UNIVERSITÉ DU QUÉBEC À MONTRÉAL

PRÉVISION DE L'ACTIVITÉ DU MARCHÉ DU TRAVAIL AUX ÉTATS-UNIS À L'AIDE DES DONNÉES DE GOOGLE TRENDS

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN ÉCONOMIQUE

PAR

HUGO COUTURE

AOÛT 2020

REMERCIEMENTS

Je voudrais tout d'abord adresser toute ma gratitude à mon directeur de mémoire, M. Dalibor Stevanovic, pour sa disponibilité, sa patience et surtout ses conseils éclairés qui ont contribué à alimenter ma réflexion.

Je tiens également à remercier ma famille et mes amis qui ont su m'appuyer tout au long de ce cheminement. Ce soutien continu m'a permis de réaliser le présent mémoire.

Je témoigne également ma reconnaissance aux membres du personnel administratif du département d'économie qui ont su trouver réponse à mes questionnements tout au long de mon processus.

TABLE DES MATIÈRES

LISTE DES TABLEAUX v
LISTE DES FIGURES vi
RÉSUMÉ vii
INTRODUCTION 1
CHAPITRE I
REVUE DE LITTÉRATURE 5
CHAPITRE II
DONNÉES 9
2.1 Données macroéconomiques 9
2.1.1 Indice du marché du travail 11
2.2 Données de Google Trends 13
2.2.1 Méthodologie de l'indice Google Trend 13
2.2.2 Recherche des mots-clés 15
2.2.3 Méthode de récupération des données de Google 17
2.2.4 Pré-sélection des mots-clés de Google Trends 19
2.3 Transformation des données 21
2.3.1 Données macroéconomiques 21
2.3.2 Données de Google Trends 23
CHAPITRE III
MÉTHODOLOGIE 26
3.1 Modèle à facteurs 26
3.2 Modèles de prévision 28
3.2.1 Modèle MIDAS-AR 28
3.2.2 Modèle LASSO 30
3.2.3 Modèle de référence 31

3.3 Approche prévisionnelle 32
3.4 Critère d'évaluation des prévisions 34
CHAPITRE IV
RÉSULTATS 36
4.1 Résultats des prévisions 36
4.2 Indices de l'activité du marché du travail 39
CONCLUSION 46
APPENDICE A
DONNÉES MACROÉCONOMIQUES 48
A.1 Provenance et transformation des séries 48
A.2 Indices du l'activité du marché du travail 49
APPENDICE B
RÉSULTATS DES ESTIMATIONS 52
B.1 Graphiques des prévisions 52
B.2 Graphiques des EQM dans le temps 59
RÉFÉRENCES 65

LISTE DES TABLEAUX

Tableau Page

2.1 Variables sélectionnées avec la méthode de pré-sélection 21

2.2 Test Dickey-Fuller augmenté avant transformation 23

2.3 Test Dickey-Fuller augmenté après transformation 24

4.1 Ratio des erreurs quadratiques moyennes 37

4.2 Résultats pour le test de signe : Indice construit avec les prévisions des séries 42

4.3 Ratio des erreurs quadratiques moyennes pour la prévision directe de l'indice 43

4.4 Résultats pour le test de signe : Indice construit avec la prévision directe 45

A.1 Variables utilisées pour la construction de l'indice 48

A.2 Résulats de l'analyse par composante principale 49

A.3 Variables sélectionnées avec la méthode de pré-sélection 51

LISTE DES FIGURES

Figure		Page
2.1	Indices du niveau d'activité du marché du travail construites par la KC Fed	11
2.2	Nouvel indice du niveau d'activité du marché du travail	12
4.1	Méthode avec les composantes : Indice résultant pour les meilleurs modèles pour H = 0	41
4.2	Méthode avec les composantes : Indice résultant pour les meilleurs modèles pour H = 1	41
4.3	Méthode par prévision directe : Indice résultant pour les meilleurs modèles pour H = 0	44
4.4	Méthode par prévision directe : Indice résultant pour les meilleurs modèles pour H = 1	45
A.1	Top dix des variables qui contribuent le plus à l'indice	50
B.1	Prévisions des meilleurs modèles pour chaque semaine (*nowcasting*)	55
B.2	Prévisions des meilleurs modèles pour chaque semaine (H = 1)	58
B.3	Moyennes des EQM des meilleurs modèles pour chaque semaine (*nowcasting*)	61
B.4	Moyennes des EQM des meilleurs modèles pour chaque semaine (H = 1)	64

RÉSUMÉ

Ce mémoire avait pour objectif d'estimer un indice décrivant les performances du marché du travail aux États-Unis. Ainsi, un indice comparable à celui fait par Willis *et al.* (2014) de la Réserve Fédérale de Kansas City a été construit. Celui-ci est une moyenne pondérée de 21 séries mensuelles représentant les divers aspects de ce marché. Cependant, sa particularité réside dans le fait qu'il est estimé à l'aide des données provenant de Google Trends. Ces données sont accessibles depuis peu et représentent les requêtes faites par les usagers du moteur de recherche de Google. Les données du géant du Web étant disponibles en temps réel selon plusieurs fréquences, le nouvel indicateur est construit à l'aide des données de recherches hebdomadaires. L'indice conçu provient ainsi de l'agrégation des prévisions de ses composantes faites à l'aide de modèles à fréquences mixtes. Ainsi, les modèles MIDAS, U-MIDAS et LASSO sont ceux privilégiés. L'utilisation de données intramensuelles offre donc la possibilité de mettre à jour les prévisions au fur et à mesure que de nouvelles données sont disponibles pour le mois en cours. Les résultats indiquent que les données hebdomadaires de Google aident effectivement à la prévision de très court terme, bien que leurs apports soient parfois modestes. Ces réalisations sont traduites dans la performance du nouvel indice qui réussit à répliquer assez bien les mouvements de celui inspiré par la Réserve Fédérale de Kansas City.

Mots-clés : Marché du travail, prévision, fréquence mixte, LASSO, Google Trends, États-Unis

INTRODUCTION

Internet fait partie intégrante de la vie de nombreuses personnes, du moins dans les pays développés du monde. En effet, en 2017, aux États-Unis, plus de 87,27% de la population utilisait l'autoroute de l'information et le nombre est toujours croissant. [1] Puisque l'Internet est si largement présent dans notre quotidien, la question se pose à savoir si à partir de ces données de navigation nous sommes en mesure de générer des connaissances sur l'activité macroéconomique. Généralement, ce genre de données est de nature privée. Cependant, depuis 2006, le géant du service technologique Google a rendu publiques des données sur le volume relatif de requêtes faites sur son moteur de recherche pour un terme en particulier à travers le temps. Ce qui est intéressant avec ces requêtes, c'est qu'elles peuvent en quelque sorte représenter l'intérêt ou l'attention que porte la population à un sujet précis, chose particulièrement difficile à mesurer. Ainsi, le web est devenu une source importante d'acquisition d'information pour les agents économiques. Reste à savoir si cette prise d'information se traduit par la suite en prise de décisions.

L'utilisation des données de Google Trends est très diversifiée et touche divers secteurs de recherche. Parmi les premiers à les avoir utilisées, Ginsberg *et al.* (2009) ont publié une recherche qui utilise le volume de termes de recherche en rapport avec la grippe et ses symptômes afin de surveiller la progression de celle-ci en temps réel. Les auteurs surveillent le comportement de recherche qui ressemble à ceux de la grippe qui sont adressés sur le moteur de recherche de Google. Ainsi,

---

1. Federal Reserve Bank of Economic Data. (2019) *Internet users for the United States.* Récupéré de : https://fred.stlouisfed.org/series/ITNETUSERP2USA/

ils ont réussi à estimer le niveau d'activité grippale hebdomadaire dans chaque région des États-Unis, avec un décalage d'environ un jour. En économie, Choi et Varian (2009) sont également parmi les premiers utilisateurs des données sur les requêtes de Google. De leur côté, elles servent à prévoir plusieurs indicateurs économiques en temps réel aux États-Unis, notamment les demandes initiales de chômage, la demande d'automobile ou encore la vente totale de maisons. En effet, ils utilisent les données de Google Trends de la première semaine pour un mois donné afin de prévoir la valeur de la variable pour ce même mois. Les auteurs ont observé que l'ajout de ces variables à de simples modèles $AR(p)$ pouvait en améliorer le pouvoir prédictif afin de prévoir « le présent » (Choi et Varian, 2009). De la même façon, Askitas et Zimmermann (2009) ont conclu dans leur recherche que l'utilisation des requêtes faites sur Google pouvait être un bon ajout dans les modèles servant à prédire le taux de chômage en temps réel en Allemagne. Depuis les résultats des recherches démontrent que l'utilisation des requêtes faites sur Internet peut représenter un apport important dans les modèles de prévisions économiques. De ce fait, les résultats des différents auteurs rapportent que l'action de chercher de l'information sur Internet peut représenter une étape préliminaire à la prise de décisions des agents économiques. En effet, les recherches effectuées par la passée apportent de l'information pertinente sur les résultats d'aujourd'hui.

Un autre aspect invitant de ce type de données est la fréquence de publication de celles-ci. Ces données sont disponibles sur une base journalière, hebdomadaire et mensuelle, ce qui rend l'analyse en temps réel intéressante. De plus, elles ne sont pas assujetties à des révisions et elles sont disponibles très rapidement. Traditionnellement chez les séries économiques publiées par les agences statistiques telles que la *Federal Reserve Bank of Economic Data* ou Statistique Canada, les données pour un mois donné sont publiées uniquement durant les mois suivants. Aussi, ces dernières peuvent être révisées vu la qualité des données accessibles au moment

de leur première sortie, ce qui peut complexifier l'analyse en temps réel. Ceci peut aussi notamment poser quelques problèmes pour les décideurs politiques ou les banques centrales qui désirent avoir une vision de l'état actuel de l'économie. Ainsi, le présent travail de recherche aura pour but de vérifier si les requêtes faites par les utilisateurs de Google peuvent effectivement être utiles afin de fournir une vision globale de l'état du marché du travail aux États-Unis. Les conditions du marché du travail sont importantes pour évaluer le bien-être économique et pour orienter les prises de décisions notamment de la politique monétaire. En effet, aux États-Unis les résultats sur le marché du travail constituent un élément explicite du double mandat de la Réserve fédérale, qui est de parvenir à un maximum d'emploi et à la stabilité des prix. Ainsi, un indicateur mensuel de l'activité réelle sur le marché du travail sera construit à l'aide des données provenant des requêtes faites sur le moteur de recherche de Google. Un modèle statistique, bien que ne pouvant pas se substituer à un examen judicieux, peut être utile, car il fournit un moyen relativement dépourvu de jugement pour résumer les informations provenant de nombreux indicateurs.

