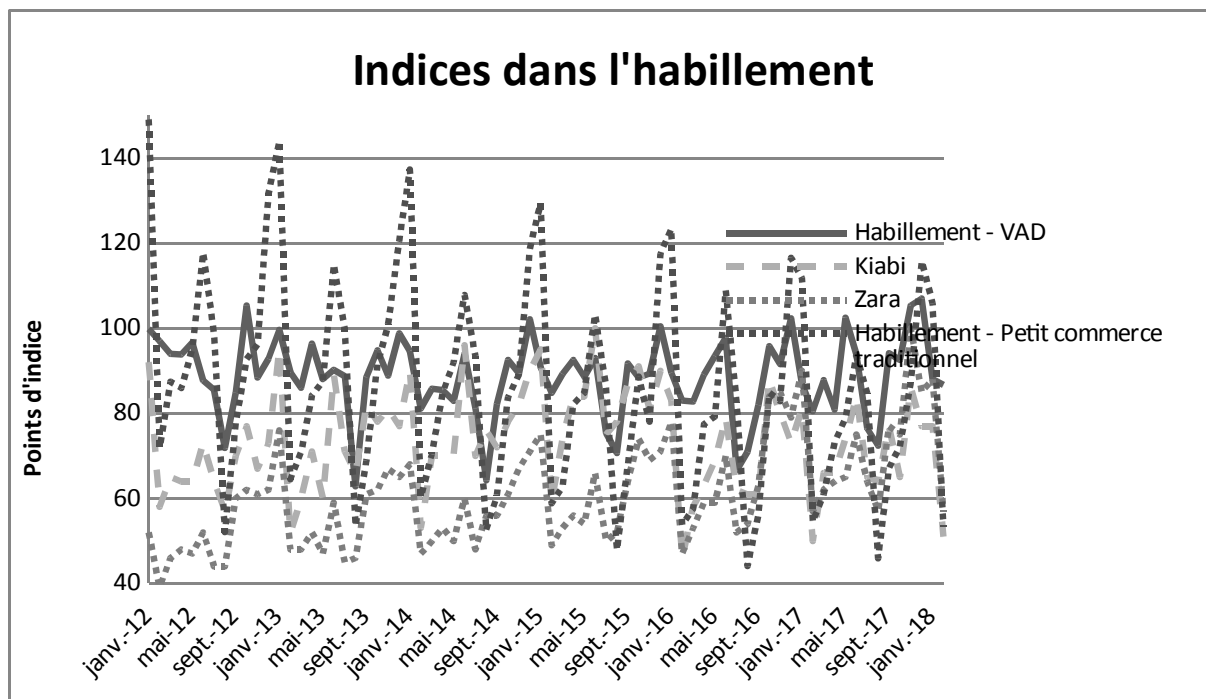
2. Modélisation

2.1. Traitement de la stationnarité et saisonnalité

La plupart des séries ne sont pas stationnaires mais plutôt intégrées d'ordre 1 : la différenciation s'impose. Cette opération classique participe à éviter les régressions fallacieuses (cf. Phillips (1986)), un phénomène fréquent lors de régressions entre séries temporelles, traduit par des résultats trop optimistes signifiés par un R^2 anormalement élevé (cf. Granger et Newbold (1974)). L'introduction d'une variable mesurant la tendance (Phillips et Perron (1988)) ou de termes autorégressifs y participent aussi.

Dans le but de mieux mesurer la tendance du e-commerce, il a été question de travailler avec les séries corrigées des variations saisonnières (CVS). Cette solution n'a pas été retenue du fait de l'instabilité de la dessaisonalisation sur les derniers points. D'une part, les historiques courts des séries n'assurent pas une dessaisonalisation de qualité sur toutes les séries¹⁷, surtout lors des premières estimations (36 points à la première itération ; plus de 70 maintenant) ; d'autant que, du fait de son émergence, le e-commerce connaît des changements de saisonnalité :

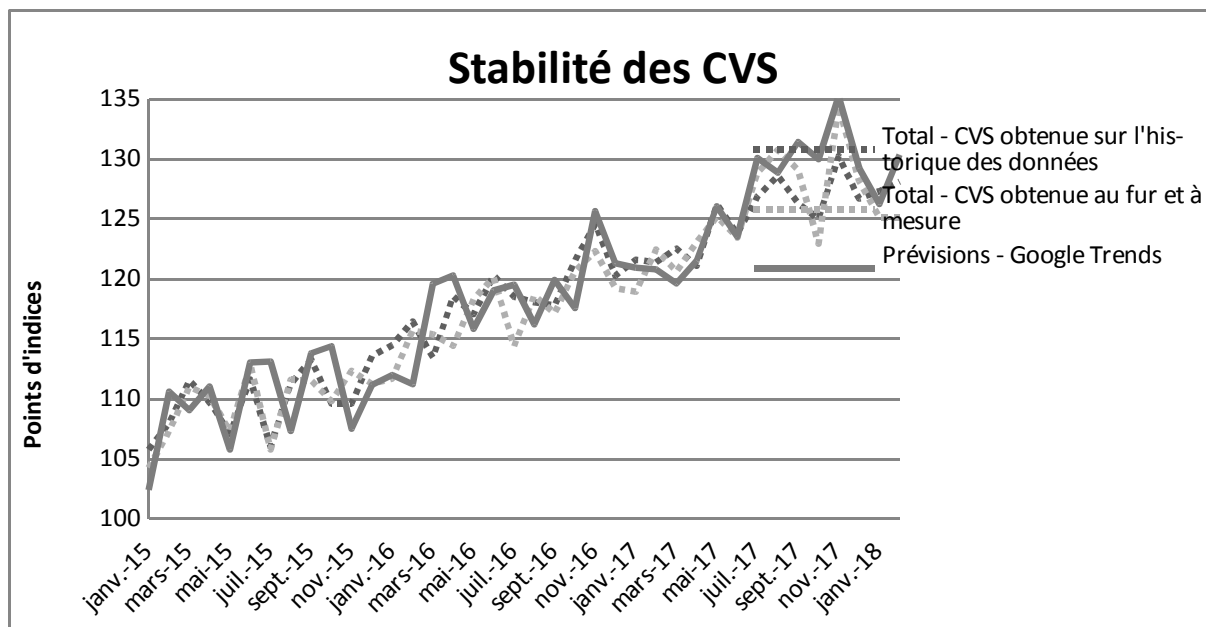
¹⁷ Plus de 150 séries réparties sur les 6 produits.



Graphique 6: Source Google Trends - FEVAD - Banque de France DGS SEEC

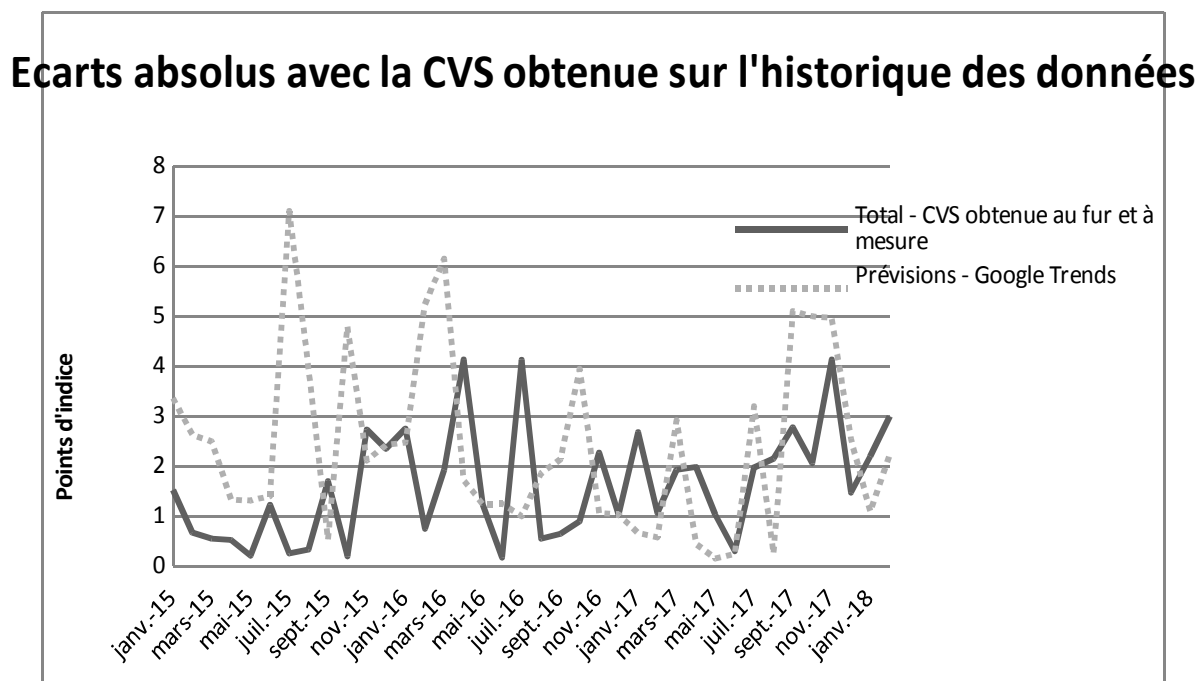
Sur le graphique ci-dessus sont représentées les séries brutes deux indices de CA de l'habillement (VAD et Petit commerce traditionnel) et deux requêtes Google Trends associées au produit : « Kiabi » et « Zara ». L'indice de la VAD connaît des changements de saisonnalités : par exemple, dans les premières années, le mois de juillet est nettement au-dessus du mois d'août. En 2015, l'écart entre les deux mois se réduit et, en 2016, le mois de juillet est plus bas. L'allure générale de la série illustre bien les changements de saisonnalités. Ce phénomène est commun aux indices Google Trends. Par exemple, pour Zara, le maximum annuel est atteint au mois de janvier pour les années allant de 2013 à 2016 ; hors, la valeur du mois de novembre 2017 est supérieure à celles des mois de janvier 2017 et 2018. De même, la série de Kiabi ne présente pas une stabilité saisonnière remarquable. Dans ces conditions, la dessaisonalisation de nombreuses séries est de qualité incertaine. En revanche, les saisonnalités de l'indice du Petit commerce traditionnel sont stables. Les changements sont plus rares pour des séries bien établies (l'indice a démarré en 1990).

