Apports de Google Trends pour prévoir la conjoncture française : des pistes limitées

Clément Bortoli

Département de la Conjoncture

Stéphanie Combes

Département des méthodes statistiques

Représentant plus de la moitié du PIB, la consommation des ménages est le poste le plus important de la demande finale. Disposer d'une information la plus à jour possible sur son évolution est donc essentiel pour établir et prévoir l'activité en France. La première donnée quantitative disponible relative aux dépenses des ménages est l'indice mensuel de consommation en biens publié sous un délai d'un mois. Les indices de chiffres d'affaires, renseignant notamment sur les dépenses en services, sont, quant à eux, disponibles sous un délai de deux mois. Enfin, une première estimation des dépenses trimestrielles de l'ensemble des biens et services est publiée au milieu du trimestre suivant.

Afin d'estimer les dépenses en temps réel ou avant la publication de ces chiffres, les modèles usuels de prévision intègrent, le plus souvent, des variables issues des enquêtes qualitatives de conjoncture. Ces données, disponibles sous moins d'un mois, constituent des indicateurs précoces d'un certain nombre de variables macroéconomiques.

Depuis quelques années, les conjoncturistes s'intéressent également aux données issues d'Internet et en particulier aux tendances des recherches des utilisateurs de Google comme sources d'informations susceptibles d'améliorer leurs prévisions. Outre leur gratuité, la vitesse à laquelle ces données peuvent être mobilisées les rendent attrayantes pour la conjoncture : disponibles à la fin de chaque semaine, elles permettent d'apprécier la popularité des requêtes des internautes pratiquement en temps réel. De plus, le volume des requêtes que les utilisateurs collectent sur des produits particuliers via le moteur de recherche pourrait refléter le volume potentiel des ventes de ces produits. Ces données pourraient donc être considérées comme des indicateurs d'intention d'achat des consommateurs, tant pour les produits manufacturés que pour les services.

Google met à disposition les tendances de recherche de ses utilisateurs par le biais de l'outil Google Trends. Les catégories proposées par cet outil regroupent les requêtes par thème. Celles qui pourraient se révéler pertinentes pour prévoir la consommation des ménages sont très nombreuses. Les utiliser au sein de modèles économétriques requiert d'être à même de juger de la qualité de leur contribution à la prévision. Ce dossier s'appuie sur les méthodes de combinaison de modèles qui permettent de tirer parti de l'ensemble de l'information disponible et de prendre en compte des prévisions obtenues par différents modèles aux performances proches, grâce à une approche « bayésienne ».

D'après les différentes modélisations testées, les tendances de recherche Google ne permettent d'améliorer la prévision des dépenses mensuelles des ménages que de façon limitée. Plus précisément, leur prise en compte ne permet pas d'améliorer la prévision globale des dépenses de consommation mensuelles des ménages en biens ou en services, en raison de leur forte hétérogénéité. En revanche, les résultats obtenus pour les achats de certains biens (habillement et équipement de la maison, notamment) sont plus positifs, certaines catégories Google Trends permettant d'améliorer la prévision de ces postes. Cependant, même dans ce cas, l'amélioration est modeste.

Apports de Google Trends pour prévoir la conjoncture française : des pistes limitées

L'évolution des dépenses des ménages, boussole de l'économie française

La consommation des ménages représente le premier poste de la demande intérieure finale française

En France, la consommation finale des ménages en biens et services est le principal poste de la demande finale intérieure : depuis le début des années 1980, la consommation des ménages en représente environ la moitié en euros courants, l'autre moitié se partageant entre d'une part les autres types de dépenses (consommations individualisable et collective des administrations publiques et des institutions sans but lucratif au service des ménages), d'autre part la formation brute de capital fixe, pour un quart chacune. La part en valeur de la consommation des ménages dans le PIB est relativement stable, variant entre 52 % et 56 % depuis le début des années 1980.

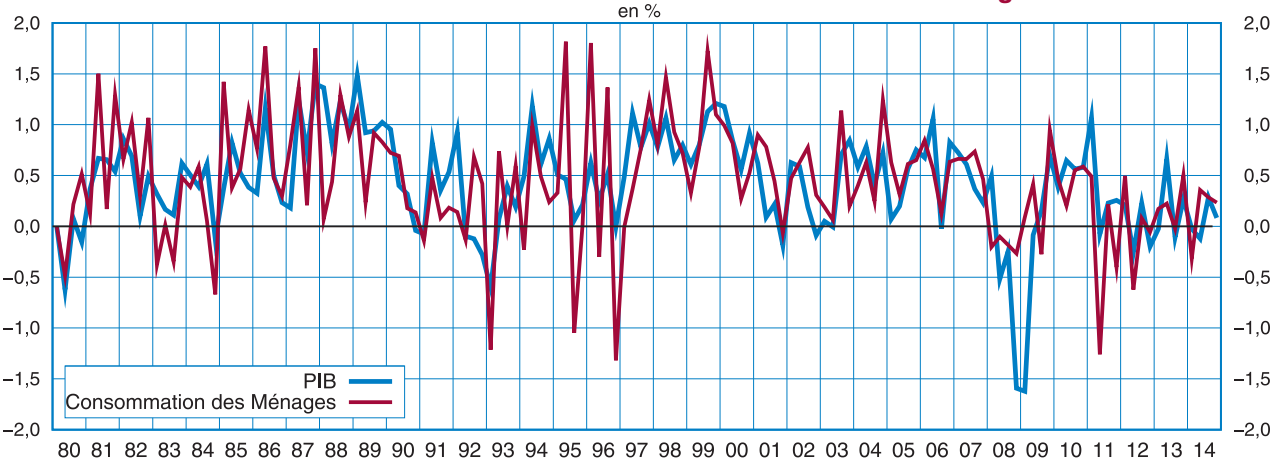
Les dépenses de consommation des ménages jouent un rôle essentiel dans les évolutions conjoncturelles de l'économie

Les différentes équations macroéconomiques utilisées pour modéliser le comportement de consommation retiennent des facteurs multiples (*Faure et al., 2012*). En premier lieu, la consommation des ménages réagit aux fluctuations de leur pouvoir d'achat. D'autres éléments ont une influence sur leurs dépenses, comme le chômage, en raison de comportements « d'épargne de précaution », les fluctuations de prix ou encore le niveau des taux d'intérêt. En second lieu, des éléments ponctuels peuvent provoquer des à-coups dans les dépenses de consommation. Ainsi, des températures éloignées des normales saisonnières en automne ou en hiver peuvent entraîner des écarts importants des dépenses de chauffage (contribuant par exemple pour -0,2 point à la hausse des dépenses totales de consommation au quatrième trimestre 2014, et pour -0,1 point à la croissance du PIB). De même, les dispositifs de primes à la casse ainsi que les modulations successives du bonus-malus écologique introduit en 2008 provoquent de forts à-coups sur la consommation d'automobiles (contribuant par exemple pour -0,8 point à la croissance totale des dépenses de consommation au deuxième trimestre 2011). Au total, la consommation des ménages a souvent un profil heurté. En raison de son poids dans l'activité française, elle constitue l'un des principaux facteurs de la variance de la croissance trimestrielle (« volatilité ») du PIB *(graphique 1)*. Entre 1980 et 2014, elle contribue ainsi pour 37 % à cette variance en moyenne, soit moins que l'investissement (40 %), mais plus que la variation des stocks (22 %).