En effet, l'un des défis de l'évaluation de la situation du marché du travail réside dans la publication mensuelle de diverses données qui mesurent chacune une dimension spécifique de ce marché et qui peuvent être susceptibles de donner des signaux contradictoires sur la santé de celui-ci. De plus, elles ont souvent un délai de publication du un ou de plusieurs mois. Comme mentionné, ces retards peuvent être problématiques pour les décideurs économiques au niveau de leurs prises de décisions concernant les politiques à adopter. Il existe tout de même déjà quelques indices qui fournissent une vision globale de ce marché. Il est question notamment ici de l'indice de la Réserve Fédéral de Kansas City le *Labor Market Conditions Indicators* (LMCI) (Willis *et al.*, 2014). L'indice est construit à partir de 24 séries macroéconomiques à fréquence mensuelle. Le but de ce présent mémoire consiste

donc à recréer un indice semblable à partir de prévisions faites sur un ensemble des séries utilisées pour la conception de l'indice original. Ainsi, la pertinence de l'ajout des données de Google sera vérifiée à partir de prévisions faites sur un large éventail de séries qui décrivent chacune à leur façon le marché du travail américain. La particularité réside dans le fait que ces prévisions seront effectuées à l'aide de modèles à fréquences mixtes en utilisant les données hebdomadaires de Google Trends. L'utilisation de données intramensuelles offre donc la possibilité de mettre à jour les prévisions au fur et à mesure que de nouvelles données sont disponibles pour le mois en cours. La publication plus fréquente des données de Google permet d'avoir une idée plus rapidement de la situation à l'aide de l'indice sans avoir à attendre la publication de l'indice original.

Le présent travail de recherche sera divisé comme suit. Au chapitre 1, on présentera une revue de littérature concernant les recherches faites en science économique qui utilisent les données de Google Trends. Le choix et l'utilisation de ces données seront explicités dans le chapitre 2. Par la suite, au chapitre 3 il sera question de la méthodologie et des modèles utilisés afin de construire l'indice. Finalement, le dernier chapitre sera consacré aux résultats obtenus.

CHAPITRE I

REVUE DE LITTÉRATURE

L'utilisation des données portant sur les recherches Internet n'est que très récente dans la littérature économique. Cela provient naturellement du fait qu'elles ne sont disponibles que depuis 2006. Les recherches sur la qualité prévisionnelle des données de Google Trends révèlent en effet que ce type de variable peut améliorer la précision des prévisions effectuées, et ce pour une large gamme de séries économiques. Le présent chapitre dresse un portrait des recherches qui ont tentées d'exploiter le potentiel de ces données pour la prévision de séries macroéconomiques.

Choi et Varian (2009) ainsi que Askitas et Zimmermann (2009) sont parmi les premiers à avoir utilisé les données de Google Trends en science économique. Les deux duos d'auteurs ont très rapidement su profiter des délais de publication très courts de ces nouvelles variables afin d'effectuer des prévisions pour la période en cours (*nowcasting*) de variables économiques. Les premiers ont fait des prévisions en temps réel au niveau du secteur des ventes immobilières et au détail et de l'emploi aux États-Unis. Ceux-ci ont ainsi montré que l'ajout des variables de recherches de Google peut améliorer la performance de simple modèle $AR(p)$. De la même manière, les seconds ont fait la prévision du taux de chômage en Allemagne à l'aide des données sur les recherches des usagers du moteur de recherche de

Google. Les auteurs ont utilisé un modèle à correction d'erreur (ECM) afin de démontrer la forte corrélation entre ces recherches et les changements dans le taux de chômage. Ainsi, les résultats de ces études ont contribué à mettre de l'avant l'utilisation de ce type de données dans la prévision à très court terme en science économique. En effet, toujours en 2009, D'Amuri (2009) a démontré l'intérêt de ces variables pour la prévision du taux de chômage en Italie, puis Suhoy (2009) l'a également fait pour l'Israël. En somme, ces articles ont ouvert la porte à l'utilisation des données issues de Google Trends afin de prédire une multitude de nouvelles variables économiques. En effet, ils ont su montré, en quelque sorte, que les comportements de recherche sur Internet des agents économiques peuvent concrètement se traduire en prises de décisions.

Au sujet des variables liées à la consommation, les données de Google Trends peuvent également s'avérer assez utiles. Leur disponibilité rapide est encore un atout très important : c'est ce qui permet d'effectuer des prévisions pour la période en cours de la consommation privée aux États-Unis (Choi et Varian, 2009, 2012). Traditionnellement, aux États-Unis, des indices basés sur les résultats de sondages sur les sentiments des consommateurs comme le *Michigan University's Consumer Sentiment Index* ou encore le *Conference Board's Consumer Confidence Index* sont utilisés comme variables pour la prévision de la consommation privée (Vosen et Schmidt, 2009). Ces indicateurs tentent de tenir compte de l'aspect économique et psychologique des comportements des individus en leur demandant, à travers des sondages, leurs opinions sur l'état présent et futur de l'économie. Cependant, Croushore (2005) démontre que l'utilisation de ces indices apporte, au final, peu aux prévisions pour la consommation. Ce phénomène est potentiellement dû au fait que les résultats de ces sondages ont de la difficulté à traduire les réelles intentions d'achats des répondants (Ludvigson, 2004). Ainsi, les recherches faites sur Google pourraient traduire ces intentions des consommateurs. Tout d'abord, parce

que les achats faits en ligne deviennent toujours de plus en plus importants. Également parce que les recherches faites sur Internet peuvent représenter une prise d'information du consommateur sur les biens ou services qu'il désire consommer prochainement. De plus, ces données ont le potentiel de rejoindre une plus grande partie de la population comparativement à celles des sondages. Avec ces atouts en tête, les chercheurs ont tenté de prédire des variables en lien avec la consommation privée. Ainsi,Vosen et Schmidt (2009) et Kholodilin *et al.* (2010) ont montré à l'aide d'un modèle ADL(p,q) que l'indice de Google Trends possède un meilleur pouvoir prédictif que les indices basés sur des sondages. De l'autre côté, les chercheurs français Combes et Bortoli (2015) ont démontré que l'ajout des données de Google Trends au modèle ARMA(p,q) sur la consommation privée agrégée en France n'améliore pas nécessairement les prévisions. Cependant, l'ajout de ces variables aux modèles univariés basés sur la consommation de différents biens et services en particulier (achat de voitures, consommation en énergie, etc) améliore la qualité des prévisions (Combes et Bortoli (2015) et Chamberlin (2010)). Les indices de sondages tels que *Michigan University's Consumer Sentiment Index* sont traditionnellement également utilisés dans la prévision de l'inflation aux États-Unis. Guzman (2011) a démontré que l'utilisation des données de Google Trends peut offrir une meilleure performance que les modèles qui utilisent les résultats de ces sondages. Seabold et Coppola (2015) ont également montré que ces données peuvent aussi être utiles pour prévoir les variations de prix des biens de consommation, mais cette fois pour les pays d'Amérique du Sud.

Du côté du marché immobilier, l'ajout des données portant sur les recherches sur Internet aux modèles de prévisions de base améliore également la performance pour ce type de variable. En effet, Wu et Brynjolfsson (2015) et Kulkarni *et al.* (2009) en démontrent les avantages pour la prévision du prix et de la vente immobilière aux États-Unis. Des économistes de la Banque Centrale d'Angleterre ont

montré que l'ajout des données de Google Trends à un modèle AR(p) portant sur la prévision du prix des maisons améliore les prévisions en Angleterre (McLaren et Shanbhogue, 2011).

Comme c'est le cas pour le présent mémoire, les données de Google ont également servis à la construit d'indices représentant diverses marchés. Notamment, Chauvet *et al.* (2016) ont développés un indice en temps réel du risque de défaillance des remboursements des prêts hypothécaires. À l'aide de mots-clés comme *foreclosure help* ou *government mortage help*, les auteurs vont tirer l'information directement auprès des personnes cherchant de l'aide via une recherche sur Internet à propos des problèmes de défaut de paiement et de saisie immobilière. De manière générale, les auteurs ont démontré que l'utilisation des données de Google Trends fournit de l'information de qualité pour ce qui est des défauts de paiements de prêts hypothécaires. Dans un autre ordre d'idée, Baker et Fradkin (2011) ont mis au point un indice journalier de l'activité de recherche d'emploi aux États-Unis construit à partir de données de recherche sur Internet du géant de la Silicon Valley. Ils ont ainsi démontré que les fluctuations dans les recherches d'emplois aux États-Unis suivent de près les activités de recherches sur le moteur de Google. Da *et al.* (2011), de leur côté proposent une mesure de l'attention des investisseurs pour des produits financiers en utilisant différentes fréquences de recherche. En effet, selon les chercheurs « *search is a revealed attention measure : if you search for a stock in Google, you are undoubtedly paying attention to it* ». Ainsi, ces auteurs ont montré que les recherches agrégées peuvent représenter une mesure objective d'attention directe pour ces produits et donc, potentiellement, une mesure d'intérêt pour la population à propos de divers sujets.

CHAPITRE II

DONNÉES

Dans ce chapitre, les données utilisées pour la construction de l'indice de l'activité économique seront abordées. Puisque l'indice n'est pas directement construit à l'aide de données macroéconomiques, il sera question de deux groupes de variables distincts. Dans le premier groupe on retrouve les variables macroéconomiques qui servent de variables de référence pour la construction de l'indice. Puis, une importante partie de ce mémoire consiste en la sélection des données du second groupe. Celui-ci contient les données tirées de Google Trends, soit les mots-clés utilisés par les usagers du moteur de recherche du géant de la Sillicon Valley. Leurs différentes spécificités feront en sorte que les deux groupes seront traités séparément dans ce chapitre.

2.1 Données macroéconomiques

L'objectif est de construire un indice représentant les performances du marché du travail américain de la même manière que le fait la Réserve Fédérale de Kansas City à l'aide du leur le *Labor Market Conditions Indicators* (LMCI) (Willis *et al.*, 2014). Cependant, la particularité de ce nouvel indice résidera dans le fait que celui-ci sera construit à l'aide des données provenant de Google Trends. Ceci offrira donc la possibilité que l'indice soit mis à jour à une fréquence hebdomadaire, soit

près d'un mois avant la sortie officielle de l'indice réel.

L'indice de la Réserve Fédérale de Kansas City, représenté à la Figure 2.1, est une moyenne pondérée de 24 séries du marché du travail américain construite à l'aide de l'analyse par composantes principales. Les variables sont choisies de manière à représenter les grandes catégories du marché du travail, soit le chômage et le sous-emploi, le niveau d'emploi, les heures de travail, les salaires, les postes vacants, les embauches, les licenciements, les démissions et les enquêtes auprès des consommateurs et des entreprises. Ainsi, pour analyser le marché du travail, en plus de se fier à quelques variables prises séparément, l'information mise en commun de ces 24 variables fournit un indicateur d'activité plus large. Il représente essentiellement la santé générale du marché du travail et peut être utile pour identifier les principaux changements dans les tendances de ce marché. De plus, l'indice peut s'avérer utile pour déterminer la sévérité des récessions. Il est construit de manière que la moyenne soit zéro et que l'écart-type soit d'un. La valeur de zéro représente donc la moyenne historique de l'échantillon. De cette façon, il mesure l'écart entre le niveau d'activité du marché du travail et la moyenne historique de l'échantillon. Comme la valeur de l'indice est recalculée à chaque ajout de nouvelles données, le niveau en lui-même n'a pas d'interprétation. Cependant, l'ampleur des variations de l'indice entre les différentes périodes indique la performance d'une période comparativement à une autre. À ce moment, la valeur de cette variation est pertinente pour des fins de comparaisons interpériodiques.