D'autre part, les derniers points d'une série CVS sont les plus susceptibles d'être modifiés par l'apparition de nouveaux points : à chaque livraison de la FEVAD, lorsqu'il est possible d'évaluer la prévision précédente, les derniers points de la série CVS changent ; pouvant influencer fortement le modèle.



Graphique 7 : Source Google Trends - FEVAD - Banque de France DGS SEEC

Les deux séries en pointillés sont les indices CVS de CA du total des VAD : le premier (dans l'ordre de la légende) est obtenu aujourd'hui, c.-à-d. en disposant de l'ensemble des données. Le second correspond à la série des derniers points obtenus avec les données alors disponibles. Par exemple, le point de janvier 2017 est le dernier point de l'indice CVS obtenu avec les données allant jusqu'à janvier 2017. Ces deux indices diffèrent dans des proportions comparables aux écarts de prévision des modèles : à titre illustratif, l'indice des prévisions du modèle Google Trends en rend compte. L'évolution des écarts avec la série CVS la plus précise, *i.e.* obtenue sur l'historique des données, le montre plus clairement :



Graphique 8 : Source Google Trends - FEVAD - Banque de France DGS SEEC

Note de lecture : avec les données disponibles aujourd'hui, le point de juillet 2016 bouge de 4,1 points d'indice. La prévision Google Trends de juillet 2016 est même plus proche de la valeur de l'indice CVS obtenu aujourd'hui que celle obtenue avec les données disponibles en juillet 2016.

Les changements de saisonnalités – questionnant la qualité des CVS – et l’instabilité des CVS sur les derniers points – les plus importants dans le cadre d’un exercice de *nowcasting* – conduit à privilégier une modélisation des données brutes différenciées.

2.2. Processus d’estimation et d’évaluation de la performance

2.2.1. Modèles

Jusqu’ici, un modèle SARIMA était utilisé pour chaque produit. Faisant office de référence par la suite, il est toujours utilisé. Par ailleurs, l’adaptive lasso est utilisé dans deux modèles, implémentés sur chaque produit :

- Modèle « Google Trends », utilisant les Google Trends (cf. Tableau 4)
- Modèle « CD », basé sur les indices quantitatifs conjoncturels issus de l’enquête CD¹⁸ (cf. Tableau 1)

Outre les variables exogènes, une tendance et une composante autorégressive font aussi partie du jeu de variables initiales. Le e-commerce étant en pleine croissance (*a priori* non linéaire), il a été décidé d’intégrer une variable de tendance dans la liste du jeu de variables initial. Et, plus qu’une composante autorégressive, il s’agit de la modélisation SARIMA de l’indice de CA. Concrètement, l’estimation se fait en deux étapes : la première consiste à modéliser l’indice par un processus autorégressif (SARIMA) ; cette estimation est une variable potentiellement sélectionnée par l’algorithme de l’adaptive lasso, au même titre que la tendance et les variables exogènes (indices Google Trends et/ou indices quantitatifs conjoncturels).

À ce stade, trois modèles fournissent une prévision distincte : le modèle SARIMA seul et les deux modèles d’adaptive lasso, Google Trends et CD.

Ensuite, ces trois modèles sont utilisés dans l’agrégation bayésienne (« Modèle agrégé » dans la suite) ; donnant lieu à une quatrième prévision.

Un cinquième modèle, appelé modèle global, permet de contribuer au débat sur la combinaison d’information. La sélection de variables est alors appliquée à toutes les variables, à savoir les données Google Trends, celles collectées par la Banque de France, la tendance et la composante autorégressive. Ce modèle peut, à certaines itérations, être identique à l’un des modèles individuels (*e.g.* si aucune requête Google Trends n’est sélectionnée).

1.1. Protocole de test

À chaque itération du protocole de test, c.-à-d. chaque mois, les conditions réelles sont répliquées. Plus précisément, les valeurs des données Google Trends et des indices quantitatifs conjoncturels issus de l’enquête du mois M sont connues, contrairement aux données FEVAD.

La première étape est l’estimation SARIMA. Outre l’obtention de sa propre prévision, cette opération sert aussi à déterminer la variable utilisée dans les modèles d’adaptive lasso. Dans un second temps, ces trois modèles (Google Trends, CD et global) sont élaborés. Le modèle agrégé ne peut être construit qu’après les modèles SARIMA, Google Trends et CD.

¹⁸ Comme dit précédemment, dans le cas du total, les indices de CA VAD sont aussi utilisés

La qualité de chaque modèle est jugée après livraison des données FEVAD puisque le critère de jugement, dans le cadre d'un exercice de *nowcasting*, est la capacité de prévision. L'indicateur est donc la RMSFE (*Root Mean Squared Forward Error*), soit l'écart-type des erreurs de prévision, mesure de l'erreur *Out Of Sample*. Sont aussi présentées les RMSE (*Root Mean Squared Error*), mesure de l'erreur *In Sample*, car elles aident à comprendre les pondérations dans le modèle agrégé.

Par ailleurs, chaque mois, la fenêtre de calibration des modèles s'agrandit d'un point. Du fait des courts échantillons à disposition, travailler avec une fenêtre extensible plutôt qu'avec une fenêtre glissante pour l'échantillon d'estimation participe à la stabilisation des modèles.

Les données FEVAD sont livrées depuis janvier 2012. La différenciation des données mène à février 2012. Avec un historique minimal de 3 ans pour assurer la robustesse de l'estimation, la première prévision est celle de février 2015.

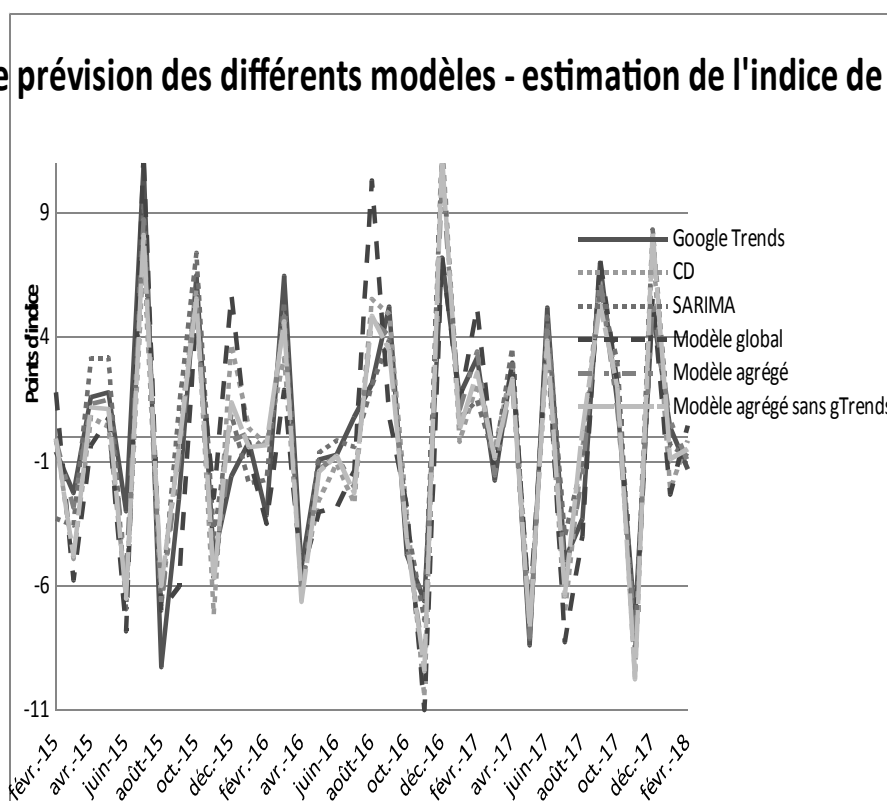
2. Résultats

Les résultats pour le total seront détaillés. Ceux pour les produits sont exposés de façon plus synthétique.

2.1. Total

Conformément à l'objectif de l'étude, les erreurs de prévision (*Out Of Sample*) constituent un résultat important :

Erreurs de prévision des différents modèles - estimation de l'indice de CA du total



Graphique 9 : Source Google Trends - FEVAD - Banque de France DGS SEEC

Note de lecture : L'erreur de prévision pour le mois de décembre 2015 du modèle global est de 5,7 points d'indice : après livraison des données FEVAD, la valeur de l'indice de CA du total était supérieure (de 5,7 points) à la prévision du modèle global.

