Se servir de Google pour prévoir la conjoncture française ?

Des pistes limitées





Clément Bortoli
Stéphanie Combes

Insee
Mesurer pour comprendre

08/04/2015

Introduction

- Depuis quelques années, l'usage des données issues d'Internet connait un grand essor.
- Google Trends permet par exemple de connaître les tendances de recherche des utilisateurs de Google depuis janvier 2004.
- Intérêt pour le conjoncturiste : des données mobilisables plus rapidement que la plupart des indicateurs quantitatifs classiques.
- Ces données pourraient être considérées comme des indicateurs d'intention d'achat : s'en servir pour prévoir la consommation des ménages apparaît donc pertinent.
- Or, cette dernière joue un rôle central dans l'économie :
  - 54 % du PIB en moyenne depuis 1980 ;
  - 37 % de la variance de la croissance du PIB.

Principaux messages

- En pratique, l'intérêt des données Google Trends pour prévoir la consommation des ménages est limité :

  - Leur utilisation ne permet pas d'améliorer la prévision globale des dépenses mensuelles des ménages en biens ou en services ;

  - Elle permet en revanche d'améliorer la prévision de certains sous-postes (habillement et équipement du logement, notamment).

Plan de la présentation

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

II. Google Trends, de quoi s'agit-il ?

III. Combinaison de modèles et sélection de variables : présentation de la méthodologie utilisée

IV. Présentation des résultats : une amélioration modeste de la prévision des achats de certains produits

Consommation des ménages, les chiffres-clés

Depuis 1980, la consommation des ménages c'est…

- Le principal poste de la demande finale intérieure (environ la moitié).
- Entre 52 % et 56 % du PIB (en euros courants).
- 37 % de la variance de la croissance trimestrielle (« volatilité ») du PIB, soit moins que l'investissement (40 %), mais plus que la variation des stocks (22 %).

La consommation des ménages en biens

La consommation des ménages en biens explique en grande partie la volatilité de leurs dépenses totales (84 % depuis 1980).

	Poids dans la consommation en biens (en euros courants)	Contribution à la volatilité 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 %**

La consommation des ménages en services

Le poids des services dans les dépenses des ménages n'a cessé d'augmenter depuis le début des années 1980 : de 40 % à 53 % (en euros courants).

	Poids dans la consommation en services (en euros courants)	Contribution à la volatilité 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 %**

Les recherches en ligne pourraient apporter une information précoce sur la consommation

- Les indicateurs de consommation sont disponibles progressivement

- Internet joue un rôle croissant dans les achats effectués par les ménages, notamment en biens.

Part d'Internet dans les dépenses…	2006	2011
… alimentaires	**0,3 %**	**0,6 %**
… d'habillement	**0,7 %**	**4,1 %**
… de biens durables	**2,3 %**	**8,6 %**

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

- ***Conclusion*** : les statistiques de recherche en ligne sont susceptibles d'apporter une information précoce sur la consommation, ou du moins sur les intentions d'achat.

Plan de la présentation

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

II. Google Trends, de quoi s'agit-il ?

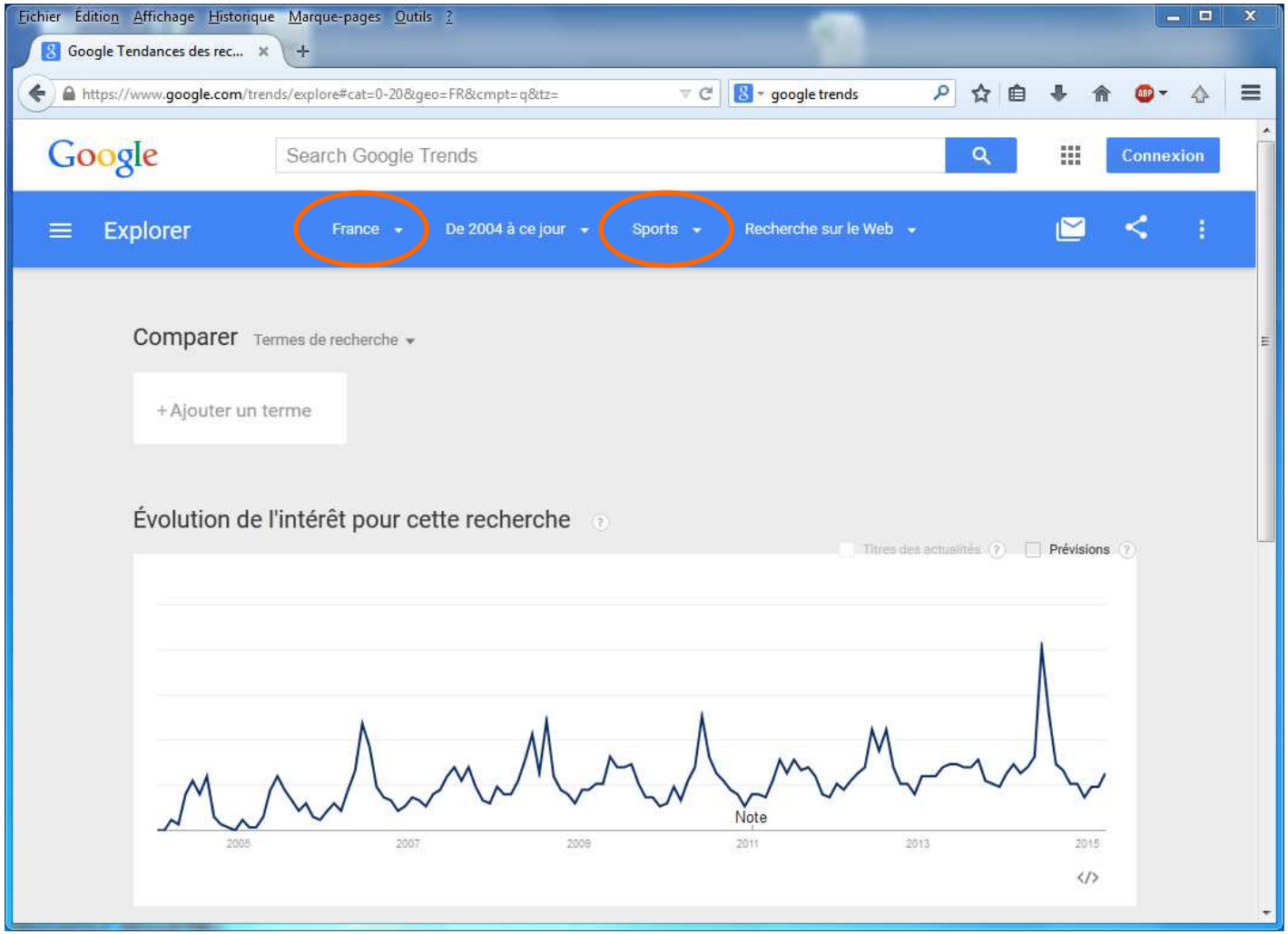
III. Combinaison de modèles et sélection de variables : présentation de la méthodologie utilisée