À cause de leur indisponibilité, certaines variables de l'indice d'origine ont été mises de côté. Ainsi, 21 des séries qui ont été utilisées sont sélectionnées. Inspiré par la méthodologie que propose la FED de Kansas City, l'indice alternatif sera également construit à l'aide de l'analyse par composantes principales (ACP). La plupart des séries proviennent de la Réserve fédérale de St. Louis et quelques-unes proviennent de l'Université du Michigan et du *National Federation of Inde-*

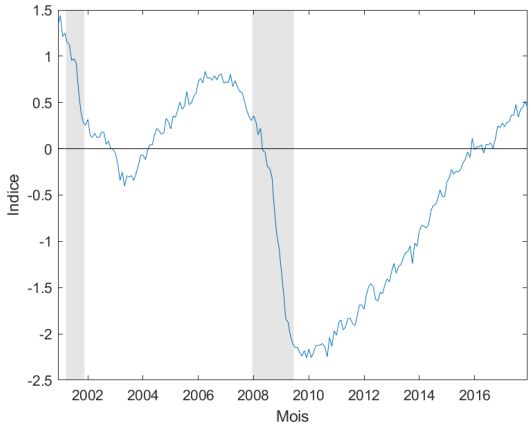


Figure 2.1: Indices du niveau d'activité du marché du travail construites par la KC Fed

Note : Les zones ombragées représentent les périodes de récessions aux États-Unis tel que l'indique de récession du pic jusqu'au creux du NBER.

pendent Business (NFIB). Celles-ci débutent en décembre 2000 et se terminent en décembre 2017, puisque les séries du taux de démissions et du taux d'embauche débutent à partir de cette date. De plus, elles ont été récupérés préalablement désaisonnalisées.

2.1.1 Indice du marché du travail

Pour la construction des indices, les variables utilisées sont transformées selon la méthodologie de Willis *et al.* (2014). Les auteurs utilisent ainsi comme indices les deux premières composantes principales estimé par ACP, qui, dans le cas-ci, expliquent plus de 73% de la variance entre les séries utilisées. Cependant, dans le présent travail de recherche, il sera uniquement question de la première composante qui explique plus de 51% de la variance. Comme pour Willis *et al.* (2014), les variables qui sont le plus corrélées avec la première composante sont celles qui sont

associées au niveau d'activité du marché de l'emploi. Des variables telles que les différentes mesures du taux de chômage ainsi que les différents taux d'embauche, de licenciement et démissions sont les variables qui contribuent le plus fortement. [2]

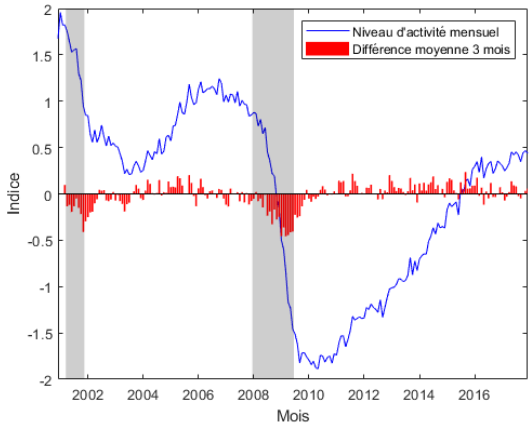


Figure 2.2: Nouvel indice du niveau d'activité du marché du travail

Note : Les zones ombragées représentent les périodes de récessions aux États-Unis tel que l'indique de récession du pic jusqu'au creux du NBER.

Ce qui est intéressant avec cet indice, c'est qu'il peut être analysé selon plusieurs horizons. Pour le court terme, les variations de la moyenne sur trois mois montrent bien la condition relative du marché du travail dans le temps. Les valeurs sous zéro signifient une contraction du niveau d'activité du marché du travail en moyenne dans les trois derniers mois et la logique est inversée pour l'augmentation du niveau d'activité. Pour ce qui est du long terme, l'indice en niveau montre la tendance que prend le marché du travail dans le temps ainsi que son niveau d'activité par rapport à la moyenne historique de l'échantillon. On remarque effectivement dans la Figure 2.2 que lors des périodes de contraction de l'économie, la valeur

2. Les séries contribuant le plus un premier indice sont montrées dans l'appendice A à la Figure A.1.

des indices diminue, puis reprend une tendance positive durant les phases d'expansions. Ainsi, la mise à jour de cet indice à l'aide des données des recherches de Google pourrait permettre de prévoir et de reconnaître l'état du marché du travail à l'avance.

2.2 Données de Google Trends

Comme il a été présenté dans le chapitre précédent, l'utilisation des données de Google Trends en science économique n'est que très récente. Plusieurs chercheurs s'y sont intéressés, notamment à cause de la rapidité à laquelle on peut y avoir accès et également parce qu'elles ne sont pas soumises à des révisions. Ceux-ci s'y sont aussi intéressés compte tenu de ce qu'elles peuvent représenter, soit l'intérêt ou l'attention que porte la population à un sujet bien particulier, chose qui est difficile à mesurer. Les paragraphes qui suivent présenteront les données de Google Trends et justifieront également le choix des mots-clés.

2.2.1 Méthodologie de l'indice Google Trend

En 2006, Google a rendu publique une partie de ses données à propos des requêtes faites sur son moteur de recherche. Ainsi, on peut gratuitement avoir accès à des données datant de 2004 jusqu'à aujourd'hui portant sur les recherches pour un mot-clé en particulier à travers le temps. En fait, ce que Google met à notre disposition est un indice de popularité relative selon la région géographique et la période pour lesquelles on désire recueillir les données. Par conséquent, ce n'est pas le volume de recherche absolue qui est directement disponible. Ce sont ces données qui sont utilisées dans ce travail de recherche. C'est notamment à partir du nombre de recherches total et du nombre de recherches faites pour un mot-clé particulier que Google construit son indice. Malheureusement, la méthodologie

complète n'est pas explicitée par l'entreprise. Une description sommaire sera donc développée. Ainsi, chaque point de l'indice est construit de la manière suivante :

Chaque point (nombre de recherches faites pour un mot-clé) est divisé par le nombre total de recherches faites dans la région géographique d'intérêt pour une fréquence donnée (journalière, hebdomadaire ou mensuelle), afin d'effectuer un comparatif de la popularité relative. Ensuite, les résultats sont normalisés sur une échelle de 0 à 100 selon la proportion du sujet de recherche par rapport à toutes les requêtes portant sur tous les sujets. [3]. La normalisation de ces données est importante puisque le nombre de personnes cherchant sur Google change constamment. Ceci affecte ultimement le nombre absolu de recherches et cette modification permet donc de comparer l'importance des recherches effectuées dans le temps. Il existe deux types de données : les données en temps réel, qui sont un échantillon aléatoire de recherches effectuées au cours des sept derniers jours, et les données en temps non réel, qui sont un échantillon aléatoire de recherche Google pouvant remonter jusqu'à 2004 et jusqu'à 36 heures avant la recherche d'un mot-clé. Autrement, sans cet échantillonnage, l'analyse sur l'ensemble des données serait trop longue selon Google. Dans le cas de cette recherche, ce sont les données en temps non réel qui nous intéressent. Ainsi, les valeurs de l'indice peuvent varier selon le moment et l'endroit où les données ont été recueillies. Cependant, Chauvet *et al.* (2016) n'ont pas trouvé de différences substantielles entre les données téléchargées sur plusieurs jours. De même pour D'Amuri et Marcucci (2017), qui n'ont jamais trouvé des corrélations inférieures à 0.99 entre les mots-clés des différents échantillons.

---

3. Méthodologie disponible à : https://support.google.com/trends/answer/4365533?hl=frref$_topic$ = 6248052

Selon la fréquence désirée, la fenêtre de disponibilité des données change. À l'heure actuelle, voici les différents horizons disponibles selon la fréquence :

- Mensuelle : les données sont disponibles pour tout l'horizon, soit de janvier 2004 à aujourd'hui.
- Hebdomadaire : les données sont uniquement disponibles sur un horizon de cinq ans. Par exemple, à partir du 1er janvier 2005 au 1er janvier 2009, puis si la fenêtre dépasse cet horizon, les données sont mensuelles.
- Journalière : les données sont disponibles pour un horizon de huit mois. Par exemple, à partir du 1er avril 2006 au 1er décembre 2006. Au-delà de cet horizon de huit mois, les données sont hebdomadaires.

2.2.2 Recherche des mots-clés

Le choix des mots-clés découlant des recherches faites sur Internet est une partie importante de ce travail de recherche. En effet, le but ici est de trouver les combinaisons de mots-clés qui permettront, le mieux possible, d'expliquer les mouvements de chacune des séries macroéconomiques présentées précédemment. Il faut, dans un premier temps, tenter de définir des mots-clés pertinents liés aux recherches faites par les usagers. Par exemple, Ross (2013) tente de prédire le taux de chômage en Angleterre avec des termes tels que *job center*, *job seekers allowance* ou encore *employment support*. La logique derrière ce type de recherches est la suivante : plus les mouvements dans l'emploi seront importants, plus les recherches liées à l'aide à l'emploi fluctueront également. Dans ce cas-ci, une augmentation relative du volume de ce type de recherches pourrait représenter une possible faiblesse dans le marché de l'emploi. Le choix des mots-clés peut naturellement provenir de l'intuition du chercheur, mais cela pourrait venir biaiser la sélection des mots. Aussi, le chercheur pourrait avoir omis une combinaison de mots-clés qui améliorerait grandement la prévision. Il pourrait également avoir de

fausses croyances qui le mèneraient sur de mauvaises pistes.

Ross (2013) propose donc la méthode de l'induction à rebours qui repose sur le fait que les meilleurs mots-clés ont déjà été choisis. En effet, les utilisateurs du moteur de recherche les ont eux-mêmes déjà sélectionnés à travers leurs requêtes. Ainsi, pour le marché du travail, il serait possible, par exemple, de sélectionner les mots-clés les plus communs en lien avec le mot « travail ». Bien sûr, il ne faut pas totalement exclure l'intuition, même avec cette méthode : les mots-clés pourraient se retrouver parmi les recherches populaires, mais n'auront pas nécessairement de liens avec la variable que l'on désire prédire. Le choix de mauvaises combinaisons pourrait alors créer des corrélations fictives et donc biaiser les résultats. Une sélection basée davantage sur l'intuition gagnerait donc à être privilégiée. Il faudra également tenir compte du fait que les recherches faites pour un sujet en particulier sur le moteur de recherche peuvent changer dans le temps, autrement l'indice pourrait être représentatif pour une période et ne pas l'être pour une autre.

Les mots-clés potentiels ont été sélectionnés à l'aide de la plateforme Adwords de Google. Celle-ci a pour principale fonction d'afficher des bannières publicitaires soit sur Google, soit sur d'autres sites qui sont ciblées selon les mots-clés que tape l'internaute ou en fonction de son comportement de navigation. Les usagers de Adwords paient lorsque l'internaute clique sur la publicité affichée selon un système d'enchère et de qualité. C'est-à-dire que plus l'annonce sera pertinente pour l'internaute, plus le prix au clic sera bas et plus la publicité sera en évidence. La plateforme met ainsi à la disposition de ses usagers des mots-clés qui peuvent être pertinents pour la présente analyse. Son utilisation permet donc notamment d'avoir à porté une grande quantité de mots-clés potentiellement pertinents que la méthode par intuition aurait pu omettre. Comme il existe de nombreuses façons de formuler l'idée pour une même intention de recherche, une large quantité de mots-clés permettra de mieux englober ces différentes possibilités.