Le graphique précédent montre les erreurs de prévision des différents modèles (en annexe, le graphique des valeurs prédites par les différents modèles *versus* les vraies valeurs). Graphiquement, les résultats sont proches. Les RMSFE et les moyennes des erreurs de prévision (en absolu) permettent de mieux appréhender les résultats :

Total	Google Trends	CD	SARIMA	Modèle global	Modèle agrégé	Modèle agrégé sans gTrends ¹⁹
RMSFE	4,8	5,2	5,0	5,5	4,8	5,0
Moyenne des erreurs de prévision absolues	3,9	4,0	3,9	4,5	3,8	3,9

Tableau 5 : Source Google Trends - Banque de France DGS SEEC

Au sens de la RMSFE, qui reste l'indicateur privilégié, le modèle Google Trends est le plus performant avec le modèle agrégé (4,8). Ici, les moins bonnes performances du modèle agrégé sans les données Google justifie l'apport de Google Trends. Le modèle agrégé est aussi meilleur au sens de la moyenne des erreurs absolues. Cette mesure de l'erreur est intéressante car l'un des objectifs de l'agrégation est aussi de minimiser les gros écarts de prévision. Les résultats des modèles individuels sont relativement proches.

Avant de comparer les deux modèles construits pour la prévision finale, c.-à-d. les modèles disposant de toute l'information (le modèle agrégé et le modèle global), les résultats des modèles individuels méritent d'être détaillés ; particulièrement le modèle Google Trends qui doit apporter d'autres garanties, en termes de parcimonie et de stabilité.

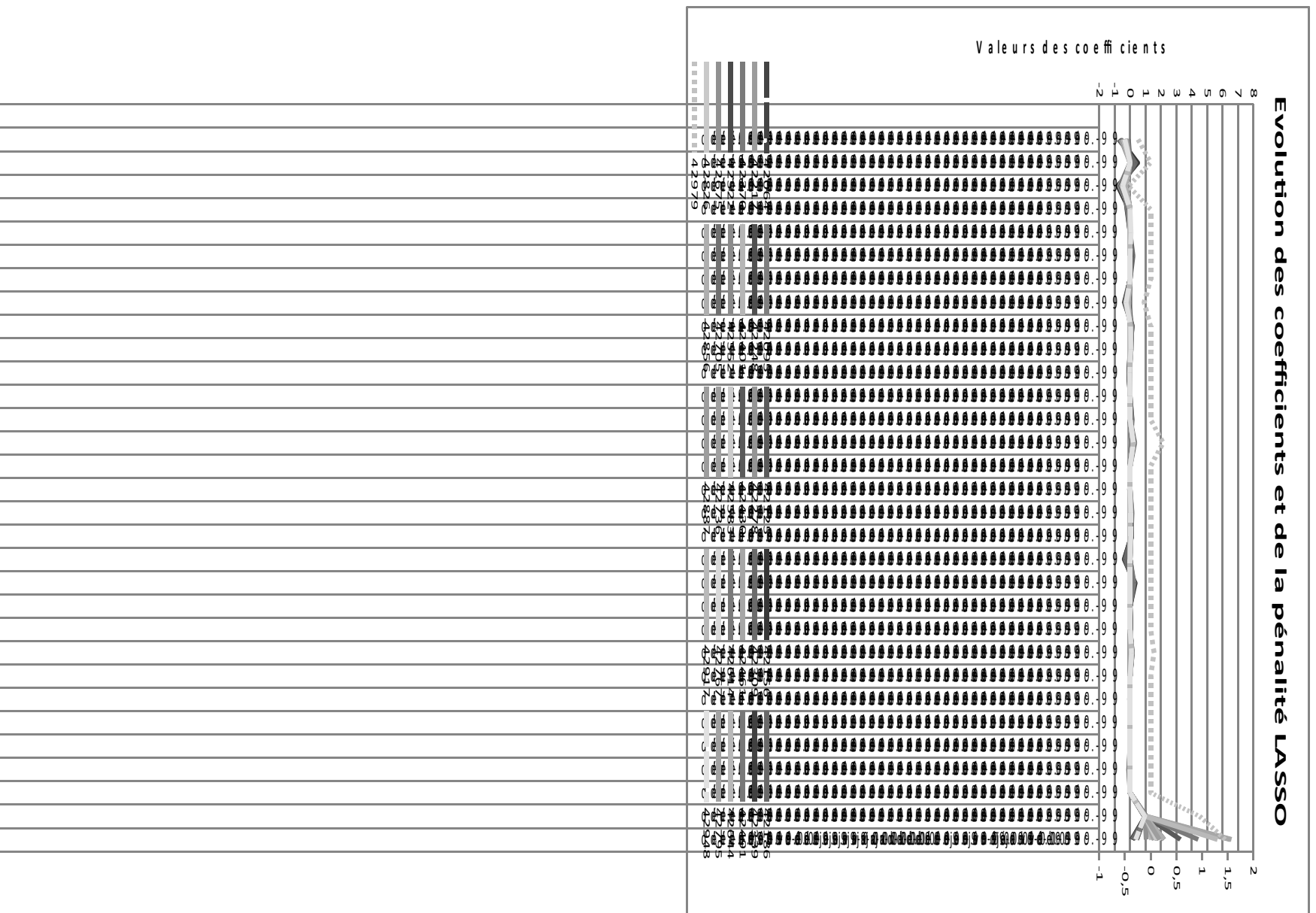
5.1.2. Modèle SARIMA

Faisant office de référence comme habituellement dans la littérature, le modèle SARIMA présente de bonnes performances. Cependant, il lui arrive d'être moins bon que les autres modèles. Par exemple, en août 2016, les données exogènes apportent une réelle information.

5.1.3. Modèle Google Trends

Conformément au protocole de test, la sélection de variables *via* l'adaptive lasso s'effectue à chaque itération. Donc, les coefficients du modèle évoluent dans le temps :

¹⁹ Le modèle agrégé sans gTrends (Google Trends) correspond à l'agrégation du modèle CD et du modèle SARIMA. Il permet notamment de juger l'apport des données Google Trends. Cependant, la variable SARIMA étant présente dans tous les modèles CD, l'agrégation perd de son sens ; ainsi, il ne sera pas présenté lors des résultats obtenus sur les produits.



Graphique 10 : source Google Trends - Banque de France DGS SEEC

Le graphique précédent montre l'évolution temporelle des coefficients. En axe secondaire, l'évolution de la pénalité lasso. Seules la variable SARIMA (proche de 1), la constante et la pénalité lasso ne sont pas en trait plein. S'il n'est pas courant de regarder l'évolution de la pénalité lasso au

cours du temps, puisqu'il s'agit d'une optimisation différente à chaque itération, celle-ci permet d'expliquer l'évolution du nombre de variables retenues : plus elle est faible, plus le nombre de requêtes Google Trends retenues est élevé. Ensuite, concernant la variable SARIMA, il était attendu que son coefficient soit proche de 1 puisqu'elle correspond à la modélisation autorégressive de la variable. Du reste, malgré la lisibilité restreinte du reste du graphique, la stabilité des coefficients des variables Google Trends – mise en relief avec l'évolution de la pénalité du lasso – est encourageante car elle signifie que ces variables modélisent une partie de l'information non captée par la composante SARIMA. Ci-dessous, le tableau des valeurs moyennes, minimums et maximums obtenues pour chaque variable :

	Moyenn e	Minimu m	Maximu m
Amazon	0,07	0,00	0,48
eBay	-0,43	-0,80	-0,24
Vente.privee.com	-0,02	-0,21	0,00
Cdiscount	0,01	0,00	0,07
FNAC	0,01	0,00	0,18
Groupe.Fnac.Darty	0,00	0,00	0,00
PriceMinister	-0,21	-0,37	0,00
Leroy.Merlin	0,05	0,00	0,13
Union.des.groupements.d.achats.public s	0,01	0,00	0,06
Castorama	-0,02	-0,09	0,00
Boulangier	0,00	0,00	0,00
Carrefour	0,01	0,00	0,12
Showroomprive.com	0,12	0,00	0,29
E.Leclerc	0,00	0,00	0,00
La.Redoute	0,00	-0,04	0,00
Auchan	0,01	0,00	0,11
Raja	0,02	0,00	0,10
Rue.du.Commerce	-0,05	-0,38	0,00
X3.Suisses	0,06	0,00	0,33
Promos + soldes + blackfriday	0,00	0,00	0,00
Alibaba.Group	0,00	0,00	0,05
Groupon	0,08	0,00	0,14
PhotoBox	0,00	0,00	0,00
Galleries.Lafayette	0,00	-0,02	0,00
Yves.Rocher	0,00	0,00	0,00
Sephora	0,00	0,00	0,04
Decathlon	0,00	-0,10	0,00
trend	0,00	0,00	0,00
SARIMA	0,97	0,93	1,04