La consommation des ménages en biens explique une très grande partie de la volatilité de leurs dépenses totales

La volatilité de la consommation totale des ménages résulte essentiellement de leur consommation de biens, qui contribue en moyenne pour 84 % à cette volatilité totale depuis 1980. La structure de la consommation en biens a peu évolué depuis trente ans. Elle se répartit entre achats de produits alimentaires (36 % de la consommation en biens en moyenne), consommation d'énergie

1 - Taux de croissance du PIB et de la consommation des ménages



Source : Insee

(17 % en moyenne, qui recouvre à la fois les dépenses de chauffage et les achats de carburants) et achats de biens fabriqués (47 % en moyenne depuis 1980), en particulier d'automobiles, de biens d'équipement du logement (produits informatiques et électroniques, électroménager et meubles) et d'habillement. Les achats d'automobiles évoluent de façon particulièrement heurtée en raison notamment des dispositifs de primes à la casse et de bonus-malus écologiques (*cf. supra*) ; sur la période considérée, ils ont contribué à 49 % de la volatilité de la consommation en biens (*tableau 1*), soit 42 % de la volatilité de la consommation totale des ménages.

Le poids des services dans les dépenses des ménages n'a cessé d'augmenter depuis le début des années 1980

Depuis le début des années 1980, la part des services dans la consommation des ménages n'a cessé de croître. La consommation des ménages en services ne représentait que 40 % de leurs dépenses totales en 1980 (en euros courants), son poids est désormais de 53 %. Cette dynamique est également perceptible en euros constants, bien qu'un peu moins marquée : sur la même période, la part des services est passée de 46 % à 54 % en volume. La majeure partie de la volatilité de la consommation en services est expliquée par un nombre relativement réduit de postes (*tableau 2*) : en effet, les dépenses en hôtellerie-restauration, en transport et en services d'information-communication (qui incluent principalement les dépenses en téléphonie et forfaits internet) contribuent pour 56 % à la volatilité de la consommation des ménages en services, alors qu'elles pèsent moins de 30 % du total en moyenne depuis 1980. À l'inverse les loyers réels et imputés représentent 33 % des services en valeur, mais du fait de leur inertie, seulement 8 % de la volatilité des dépenses en services – soit moins de 1 % de la volatilité des dépenses totales des ménages.

Tableau 1
Poids dans la consommation en biens et contribution à la variance des évolutions de la consommation en biens des dépenses par type de produit

	Poids dans la consommation en biens	Contribution à la variance de la croissance de la consommation en biens
Alimentation	36 %	11 %
Biens fabriqués	47 %	68 %
dont : Automobiles	*12 %*	*49 %*
Équipement du logement	*8 %*	*7 %*
Habillement	*12 %*	*8 %*
Énergie	17 %	21 %
Total des biens	100 %	100 %

Lecture : entre 1980 et 2014, les dépenses en produits alimentaires ont représenté en moyenne 36 % de la consommation totale des ménages en biens (en euros courants). Sur la même période, la croissance de ces dépenses a été responsable de 11 % de la variance des évolutions trimestrielles de la consommation totale des ménages en biens.
Source : Insee

Tableau 2
Poids dans la consommation en services et contribution à la variance des évolutions de la consommation en services des dépenses pour certains produits

	Poids dans la consommation en services	Contribution à la variance de la croissance de la consommation en services
Services de logement	33 %	8 %
Hôtellerie et restauration	13 %	26 %
Information et communication	10 %	17 %
Services de transport	6 %	13 %
Autres services	38 %	36 %
Total des services	100 %	100 %

Source : Insee

Apports de Google Trends pour prévoir la conjoncture française : des pistes limitées

Les recherches en ligne sont susceptibles d'apporter une information rapide sur la consommation des ménages

Les indicateurs de consommation sont disponibles progressivement

Ainsi, les dépenses de consommation finale des ménages en biens et services jouent un rôle central dans les développements conjoncturels de l'économie française. Or, les données concernant les dépenses des ménages ne sont pas disponibles instantanément. Seuls quelques indicateurs spécifiques sont disponibles quasiment en temps réel : les immatriculations d'automobiles sont connues dès le premier jour ouvré du mois suivant ; les données portant sur la consommation d'électricité sont publiées quotidiennement quasiment en temps réel. La première donnée agrégée disponible relative aux dépenses des ménages est l'indice mensuel de consommation en biens, publié près de 30 jours après le mois considéré. Les chiffres d'affaires dans les services, renseignant notamment sur les dépenses en services sont, quant à eux, publiés près de 60 jours après le mois considéré. D'autres données sur les services sont connues plus tard encore, par exemple les données de téléphonie publiées par l'Arcep environ 70 jours après la fin du trimestre considéré. Enfin, c'est environ 45 jours après la fin d'un trimestre qu'une première estimation des comptes trimestriels de dépenses des ménages est diffusée (les comptables nationaux doivent donc extrapoler certains indicateurs non encore connus à cette date).

Les statistiques de recherche en ligne sont susceptibles d'apporter une information précoce sur la consommation

Poser un diagnostic précis plus rapide sur les évolutions conjoncturelles des dépenses de consommation des ménages demande donc de disposer d'indicateurs précoces, quasiment en temps réel. Dans cette optique, l'emploi de statistiques sur les recherches en ligne paraît prometteur, ne serait-ce que parce qu'Internet joue un rôle croissant dans les achats effectués par les ménages. Ainsi, entre 2006 et 2011, la part des achats de biens durables par internet est passée de 2 % à 9 %, et s'élève même à 11 % pour les biens culturels (*Krankadler, 2014*). Internet joue également un rôle croissant dans la consommation de services (transports, commerce, services financiers, etc.). Par ailleurs, même si l'achat n'a pas lieu sur Internet, les ménages peuvent se renseigner au préalable par le biais de moteurs de recherche. Ces requêtes peuvent alors contenir des intentions d'achat révélatrices.

Internet est déjà utilisé pour prévoir des indicateurs socio-économiques depuis plusieurs années

Or, en 2006, Google a lancé Google Trends, un outil mettant gratuitement à disposition des séries dont l'évolution reflète l'intérêt des internautes pour une requête ou un ensemble de termes de recherche sémantiquement proches. Cette application médiatise notamment les recherches les plus populaires des utilisateurs de son moteur de recherche et ce, pratiquement en temps réel. En 2009, le groupe a publié une analyse de l'intérêt de l'utilisation de ces séries pour la prévision d'indicateurs socio-économiques (*Choi* et *Varian*, 2009). D'après cette étude sur données américaines, la prévision des achats automobiles, des ventes de détail et des acquisitions de logements pouvait être améliorée en introduisant ce type de séries dans des modèles simples utilisant la dynamique de la série à prévoir (modèle autoregressif). L'utilisation des données de Google Trends dans des domaines variés et leur intégration dans des modélisations économétriques plus complexes ont également été testées par la suite.

Le plus célèbre exemple d'utilisation est l'application Google Flu développée par Google pour prévoir la progression de l'épidémie de grippe en temps réel à partir des requêtes des utilisateurs. En économie, *Askitas et Zimmermann (2009)* utilisent la fréquence de l'utilisation de certains termes de recherche pour prévoir le taux de chômage en Allemagne ; *Kulkarni et al. (2009)* suggèrent un lien entre la fréquence de plusieurs termes de recherche et les prix du logement aux États-Unis ; *Vosen et Schmidt (2011)* utilisent également ce type de séries pour prévoir les dépenses des ménages aux États-Unis.