Ainsi, pour une idée de mot-clé donnée la plateforme en identifie d'autres pouvant être pertinents. Les recherches *job*, *unemployment*, *employ*, *work*, *salary*, *wage*, *resume*, *hire*, *getting fired* et *occupation* ont été sélectionnées afin d'obtenir des mots-clés potentiels. Elles ont été choisi notamment parce qu'elles rendent compte du processus de recherche d'emploi et également de celui de la perte d'emploi. Par exemple, pour la recherche *job* la plateforme propose d'autres mots-clés tels que *job search*, *find a job*, *hiring* ou encore *job seeker*. La plupart des mots-clés proposés traduisent la recherche d'emploi par un individu ou encore des employeurs qui recherchent des candidats. De la même façon, pour le second mot-clé *unemployment*, Adwords propose des mots tels que *unemployment claims*, *unemployment office*, *unemployment edd*, *unemployment sign in* et *unemployment information*. Ainsi, on peut penser que les mots-clés proposés rendent compte du fait que ceux qui recherchent ces mots sont soit en processus d'assurance emploi, soit déjà sur cette assurance. Tous les mots-clés proposant des domaines de sites Web n'ont pas été retenus. En effet, les sites Web utilisés pour chercher un emploi ou pour trouver un employé sont susceptibles de changer assez souvent dans le temps. Un filtrage manuel a également été effectué afin de réduire le nombre de mots-clés non pertinents. Au total, plus de 3 400 mots-clés ont été sélectionnés à l'aide de cet outil.

2.2.3 Méthode de récupération des données de Google

Comme mentionné précédemment, la méthode utilisée par Google pour la standardisation de leurs données n'a pas été explicitée, ce qui rend la tâche particulièrement difficile lorsque l'on veut utiliser ce type de données à des fréquences plus élevées que mensuelles. En effet, ces données peuvent être directement récupérées depuis leur première disponibilité jusqu'à aujourd'hui, ce qui n'est pas le cas pour celles qui ont des fréquences plus élevées. Ainsi, pour obtenir l'ensemble des don-

nées disponibles pour les autres fréquences, il faut faire se chevaucher différentes sous-périodes. Une simple juxtaposition de ces sous-périodes ne représente pas bien les données pour l'ensemble. Effectivement, les indices entre les différentes sous-périodes ne sont pas sur le même niveau puisqu'ils n'utilisent pas le même échantillon lors de leur construction par Google. Ainsi, une méthode de mise en commun inspirée par Bleher et Dimpfl (2019), qui ont développé un algorithme permettant de régler le problème de chevauchement des données de Google, est utilisée.

De cette façon, prenons le cas de deux sous-périodes qui se chevauchent pour un certain laps de temps, disons v_t^a et v_t^b. Comme il a été mentionné précédemment, la valeur de l'indice de Google est dépendante de la période à laquelle elle appartient, donc les deux sous-périodes devraient avoir pour le même temps t une échelle différente. Par conséquent, l'algorithme permet de mettre les deux sous-périodes à la même échelle. Celui-ci repose sur une approximation de la relation linéaire entre les deux sous-périodes, soit l'équation suivante :

$$v_t^b = \beta v_t^a + \epsilon_t \tag{2.1}$$

Les auteurs suggèrent donc la construction d'un indice à l'aide de plusieurs sous-périodes qui se chevauchent, puis d'utiliser cette période de chevauchement afin d'estimer les paramètres de l'équation (2.1) par moindre carré ordinaire. Cependant, comme ceux-ci utilisent des données journalières l'algorithme a été quelque peu modifié pour tenir compte de la différence de fréquence dans les données utilisées. Soit l'algorithme suivant :

1. Recueillir des sous-périodes de cinq ans pour l'entièreté de la période d'intérêt. Chaque sous-période doit contenir au moins une période d'un an qui se chevauche entre elles.
2. Débuter avec les deux sous-périodes les plus vieilles et identifier les obser-

vations qui se chevauchent. La sous-période la plus vieille (A) est nommée v_t^a et l'autre (B) est nommé v_t^b.

3. Estimation des paramètres de l'équation (2.1) par moindre carré ordinaire en utilisant les observations se chevauchant.
4. Calculer les nouvelles valeurs de l'indice de Google $\hat{v}_t^b$ en utilisant l'estimateur $\hat{\beta}$.
5. Mettre en commun les valeurs d'origines de v_t^a et les valeurs estimées $\hat{v}_t^b$ dans un sous-ensemble. Ce nouveau sous-ensemble prend la place de A et B est remplacé par la prochaine sous-période.
6. Répétition des étapes 2 à 5 jusqu'à ce qu'il ne reste plus de sous-période.

2.2.4 Pré-sélection des mots-clés de Google Trends

La méthode utilisée est la recherche massive de mots-clés afin de trouver ceux qui performent le mieux en matière de prévisions. Cependant, tous les mots-clés, bien qu'en lien avec la variable à prévoir, n'apporte possiblement pas toute une plus-value au niveau de la prévision. Alors, une méthode de réduction de dimension sera utilisée afin de restreindre leur nombre dans le but de conserver uniquement ceux qui sont potentiellement utiles à cette fin. Inspiré par la méthode de l'induction à rebours de Ross (2013), une approche basée sur la significativité statistique des coefficients sera privilégiée. C'est-à-dire qu'au lieu de supposer la pertinence de certains mots-clés, ceux-ci seront sélectionnés à l'aide du lien statistique qu'ils entretiennent avec les variables à prévoir. En effet, la significativité statistique des coefficients pour l'intégralité des mots-clés sera testée sur l'ensemble des séries utilisées pour la construction de l'indice.

C'est en fait la significativité du coefficient $\hat{\beta}$ qui sera testé à l'aide de l'équation suivante :

$$y_t = \alpha + \rho y_{t-1} + \beta x_{t-h} + \epsilon_t \tag{2.2}$$

y_t est la série du marché du travail stationnarisée et x_t une série mensuelle de Google Trends. Ainsi, pour chaque série, tous les mots-clés recueillies seront inclus l'un à la suite de l'autre afin de tester si leur coefficient est statistiquement différent de zéro. Les données mensuelles seront d'abord utilisées à la place des données hebdomadaires en vue de sélectionner ces mots-clés puisqu'elles contiennent toutes l'information de recherches faites pour un mois. Le modèle sera estimé pour le mois en cours (h = 0) et pour un mois à l'avance (h = 1) afin de tester le pouvoir prédictif des données de Google à très court terme. Comme les données mensuelles de Google Trends sont toujours disponibles avant la nouvelle sortie des séries du marché du travail, un tel exercice est alors possible. Afin que cette méthode de *filtrage* soit plus efficace, seuls les mots-clés dont les coefficients pour les deux horizons sont significatifs à un seuil de 1% seront sélectionnés. Cependant, pour certaines séries, aucun mot-clé n'était significatif à ce seuil. La même démarche a alors été faite, mais avec un seuil de 5%. L'exercice est intéressant dans la mesure où ce sont les données qui vont en elles-mêmes dicter les mots-clés potentiellement pertinents.

Le tableau 2.1 indique le nombre de mots-clés sélectionnés pour chaque série du marché du travail. Un même mot-clé peut être présent pour plusieurs séries. Ainsi, la méthode en a sélectionné plus de 340 différents sur les 3 400 proposés au départ. Ceux-ci seront par la suite utilisés pour récupérer des données hebdomadaires afin d'estimer des modèles à fréquences mixtes. Le nombre de mots-clés pertinents varie beaucoup d'une variable à l'autre. Certaines séries, comme c'est le cas notamment pour les différentes mesures de chômage, ont plus de 200 mots-clés dont les coef-

Tableau 2.1: Variables sélectionnées avec la méthode de pré-sélection

Variables	Nombre de mots-clés significatifs
Civilian Unemployment Rate	163
Total unemployed, plus all marginally attached workers plus total employed part time for economic reasons	204
Labor Force Flows Unemployed to Employed : 16 Years and Over	0
Quits : Total Private	2
Civilian Employment-Population Ratio	181
Employment Level : Part-Time for Economic Reasons, All Industries	92
Job Leavers as a Percent of Total Unemployed	5
Of Total Unemployed, Percent Unemployed 27 Weeks and over	221
Job Losers as a Percent of Total Unemployed	7
Average Hourly Earnings of Production and Nonsupervisory Employees : Total Private	2
All Employees : Total Private Industries	6
Indexes of Aggregate Weekly Payrolls of Production and Nonsupervisory Employees : Total Private	209
All Employees : Professional and Business Services : Temporary Help Services	4
Civilian Labor Force Participation Rate	12
Index of Aggregate Weekly Hours : Production and Nonsupervisory Employees : Total Private Industries	102
Hires : Total Private	11
Monthly average Initial Claims	25
Manufacturing Employment Index	3
Percent of firms planning to increase employment	1
Expected job availability	1
Percent of firms with positions not able to fill right now	6

ficients étaient statistiquement significatifs, alors que d'autres n'en ont que très peu. Ces résultats refléteront peut-être la vertu des recherches Google pour les prévisions des séries où le nombre de mots-clés significatifs est plus important.

2.3 Transformation des données

2.3.1 Données macroéconomiques

Avant de débuter les estimations, on doit vérifier si les séries à l'étude suivent un processus stationnaire. Le test Dickey-Fuller augmenté (Dickey et Fuller, 1979) est largement rependu et sert à tester la non-stationnarité d'une série temporelle.

$$\Delta Y_t = \beta_0 + \beta_1 t + \alpha Y_{t-1} + \sum_{i=1}^{\rho} \delta_i \Delta Y_{t-i} + \epsilon_t \tag{2.3}$$

Le test consiste à estimer les paramètres de l'équation 2.3 par moindre carré ordinaire et à faire un test avec l'aide d'une statistique t sous les hypothèses suivantes :

$$H_0 : \alpha = 0 \quad \text{(Non-stationnaire)}$$

$$H_1 : \alpha < 0 \quad \text{(Stationnaire)}$$

Afin d'effectuer correctement le test, il est important de sélectionner le bon ordre autorégressif. Si l'on choisit trop peu de retards, alors le test n'aura pas le bon niveau puisqu'il existera toujours une dynamique dans le terme d'erreur ϵ_t et $\hat{\alpha}$ sera donc biaisé. D'un autre côté, si le test contient trop de retards, on aura une perte au niveau de la puissance du test (perte de degré de liberté). La méthode de Campbell et Perron (1991) a été utilisée afin de sélectionner le nombre de retards. Celle-ci consiste à fixer un nombre maximal de retards et de tester la significativé statistique du coefficient sur le dernier retard. Lorsque celui-ci n'est pas significatif à un seuil de 5%, on le retranche, puis on estime à nouveau. On teste ainsi de suite la significativité du dernier retard jusqu'à ce que celui-ci soit significatif. À cause du nombre restreint d'observations, le nombre de retards maximal a été fixé à huit. Le test sera d'abord effectué sur les séries en niveau, puis, si nécessaire, sur les séries transformées par la suite afin d'en vérifier la stationnarité.