Tableau 6 : Source Google Trends - FEVAD - Banque de France DGS SEEC

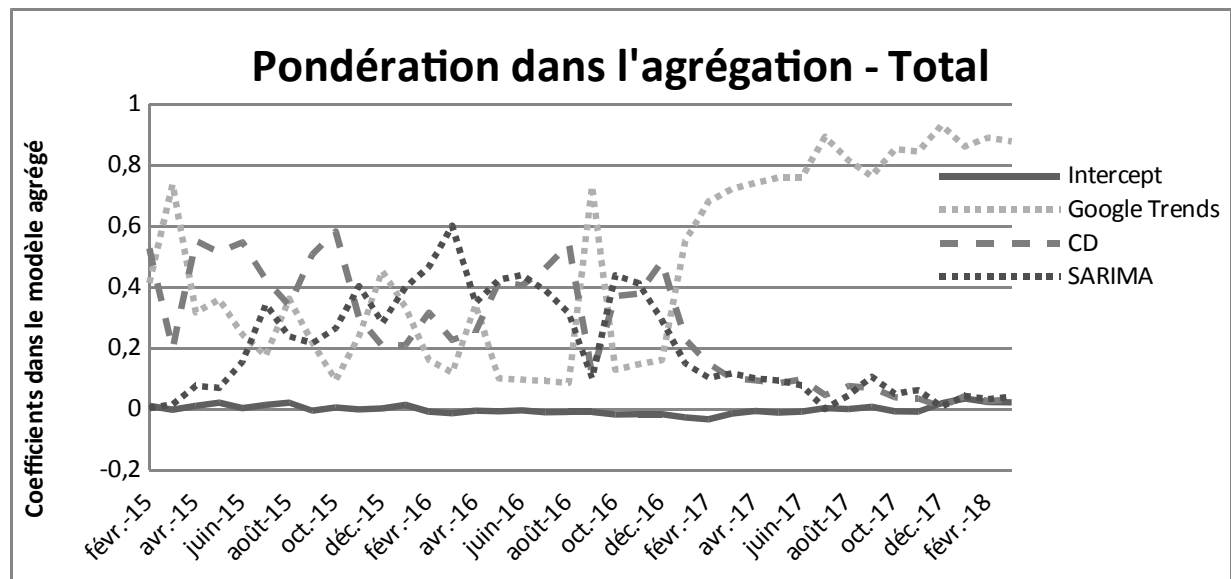
Concernant la sélection, près de 9 variables sont retenues en moyenne à chaque itération ; ce qui est acceptable au vu de la taille des échantillons (36 points à la première itération, 72 à la dernière). Les variables Google Trends les plus retenues sont eBay, PriceMinister, Groupon, Showroomprivé et Leroy Merlin (cf. tableau du nombre de sélections des variables en annexe).

5.1.4. Modèle Commerce de détail

Pour rappel, les variables utilisées sont l'indice de l'ensemble des ventes physiques et les indices de CA VAD des cinq produits prolongés au dernier point par un SARIMA. L'évolution des coefficients est donnée en annexe. Le modèle est parcimonieux, sélectionnant une à deux variables en plus de la composante SARIMA. L'indice de CA VAD de l'habillement est systématiquement et logiquement – en moyenne, le montant de CA de l'habillement représente 22% du total, soit le plus gros des cinq produits – sélectionné. En termes de prévision, il est moins bon que les modèles Google Trends et SARIMA.

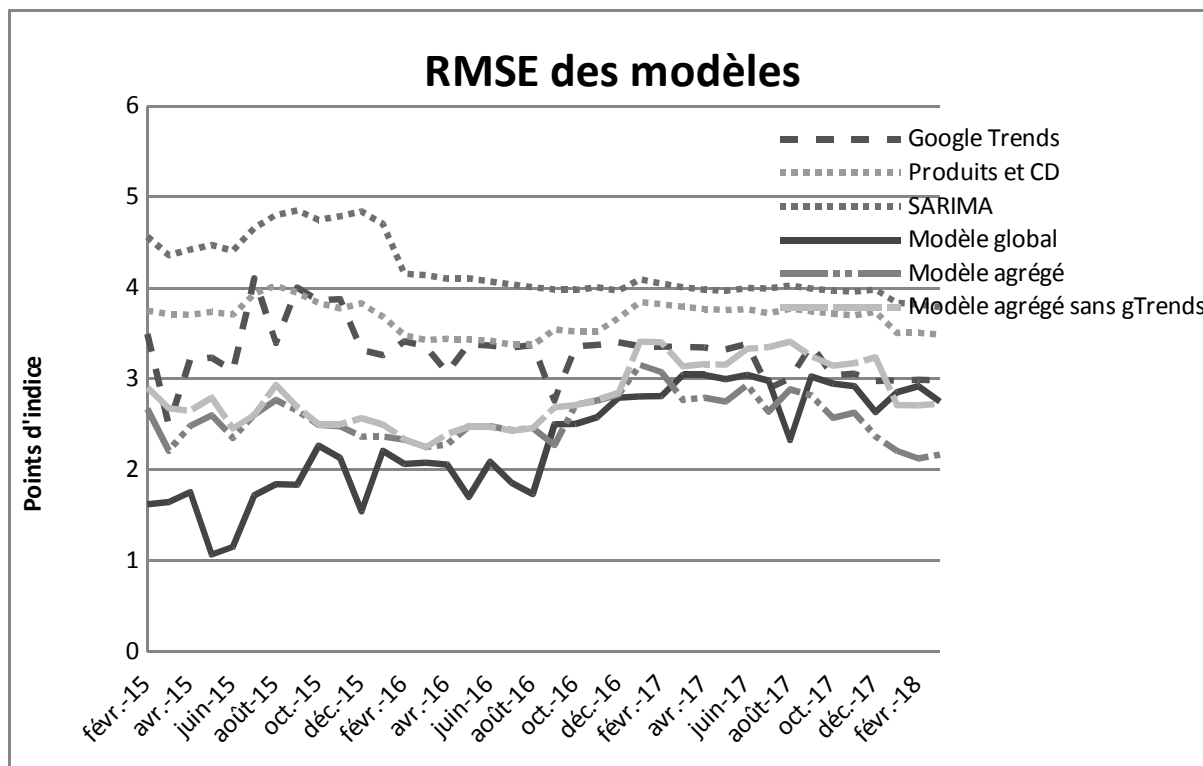
5.1.5. Modèle agrégé

Le modèle agrégé présente les meilleures performances prédictives en moyenne selon les deux indicateurs, la RMSFE et l'erreur moyenne absolue. Au cours du temps, il n'a jamais la moins bonne prévision. Les poids des modèles permettent de juger sa stabilité :



Graphique 11 : Source Google Trends - FEVAD - Banque de France DGS SEEC

L'augmentation des poids du modèle Google Trends depuis la fin 2016 est un élément encourageant. En moyenne, son poids (0,48) est supérieur à celui des deux autres modèles, le SARIMA (0,26) et CD (0,20). L'évolution des erreurs de prévision des modèles Google Trends, CD et SARIMA éclaire celle des poids. La proximité des erreurs de prévision des trois modèles s'explique principalement par la présence de la variable SARIMA dans les modèles Google Trends et CD. Néanmoins, certaines différences méritent une attention particulière. Par exemple, au mois d'août 2016, le modèle CD joue le rôle principal dans l'agrégation, avec un poids de 0,54, contre 0,32 pour le modèle SARIMA et 0,09 pour le modèle Google Trends. Lors de la livraison des données FEVAD fin septembre 2016, il est possible de confronter les prévisions avec la valeur réelle. Le graphique des erreurs de prévision montre que le modèle CD est le moins bon des trois avec une erreur de 5,5 points d'indice, contre 2,1 pour les modèles SARIMA et Google Trends. Une fois l'erreur apprise, le mois suivant, la pondération change drastiquement : le poids du modèle CD passe à 0,11, contre 0,10 pour le modèle SARIMA et 0,73 pour le modèle Google Trends. L'agrégation bayésienne fait partie des algorithmes d'apprentissage statistique. Par ailleurs, apprendre de ses erreurs n'est pas le seul critère de détermination des poids dans l'agrégation : l'erreur *In Sample* y contribue aussi. Ci-dessous, le graphique de l'évolution des RMSE :



Graphique 12 : Source Google Trends - Banque de France DGS SEEC

Note de lecture : lors de la calibration des modèles pour la prévision de décembre 2015, la RMSE la plus faible était celle du modèle global (1,5). Le modèle SARIMA présentait alors la plus grosse erreur *In Sample* (4,8).

À chaque itération, il est possible de calculer la RMSE réalisée sur l'échantillon de calibration du modèle. Le modèle SARIMA présente la plus mauvaise erreur *In Sample*, contrastant avec ses bonnes capacités prédictives. Ce graphique permet aussi de voir que l'agrégation d'information permet de diminuer l'erreur sur l'échantillon de calibration du modèle, par rapport à ses composantes. Ensuite, si les évolutions des erreurs *In Sample* des modèles agrégés (avec et sans Google Trends) sont semblables, la performance prédictive est meilleure avec l'apport des données Google Trends. Enfin, le modèle Google Trends présente des erreurs proches des autres modèles, confortant l'idée que le nombre de variables sélectionnées par l'adaptive lasso est convenable et qu'il n'y a pas de sur-apprentissage. Pour s'en assurer, les erreurs *In Sample* (RMSE) peuvent être mises en perspective avec celles réalisées *Out Of Sample* (RMSFE) :

	Google Trends	CD	SARIMA	Modèle global	Modèle agrégé	Modèle agrégé sans gTrends
Moyenne des RMSE	3,29	3,67	4,20	2,31	2,56	2,81
RMSFE	4,79	5,19	5,00	5,47	4,80	5,00

Tableau 7 : Source Google Trends - Banque de France DGS SEEC

Logiquement, les erreurs de prévisions sont plus grandes. Ensuite, les classements des modèles sont respectés lors du passage de la RMSE à la RMSFE, sauf pour le modèle global : contrairement aux autres, son erreur fait plus que doubler hors échantillon.

5.1.6. Modèle global

Ce phénomène s'explique probablement par un sur-apprentissage. Effectivement, bien que la procédure de l'adaptive lasso est la même que pour les modèles Google Trends et CD, le modèle global est moins parcimonieux : en moyenne, 13 variables sont sélectionnées ; ce qui est probablement trop, en tous cas au début (36 points à la première itération). Il retient plus de variables que les modèles Google Trends et CD cumulés. Plus précisément, sur la période du protocole de test, 82% des variables sélectionnées dans le modèle global le sont dans l'un des deux autres modèles ; 12% ne sont sélectionnées que dans le modèle global et les 6% restant concernent les variables sélectionnées par les modèles Google Trends ou CD et non sélectionnées par le modèle global. En somme, la sélection de variables du modèle global est trop large ; engendrant un phénomène de sur-apprentissage. Du reste, l'évolution des coefficients (cf. graphique en annexe) est moins stable.