IV. Présentation des résultats : une amélioration modeste de la prévision des achats de certains produits

Google Trends, de quoi s'agit-il ?

- Outil mettant gratuitement à disposition des séries reflétant les tendances de recherche des utilisateurs de Google.
- Séries hebdomadaires, commençant en 2004.
- Possibilité de filtrer l'origine géographique des requêtes utilisées.
- Les données de Google Trends sont déjà utilisées pour prévoir des indicateurs socio-économiques depuis plusieurs années :
  - Application Google Flu (2008 aux États-Unis, 2009 en Europe) ;
  - Choi et Varian (2009) ;
  - Askitas et Zimmermann (2009) ;
  - Kulkarni et al. (2009)
  - Vosen et Schmidt (2011).

Google Trends, de quoi s'agit-il ?



Source : Google Trends

Les catégories Google Trends couvrent certains postes de dépenses des ménages

Certaines catégories semblent *a priori* pertinentes pour prévoir la consommation des ménages…

- …en biens :

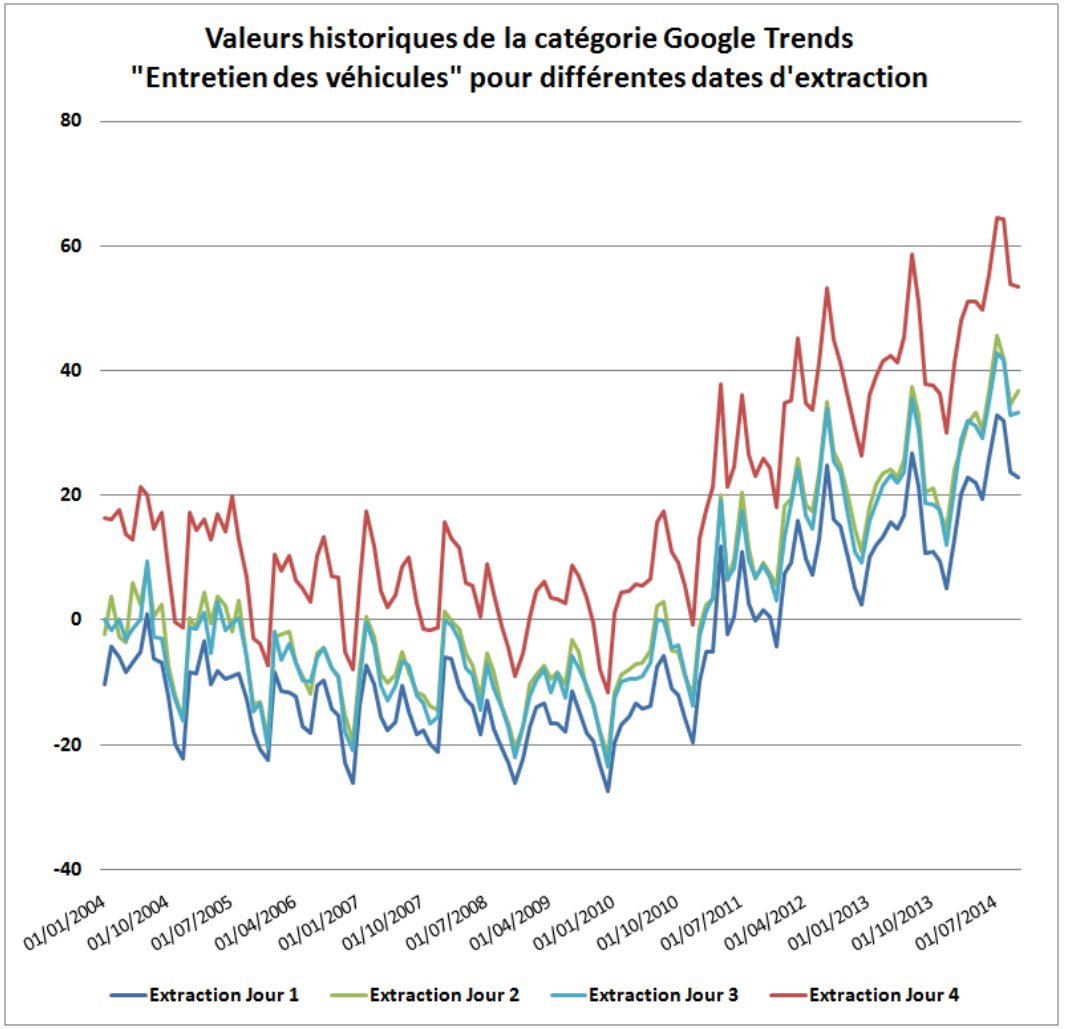
Alimentaire	Automobiles	Équipement du logement	Habillement
• Produits du tabac • Boissons alcoolisées • Boissons non alcoolisées • Épiceries et magasins d'alimentation • Grands magasins et hypermarchés	• Automobiles et véhicules • Achats de véhicules • Entretien des véhicules • Marques automobiles • Pièces et accessoires pour véhicules • Motos • Scooters cyclomoteurs	• Informatique et électronique • Internet et télécoms • Électronique grand public • Appareils mobiles et sans fil • Ordinateurs portables et notebooks • Appareils ménagers • Mobilier de maison • Maison et jardinage • Sports	• Habillement • Articles de sport