Apports de Google Trends pour prévoir la conjoncture française : des pistes limitées

Les données Google Trends sont mobilisables rapidement

Le principal attrait des données Google Trends pour la conjoncture réside dans leur capacité à être mobilisables rapidement et à une fréquence plus élevée que la plupart des séries économiques traditionnelles, les données étant publiées à la fin de chaque semaine. De plus, il est possible d'obtenir des données par origine géographique : on peut ainsi se limiter à l'étude des requêtes effectuées depuis la France pour essayer d'en tirer un diagnostic sur l'évolution de la conjoncture en France. Les séries brutes correspondant à la fréquence réelle d'utilisation d'un terme de recherche ne sont pas publiques. En effet, les données mises à disposition sont corrigées d'une tendance qui résulterait d'un gain de popularité du moteur de recherche lui-même, et se présentent sous la forme de séries temporelles de nombres entiers, normalisées *in fine* de telle sorte que leur maximum soit égal à 100 (*graphique 2*).

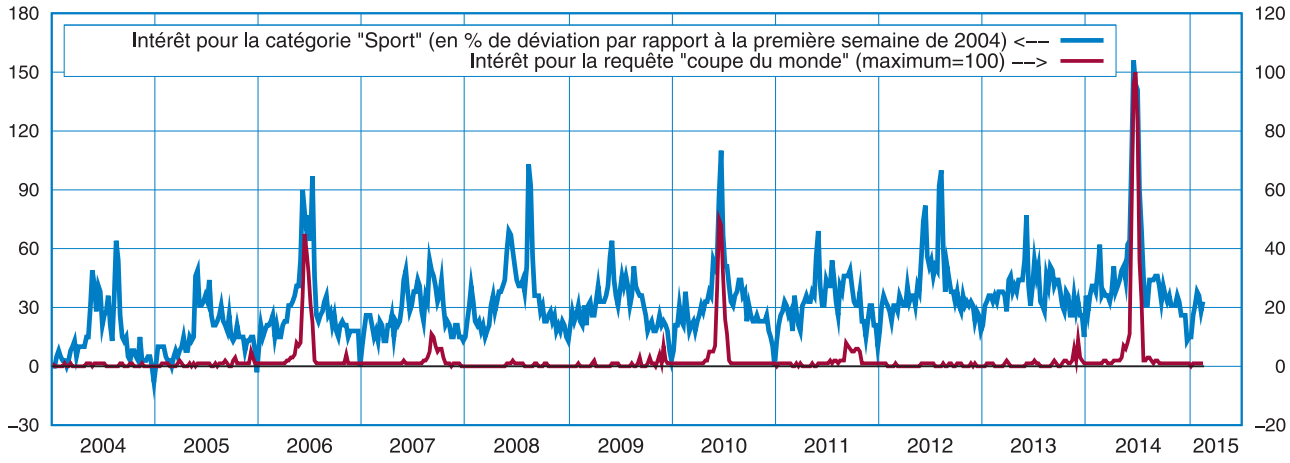
Les catégories de Google Trends agrègent des termes proches

Le sens d'un terme de recherche pouvant évoluer au fil du temps, il est préférable de travailler sur des catégories ou concepts plutôt que sur des termes précis. Google Trends met ainsi à disposition de ses utilisateurs des statistiques sur des ensembles de termes sémantiquement proches, appelés « catégories ». La normalisation appliquée par Google aux séries relatives aux catégories Google Trends diffère de celle appliquée aux requêtes simples : la fréquence atteinte par la catégorie la première semaine de 2004 est utilisée comme référence, les points suivants de la série étant exprimés en points de déviation par rapport à ce niveau. La catégorie « Sport » agrège par exemple l'ensemble des termes de recherche liés au domaine sportif. Les utilisateurs français de Google y ont montré un regain d'intérêt les étés des années paires. Plus précisément, les recherches sur le sport sont en recrudescence marquée lors des Coupes du Monde de football de 2006, 2010, 2014, des championnats d'Europe de football et des Jeux Olympiques d'été de 2004, 2008 et 2010. Les achats de téléviseurs augmentant en général fortement au moment des grands événements sportifs, l'utilisation de la catégorie « Sport » peut alors permettre de mesurer le degré d'intérêt que suscite un événement sportif chez les consommateurs français et donc de quantifier en temps réel cette consommation supplémentaire de biens d'équipement du logement.

L'outil Google Trends présente néanmoins plusieurs faiblesses

Cependant, la façon dont Google élabore les séries diffusées dans Google Trends manque de transparence, ce qui constitue une faiblesse de l'outil. Ainsi, les traitements et normalisations appliqués aux requêtes et catégories ne sont pas précisés. De même, la gestion dans le temps des catégories et de leur composition, notamment en cas d'apparition d'une nouvelle requête populaire à une date donnée, ne sont pas documentées. De plus, l'outil présente des limites

2 - Intérêt pour la requête « coupe du monde » et pour la catégorie « Sport » de Google Trends en France



Source : Google Trends

Apports de Google Trends pour prévoir la conjoncture française : des pistes limitées

Encadré 1 - La pérennité des séries obtenues sous Google Trends est incertaine

Produire une statistique régulière requiert que les données utilisées soient pérennes et de qualité. Les données Google Trends sont à analyser à cette aune. En particulier, les séries fournies résultent d'un échantillonnage aléatoire et peuvent donc différer d'une extraction à l'autre. Plus précisément, à chaque extraction des données, l'outil fournit une série construite à partir du décompte des requêtes réalisées sur l'historique d'un échantillon d'utilisateurs. Des échantillons différents peuvent donc mener à la production de séries non coïncidentes pour une même requête. Ce phénomène a tendance à s'amplifier lorsque la requête est peu courante, mais la fréquence suffisante d'une requête pour que la série soit relativement robuste à l'échantillonnage n'est pas communiquée. On peut faire l'hypothèse que l'utilisation des catégories proposées par Google Trends, qui regroupent un grand nombre de requêtes sous le même label, permet de minimiser le bruit.

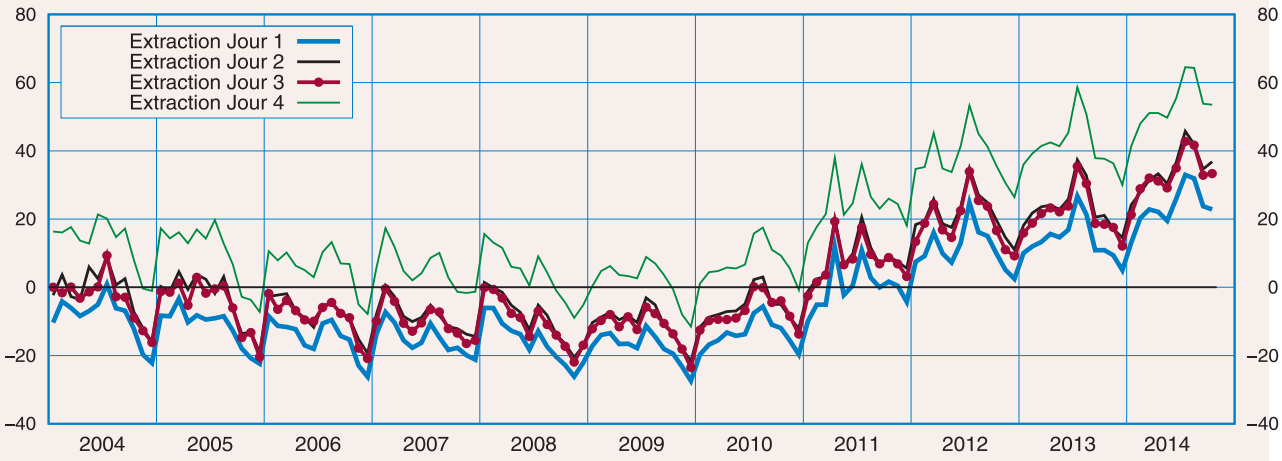
La question de l'avenir de l'outil lui-même se pose également. La production pérenne d'un indicateur à partir des tendances de requête Google serait aussi menacée par l'évolution de la stratégie commerciale du groupe que par les évolutions technologiques du moteur de recherche. L'outil Google Flu, autre application de Google, qui s'était d'abord montrée performante dans la prévision de la progression de l'épidémie de grippe, a ensuite rencontré des limites que l'on attribue en partie à ces deux évolutions.