Dans le cas du total, le modèle global est moins performant que le modèle agrégé. En outre, le modèle agrégé est lisible et ses composantes sont des modèles stables ; contrairement au modèle global dont l'évolution des coefficients le rend moins interprétable que le modèle agrégé. L'utilisation du modèle agrégé est donc favorisée. Si les résultats obtenus pour le total ont été largement détaillés, ceux des produits sont présentés plus succinctement.

2.2. Produits

2.2.1. Parcimonie

L'adaptive lasso vise la parcimonie des modèles. Ci-dessous, pour chaque produit, le tableau du nombre moyen de variables retenues par modèle (concerné par la sélection de variables) :

	Google Trends	CD	Modèle global
Chaussures	9,7	2,2	10
Meubles	10,3	2,4	10,9
Électroménager	9	2,1	5,9
EGP	8,8	2	10,5
Habillement	12,8	3,6	8,7

Tableau 8 : Source Google Trends - Banque de France DGS SEEC

Les modèles CD sont (par construction) les plus parcimonieux ; le plus souvent, une variable issue de l'enquête est sélectionnée, en plus de la composante SARIMA. Les modèles Google Trends sont moins parcimonieux ; le nombre de variables sélectionnées reste correct au vu de la taille des échantillons, à part peut-être pour l'habillement.

Dans les modèles Google Trends des produits, outre la composante SARIMA et la constante (systématiquement sélectionnées), les cinq variables les plus retenues (sur 38 itérations) sont :

Chaussures	Spartoo (38)	Sarenza (36)	Converse (36)	Chaussures de ville (32)	Chaussures de foot (28)
Meubles	Cinna (38)	Roset (33)	Meuble en bois (32)	Buffet + commode + vaisselier (31)	IKEA (26)
Électroménager	Machine à laver le linge (37)	Four (28)	Cuisinière (28)	Conforama (26)	GrosBill (24)
EGP	Appareil photo reflex numérique	Télévision (35)	JBL (35)	Sony (27)	Samsung Electronics (23)

	(37)				
Habillement	Costume (34)	Décoration (29)	Jennyfer (28)	Lingerie (27)	3 Suisses (25) et Vêtements femme (25)

Tableau 9 : Source Google Trends - FEVAD - DGS SEEC Banque de France

Le tableau précédent illustre l'hétérogénéité des requêtes Google le plus souvent sélectionnées dans les modèles Google Trends : articles (four, télévision), marque (Cinna, Samsung), requête générale (vêtements femme, chaussures de foot), pure player (Spartoo, GrosBill) et spécialiste de la VAD (3 Suisses)... La diversité des comportements des utilisateurs du moteur de recherche Google est bien retranscrite ici. À noter que la variable de tendance n'est jamais sélectionnée.

Contrairement au cas du total, pour chacun des produits, le modèle global est plus parcimonieux que le modèle Google Trends, réduisant ainsi le risque de sur-apprentissage. En premier lieu, le tableau des moyennes des RMSE :

Moyenne des RMSE	Google Trends	CD	SARIMA	Modèle global	Modèle agrégé
Chaussures	8,2	10,5	10,9	7,8	7,6
Meubles	6,0	7,3	7,4	5,7	5,5
Électroménager	6,1	6,9	7,2	6,4	5,5
EGP	5,8	7,4	7,7	5,5	7,2
Habillement	5,3	6,0	6,3	5,9	4,3

Tableau 10 : Source Google Trends - Banque de France DGS SEEC

Conformément aux attentes, les modèles disposant de toute l'information sont globalement mieux calibrés que les modèles à une seule source d'information (Google Trends, CD, SARIMA). Le second enseignement de ce tableau est que le modèle Google Trends est systématiquement meilleur que le modèle CD et le modèle SARIMA ; ce qui peut s'expliquer par un nombre plus grand de variables retenues.

2.2.2. Capacités prédictives

RMSFE	Google Trends	CD	SARIMA	Modèle global	Modèle agrégé
Chaussures	13,2	12,7	13,6	13,8	13,4
Meubles	11,9	12,3	12,0	13,2	11,8
Électroménager	11,7	10,4	10,2	12,3	11,2
EGP	15,5	15,3	16,4	11,5	13,1
Habillement	9,8	10,1	9,2	15,2	9,7

Tableau 11 : Source Google Trends - Banque de France DGS SEEC

Si le modèle Google Trends était, en moyenne, systématiquement mieux calibré que les modèles CD et SARIMA, il n'est pas meilleur pour la prévision. Ses performances prédictives sont globalement du même ordre de grandeur que celles des modèles CD et SARIMA.

Plus généralement, les résultats obtenus sur les différents produits sont mitigés et contrastés. L'apport des données exogènes – que ce soit les données Google Trends ou les indices conjoncturels quantitatifs – n'est pas la panacée.

Le modèle le plus performant pour les chaussures est le modèle CD ; le modèle Google Trends est légèrement meilleur que le modèle SARIMA. Concernant les meubles, l'apport de Google Trends est net. L'EGP progresse (par rapport au modèle SARIMA) également avec les données exogènes. En

revanche, dans les cas de l'électroménager et de l'habillement, leur apport n'améliore pas les résultats (par rapport au modèle SARIMA).

Ensuite, vis-à-vis de la combinaison de l'information, les résultats sont aussi mitigés. D'une part, le modèle agrégé est meilleur que le modèle global, excepté pour l'EGP. D'autre part, la combinaison d'information n'apporte pas les résultats escomptés. Le modèle agrégé n'est meilleur que dans le cas des meubles, seul produit pour lequel le modèle Google Trends surperforme les modèles CD et SARIMA. De fait, les performances *In Sample* influent sur la pondération des modèles dans l'agrégation. Et, le modèle Google Trends étant mieux calibré (selon la moyenne des RMSE), son poids dans l'agrégation est important :

Poids moyen de chaque modèle dans le modèle agrégé	Google Trends	CD	SARIMA
Chaussures	0,55	0,21	0,19
Meubles	0,71	0,09	0,13
Électroménager	0,47	0,14	0,33
EGP	0,48	0,20	0,27
Habillement	0,58	-0,50	0,80

Tableau 12 : Source Google Trends - Banque de France DGS SEEC

Comme pour les tableaux des moyennes des RMSE, celles ces poids dans l'agrégation sont établis sur la période du protocole de test. L'habillement est le seul produit pour lequel le poids du modèle Google Trends n'est pas le plus important.

Pour plus de détails, des graphiques analogues à ceux présentés pour le total sont en annexe. Ils concernent les chaussures.

3. Conclusion

Le e-commerce est un phénomène en pleine expansion. De fait, les ventes réalisées sur internet prennent davantage de poids dans la consommation des ménages, et donc dans l'enquête mensuelle de conjoncture Commerce de détail de la Banque de France. Dans ce contexte, l'estimation des chiffres d'affaires livrés (tardivement) par la FEVAD devient un sujet de premier plan.

Jusqu'ici, elle était le fruit d'un modèle autorégressif. Ces travaux étudient l'apport des données exogènes que sont les indices du commerce de détail traditionnel (issus de l'EMC) et les indices Google Trends. Chaque source apporte une information qui lui est propre. L'avantage commun de ces sources de données, à savoir être disponible avant les livraisons de la FEVAD, est idoine pour l'exercice de *nowcasting*.

Cependant, une nouvelle source de données (Google Trends) doit être utilisée avec précaution. Des tests de robustesse préalables à son utilisation ont été nécessaires ; au même titre que la mise en place du traitement systématique des valeurs aberrantes, parfois dues à des changements méthodologiques opérés par Google (malheureusement sans information suffisante). Enfin, il a fallu coupler l'immense champ des variables Google possibles avec la faible profondeur de l'historique des données FEVAD (livraisons mensuelles depuis 2012). Cette contrainte duale trouve sa solution dans le domaine du machine learning, avec l'adaptive lasso (Zou (2006)). La sélection de variables opère à chaque itération, palliant ainsi le risque lié au dynamisme du e-commerce et à la possible instabilité des mots-clés correspondant puisqu'il est possible de rétopoler les résultats avec d'autres jeux de variables initiaux. Ainsi, le modèle est souple et présente une forte capacité d'adaptation, nécessitée par la mouvance du phénomène modélisé.

S'ensuit la question d'exploiter la complémentarité des différentes sources de données. Dans cette étude, l'agrégation bayésienne de modèles simples apporte de meilleurs résultats que le modèle global (adaptive lasso appliqué à tous les régresseurs simultanément). La petite taille des échantillons de calibration des modèles peut contribuer à défavoriser un modèle avec beaucoup de variables. Par exemple, dans le cas du total, un phénomène de sur-apprentissage est détecté pour le modèle global. De plus, l'agrégation apporte de la lisibilité au modèle, utile en production.

Néanmoins, l'apport des données exogènes reste mitigé. Il est plus clair dans le cas du total. Or, les livraisons de la FEVAD sont établies sur l'échantillon de ses 70 déclarants les plus gros (en termes de CA). Le nombre de déclarants est donc inférieur pour les produits, pour lesquels les résultats sont beaucoup plus heurtés et donc plus difficiles à appréhender. D'ailleurs, l'ordre de grandeur des RMSFE obtenues sur le total est deux à trois fois plus faible que sur les produits.