Les résultats du test inscrits dans le Tableau 2.2 suggèrent que les séries en niveau ne sont pas stationnaires autour d'une tendance linéaire déterministe. En effet, aucune série n'est statistiquement significative à un seuil de 5%. Il faudra alors les transformer afin d'éliminer cette tendance. Les transformations appliquées aux différentes séries sont présentées dans l'annexe A au tableau A.1.

Tableau 2.2: Test Dickey-Fuller augmenté avant transformation

Variable	Nombre de retards	Statistique-t	Valeur-p
Civilian Unemployment Rate	5	-1.7144	0.42248
Total unemployed, plus all marginally attached workers plus total employed part time for economic reasons	5	-1.6878	0.43415
Labor Force Flows Unemployed to Employed : 16 Years and Over	4	-1.1181	0.68461
Quits : Total Private	6	-1.3102	0.60015
Civilian Employment-Population Ratio	7	-1.7541	0.40502
Employment Level : Part-Time for Economic Reasons, All Industries	5	-1.5121	0.51138
Job Leavers as a Percent of Total Unemployed	1	-1.158	0.66708
Of Total Unemployed, Percent Unemployed 27 Weeks and over	4	-1.0964	0.69412
Job Losers as a Percent of Total Unemployed	8	-1.8048	0.38276
Average Hourly Earnings of Production and Nonsupervisory Employees : Total Private	6	-1.9181	0.6305
All Employees : Total Private Industries	2	-2.0992	0.54199
Indexes of Aggregate Weekly Payrolls of Production and Nonsupervisory Employees : Total Private	8	-1.3824	0.86211
All Employees : Professional and Business Services : Temporary Help Services	7	-1.8701	0.65396
Civilian Labor Force Participation Rate	1	-1.9683	0.60608
Index of Aggregate Weekly Hours : Production and Nonsupervisory Employees : Total Private Industries	8	-1.8136	0.68159
Hires : Total Private	6	-1.6276	0.46064
Monthly average Initial Claims	3	-1.5378	0.50011
Manufacturing Employment Index	3	-2.8318	0.18827
Percent of firms planning to increase employment	1	-1.5389	0.49962
Expected job availability	0	-3.0894	0.11288
Percent of firms with positions not able to fill right now	2	-0.69368	0.9713

Note : La valeur critique du test est sélectionnée selon la représentation des termes déterministe propres à chaque série. Le nombre de retards maximales est de 8 et le seuil de significativité pour le retranchement des retards est de 5%. Ainsi, si la valeur-p du test d'hypothèse sur le coefficient du dernier retard est supérieur à ce seuil celui-ci est retranché.

Le test dans le cas de données transformées suggère des résultats plus probants. Effectivement, les séries rejettent toutes l'hypothèse nulle de la non-stationnarité à un seuil de 5%. Ainsi, ce sont ces données transformées qui seront utilisées afin d'en faire la prévision. Cependant, il faudra d'abord s'assurer que les données de Google ne contiennent pas elles aussi de racines unitaires.

2.3.2 Données de Google Trends

Les données macroéconomiques ont été récupérées préalablement désaisonnalisées, toutefois les données de Google Trends ne le sont pas. Ainsi, certaines séries de mots-clés semblent contenir des périodes de saisonnalité assez claires, particuliè-

Tableau 2.3: Test Dickey-Fuller augmenté après transformation

Variable	Nombre de retards	Statistique-t	Valeur-p
Civilian Unemployment Rate	8	-4.3631	0.001
Total unemployed, plus all marginally attached workers plus total employed part time for economic reasons	8	-4.2837	0.001
Labor Force Flows Unemployed to Employed : 16 Years and Over	0	-10.5968	0.001
Quits : Total Private	0	-9.6847	0.001
Civilian Employment-Population Ratio	8	-4.5372	0.001
Employment Level : Part-Time for Economic Reasons, All Industries	8	-6.1978	0.001
Job Leavers as a Percent of Total Unemployed	0	-8.7836	0.001
Of Total Unemployed, Percent Unemployed 27 Weeks and over	0	-6.0435	0.001
Job Losers as a Percent of Total Unemployed	8	-6.4158	0.001
Average Hourly Earnings of Production and Nonsupervisory Employees : Total Private	8	-14.5017	0.001
All Employees : Total Private Industries	0	-8.8901	0.001
Indexes of Aggregate Weekly Payrolls of Production and Nonsupervisory Employees : Total Private	8	-3.5402	0.0085018
All Employees : Professional and Business Services : Temporary Help Services	8	-3.028	0.034503
Civilian Labor Force Participation Rate	0	-8.7398	0.001
Index of Aggregate Weekly Hours : Production and Nonsupervisory Employees : Total Private Industries	8	-3.5939	0.0073107
Hires : Total Private	0	-10.4956	0.001
Monthly average Initial Claims	8	-5.8403	0.001
Manufacturing Employment Index	0	-6.7668	0.001
Percent of firms planning to increase employment	0	-9.6764	0.001
Expected job availability (U of Michigan)	0	-9.4479	0.001
Percent of firms with positions not able to fill right now	0	-9.772	0.001

Note : La valeur critique du test est sélectionnée selon la représentation des termes déterministe propres à chaque série. Le nombre de retards maximales est de 8 et le seuil de significativité pour le retranchement des retards est de 5%. Ainsi, si la valeur-p du test d'hypothèse sur le coefficient du dernier retard est supérieur à ce seuil celui-ci est retranché.

rement entre les mois de novembre et de décembre. Afin de remédier à ce problème, les méthodes plus conventionnelles basées sur le modèle ARIMA ne sont pas applicables pour les données à fréquence hebdomadaire, puisque le nombre de semaines ainsi que les périodes de l'année qui les composent ne sont pas stables dans le temps. De cette façon, ces données ont été désaisonnalisées à l'aide de l'algorithme *Seasonal-Trend decomposition procedure using Loess* (STL) (Cleveland *et al.*, 1990). Cette procédure est capable de décomposer de manière flexible une série temporelle à haute fréquence en composantes linéaires de tendance (T), de saisonnalité (S) et de reste (R) basées sur l'algorithme de régression locale pondérée Loess. Pour chaque semaine t, une série temporelle peut être réécrite de la

manière suivante :

$$Y_t = T_t + S_t + R_t \tag{2.4}$$

Donc, pour chaque série d'un mot-clé, la tendance saisonnière S_t a été soustraite. Par la suite, afin de stationnarisé les données, tout comme l'on fait Niesert *et al.* (2019), l'ensemble de celles-ci sont prises en première différence.

CHAPITRE III

MÉTHODOLOGIE

Dans ce chapitre, il sera question de la méthodologie utilisée pour la sélection des mots-clés pertinents et des modèles utilisés pour arriver à cette fin. Dans un premier temps, la technique utilisée afin de réduire la dimension des données de Google y sera abordée. Les différents modèles de prévisions à fréquences mixtes ainsi que leurs critères d'évaluation seront vus par la suite. Ceux-ci permettront la sélection des mots-clés pertinents, c'est-à-dire ceux qui reflètent le mieux les mouvements des variables macroéconomiques utilisées pour la construction de l'indice. Pour finir, l'approche prévisionnelle sera explicitée.

3.1 Modèle à facteurs

Malgré la présélection effectuée précédemment, le nombre de prédicteurs potentiels reste assez élevé soit 340 séries de mots-clés de Google Trends, ce qui représente un nombre supérieur au nombre d'observations mensuelles. Afin de régler ce problème de dimension, les séries seront modélisées en un nombre réduit r de facteurs latents non observés. Ces variables latentes issues du modèle à facteurs permettent de résumer l'information contenue dans un grand nombre de variables en un nombre plus restreint de facteurs communs. Les mots-clés de Google ont été sélectionnés selon cinq catégories, soit (1) Recherche de travail (2) Recherche pour

aide financière pour les sans-emploi (3) Recherche sur les salaires (4) Recherche concernant les *curriculum vitae* et (5) Recherche concernant les mises à pied. Ce modèle permettra donc de déceler les tendances et similarités dans les recherches pour ces catégories de mots-clés. Avant de procéder à l'estimation des facteurs, les séries sont normalisées afin que les mots-clés où la variance est plus faible ne soient pas pénalisés. Le modèle à facteur statique suivant sera considéré :

$$X_t = \Lambda F_t + \epsilon_t \tag{3.1}$$

X_t est la matrice T x K de l'ensemble des données de Google Trends, F_t sont les r facteurs latents non observés ($r < T$) et Λ est une matrice de poids. On tente alors d'expliquer X_t dans un modèle linéaire avec l'aide de r facteurs. Il existe plusieurs façons d'estimer les matrices de pondération Λ, mais l'analyse par composante principale (ACP) sera la méthode privilégiée, soit la minimisation à l'aide de la méthode des moindres carrés non linéaires de la fonction objective suivante écrite en fonction des valeurs hypothétiques :

$$V(\widehat{F}, \hat{\Lambda}) = (NT)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} (x_{i,t} - \hat{\lambda}_i \widehat{F}_t)^2 \tag{3.2}$$

Où $\hat{\lambda}_i$ est la valeur à la $i^{\text{ème}}$ ligne de la matrice de poids $\hat{\Lambda}$ et $\widehat{F} = (\widehat{F_1}, \widehat{F_2}, ...\widehat{F_t})$.

Comme démontré par Stock et Watson (2002) qui font de la prévision à l'aide de facteurs, la minimisation de la fonction (3.2) est l'équivalent de maximiser $tr[\hat{\Lambda}' X' X \hat{\Lambda}]$ sujet à $\hat{\Lambda}'\hat{\Lambda}/N = I$. Le problème est ainsi résolu en égalisant $\hat{\Lambda}$ aux vecteurs de valeurs propres de la matrice $X'X$ qui correspondent alors aux r valeurs propres les plus élevées. L'estimateur des composantes principales de F est donc :

$$\widehat{F} = X'\hat{\Lambda}/N \tag{3.3}$$

Ainsi, $\hat{F}$ est la matrice qui contient les facteurs qui seront utilisés dans les modèles qui vont suivre.

3.2 Modèles de prévision

Les modèles sont appliqués sur l'ensemble des séries macroéconomiques préalablement stationnarisées. Les modèles MIDAS utilisent les facteurs estimés par ACP des séries de Google et le modèle LASSO utilise directement les séries de mots-clés pour leurs estimations.

3.2.1 Modèle MIDAS-AR

Le modèle a été introduit par Ghysels *et al.* (2004) et permet de modéliser la réponse de la variable dépendante aux variables explicatives de fréquence supérieure sous la forme d'un modèle à retards distribués extrêmement parcimonieux. L'estimation d'un vecteur de poids associée aux variables à plus haute fréquence permet ainsi d'empêcher la prolifération de paramètres qui pourraient en résulter. En plus de l'utilisation des données à fréquence supérieure, comme mentionné, on ajoute au modèle une portion autorégressive qui tient compte de l'autocorrélation de la série à prévoir dans le temps.

$$Y_{t+h}^{(m)} = \mu + \sum_{p=1}^{p_y^{(m)}} \alpha_p Y_{t-p}^{(m)} + \sum_{k=1}^{K} \beta_k \sum_{j=0}^{p_x^{(w)}} b_k\big(j;\theta\big)\hat{F}_{k,t-j}^{(w)} + \epsilon_{t+h}^{(m)} \tag{3.4}$$

C'est le vecteur de poids $b\big(j;\theta\big)$ qui permet d'agréger les retards de la variable à plus haute fréquence et ainsi réduire le nombre de paramètres à estimer. Les poids sont construits de sorte que les éléments de $b\big(j;\theta\big)$ soit compris entre $[0,1]$ et que la somme $\sum_{j=0}^{p_x^{(w)}} b\big(j;\theta\big) = 1$, ce qui permet d'identifier le coefficient $\hat{\beta}$. Il existe différentes façons de paramétrer le vecteur de poids, mais seulement deux seront retenus pour le présent travail.