- …en services :

Transport	Hébergement et Restauration	Information et Communication	Services financiers
• Voyages • Location de voitures et taxis • Voyages aériens • Bus et trains	• Restaurants • Hôtels et hébergements	• Internet et télécoms • Fournisseurs d'accès et opérateurs • Appareils mobiles et sans fil • Livres et littérature	• Banque • Assurance

Google Trends présente certaines faiblesses

- Le traitement des données manque de transparence.
- Les valeurs prises par les séries peuvent varier d'une extraction à l'autre.
- Les algorithmes utilisés par le moteur de recherche peuvent évoluer et entraîner un changement de la manière dont les usagers l'utilisent.



Source : Google Trends

Plan de la présentation

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

II. Google Trends, de quoi s'agit-il ?

III. Combinaison de modèles et sélection de variables : présentation de la méthodologie utilisée

IV. Présentation des résultats : une amélioration modeste de la prévision des achats de certains produits

Le nombre élevé de régresseurs potentiels nécessite une stratégie de sélection

- Une cinquantaine de catégories Google Trends ont été ciblées pour la prévision.
- On retient également leur premier retard
- Les variables explicatives sont donc au nombre d'une centaine… pour environ 130 observations (séries mensuelles, commençant en janvier 2004).
- Ainsi, il est nécessaire de procéder à une sélection en utilisant :
  - Des dires d'experts
  - Un algorithme itératif (*stepwise*, *pc-gets*…)
  - Une fonction objectif à minimiser pénalisant le manque de parcimonie (critère d'information, Lasso…)
  - Une analyse en composante principale
  - Une combinaison de modèles

Combinaison de modèles par approche bayésienne : le principe

On fixe une probabilité *a priori* sur un modèle M_i et on en déduit une probabilité *a posteriori* en fonction des observations de la variable à expliquer ***y*** et des ***k*** variables explicatives ***X***.