Google Flu a été lancé en 2008 aux États-Unis et étendu l'année d'après à une douzaine de pays européens, dont la France. Pour produire son indicateur, Google a cherché à identifier parmi des milliers de mots-clé les plus corrélés aux évolutions des indicateurs officiels fournis par des organismes de veille sanitaire tels que les statistiques établies par le réseau Sentinelles de l'Institut national de la santé et de la recherche médicale (Inserm) dans le cas de la France. Les termes de recherche qui connaissaient des pics de fréquence d'utilisation identiques à ceux de la progression de la grippe saisonnière ont été sélectionnés. Lorsqu'il a été lancé aux États-Unis, cet outil apparaissait prometteur, puisqu'il parvenait à fournir un indicateur avancé d'une à deux semaines par rapport aux publications officielles, tirant parti de la disponibilité quasi immédiate des requêtes *Google* (*Ginsberg et al., 2009*). L'outil a été mis en défaut une première fois en 2009 lorsqu'il n'a pas été capable de prévoir l'épidémie non saisonnière de grippe porcine (H1N1) aux États-Unis. Bien que corrigé à la suite de cet épisode, Google Flu a également connu des défaillances lors des hivers 2011-2012 et 2012-2013, en surestimant largement les épidémies de grippe aux États-Unis. Cette perte de performance a été analysée par *Lazer et al. (2014)*, qui ont avancé plusieurs explications :

- le comportement de recherche des internautes peut être modifié par la couverture médiatique des épidémies considérées ;
- les performances du moteur de recherche et les algorithmes utilisés peuvent évoluer et entraîner un changement de la manière dont les usagers l'utilisent.

Ainsi, les performances d'un indicateur fondé sur cet outil dépendent étroitement des habitudes d'utilisation d'Internet. Par exemple, la part croissante que prennent les applications pour smartphones dans l'utilisation d'Internet pourrait conduire à terme à réduire le rôle joué par les moteurs de recherche, si les internautes privilégient des applications dédiées aux achats : la capacité des tendances de recherche à capturer le comportement des consommateurs pourrait alors s'en trouver diminuée. ■

Valeurs historiques de la catégorie Google Trends « Entretien des véhicules »



Lecture : les données hebdomadaires Google Trends sont ici mensualisées
Source : Google Trends

dommageables à une production statistique pérenne (*encadré 1*). D'abord, les séries fournies sont le résultat d'un échantillonnage aléatoire et peuvent donc différer d'une extraction à l'autre. Ensuite, l'application Google Trends est, par construction, dépendante de l'évolution de la stratégie commerciale du groupe et des évolutions technologiques du moteur de recherche lui-même, adapté pour répondre aux attentes de ses usagers. Ainsi, depuis sa création, l'outil et l'étendue des séries auxquelles l'utilisateur a accès ont été considérablement modifiés. Enfin, le moteur de recherche a beaucoup évolué : ces modifications seraient à l'origine d'une dégradation marquée de la performance de l'outil Google Flu à partir de 2009.

Les catégories Google Trends couvrent largement les différents postes de dépenses des ménages

Parmi les catégories proposées par Google Trends, une cinquantaine constituent des candidats potentiellement intéressants pour la prévision de la consommation de biens ou de services. Ces catégories semblent, en outre, couvrir l'ensemble des postes de dépenses (*tableaux 3 et 4*).

Tableau 3
Exemples de catégories *a priori* pertinentes pour prévoir la consommation des ménages en biens

Alimentaire	Biens fabriqués				
	Autres biens fabriqués	Biens durables			Habillement
		Autres biens durables	Automobiles	Equipement du logement	
Produits du tabac	Jeux		Automobiles et véhicules	Informatique et Electronique	Habillement
Boissons alcoolisées	Jouets		Achat de véhicules	Internet et Télécoms	Articles de sport
Boissons non alcoolisées	Santé		Entretien des véhicules	Electronique grand public	
Epiceries et magasins d'alimentation	Shopping		Marques automobiles	Appareils mobiles et sans fil	
Grands magasins et hypermarchés	Soins du corps et remise en formes		Pièces et accessoires pour véhicule	Ordinateurs portables et notebooks	
	Sports		Motos	Appareils ménagers	
	Articles de sport		Scooters et cyclomoteurs	Mobilier de maison	
	Articles de cuisine et de repas			Maison et jardinage	
	Livres et littérature			Sports	
	Maquillage et cosmétique				
	Bagages et accessoires de voyage				

Tableau 4
Exemples de catégories *a priori* pertinentes pour prévoir la consommation des ménages en services

Transport	Hébergement et restauration	Information et communication	Services financiers	Services immobiliers	Services aux ménages
Voyages	Restaurants	Internet et télécoms	Assurance	Location de maisons et d'appartements	Arts et divertissements
Location de voitures et taxis	Hôtels et hébergements	Fournisseurs d'accès et opérateurs	Banque		Hobbies et loisirs
Voyages aériens		Appareils mobiles et sans fil			Soins du corps et remise en forme
Bus et trains		Livres et littérature			Sports
		Portails d'achats en ligne et moteurs de recherche			

Apports de Google Trends pour prévoir la conjoncture française : des pistes limitées

Combiner des modèles permet de tenir compte de la diversité des séries qui peuvent être mobilisées pour la prévision

Le nombre élevé de variables explicatives potentielles nécessite une stratégie de sélection ...

Outre leur disponibilité rapide, qui peut permettre d'évaluer l'activité économique réelle avec un délai moindre que la plupart des indicateurs économiques, les séries de Google Trends sont nombreuses et représentent donc un important gisement d'information *a priori*. La modélisation pourrait même intégrer les séries retardées en faisant l'hypothèse que les utilisateurs se renseignent jusqu'à un mois à l'avance avant de réaliser un achat. Une cinquantaine de séries Google Trends ayant été ciblées, les variables explicatives potentielles sont au nombre d'une centaine, pour environ 130 observations mensuelles, de janvier 2004 à aujourd'hui. Le nombre de régresseurs potentiels est donc élevé comparativement au nombre d'observations. En pratique, en prévision, il est préférable de se concentrer sur les modèles les plus parcimonieux, c'est-à-dire qui n'utilisent qu'un nombre limité de variables. Le modélisateur dispose de plusieurs stratégies pour sélectionner les variables les plus pertinentes.

... qui, ici, ne peut s'appuyer sur un dire d'expert

La première consiste à sélectionner *a priori* les variables les plus pertinentes pour l'étude, sur la base de dire d'experts. Pour les séries correspondant aux catégories de Google Trends, cette expertise est à construire. Si certaines catégories semblent naturellement intéressantes (par exemple les catégories « voyages », « voyages aériens » ou encore « bus et train » pour les dépenses en services de transport), on peut préférer ne pas exprimer d'*a priori* sur la sélection de l'ensemble des variables pertinentes.