Bibliographie

Aiofli M. & Timmerman A. (2006), Persistence of forecasting performance and combination strategies, *Journal of Econometrics*, 135(1-2), 31-53

Askistas N. & Zimmerman K. F. (2009), Google Econometrics and Unemployment Forecasting, *Applied Economics Quarterly* 55(2), 107-120

Bates J. & Granger C. (1969), The combination of forecasts, *Operational Research Quarterly*, 20(4), 451-468

Bec F. & Mogliani M (2015), Nowcasting French GDP in real-time with surveys and « blocked » regressions: Combining forecasts or pooling information?, *International Journal of Forecasting* 31 1021-1042

Bortoli C. & Combes S. (2015), Apports de Google Trends pour prévoir la conjoncture française : des pistes limitées. Note de conjoncture Insee, mars 2015

Breiman (1996), Stacked Regressions, *Machine Learning*, 24, 49-64

Choi H. & Varian H. (2009), Predicting initial claims for Unemployment Benefits, Technical Report, Google

Choi H. & Varian H. (2011), Predicting the Present with Google Trends, Technical Report, Google

Diebold F. (1989), Forecast combination and encompassing: reconciling two divergent literatures, *International Journal of Forecasting*, Vol. 5, pp 589-592

De Gooijer J. & Hyndman R. (2006), 25 years of time series forecasting, *International Journal of Forecasting*, Vol. 22, pp 443-473

Efron B., Hastie T., Johnstone I. & Tibshirani (2004), Least angle regression, *The Annals of Statistics*, Vol. 32, pp 407-499

Elliott G., Rothenberg T. and Stock J. (1996), Efficient tests for an autoregressive unit root, *Econometrica*, Vol 64, No. 4, 813-836

Ettredge M., Gerdes J. & Karuga G. (2005), Using Web-based Search Data to Predict Macroeconomic Statistics, *Communications of the ACM*, Vol. 48, No. 11, pp 87-92

Clements M. & Galvão A. (2006), Combining predictors and combining information in modelling: forecasting US recession probabilities and output growth, *Contributions to Economic Analysis*, Vol. 276, pp 55-73

Clements M. & Galvão A. (2008), Macroeconomic Forecasting With Mixed-Frequency Data: Forecasting Output Growth in the United States, *Journal of Business & Economic Statistics*, Vol. 26, No. 4, pp 546-554

Granger C. & Newbold P. (1974), Spurious Regressions in Econometrics, *Journal of Econometrics*, 2, 111-120

Hoerl A. & Kennard R. (1970), Ridge Regression: Biased Estimation for Nonorthogonal Problems, *Technometrics*, Vol. 12; No. 1, pp 55-67

Hoeting J., Madigan D., Raftery A. & Volinsky C. (1999), Bayesian Model Averaging: A Tutorial, *Statistical Science*, Vol 14, No. 4, 382-417

Huang H. & Lee T. (2010), To Combine Forecasts or to Combine Information? *Econometric Reviews*, Vol. 29, No. 5-6, pp 534-570

Hyndman R. & Athanasopoulos G. (2018), Forecasting: principles and practice (Online textbook), 3.3 Residuals diagnostics, 3.4evaluating forecast accuracy <https://otexts.org/fpp2/>

Kozicki S. & Hoffman B. (2004), Rounding Error: A Distorting Influence on Index Data, *Journal of Money, Credit and Banking* Vol. 36, No. 3 (June 2004, Part 1) pp 319-338

Kuzin V., Marcellino M. & Schumacher C. (2013), Pooling versus model selection for nowcasting GDP with many predictors: Eempirical evidence for six industrialized countries, *Journal of Applied Econometric*, Vol. 28, No. 3, pp 392-411

Marin J. & Robert C (2010), Les bases de la statistique bayésienne. Rapport des Universités Montpellier II et Dauphine – Crest Insee

McLaren N. & Shanbhogue R. (2011), Using internet search data as economic indicators, *Quarterly Bulletin* (Bank of England) 2011 Q2, 134-140

Phillips P. (1986), Understanding spurious regression in econometrics, *Journal of Econometrics*, 33, 311-340

Phillips P. & Perron P. (1988), Testing for a unit root in time series regression, *Biomètrika*, 75, 2, pp335-346

Tibshirani R. (1996), Regression shrinkage and Selection via the Lasso, Journal of the Royal Statistical Society. Series B (Methodological), Vol. 58, pp 267-288

Zeugner S. (2011), Bayesian Model Averaging with BMS

Zou H. (2006), The Adaptive Lasso and Its Oracle Properties, Journal of the American Statistical Association, Vol. 101, No. 476

Annexes

Capture d'écran Google Trends

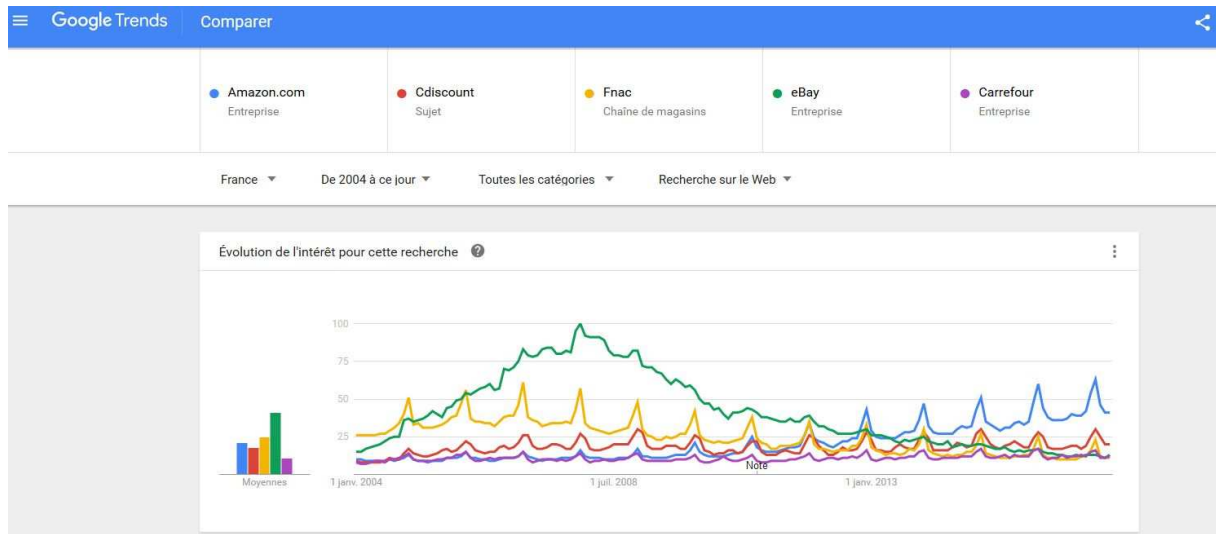
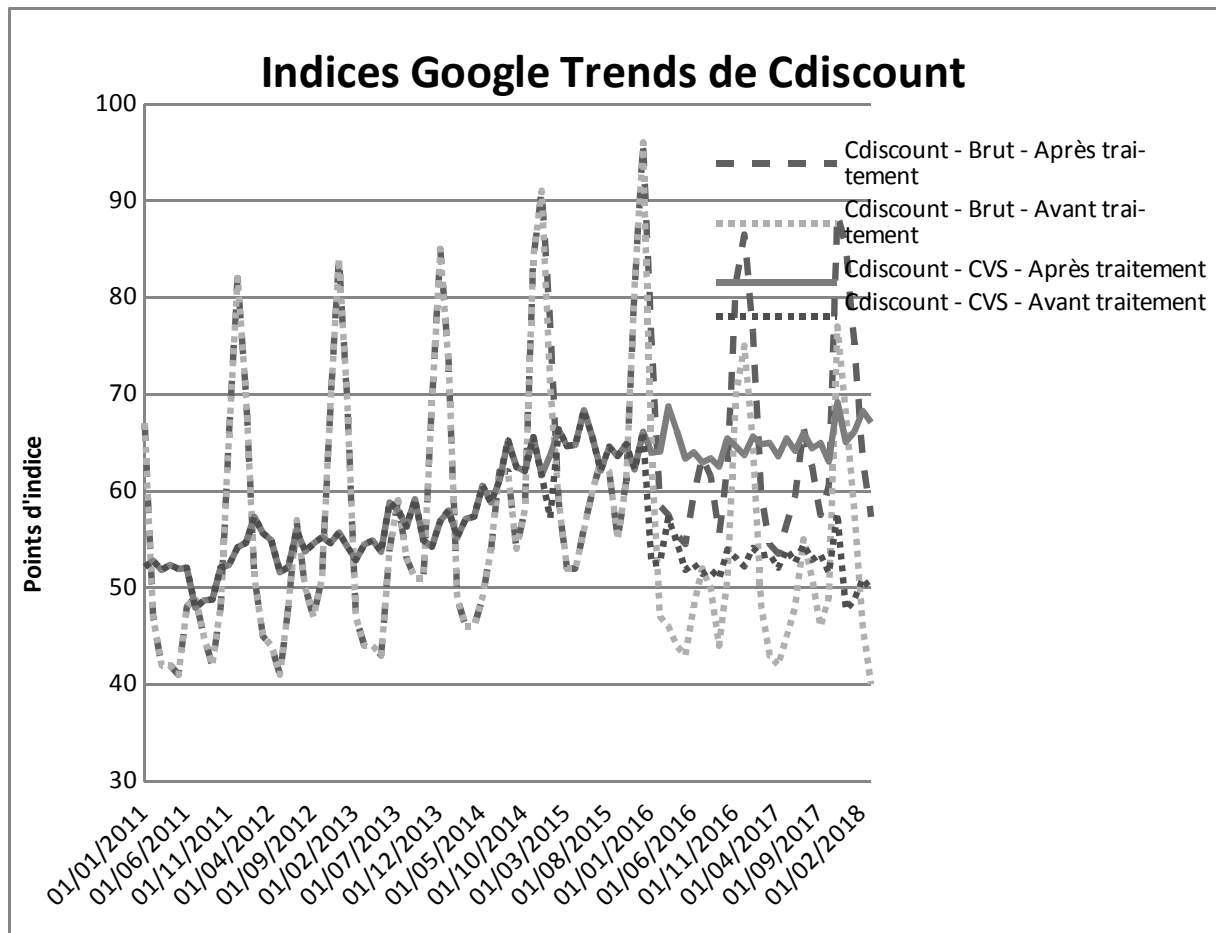
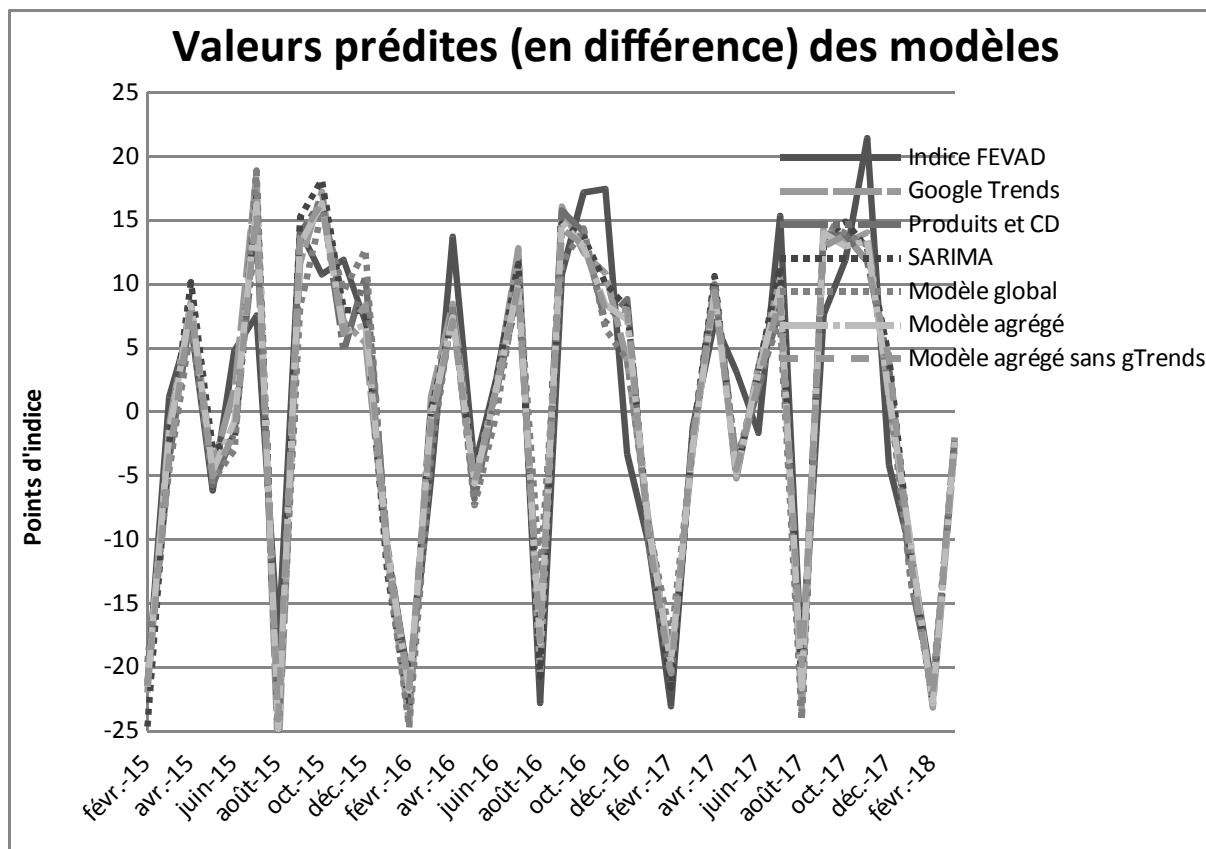