Probabilité *a posteriori* du modèle M_i

Vraisemblance marginale du modèle M_i

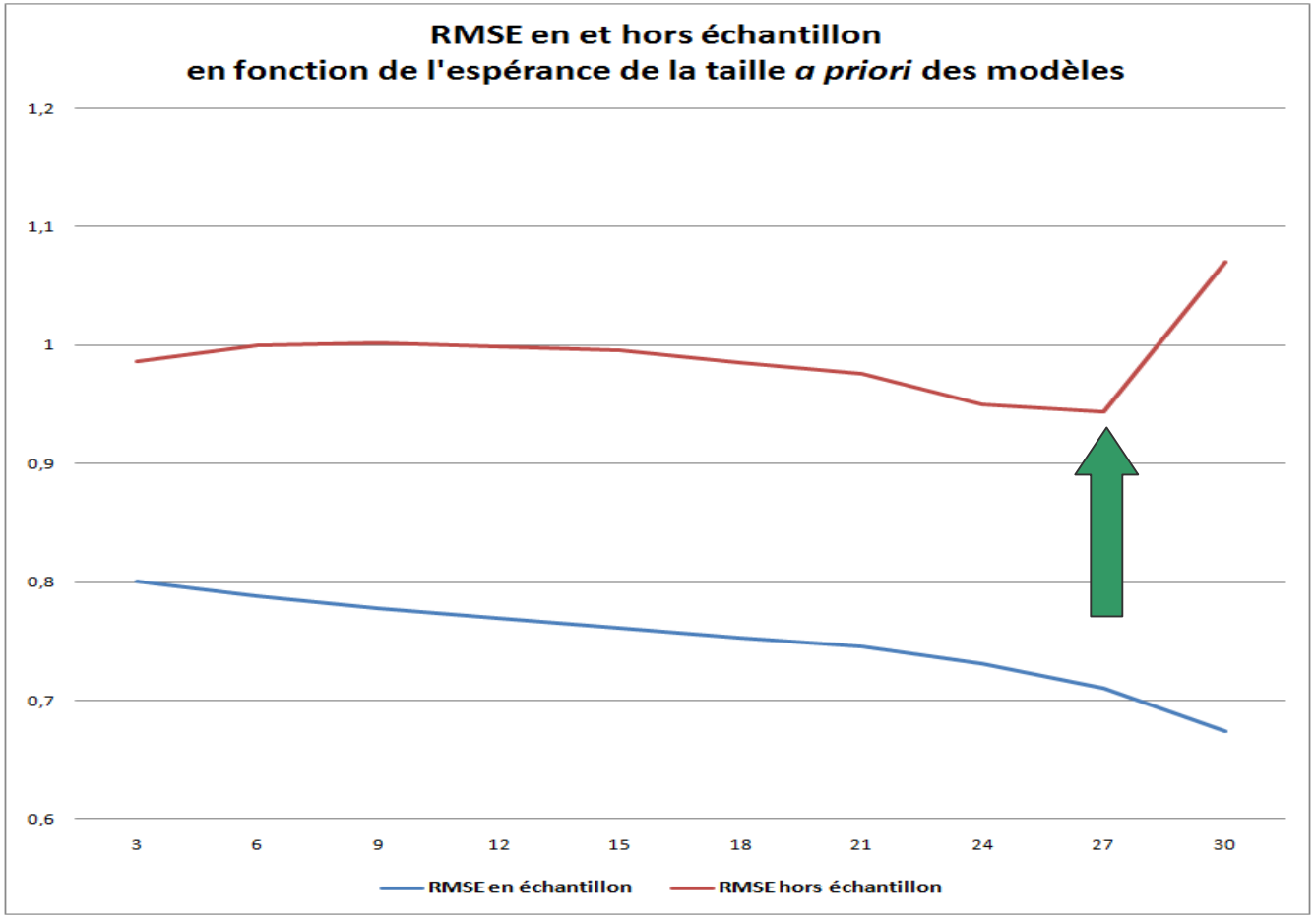
Probabilité *a priori* du modèle M_i

$$p(M_i|y,X) = \frac{p(y|M_i,X).p(M_i)}{p(y|X)} = \frac{p(y|M_i,X).p(M_i)}{\sum_{l=1}^{2^k} p(y|M_l,X).p(M_l)}$$

- La prévision de ***y*** sera donc la combinaison des prévisions obtenues à l'aide des différents modèles pondérés par leur probabilité *a posteriori*.

$$\hat{y}_{t+1} = \sum_{l=1}^{2^k} p(M_l|\underline{y_t},\underline{X_t}).\hat{y}_{l,t+1} \approx \frac{\sum_{l=1}^{L} p(M_l|\underline{y_t},\underline{X_t}).\hat{y}_{l,t+1}}{\sum_{l=1}^{L} p(M_l|\underline{y_t},\underline{X_t})}$$

Choix du paramétrage : contrôle du surapprentissage par validation croisée



- La valeur finalement retenue pour la taille *a priori* du modèle sera celle qui minimise le RMSE hors échantillon.

Mise en pratique de la combinaison de modèles pour les estimations

- Variables à prévoir : croissance mensuelle de la consommation des ménages en biens (en volume) publiée par l'Insee à la fin du mois suivant le mois d'intérêt.

- Variables explicatives : les quatre premiers retards de la variable à prévoir, les taux de croissance mensuels d'une cinquantaine de catégories Google Trends (qui ont été au préalable mensualisées et désaisonnalisées) ainsi que leur premier retard.

- Période d'estimation : mars 2004 – décembre 2011.

- La qualité de la prévision (RMSE hors échantillon) est ensuite mesurée sur la période janvier 2012 – décembre 2014.