1. Polynôme exponentiel d'Almon

Le fait que ce polynôme soit assez flexible et parcimonieux fait en sorte qu'il très populaire afin d'estimer les modèles MIDAS. En effet, il n'a que deux paramètres θ_1 et θ_2 à estimer. La fonction exponentielle permet de produire des formes de « bosse », ce qui assure un déclin de la valeur des poids dans le temps. Cette perte de poids est garantie tant que $\theta_2 \leq 0$. L'importance associée à chacun des retards peut diminuer lentement ou rapidement selon le nombre ajouté au modèle. De plus, il est important de souligner que le taux de décroissance détermine le nombre de retards inclus. En effet, une fois la forme fonctionnelle de $b\big(j;\theta\big)$ spécifiée, la sélection du nombre de retards optimal est exclusivement conduite par les données. C'est-à-dire que les retards qui ont une influence moindre dans l'estimation du polynôme voient l'importance de leur contribution tendre vers zéro.

$$b\big(j;\theta\big) = b\big(j;\theta_1,\theta_2\big) = \frac{exp(\theta_{1,j} + \theta_{2,j}^2)}{\sum_{j=1}^{N} exp(\theta_{1,j} + \theta_{2,j}^2)} \tag{3.5}$$

Où N est le nombre de retards maximal pour la variable à haute fréquence.

Vu la forme non linéaire de $b\big(j;\theta\big)$ on ne peut procéder par moindre carré ordinaire. Effectivement, afin de trouver la valeur optimale de θ, il faudra alors estimer les paramètres par moindre carré non linéaire.

2. U-MIDAS-AR

Le modèle U-MIDAS est une variante sans restriction du modèle MIDAS et a été introduit par Foroni *et al.* (2015). Les auteurs démontrent notamment que le modèle sans restriction peut obtenir de meilleurs résultats lorsque l'écart de fréquence entre les séries utilisées est faible. Caractéristiquement, le modèle MIDAS introduit un polynôme de poids qui permet de synthétiser l'information des variables à plus hautes fréquences afin d'en réduire les coefficients à estimer. Le modèle sans restriction de son côté n'applique pas de polynôme de poids sur ces

variables. De cette manière, celui-ci constitue le cas particulier où le vecteur de poids $b(j;\theta) = 1$ pour tous les retards des facteurs de Google. Il offre donc davantage de flexibilité comparativement à la paramétrisation du polynôme exponentiel d'Almon. L'importance des variables à plus hautes fréquences est donc déterminée directement lors de l'estimation des coefficients du modèle. Voici la représentation générale d'un modèle autorégressif multivarié :

$$Y_{t+h}^{(m)} = \mu + \sum_{p=1}^{p_y} \alpha_p Y_{t-p}^{(m)} + \sum_{k=1}^{K} \sum_{j=0}^{p_x} \beta_{k,j+1} \hat{F}_{t-j}^{(w)} + \epsilon_{t+h}^{(m)} \tag{3.6}$$

K est le nombre de variables à haute fréquence utilisées et p_x est le nombre maximal de retards. Un des avantages de ce modèle est qu'il est simple à estimer. Effectivement, on peut directement trouver ses coefficients par moindre carré ordinaire. D'un autre côté, un de ses points faibles est que le nombre de paramètres peut augmenter très rapidement, augmentant de ce fait la variance des prévisions.

3.2.2 Modèle LASSO

Le modèle LASSO a été introduit par Tibshirani (1996) et peut être représenté comme un modèle de régression linéaire contenant une contrainte supplémentaire. Le LASSO est une méthode de régularisation. Dans son cas, celui-ci ajoute un élément de pénalité $\sum_{j=1}^{p} |\beta_j| < s$ à la fonction objective pour réduire l'importance de certaines variables explicatives. Le modèle est ainsi contraint à ce que la somme en valeur absolue de ses coefficients soit inférieure à une certaine valeur fixe s. Le modèle remplit son objectif en imposant que certains coefficients soient égaux à zéro. Cette caractéristique fait en sorte qu'en présence d'importante quantité de variables explicatives possibles, ce modèle s'avère utile pour en réduire la dimension. Ainsi, le modèle sera appliqué sur les séries de mots-clés de Google qui ont été préalablement sélectionnés à l'aide de la méthode de filtrage directement et non sur leurs facteurs comme se fut le cas précédemment. L'utilisation du LASSO

permettra donc de cibler les mots-clés ayant le meilleur pouvoir prédictif. Son utilisation permettra aussi de vérifier si la synthétisassions de l'information contenue dans les différents mots-clés offre de meilleurs résultats que l'utilisation des mots-clés pris individuellement. Les coefficients $\hat{\beta}_{LASSO}$ du modèle LASSO minimise la fonction objective suivante :

$$\underset{\beta}{\operatorname{argmin}}\left\{\sum_{i=1}^{T}\left(Y_i-\beta_0-\sum_{j=1}^{p}\beta_j X_{ij}\right)^2\right\} \quad \text{sujet à} \quad \sum_{j=1}^{p}|\beta_j|<s \tag{3.7}$$

Pour $s>0$, le lagrangien s'écrit :

$$\underset{\beta}{\operatorname{argmin}}\left\{\sum_{i=1}^{T}\left(Y_i-\beta_0-\sum_{j=1}^{p}\beta_j X_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j|\right\} \tag{3.8}$$

Y est un vecteur T x 1 qui contient une des séries macroéconomiques et X est une matrice T x K où $K>T$ qui contient les retards des séries de mots-clés et les retards de la variable à prévoir. Le paramètre de réglage λ sert à contrôler l'impact relatif de ces deux termes sur les estimations des coefficients de régression. La valeur de λ varie entre zéro et l'infini. Par exemple, pour une valeur de $\lambda=0$, on obtient l'estimateur des moindres carrés ordinaires sans pénalité et lorsque $\lambda=\infty$, alors tous les coefficients sont tirés vers zéro. Par conséquent, en fonction de la valeur de λ, le lasso peut produire un modèle impliquant n'importe quel nombre de variables. Il effectue ainsi un arbitrage entre biais et variance. En effet, une augmentation de λ tire davantage de coefficients vers zéro et accentue le biais, mais diminue du même coup la variance de la prévision et vice versa.

3.2.3 Modèle de référence

Ce modèle constituera le modèle de référence, soit le modèle contre lequel les performances des autres modèles seront comparées. Le modèle autorégressif $AR(p)$ fera donc office de référence. Celui-ci est souvent utilisé à cause de sa simplicité dans la littérature des prévisions effectuées avec les données de Google.

$$Y_{t+h} = \mu + \sum_{p=1}^{P_y} \alpha_p Y_{t-p} + \epsilon_{t+h} \tag{3.9}$$

Il permettra ainsi de vérifier si l'ajout des données de Google Trends aux modèles améliore les prévisions pour un horizon très court.

3.3 Approche prévisionnelle

Comme l'on montré les résultats issus de la littérature, les modèles de prévision faits avec l'aide des données de Google Trends performent bien, surtout pour de très courts horizons. Par conséquent, elles seront utilisées ici dans le but d'effectuer des prévisions pour la période en cours (*nowcasting*) et pour un horizon d'un mois. Chaque modèle sera d'abord estimé avec uniquement la première semaine du mois en cours pour le *nowcasting* ou la première semaine du mois précédant pour l'horizon d'un mois. Puis, on y ajoutera les unes après les autres les trois autres semaines du même mois [4]. L'objectif est de vérifier si l'ajout d'information intramensuelle au sujet des recherches effectuées sur Google permet d'obtenir de meilleurs résultats au niveau de la qualité des prévisions.

On procède aux estimations à l'aide d'une fenêtre récursive d'un mois, c'est-à-dire que dans un premier temps on estime les modèles avec un sous-échantillon et on effectue la prévision pour l'horizon désiré. Par la suite, on ajoute à ce même ensemble l'information de la période suivante, puis on effectue la prévision. On suit donc ce procédé jusqu'à la fin de l'échantillon. Par exemple, pour l'ensemble des modèles, l'estimation débutera avec le sous-échantillon 2004M1-2007M8 avec lequel on prédira 2007M9 et ainsi de suite jusqu'à 2017M12. Au total, ce processus

4. Pour ce qui est des mois contenant cinq semaines, la deuxième semaine de ce mois sera considérée comme la première étant donné le peu de jours contenus dans la première semaine.

produira un échantillon de 124 prévisions. De plus, pour les modèles MIDAS, lorsqu'il y a ajout d'information à l'intérieur de la fenêtre d'estimation, les facteurs issus des mots-clés de Google Trends sont réestimés à chaque fois. L'expansion de cette fenêtre permet de notamment capter les changements structurels dans le temps, puisque les paramètres sont réestimés à chaque période. Ainsi, ce procédé est intéressant dans la mesure où il permettra aux modèles d'adapter leurs paramètres tout au long dans la période de prévisions, notamment autour de la récession de 2007-2009.

Pour ce qui est des variables choisies pour les estimations, celles des modèles MIDAS sont naturellement très semblables. En effet, pour les deux spécifications, un maximum 16 retards des facteurs estimés de Google sont employés. Cependant, le nombre de retards de la variable dépendante varie entre les spécifications. Pour le modèle MIDAS avec polynôme exponentiel d'Almon, seulement un retard a été sélectionné afin de garder le modèle le plus parcimonieux possible. En revanche, la version non contrainte en contient un maximum de quatre afin de laisser un peu plus de flexibilité au modèle. Pour ce dernier, les choix du nombre de retards des facteurs de Google et de la variable dépendante sont faits simultanément selon le critère d'information bayésien (BIC). Ceci n'est pas le cas pour la paramétrisation exponentielle d'Almon puisque le polynôme de poids choisit indirectement le nombre de retards en faisant tendre la valeur de leur poids vers zéro pour ceux qui ont moins d'importance. Cette méthodologie est appliquée avec l'ajout consécutif des trois premiers facteurs de Google issus de l'analyse par ACP.

Du côté du modèle LASSO, l'estimation est effectuée avec les séries de mots-clés de Google directement et non à l'aide de leurs facteurs. On utilise dans le modèle 16 retards pour chacun des mots-clés et quatre retards de la variable à prévoir. Les différents modèles sont estimés en imposant des restrictions sur le nombre de coefficients à ne pas réduire à zéro. De cette façon, on impose aux modèles de

conserver respectivement de un à seize coefficients, puis le modèle sans restriction est également estimé. Le choix du nombre de coefficients est dépendant de la valeur du paramètre de pénalité λ sélectionné. Ainsi, les modèles sont estimés avec 1000 valeurs différentes du paramètre et le choix de sa valeur optimale est fait à l'aide de la méthode de validation croisée. Ainsi, pour chaque valeur de λ l'échantillon qui sert à l'estimation des paramètres est divisé en cinq parties égales, puis chaque partie est prédite à l'aide des quatre autres. La valeur de λ, pour laquelle la moyenne des erreurs quadratiques moyennes obtenues des cinq prévisions est la plus faible, est celle sélectionnée. Cette méthode peut aussi permettre d'éviter les problèmes de surapprentissage (*overfitting*) associés à l'utilisation d'un très grand nombre de variables explicatives dans l'estimation de modèles.