Une sélection automatique de ces séries à partir de critères statistiques peut être mise en œuvre

De nombreuses approches économétriques ou d'analyse des données ont été développées pour permettre une sélection automatique de variables explicatives à partir de critères statistiques :

- les algorithmes itératifs qui ajoutent (respectivement retirent) un certain nombre de variables à partir d'un premier modèle vide (respectivement plein) en fonction d'un critère de significativité,
- les approches visant à minimiser une fonction objectif avec une pénalité favorisant la parcimonie (par exemple le Lasso[1] ou les critères d'information),
- les approches par analyse en composantes principales qui servent à résumer l'information d'un grand nombre de variables dans un petit nombre de facteurs,
- les combinaisons de modèles.

L'approche par combinaison de modèles permet de gagner en robustesse

Les combinaisons de modèles (encadré 2) présentent l'intérêt de prendre en compte l'aléa lié à la modélisation retenue, en particulier lorsque plusieurs modèles ont des performances en prévision jugées comparables et qu'il serait arbitraire d'en privilégier un unique. De plus, ces modèles fournissent des prévisions qui peuvent significativement différer pour un horizon considéré. La combinaison des prévisions issues de ces différents modèles prend davantage en compte l'ensemble de l'information disponible, mais permet également de gagner en robustesse en cas de choc sur une variable isolée.

Les poids associés aux prévisions issues des modèles retenus dans la combinaison peuvent être calculés de diverses façons, par exemple ici avec l'approche bayésienne qui permet notamment de piloter la taille des modèles en favorisant la parcimonie.

(1) La méthode du Lasso correspond à une régression linéaire qui prend en considération à la fois la qualité de l'ajustement et la valeur absolue des coefficients. L'importance relative de ces deux objectifs est pilotée par un paramètre : si ce dernier est élevé, de nombreux coefficients seront nuls et de nombreuses variables explicatives n'entreront donc pas dans la régression.

Apports de Google Trends pour prévoir la conjoncture française : des pistes limitées

Les tendances de recherche Google permettent d'améliorer modestement la prévision de la consommation des ménages pour certains produits

L'utilisation de Google Trends améliore modestement la prévision de la consommation de certains produits

L'utilisation des tendances de recherche Google ne permet pas d'améliorer la prévision des dépenses de consommation mensuelles des ménages en biens ou en services lorsque ces dernières sont considérées à un niveau agrégé, probablement du fait de leur forte hétérogénéité. En revanche, les résultats obtenus pour les achats de certains biens (habillement, équipement de la maison, alimentaire) et de certains services (transports) sont plus positifs.

Les catégories Google Trends sont ici utilisées pour prévoir la croissance mensuelle des dépenses de consommation des ménages en biens (en volume) publiée par l'Insee à la fin du mois suivant le mois d'intérêt. Les catégories Google Trends sont mensualisées, puis désaisonnalisées ; leurs taux de croissance mensuels sont alors utilisés comme variables explicatives potentielles, ainsi que leur premier retard (c'est-à-dire la valeur prise par ce taux de croissance au mois précédent). En outre, les quatre premiers retards de la variable modélisée sont aussi retenus. Une procédure similaire est également appliquée pour tenter de prévoir la croissance mensuelle des dépenses de consommation en services. Les estimations sont effectuées sur la période mars 2004 - décembre 2011 puis la qualité de la prévision est mesurée sur la période janvier 2012 - décembre 2014 sur le critère de l'erreur quadratique moyenne (RMSE).

Les tendances de recherche Google ne permettent pas d'améliorer la prévision des dépenses de consommation en biens ou en services lorsque ces dernières sont considérées à un niveau agrégé…

Pour la consommation totale en biens, l'utilisation des données Google Trends ne permet pas d'améliorer la qualité de la prévision par rapport à un modèle simple utilisant la seule dynamique de la série (ARMA) quelle que soit l'approche utilisée. De fait, la variable la plus pertinente pour expliquer la croissance mensuelle de la consommation en biens s'avère être la croissance de cette même consommation au mois précédent. Dans l'approche bayésienne, des catégories

Encadré 2 - Sélection bayésienne des variables et combinaison de modèles

L'approche bayésienne (voir par exemple *Raftery et al. 1997*) consiste à fixer une probabilité *a priori* sur des paramètres du modèle (qui intègrent ici la composition du modèle c'est-à-dire le fait d'inclure telle ou telle variable), et d'en déduire une probabilité *a posteriori* compte tenu des observations dont on dispose. Si on note Y les observations d'une variable d'intérêt, X les observations des K régresseurs disponibles, et M un modèle caractérisé par les régresseurs qu'il comporte et des paramètres de modélisation, alors la probabilité *a posteriori* s'obtient selon la formule de *Bayes* à partir d'une probabilité *a priori* $P(M_i)$ et de la vraisemblance des données conditionnellement au modèle $P(Y|M_i)$:

$$P(M_i|Y)=\frac{P(Y|M_i)P(M_i)}{P(Y|X)}=\frac{P(Y|M_i)P(M_i)}{\sum_{l=1}^{2^k}P(Y|M_l)P(M_l)}$$

La combinaison de modèles dans une approche bayésienne consiste à combiner les prévisions de différents modèles en les pondérant par leur probabilité *a posteriori*. La prévision d'une variable d'intérêt y donnée sera donc la combinaison des prévisions obtenues à partir des différents modèles pondérés par leur probabilité *a posteriori* définie ci-dessous :

$$\hat{y}_{T+h}=\sum_{i=1}^{2^k}P(M_i|y_1,\dots y_T)\hat{y}_{T+h,i}$$

L'intérêt de cette approche réside dans le fait que le modélisateur peut fixer les distributions *a priori* des modèles et, par ce biais, favoriser les modèles parcimonieux en affectant un poids plus petit à ceux qui comportent un grand nombre de régresseurs. En effet, lorsqu'un grand nombre de variables sont disponibles, le risque est d'aboutir à un modèle qui explique bien les observations passées mais qui n'est pas très performant en prévision (on parle alors de « surapprentissage », *encadré 3*). La prévision est réalisée en combinant les L modèles les plus probables. Le nombre de modèles retenus est choisi par le modélisateur, en tenant compte de la distribution *a posteriori* des modèles : ainsi, si la distribution est plate, il est préférable de retenir un nombre de modèles L assez élevé puisque ces derniers sont équiprobables, alors que ce nombre sera plus faible si seuls quelques modèles se distinguent significativement des autres. La prévision peut s'écrire de la façon suivante :

$$\hat{y}_{T+h}=\frac{\sum_{i=1}^{L}P(M_i|y_1,\dots,y_T)\hat{Y}_{T+h,i}}{\sum_{i=1}^{L}P(M_i|y_1,\dots,y_T)}$$

De cette façon, certaines variables ont plus souvent été retenues dans les modèles combinés que d'autres : on a donc procédé indirectement à une sélection des variables. ■

Apports de Google Trends pour prévoir la conjoncture française : des pistes limitées

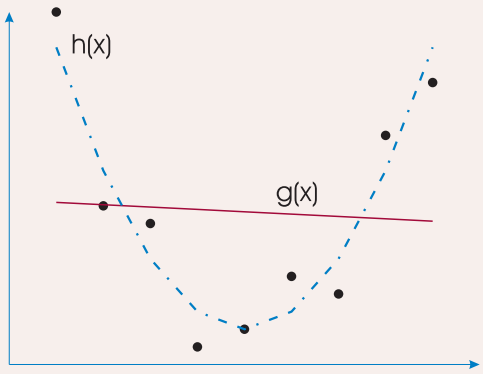
Encadré 3 - En prévision, le mieux peut être l'ennemi du bien

Pourquoi chercher à favoriser des modèles parcimonieux si l'on dispose d'un grand nombre de variables qui semblent toutes pertinentes ? L'intégration d'un grand nombre de variables permet toujours d'améliorer l'ajustement des modèles aux observations, mais elle peut aussi détériorer les propriétés prédictives de ces modèles. Un modèle avec un ajustement parfait sur une période donnée sera difficilement généralisable, c'est-à-dire qu'il pourra être beaucoup moins performant dès que l'on introduira de nouvelles observations.