Plan de la présentation

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

II. Google Trends, de quoi s'agit-il ?

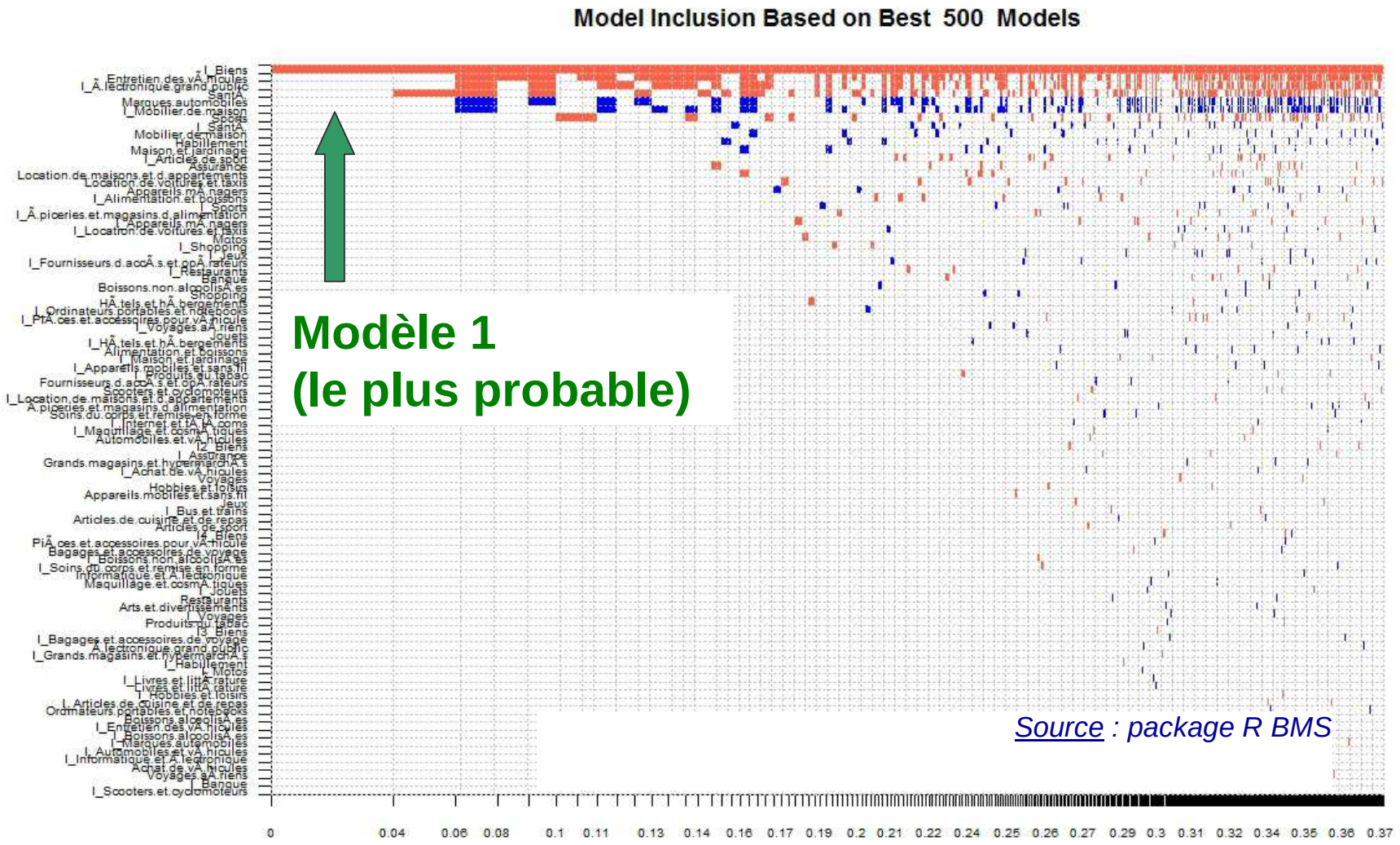
III. Combinaison de modèles et sélection de variables : présentation de la méthodologie utilisée

IV. Présentation des résultats : une amélioration modeste de la prévision des achats de certains produits

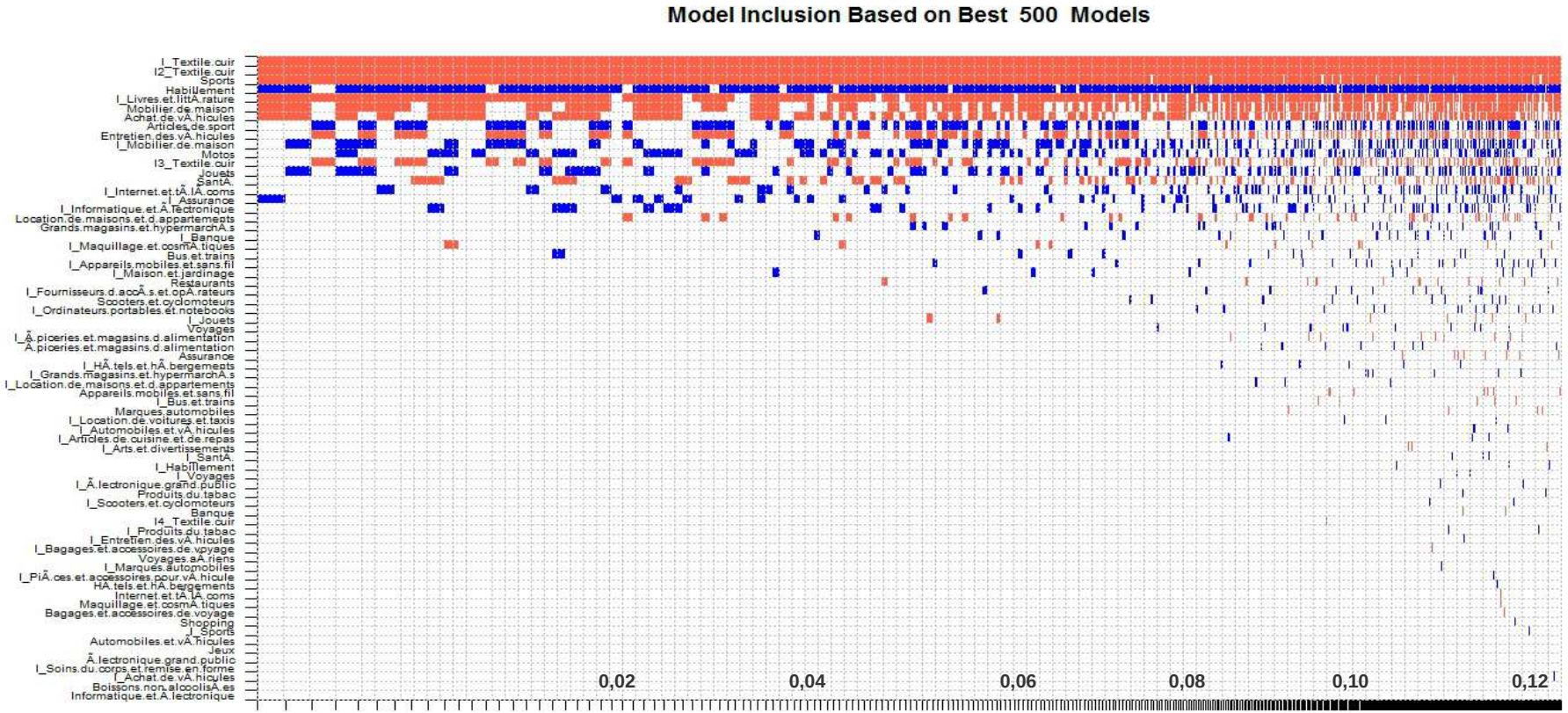
Prévision de la consommation totale des ménages en biens et en services