3.4 Critère d'évaluation des prévisions

Le critère de l'erreur quadratique moyenne (EQM) sera celui utilisé afin de comparer les résultats des modèles. Ainsi, les erreurs quadratiques moyennes calculées hors échantillon pour l'ensemble des modèles sont recueillies. Pour les deux horizons, le modèle de référence pour la prévision d'un mois en avance est utilisé, puisque le *nowcasting* n'est pas possible avec ce modèle.

$$\frac{EQM^{GT}}{EQM^{AR}} = \frac{\frac{1}{N}\sum_{i=0}^{N}(\epsilon_{t+i}^{GT})^2}{\frac{1}{N}\sum_{i=0}^{N}(\epsilon_{t+i}^{AR})^2} = \frac{\frac{1}{N}\sum_{i=0}^{N}\left(y_{t+i} - \hat{y}_{t+i}^{GT}\right)^2}{\frac{1}{N}\sum_{i=0}^{N}\left(y_{t+i} - \hat{y}_{t+i}^{AR}\right)^2} \tag{3.10}$$

Les prévisions $\hat{y}^{GT}$ sont celles obtenues à l'aide des modèles utilisant les données de Google et $\hat{y}^{AR}$ sont celles obtenues du modèle de référence. Dans le cas où le ratio des erreurs est inférieur à un, cela signifie que le modèle augmenté de l'information hebdomadaire de Google performe mieux que le modèle de référence.

Le test Diebold et Mariano (1995) sera également utilisé afin de vérifier si les

performances des prévisions des modèles présentés précédemment sont statistiquement différentes du modèle de référence $AR(p)$. Ainsi, l'hypothèse nulle de ce test est que la performance des deux modèles de prévision mis en compétition est la même. La mesure de performance est représentée par leurs erreurs de prévisions hors échantillon. Le test est fondé sur la différence des erreurs soit :

$$d_t = (\epsilon^{GT}_{t+h|t})^2 - (\epsilon^{AR}_{t+h|t})^2 \tag{3.11}$$

Où $\epsilon_{t+h} = y_{t+h} - \widehat{y}_{t+h}$. De cette façon, l'hypothèse nulle et alternative sont :

$$H_0 : E(d_t) = 0$$

$$H_1 : E(d_t) \neq 0$$

La statistique de test est la suivante :

$$DM = \frac{\bar{d}}{\sqrt{\hat{V}(\bar{d})}} \rightarrow N(0,1) \tag{3.12}$$

Où $\bar{d} = \frac{1}{T}\sum_{t=1}^{T} d_t$ et $\hat{V}(\bar{d}) = \frac{1}{T}\left(\hat{\gamma}_0 + 2\sum_{i=1}^{h-1}\hat{\gamma}_i\right)$ et $\gamma_i = cov(d_t - d_{t-i})$

DM suit une loi normale centrée à zéro avec une variance d'un. Ainsi, un ratio des erreurs quadratiques moyennes inférieur à un combiné au rejet de la nulle du test Diebold-Mariano suggèrent que les méthodes d'estimations présentées précédemment performent mieux que le modèle de référence $AR(p)$.

CHAPITRE IV

RÉSULTATS

Les résultats des estimations des différents modèles explicités précédemment seront présentés dans ce chapitre. Tout d'abord, il sera question de l'analyse des résultats des prévisions et de l'intérêt de l'ajout des variables de Google dans les modèles économétriques. Ensuite, l'indice de l'activité du marché du travail résultant de ces prévisions sera présenté et comparé avec celui originalement estimé.

4.1 Résultats des prévisions

Dans le Tableau 4.1 sont reportés uniquement les résultats des estimations pour la meilleure spécification de chaque modèle. Les résultats représentent le rapport des erreurs quadratiques moyennes entre les modèles comprenant les données de Google Trends et le modèle de référence. C'est à l'aide de ce critère que la pertinence de l'ajout des données des recherches faites sur le moteur de recherche aux modèles de prévisions sera évaluée. Ainsi, un ratio inférieur à un signifie que le modèle augmenté des données de Google performe mieux que celui de référence dans l'échantillon choisi.

Tableau 4.1: Ratio des erreurs quadratiques moyennes

		H = 0				H = 1			
Séries	Modèles	Semaine 1	Semaine 2	Semaine 3	Semaine 4	Semaine 1	Semaine 2	Semaine 3	Semaine 4
Civilian Unemployment Rate	MIDAS-AR(1)	1.0132	0.9988	**0.9527**	1.0157	0.9166	**0.8929**	0.9631	0.9494
	U-MIDAS-AR	0.9858	0.9844	0.9965	0.9725	0.9056*	0.9051	0.9428	0.9086
	LASSO-AR	1.0056	1.005	1.0097	1.0117	0.9441	0.9135	0.9444	0.9435
Total unemployed plus all marginally attached workers plus total employed part time for economic reasons	MIDAS-AR(1)	0.909	0.974	0.922	**0.907**	0.9818	0.9703	0.9222	**0.8963**
	U-MIDAS-AR	0.9528	0.9495	0.9256	0.9716	0.9292*	0.9313*	0.9333	0.9185
	LASSO-AR	1.0182	1.0181	1.0238	1.0226	1.0116	1.0291	1.0117	1.0149
Labor Force Flows Unemployed to Employed : 16 Years and Over	MIDAS-AR(1)	1.0019	0.992	1.0053	1.0063	1.3864***	1.3627***	1.4332***	1.3404***
	U-MIDAS-AR	1.005	**0.9919**	1.0136	1.0093	1.4313***	1.3711***	1.3977***	1.3962***
	LASSO-AR	1.1218*	1.1324*	1.1311**	1.1182*	1.3145***	**1.2957****	1.3169**	1.3023***
Quits : Total Private	MIDAS-AR(1)	**0.9691**	0.9736	0.9892	0.9792	1.1798**	1.131*	1.1222	1.0953
	U-MIDAS-AR	1.0008	1.0079	0.9929	1.0191	1.1269	1.1233	1.1222	1.1299*
	LASSO-AR	1.0504	1.0748	1.0645	1.0522	**1.0722**	1.0785	1.0754	1.0833
Civilian Employment Population Ratio	MIDAS-AR(1)	0.9937	1.007	1.0892	0.9939	1.0155	0.9771	0.9317	0.9265
	U-MIDAS-AR	1.0076	**0.9647**	0.9958	0.9775	0.9509	0.9387	0.9489	**0.9256***
	LASSO-AR	1.0037	1.0131	1.0086	1.0075	1.0224	1.0301	1.0199	1.0323
Employment Level : Part-Time for Economic Reasons, All Industries	MIDAS-AR(1)	**0.8961**	0.9867	0.9093	0.9485	1.1175**	0.9728	0.9112	**0.8842**
	U-MIDAS-AR	0.9263*	0.9361	0.95	0.9829	0.9279	0.9114	0.9099*	0.9677
	LASSO-AR	0.8988	0.9129*	0.9235*	0.8967*	0.9165*	0.9171*	0.9183*	0.9201*
Job Leavers as a Percent of Total Unemployed	MIDAS-AR(1)	1.0273	1.1039	0.9646	**0.9586**	1.0509	0.9914	1.0516	0.9841
	U-MIDAS-AR	1.043*	1.0215*	1.0077	1.0602*	1.0376	1.0526	0.9898	0.9961
	LASSO-AR	0.9692	0.9799	0.983	0.9748	0.9664	0.968	**0.9594**	0.9618
Of Total Unemployed, Percent Unemployed 27 Weeks and over	MIDAS-AR(1)	0.9988	0.9797	1.0245**	1.0406**	**0.9164**	0.9364	1.0339	1.0003
	U-MIDAS-AR	**0.9695****	0.9973	1.0113***	1.01	0.9449	0.9483	0.9609	0.987
	LASSO-AR	0.9763	0.9714	0.9774	0.9794	0.9706	0.9758	0.9747	0.9724
Job Losers as a Percent of Total Unemployed	MIDAS-AR(1)	1.043	1.1555**	1.009	**0.9592**	1.0936	1.0272	1.0325	1.0277
	U-MIDAS-AR	1.0511*	1.006	1.015	1.0054	1.0516	1.036	1.0354	1.0297
	LASSO-AR	0.9848	0.9795	0.9832	0.9842	0.9715	**0.9621**	0.9678	0.966
Average Hourly Earnings of Production and Nonsupervisory Employees Total Private	MIDAS-AR(1)	1.5745***	1.6328***	1.6259***	1.6078***	2.5725***	2.4559***	2.395***	2.4584***
	U-MIDAS-AR	1.0201	1.02	1.0129	**0.9943**	2.2867***	2.3134***	**2.2245*****	2.2296***
	LASSO-AR	1.5872***	1.5618***	1.633***	1.6005***	2.4229***	2.4255***	2.4143***	2.4324***
All Employees Total Private Industries	MIDAS-AR(1)	1.0128	1.0145	1.0051	0.9745	1.1682**	1.1703**	1.1834***	**1.132***
	U-MIDAS-AR	0.9854	0.9778	0.9726	**0.9675**	1.1501**	1.1631**	1.1609**	1.1762***
	LASSO-AR	1.0784	1.0743	1.0595	1.0957	1.1938**	1.1933**	1.198**	1.1681**
Indexes of Aggregate Weekly Payrolls of Production and Nonsupervisory Employees Total Private	MIDAS-AR(1)	1.2659**	1.4686***	1.2463*	1.3811*	1.2107**	1.1109	1.1356*	1.1346
	U-MIDAS-AR	1.0045	**1.0037**	1.0242	1.0162**	**0.957**	0.9602	0.9642	0.968
	LASSO-AR	1.1415	1.1622	1.151	1.1486	1.1468	1.1489	1.1492	1.145
All Employees Professional and Business Services : Temporary Help Services	MIDAS-AR(1)	1.1563	1.073	1.122	1.1399	1.2262	1.1994	1.1362	**1.089**
	U-MIDAS-AR	1.0255	**0.956**	1.0239	0.9956	1.1564*	1.098	1.1281	1.1823
	LASSO-AR	1.3168	1.3128*	1.3171	1.3253	1.3627*	1.3754	1.3694*	1.406*
Civilian Labor Force Participation Rate	MIDAS-AR(1)	1.026	0.9955	1.0313**	**0.989**	1.1024**	1.1164**	1.1225**	1.0986**
	U-MIDAS-AR	0.997	1.0098	1.0263**	1.0034	1.0907**	1.0802*	1.1024**	1.1164**
	LASSO-AR	1.0279	1.0187	1.0286	1.0306	1.0576	1.0583*	**1.0545**	1.0598*
Index of Aggregate Weekly Hours : Production and Nonsupervisory Employees Total Private Industries	MIDAS-AR(1)	1.126	1.1315	1.0664	1.0983	0.928	0.8953	0.8935	0.8439
	U-MIDAS-AR	0.8942	**0.8715**	0.893	0.8927	0.7898**	**0.7854****	0.8026**	0.799**
	LASSO-AR	0.9423	0.9241	0.9407	0.9398	0.8692	0.9098	0.8844	0.8958
Hires Total Private	MIDAS-AR(1)	0.9493	1.0107	0.9449	0.9528	1.225**	1.1917*	1.1996**	1.2157**
	U-MIDAS-AR	1.0074	1.009	0.9894	0.9848	1.1792**	1.1622*	1.1839**	1.1615*
	LASSO-AR	**0.9934**	0.9955	0.9968	1.0022	**1.1436***	1.1452*	1.15*	1.1447*
Monthly average Initial Claims	MIDAS-AR(1)	0.9977	0.9562	**0.8993**	0.9224	0.9426	**0.8889**	1.0032	1.0382
	U-MIDAS-AR	1.01	1.0393	1.0999	1.0194**	0.9251	0.928	0.9174	0.9251
	LASSO-AR	0.9785	0.9745	0.9771	0.9797	0.9533	0.9616	0.9612	0.9721
ISM Manufacturing Employment Index	MIDAS-AR(1)	1.0416**	1.0473**	1.0315	1.0262*	1.0583**	1.1058**	1.0299	1.065*
	U-MIDAS-AR	1.0058	0.9973	1.0064	1.0334*	1.0263	1.0254	1.0485*	1.0696*
	LASSO-AR	**0.9885**	0.9946	0.993	0.9945	0.9833	0.9767	0.9826	**0.9596***
NFIB Percent of firms planning to increase employment	MIDAS-AR(1)	0.9896	**0.9571**	0.9887	0.9803	1.1028	1.086	1.0667	**0.9891**
	U-MIDAS-AR	1.0375	0.9771	1.0154	1.0253	1.0348	0.9974	1.0098	1.0046
	LASSO-AR	0.9704	0.9692	0.9649	0.962	1.0069	1.0066	1.005	1.0048
Expected job availability (U of Michigan)	MIDAS-AR(1)	1.0033	0.9736	0.9751	1.0453	**0.9071****	0.9101**	0.9333**	0.9603
	U-MIDAS-AR	0.9668*	0.9943	0.9833	1.0355	0.9224**	0.931**	0.9517*	0.9344**
	LASSO-AR	**0.9005****	0.923**	0.9105***	0.9228**	0.9174**	0.9282**	0.9275**	0.9336**
NFIB Percent of firms with positions not able to fill right now	MIDAS-AR(1)	1.2312**	1.1892*	1.0724	1.1417*	1.4774***	1.4705***	1.6222***	1.4268***
	U-MIDAS-AR	1.0381	1.0162	0.9965	**0.9489****	1.451***	1.4211***	1.3971**	1.3951**
	LASSO-AR	1.2551*	1.2404*	1.2604**	1.2455**	1.3819**	**1.3811****	1.387***	1.3867***