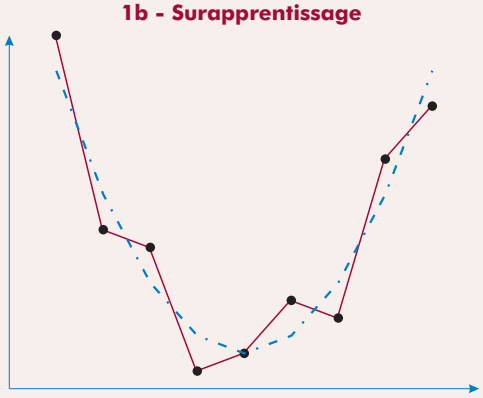
Pour s'en rendre compte, on s'intéresse aux données générées à partir d'une fonction h et d'un aléa ε : $y(x)=h(x)+\varepsilon$. En pratique, et dans cet exemple illustratif très simple, la dépendance de y en x est quadratique. Le prévisionniste, qui ignore en principe cette « vraie » relation, peut chercher un modèle le plus simple g(x) qui consiste en une dépendance simplement linéaire. Ce modèle fait intervenir peu de paramètres mais génère un biais important : l'ajustement est mauvais (*graphique 1a*). Inversement, le prévisionniste peut, en ajoutant un grand nombre de paramètres dans la modélisation, effectuer un ajustement parfait (*graphique 1 b)*. Mais cet ajustement parfait capte, à tort, l'aléa intervenant ici dans le processus de génération de ces données, par définition imprévisible. Pour éviter le « surapprentissage » illustré dans le graphique de droite, mais réduire le biais que l'on constate sur le graphique de gauche, un arbitrage doit être fait entre biais et variance. En général, on utilise des critères de performance évalués hors échantillon pour juger de cette pertinence, comme le RMSE (*root mean square error*, c'est-à-dire la racine de la moyenne des écarts au carré entre prévision et réalisé).

Cet aspect n'est pas automatiquement pris en compte dans l'approche bayésienne puisque la probabilité *a posteriori* d'un modèle repose sur la vraisemblance des observations et non sur un critère de performance en prévision. Or la performance des modèles en prévision dépend de manière cruciale de la taille moyenne des modèles *a priori* (*via* la distribution *a priori* des modèles choisis). Pour sélectionner la meilleure combinaison de modèles en fonction de ses performances en prévision, et arbitrer entre biais et variance, on procède par validation croisée en calculant les RMSE hors échantillon des modèles sélectionnés par l'algorithme pour différentes valeurs de la taille moyenne *a priori* des modèles. Le phénomène de surapprentissage se manifeste lorsque la taille a priori du modèle augmente : lorsqu'on favorise les modèles intégrant un grand nombre de variables, le RMSE en échantillon s'améliore continuellement, tandis que le RMSE hors échantillon finit par se dégrader (*graphique 2*). La valeur finalement retenue pour la taille *a priori* du modèle sera celle qui minimise le RMSE hors échantillon. ■

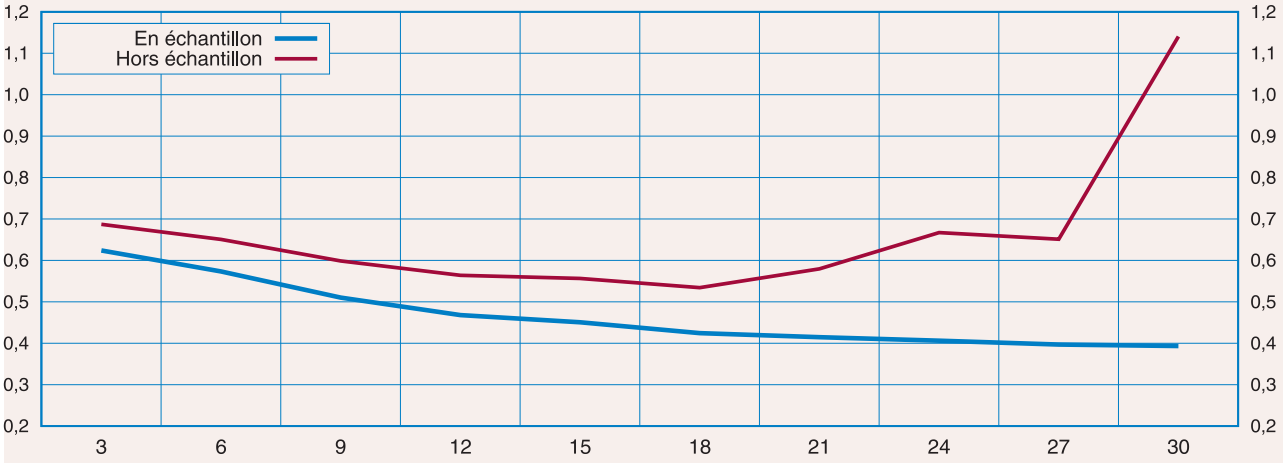
1a - Estimation linéaire



1b - Surapprentissage



2 - RMSE en et hors échantillon en fonction de l'espérance de la taille *a priori* des modèles



Apports de Google Trends pour prévoir la conjoncture française : des pistes limitées

Google Trends apparaissent certes parmi les variables explicatives qui ressortent fréquemment, mais elles présentent un caractère hétérogène (« Entretien des véhicules », « Électronique grand public », « Mobilier de maison », « Santé », etc). Une partie de la difficulté à expliquer finement par ces catégories l'évolution de la consommation de biens à un niveau agrégé peut s'expliquer en partie par le caractère hétérogène de cette consommation : la dynamique des dépenses en automobiles est *a priori* très différente de celles en produits alimentaires. Des résultats similaires sont également obtenus pour la consommation mensuelle en services : cela incite à considérer la consommation des ménages à un niveau plus désagrégé.

… mais des résultats plus favorables sont obtenus pour certains postes de dépenses en biens ou en services

Pour la prévision de la consommation par bien, il existe déjà aujourd'hui des indicateurs conjoncturels fiables pour estimer quasiment en temps réel les achats d'automobiles et les dépenses énergétiques. Trois autres postes ont donc été ciblés en priorité pour tester l'intérêt des données Google Trends : les dépenses alimentaires, en habillement et en biens d'équipement du logement. Les résultats obtenus à ce niveau fin sont plus positifs. Néanmoins, pour ces trois postes, les performances en prévision des séries Google Trends ne sont pas équivalentes.