- Pour la consommation totale en biens, les catégories Google Trends ne permettent pas d'améliorer la prévision par rapport à un simple modèle ARMA.
- La variable la plus pertinente pour expliquer la croissance de la consommation en biens est la croissance de cette même consommation au mois précédent.
- Le caractère hétérogène de la consommation totale des ménages en biens peut expliquer la difficulté à prévoir les évolutions à l'aide des catégories Google Trends.
- Le constat est similaire pour la consommation totale en services.
- Conclusion : la consommation des ménages doit être considérée à un niveau plus désagrégé.

Consommation totale en biens : les résultats détaillés de la combinaison de modèles



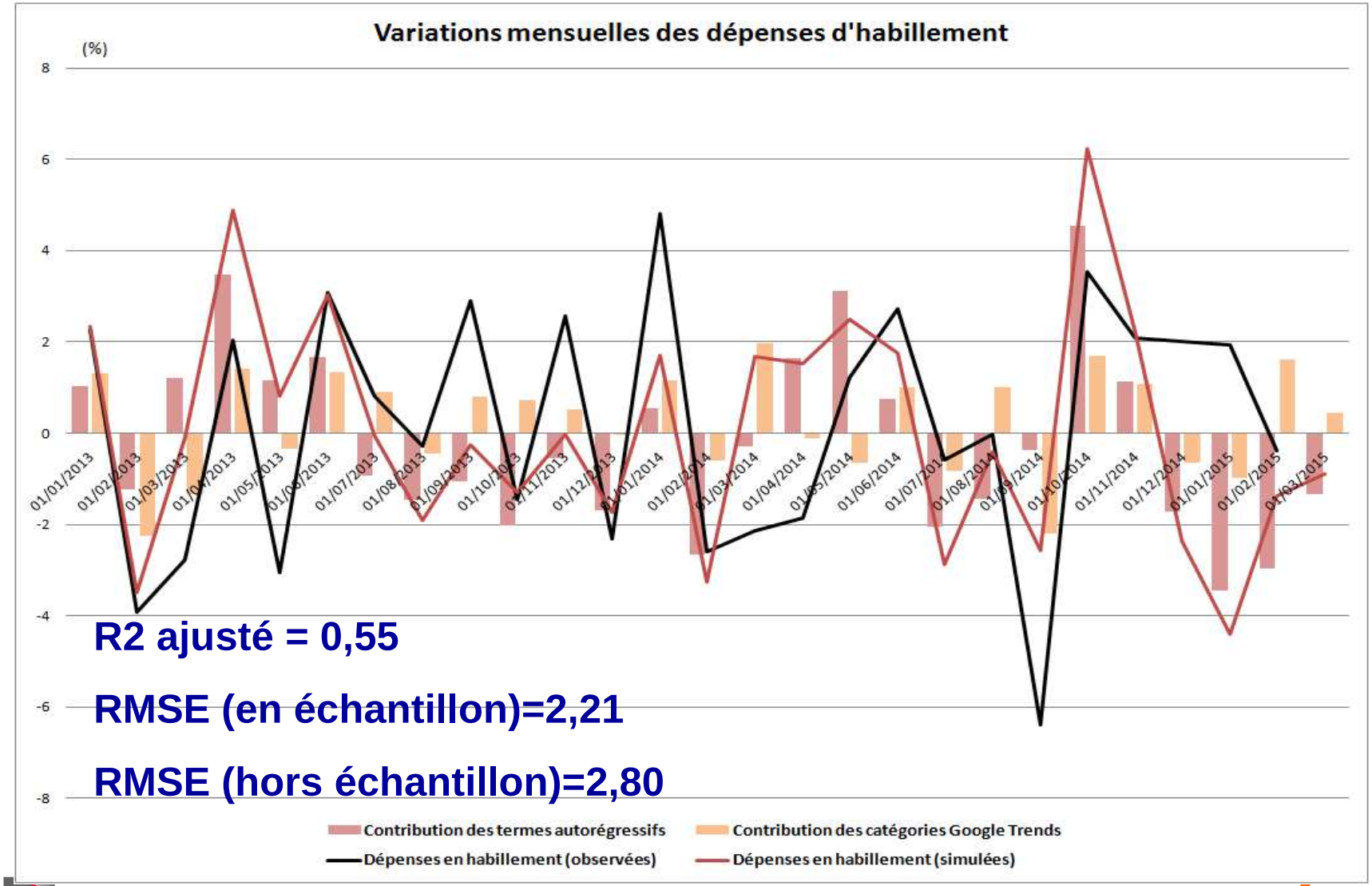
Dépenses d'habillement : les résultats détaillés de la combinaison de modèles