Figure 1: Source Google

Traitement des valeurs aberrantes – l'exemple de Cdiscount



Graphique 13 : Source Google Trends - DGS SEEC Banque de France



Graphique 14 : Source Google Trends - FEVAD - DGS SEEC Banque de France

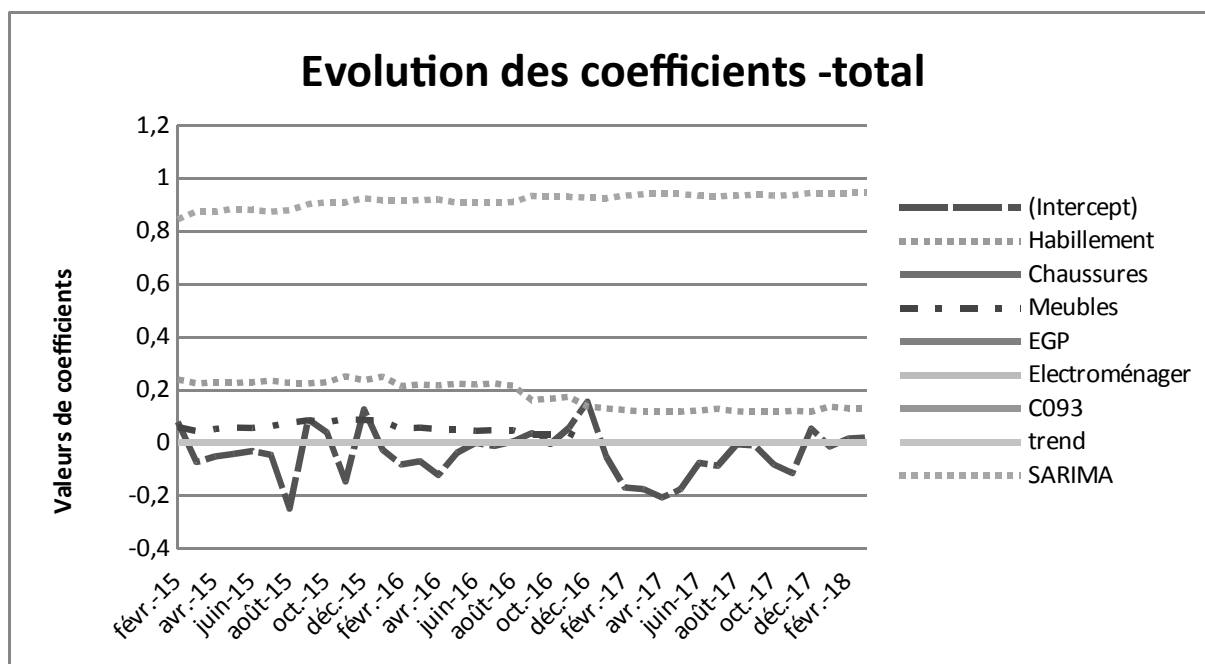
Nombre de sélections des variables dans le modèle Google Trends pour l'estimation du CA total

Variable	Nombre de sélections
(Intercept)	38
Amazon	9
eBay	38
Vente.privee.com	8
Cdiscount	5
FNAC	5
Groupe.Fnac.Darty	0
PriceMinister	37
Leroy.Merlin	32
Union.des.groupements.d.achats.publics	14
Castorama	13
Boulangier	0
Carrefour	2
Showroomprive.com	35
E.Leclerc	0
La.Redoute	1
Auchan	5
Raja	13
Rue.du.Commerce	8
X3.Suisses	15
Promos + soldes + blackfriday	0

Alibaba.Group	7
Groupon	37
PhotoBox	0
Galleries.Lafayette	4
Yves.Rocher	0
Sephora	2
Decathlon	3
trend	0
SARIMA	38