Google Trends permet d'améliorer la prévision des dépenses en habillement et en équipement du logement

Pour les dépenses en habillement, les variables explicatives qui ressortent le plus souvent dans l'approche bayésienne sont les deux premiers retards de la variable modélisée, ce qui illustre le caractère très auto-corrélé de la série. Les catégories Google Trends « Habillement » et « Articles de sport », dont les thématiques sont clairement liées au type de dépenses étudié, apparaissent également parmi les régresseurs les plus probables. L'utilisation des catégories Google Trends permet d'améliorer la prévision de ce poste de consommation : l'erreur quadratique moyenne de prévision (RMSE) est diminuée d'environ 10 % par rapport à un simple modèle autorégressif (*encadré 4*). Les résultats obtenus pour la prévision des achats de biens d'équipement du logement sont également positifs : la variable explicative qui ressort le plus souvent est la catégorie « Mobilier de maison », dont le thème est là encore étroitement lié au type d'achats considéré. L'amélioration de la prévision est cependant plus modeste, puisque le RMSE est diminué de moins de 5 % par rapport à un modèle simple utilisant la seule dynamique de la variable à prévoir (ARMA).

L'apport est plus limité pour les dépenses alimentaires…

Enfin, concernant les dépenses alimentaires, l'utilisation des tendances de recherche *Google* n'améliore pas la prévision par rapport à un modèle autorégressif. Toutefois les catégories Google Trends « Produits du tabac » et « Boissons alcoolisées » apparaissent aux côtés du premier retard de la variable modélisée parmi les régresseurs les plus souvent sélectionnés. Le premier retard de la catégorie « Sport » est également un régresseur qui ressort souvent, mais un lien direct avec les dépenses alimentaires semble plus difficile à établir.

… comme pour les services de transport

Concernant la consommation de services, les résultats les plus probants ont été obtenus sur le poste « Transport ». En effet, si les catégories Google Trends ne permettent pas d'améliorer la qualité de la prévision « hors échantillon », certaines d'entre elles apparaissent néanmoins parmi les régresseurs les plus probables dans l'approche bayésienne, comme les catégories « Voyages aériens » et « Hôtels et hébergement ». En échantillon, la catégorie « Voyages aériens » permet notamment de bien ajuster les évolutions heurtées de la consommation de services de transports en 2010, année pendant laquelle le trafic aérien a été particulièrement perturbé en France, en raison de l'éruption du volcan islandais Eyjafjallajökull au printemps et de la pénurie de glycol dans les aéroports au moment de l'épisode neigeux de décembre.

Apports de Google Trends pour prévoir la conjoncture française : des pistes limitées

Encadré 4 - La prévision de la consommation d'habillement peut-être enrichie en incluant des catégories Google Trends grâce à l'approche bayésienne

Pour certains types d'achats comme les dépenses en habillement, l'approche bayésienne incluant des catégories Google Trends peut permettre d'améliorer les prévisions, en les enrichissant notamment par rapport à un modèle simple utilisant la seule dynamique passée de la série (de type ARMA). De plus, elle présente deux caractéristiques intéressantes :

- Les probabilités *a posteriori* des différents modèles possibles s'avèrent assez proches, ce qui justifie de chercher à combiner ces derniers. Leur combinaison permet d'améliorer les performances en prévision (*graphique 1*).
- La paramétrisation de l'espérance *a priori* de la taille des modèles permet d'optimiser le modèle en fonction de ces performances en prévision ; on peut donc contrôler le risque de surapprentissage (*encadré 3*).

L'approche bayésienne peut s'interpréter comme une sélection indirecte de variables, par le biais de la sélection des modèles affectés d'une plus forte probabilité *a posteriori* car elle détermine les régresseurs ayant la plus forte probabilité d'inclusion dans les modèles retenus. L'approche bayésienne est globalement équivalente en termes de performance aux méthodes de sélection de variables par algorithme itératif classique (algorithme *stepwise* ou *pc-gets*).

Les catégories Google Trends les plus probables obtenues par l'approche bayésienne pour la prévision des dépenses en habillement sont listées dans le *tableau* suivant.

Dans cet exercice, les deux premiers retards de la variable modélisée ressortent le plus souvent, ce qui illustre le caractère très auto-corrélé de la série ; notamment, après une poussée des dépenses liée par exemple à une situation exceptionnelle de soldes, la probabilité d'un contrecoup le mois suivant est importante. Deux catégories Google Trends directement liées aux dépenses d'habillement apparaissent également parmi les variables explicatives les plus probables : les catégories « Habillement » et « Articles de sport ». Il n'est cependant pas possible de garantir la pertinence de toutes les séries sélectionnées, certaines pouvant être fortement corrélées avec les dépenses d'habillement sans qu'il n'y ait de lien de causalité avéré (par exemple « Livres et littérature » ou encore « Achat de véhicules »).

En ne conservant que les catégories *a priori* les plus pertinentes parmi les catégories les plus probables, on peut alors améliorer les prévisions mensuelles tout en privilégiant la simplicité du modèle et en évitant certains écueils tels que les problèmes de multicolinéarité. Ainsi, l'inclusion des catégories Google Trends (« Articles de sport » et « Habillement ») permet d'améliorer la prévision des dépenses d'habillement par rapport à un modèle autorégressif utilisant la seule dynamique de la série : le RMSE hors échantillon est réduit de 11 % *(graphique 2)*. ■

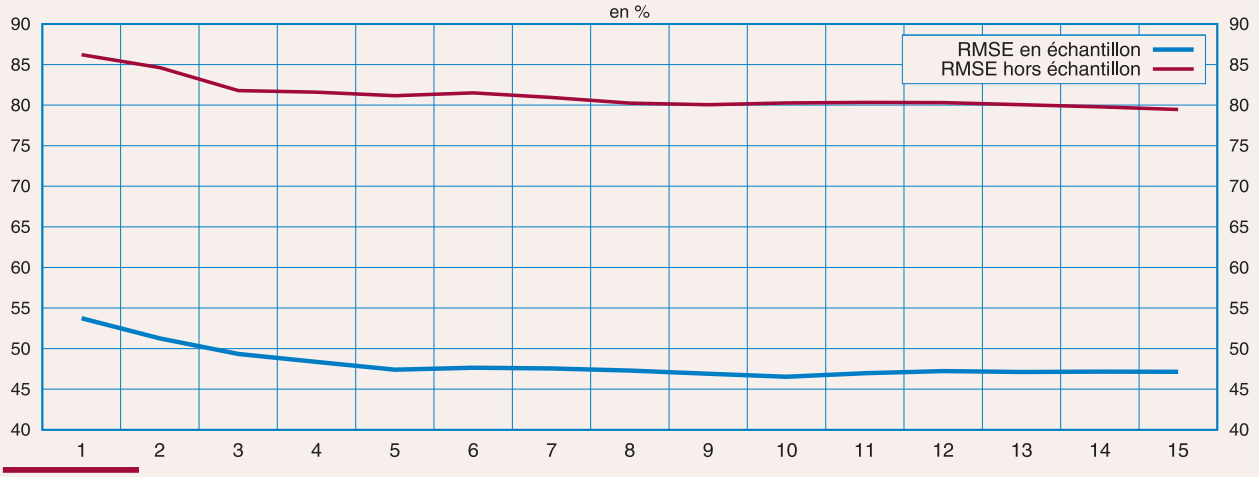
Probabilité d'inclusion et coefficient moyen des variables explicatives de la croissance des dépenses d'habillement les plus probables

Nom du régresseur	Probabilité a posteriori d'inclusion	Coefficient moyen
Premier retard de la variable modélisée	1,00	-0,56 (0,10)
Deuxième retard de la variable modélisée	0,99	-0,36 (0,11)
Sport (catégorie Google Trends)	0,96	-0,18 (0,06)
Habillement (catégorie Google Trends)	0,95	0,39 (0,15)
Livres et littérature (catégorie Google Trends, premier retard)	0,84	-0,25 (0,13)
Mobilier de maison (catégorie Google Trends)	0,58	-0,14 (0,14)
Achat de véhicules (catégorie Google Trends)	0,46	-0,10 (0,12)
Articles de sport (catégorie Google Trends)	0,44	0,14 (0,17)

Lecture : la probabilité *a posteriori* d'inclusion dans le modèle correspond à la somme des probabilités *a posteriori* des modèles dans lequel le régresseur apparaît. Elle est de 95 % pour la catégorie Google Trends « Habillement ». Le coefficient moyen correspond à la moyenne du coefficient sur le régresseur pour les modèles les plus probables, pondérée par la probabilité *a posteriori* du modèle. Les erreurs standards moyennes sont indiquées entre parenthèses.