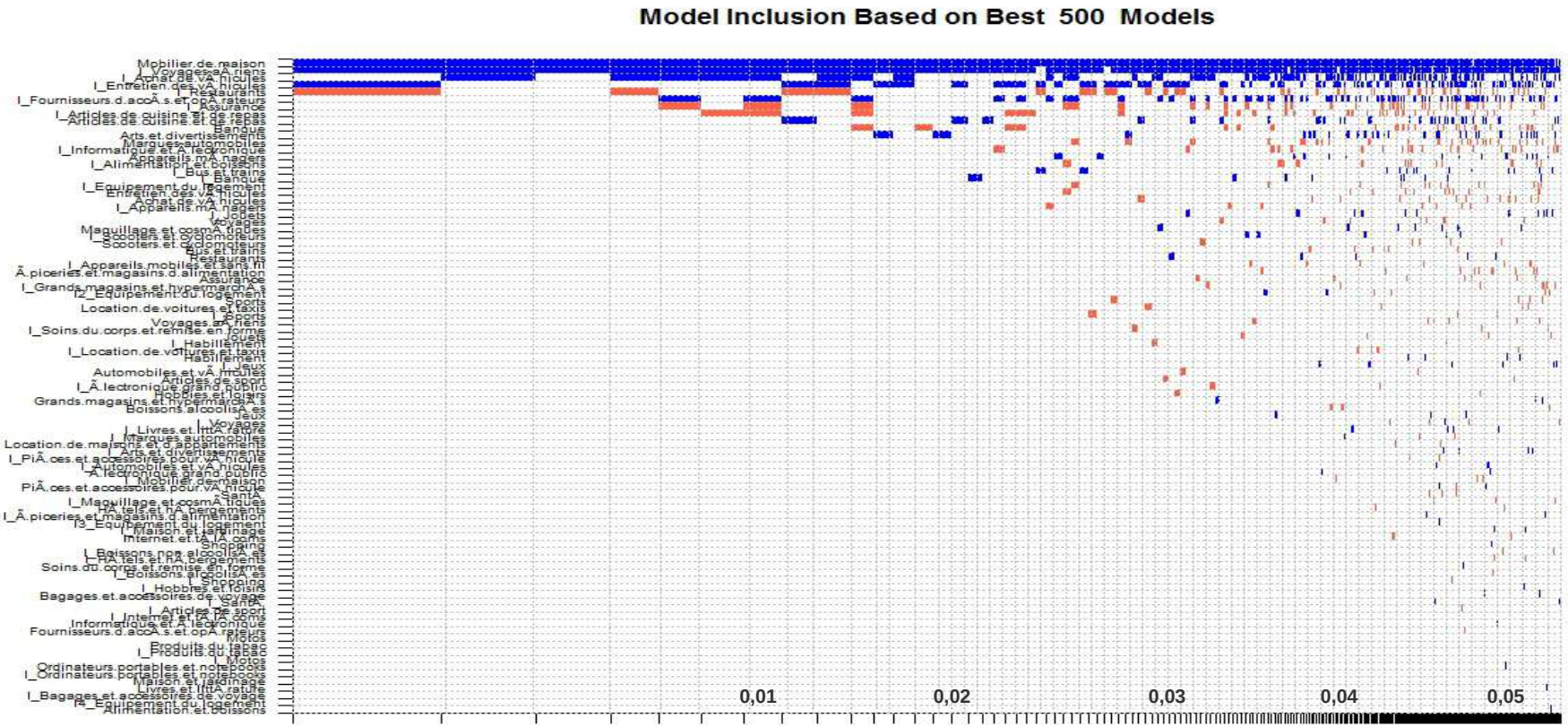
Source : package R BMS

Le RMSE hors échantillon est diminué d'environ 10 % par rapport à un simple modèle autorégressif.