Tableau 13: Source Banque de France DGS SEEC – FEVAD – Google Trends

Évolution des coefficients du modèle CD dans l'estimation du CA total



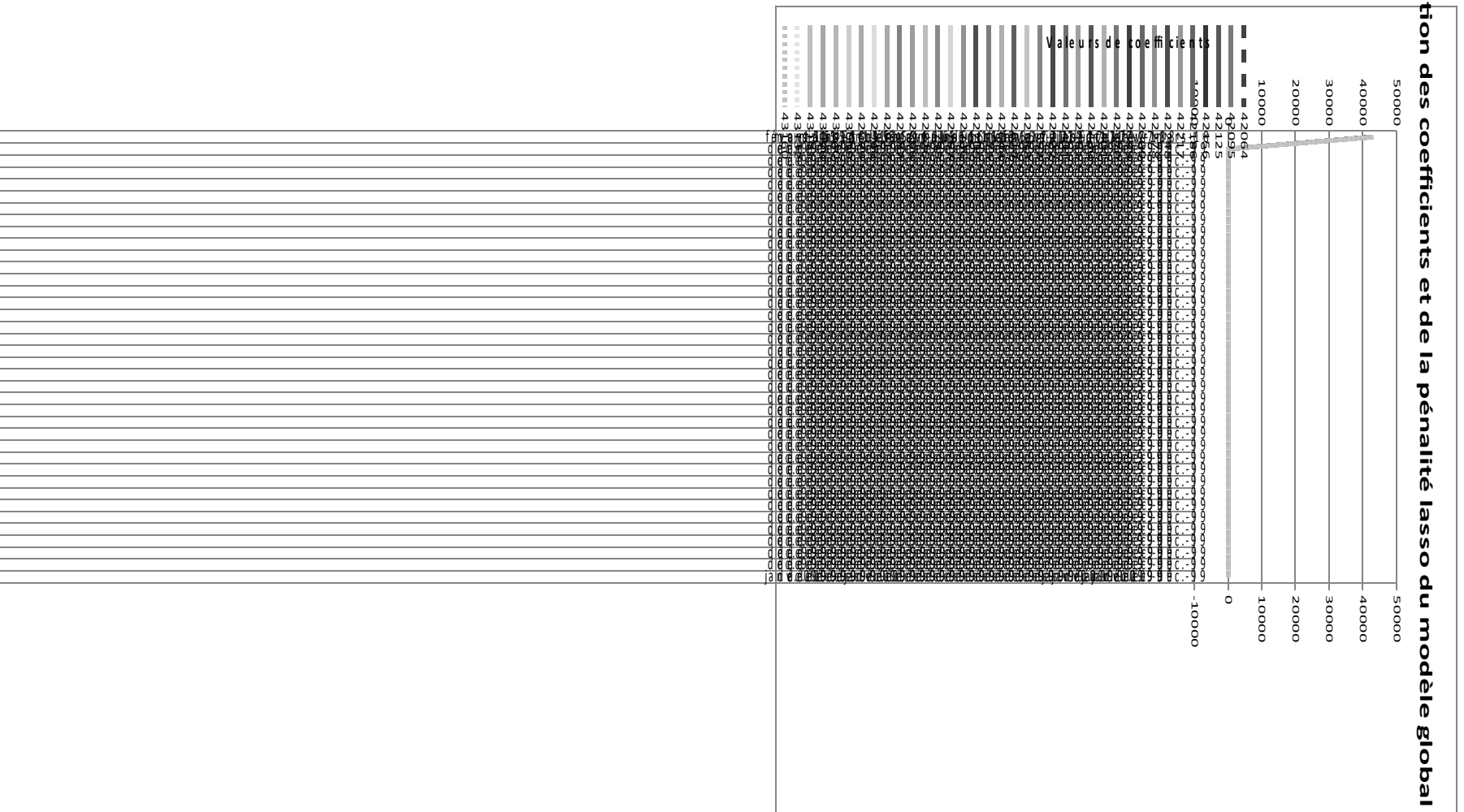
Graphique 15 : Source Banque de France DGS SEEC - FEVAD

Seules les variables sélectionnées ne sont pas en trait plein.

Stabilité du modèle global dans l'estimation du total

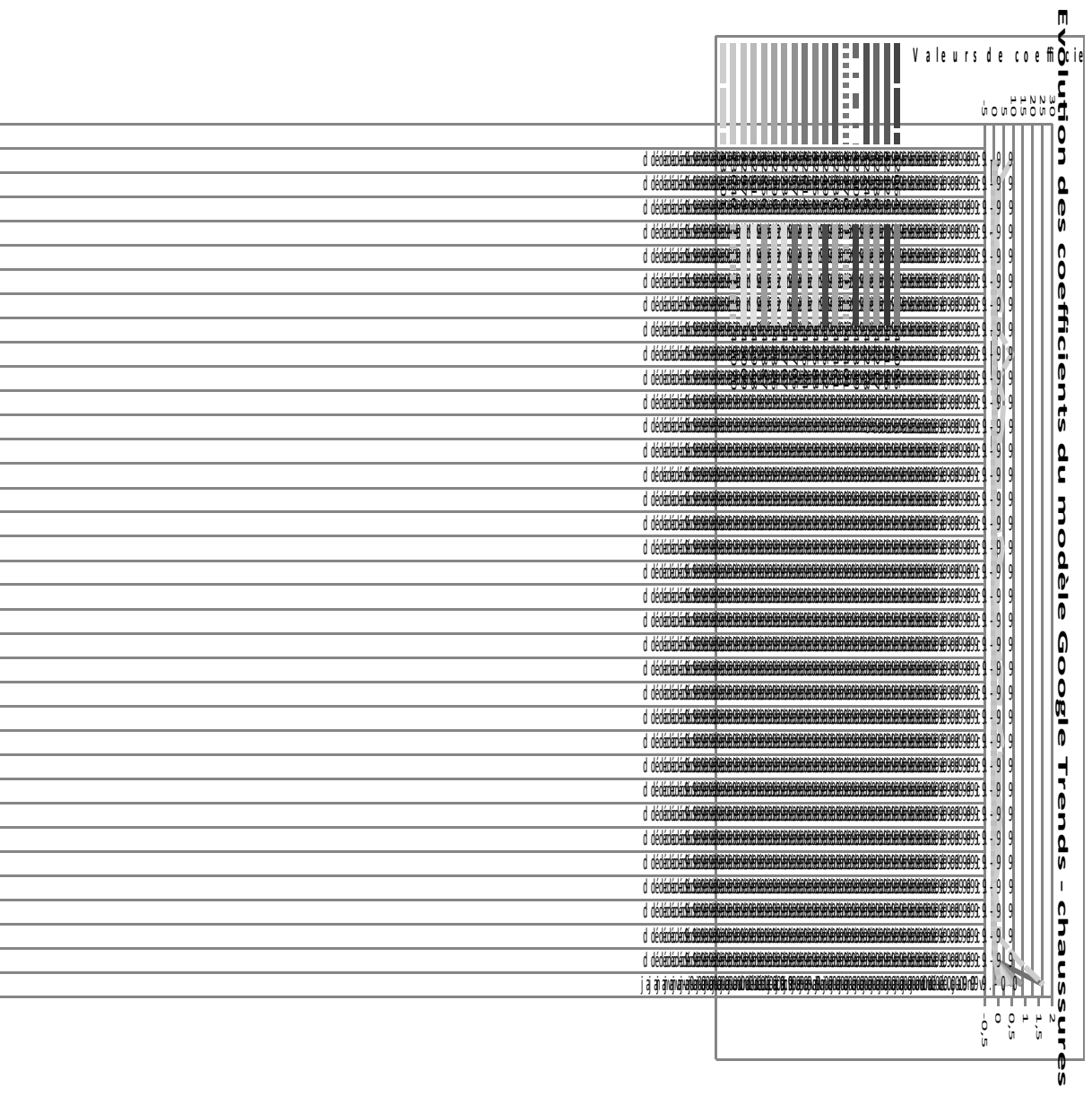
De manière analogue au graphique présentant l'évolution des coefficients du modèle Google Trends pour l'estimation de l'indice de CA total, la pénalité lasso est en axe secondaire. Les requêtes Google sont en trait plein, contrairement à la constante, la variable SARIMA et la pénalité lasso.

Evolution des coefficients et de la pénalité lasso du modèle global - Total



Graphique 16 : Source Banque de France DGS SEEC - FEVAD

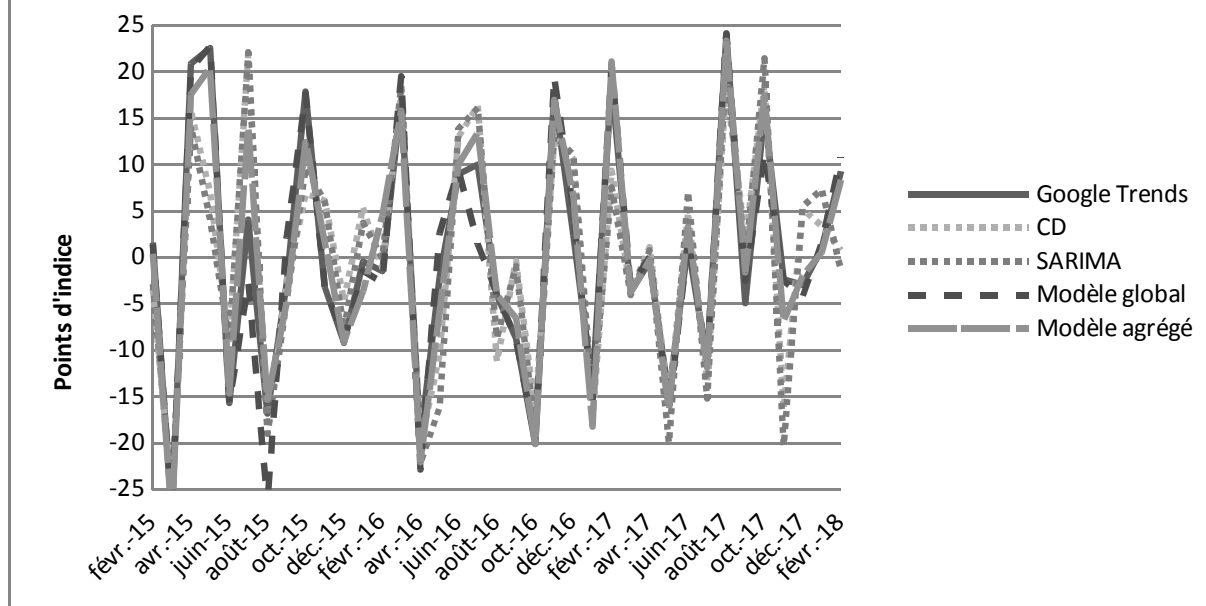
Chaussures



Graphique 17 : Source Banque de France DGS SEEC - FEVAD

Outre la constante, la variable SARIMA et la pénalité lasso (en axe secondaire), les requêtes Google les plus « importantes », c.-à-d. systématiquement sélectionnées et avec des coefficients élevés (en moyenne en absolu), ne sont pas en trait plein. Ces requêtes sont Sparto, Sarenza, soit deux des trois leaders des ventes de chaussures sur Internet, et Converse.

Erreurs de prévision - Chaussures



Graphique 18 : Source Banque de France DGS SEEC – FEVAD