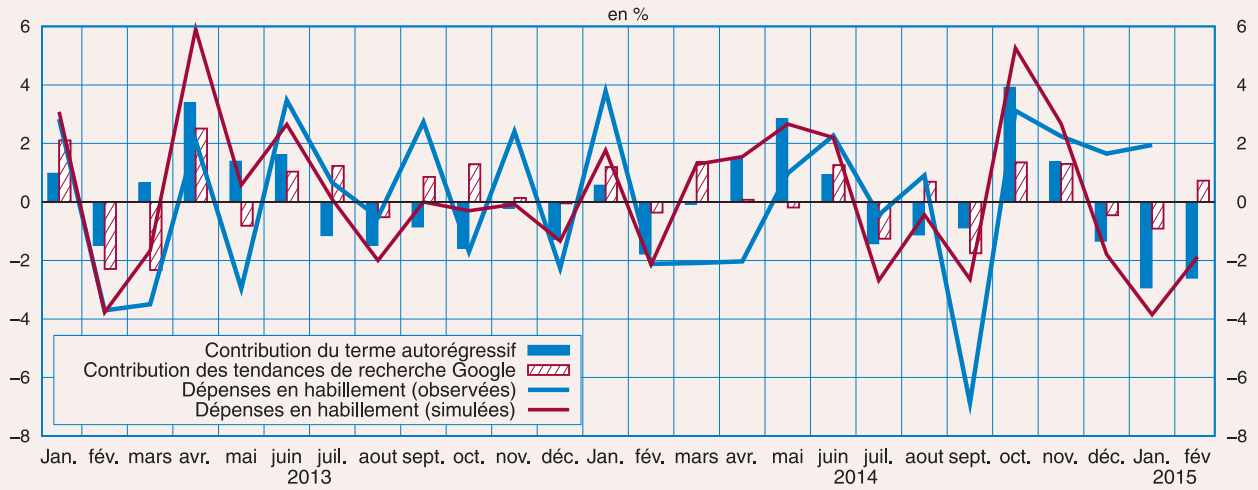
Apports de Google Trends pour prévoir la conjoncture française : des pistes limitées

1 - Dépenses d'habillement : évolution de l'erreur moyenne en fonction du nombre de modèles retenus



Lecture : Erreurs moyennes au carré exprimées en pourcentage de la variance de la série. Lorsque cinq modèles sont retenus, les erreurs moyennes en et hors échantillon sont respectivement égales à 47 % et 81 % de la variance de la série.

2 - Variation mensuelle des dépenses de consommation en habillement



Lecture : la série historique des dépenses en habillement a été comparée depuis début 2013 à un modèle simple utilisant le premier retard de la série et les tendances de recherche Google « habillement » et « articles de sport » (t-stats entre parenthèses) :

$$Dépenses\_habillement_t = \underset{(-6,4)}{-0,62}\ Dépenses\_habillement_{t-1} \underset{(-5,1)}{-0,51}\ Dépenses\_habillement_{t-2} \underset{(-2,8)}{-0,24}\ Dépenses\_habillement_{t-3} + \underset{(2,5)}{0,23}\ Google\_articles\_sport_t + \underset{(3,8)}{0,42}\ Google\_habillement_t$$

R2 ajusté = 0,55 RMSE (en échantillon) = 2,21 RMSE (hors échantillon) = 2,80 ■

Apports de Google Trends pour prévoir la conjoncture française : des pistes limitées

Conclusion : les recherches effectuées par les internautes sont informatives mais se révèlent limitées en pratique pour la prévision conjoncturelle

D'après les différentes modélisations testées, l'ajout des tendances de recherche sur *Google* ne permet d'améliorer la prévision des dépenses mensuelles des ménages que dans des cas ciblés. Plus précisément, ces séries ne permettent pas d'améliorer la prévision des dépenses de consommation mensuelles des ménages en biens ou en services lorsque ces dernières sont considérées à un niveau agrégé, du fait de la forte hétérogénéité des évolutions par produit. En revanche, les résultats obtenus pour les achats de certains biens (habillement et équipement de la maison, notamment) sont encourageants, certaines catégories Google Trends apparaissant en effet comme des variables explicatives probables. Cependant, lorsque la prévision est améliorée, elle ne l'est que modestement. En outre, le fait qu'elle ne le soit que pour quelques produits fait courir le risque que ces résultats favorables puissent être imputés à la « chance » : bien que non nul, ce risque semble faible puisque les catégories Google Trends les plus probables pour modéliser la consommation des produits étudiés sont généralement directement liées à ces derniers.

En outre, comme l'illustre l'exemple de l'indicateur Google Flu bâti sur le même principe, la pérennité de ces résultats devrait être régulièrement vérifiée. En effet, la stabilité dans le temps de leur mode de construction, sujet aux évolutions stratégiques ou technologiques de Google et de son moteur de recherche, ainsi que des comportements des internautes, paraît difficile à assurer.

Ces limites vaudraient *a fortiori* pour une utilisation de Google Trends pour produire des statistiques de consommation. ■

Bibliographie

Askitas N. et Zimmermann K. F. (2009), « Google Econometrics and Unemployment Forecasting », *Applied Economics Quarterly* 55(2), 107-120.

Choi H. et Varian H. (2009), *Predicting the Present with Google Trends,* Technical report, Google.

Faure M.-E., Kerdrain C. et Soual H. (2012), « La consommation des ménages dans la crise », *Note de Conjoncture*, juin, p. 23-37, Insee.

Ginsberg J., Mohebbi M. H., Patel R. S., Brammer L., Smolinski M. S. et Brilliant L. (2009), « Detecting influenza epidemics using search engine query data », *Nature 457*, p. 1012-1014.

Kulkarni R., Haynes K., Stough R. et Paelinck J. (2009), « Forecasting housing prices with Google econometrics », *Research Paper* 2009-10, George Mason University School of Public Policy.

Krankadler É. (2014), « Où fait-on ses courses ? », *Insee Première*, n° 1526, Insee.

Lazer D., Kennedy R,. King G. et Vespignani A. (2014), « The parable of Google Flu: traps in big data analysis », *Science*, 343, p. 1203-1205.

Raftery A.E, Madigan D. et Hoeting J.A (1997), « Bayesian Model Averaging for Linear Regression Models », *Journal of the American Statistical Association*, Vol. 92, n° 437, p. 179-191.

Vosen S. et Schmidt T. (2011), « Forecasting Private Consumption: Survey-Based Indicators vs. Google Trends », *Journal of Forecasting*, Vol. 30, n° 6, p. 565-578.■