Dépenses d'habillement : proposition d'étalonnage simple



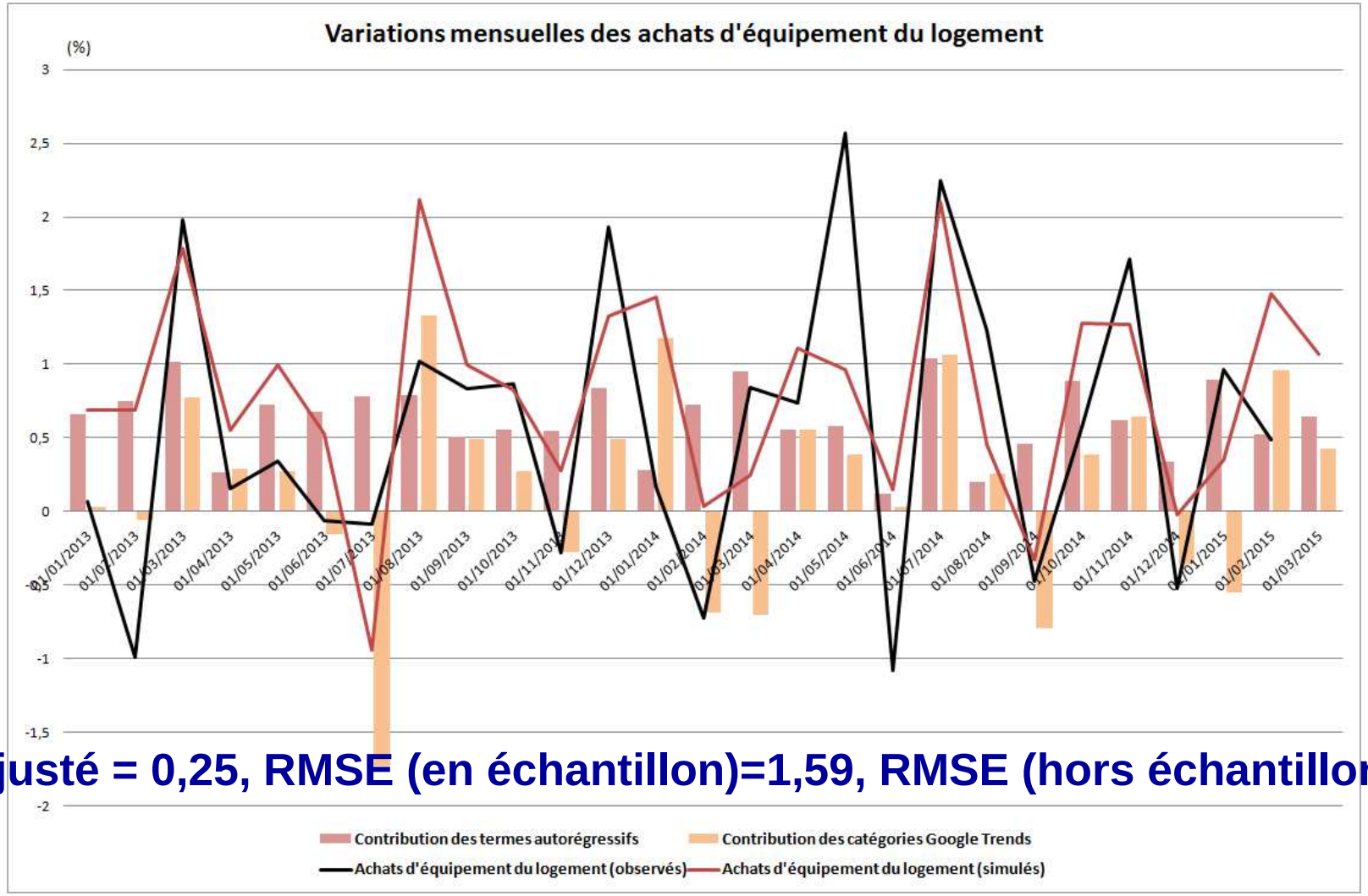
Consommation en équipement du logement : les résultats détaillés de la combinaison de modèles



Source : package R BMS

Le RMSE hors échantillon est diminué d'environ 5 % par rapport à un simple modèle autorégressif.

Achats d'équipement du logement : proposition d'étalonnage simple



R2 ajusté = 0,25, RMSE (en échantillon)=1,59, RMSE (hors échantillon)=0,91

Un apport plus limité pour les dépenses alimentaires et en services de transport

- Pour les dépenses alimentaires et en services de transport, l'utilisation des catégories Google Trends ne permet pas d'améliorer la prévision.
- Néanmoins, certaines catégories *a priori* pertinentes apparaissent dans approche bayésienne parmi les régresseurs les plus probables.
  - Pour les dépenses alimentaires, les catégories « Produits du tabac » et « Boissons alcoolisées » sont, avec le premier retard de la variable modélisée, les régresseurs les plus probables.
  - Pour les dépenses en services de transport, les régresseurs les plus probables sont les catégories « Hôtels et hébergement » et « Voyages aériens ». Cette dernière permet notamment de bien ajuster en échantillon les évolutions heurtées de ces dépenses en 2010.

Conclusion : des données informatives mais dont l'utilisation en prévision est limitée

- L'ajout des données Google Trends ne permet d'améliorer la prévision des dépenses mensuelles des ménages que dans des cas ciblés.

- Les limites des données Google Trends exigent que la pérennité de ces résultats soient régulièrement vérifiée.

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