Note : Les nombres représentent les EQM relatives entre les différents modèles présentés précédemment et le modèle de référence $AR(p)$. La meilleure spécification du modèle pour une variable donnée est indiquée en gras. Les étoiles représentent le niveau de significativé auquel le test Diebold-Mariano répond. Ainsi, ***, ** et * correspondent respectivement à un niveau de significativité de 1%, 5% et 10%.

Pour les prévisions de la période en cours, les modèles MIDAS avec les facteurs ont un avantage clair par rapport aux modèles LASSO. En effet, ceux-ci performent mieux pour 18 des 21 variables à prévoir. L'agrégation de l'information des recherches de Google semble donc beaucoup mieux performer que l'utilisation directe des mots-clés individuels recherchés. Le polynôme exponentiel d'Almon offre des performances similaires à la version non contrainte du modèle MIDAS. De plus, les modèles plus parcimonieux ne présentant qu'un seul facteur sont ceux qui produisent le plus souvent les meilleurs résultats. Cependant, pour l'horizon d'un mois la performance générale des modèles est plus variable. Bien qu'une bonne partie des modèles semblent mieux performer que le modèle de référence, la performance générale diminue légèrement par rapport au *nowcasting*, ce qui montre l'efficacité des données de Google et l'intérêt pour ce genre de prévisions. De plus, les modèles semblent utiliser plus de facteurs et de mots-clés et sont donc moins parcimonieux que pour les estimations précédentes. Pour les deux horizons, les modèles avec les données de Google semblent, notamment, avoir la capacité de bien prévoir les deux mesures du taux de chômage américain, qui sont des indicateurs clés pour l'analyse de la santé du marché du travail. Effectivement, comme les dépenses de consommation personnelles comptent pour près de 70% du produit intérieur brut américain, le niveau futur du taux de chômage fournit une idée du futur niveau des revenus et par le fait même du niveau de la demande agrégée.

Bien que les données hebdomadaires semblent aider à améliorer la performance des modèles de prévisions, l'ajout des semaines consécutives au cours du mois ne semble pas drastiquement changer l'efficacité pour les deux horizons. De plus, l'ajout consécutif des semaines au cours d'un mois n'augmente pas nécessairement le pouvoir prédictif des modèles. En effet, pour certaines variables, l'ajout de ces semaines entraine plutôt une dégradation des performances. Néanmoins, l'infor-

mation contenue uniquement dans la première semaine pour les deux horizons semble généralement contenir moins de pouvoir prédictif que les semaines qui lui suivent.

Les estimations ont débuté avant la Grande Récession afin de vérifier si les variables de Google permettraient de bien capter le début des périodes de crise. Ainsi, les Figure B.3 et B.4 de l'annexe B présentent les ratios des erreurs quadratiques moyennes entre les meilleurs modèles et le modèle de référence dans le temps. De manière générale, les modèles performent moins bien et sont plus volatiles durant la période de récession. Cependant, les résultats semblent, pour la plupart, s'améliorer et se stabiliser plus l'échantillon utilisé, pour faire l'estimation, augmente. Le problème peut naturellement provenir de l'incertitude économique entourant les épisodes de récessions, mais aussi du fait que l'échantillon pour faire les prévisions durant cette période est relativement faible, ce qui entraine de l'instabilité (variation) dans les paramètres pour les premières prévisions.

De manière générale, l'ajout des variables à fréquences hebdomadaires de Google traduit une augmentation du potentiel prédictif des modèles utilisés. Ainsi, dans la plupart des cas, ces derniers surpassent le modèle de référence. Cependant, ces améliorations demeurent souvent assez modestes et l'ajout des données semaine après semaine n'améliore pas nécessairement les résultats des prévisions. Toutefois, l'objectif du présent travail n'est pas d'uniquement vérifier la pertinence de l'ajout des données de Google dans les modèles de prévisions, mais également de vérifier si celles-ci permettent de recréer les mouvements du marché du travail.

4.2 Indices de l'activité du marché du travail

L'indice est construit à l'aide des prévisions des modèles présentant les meilleures performances au niveau du ratio des erreurs quadratiques moyennes pour chaque

variable. S'il existe une égalité au sein des EQM, le modèle le plus parcimonieux est celui qui a été sélectionné.

Les Figures 4.1 et 4.2 montrent respectivement les indices obtenus à l'aide des prévisions pour la période en cours et pour l'horizon d'un mois. Ces indices ne répliquent pas exactement les mêmes mouvements que l'indicateur réel pour chaque point dans le temps. Néanmoins, ceux-ci semblent très bien capter les périodes de creux et de reprises durant la récession de 2007-2009. Ils semblent également bien capter les tendances de long terme du marché du travail. Cependant, l'intérêt de l'utilisation des Google Trends est notamment pour l'analyse de très court terme. Or, bien que les écarts entre l'indice original et ceux qui ont été prédits soient assez minces, ce qui intéresse ici est plutôt le signe de la variation d'une période à l'autre. En effet, ce signe indique dans quel sens l'état général du marché se dirige. Une variation positive signifie que le marché du travail se porte mieux comparativement à la période précédente. La logique est inversée lorsque la variation est du côté négatif. Comme il a été mentionné, la valeur de l'indice n'a pas de signification économique en elle-même. L'aspect qui importe donc ici est en fait la capacité de l'indice de Google à prévoir le signe de cette variation, ce qui indiquera la santé relative future du marché du travail.

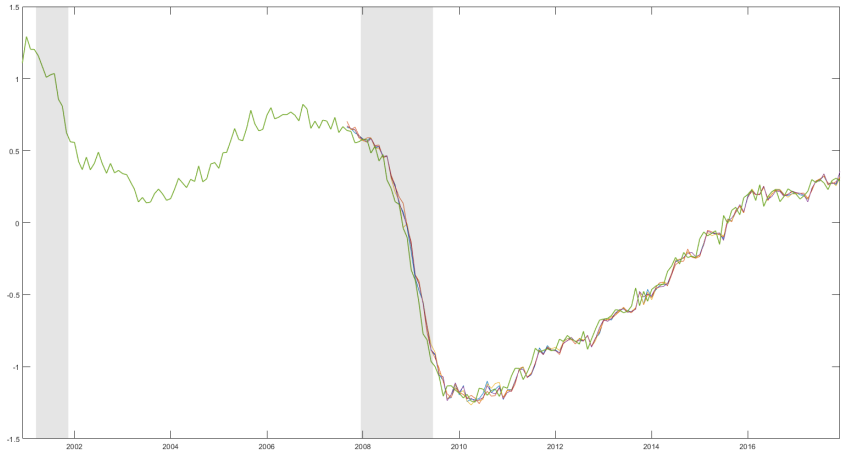


Figure 4.1: Méthode avec les composantes : Indice résultant pour les meilleurs modèles pour H = 0

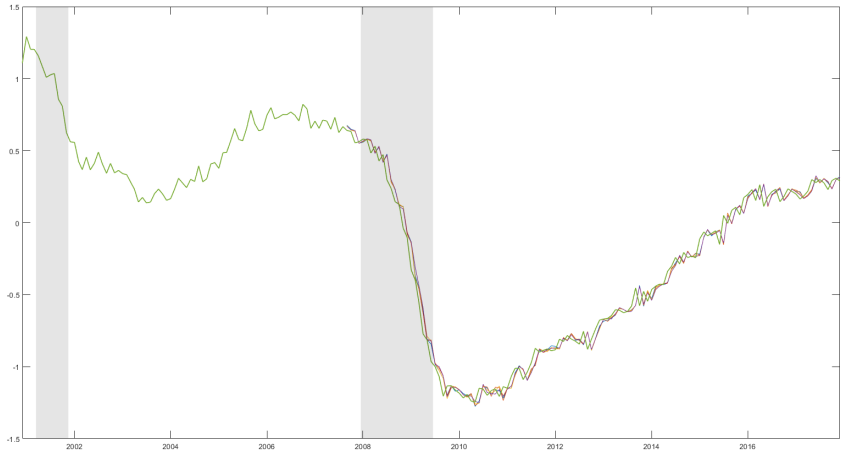


Figure 4.2: Méthode avec les composantes : Indice résultant pour les meilleurs modèles pour H = 1

Afin de comparer la performance des indices construits à partir des données de Google, ceux-ci seront soumis au test de signe de Pesaran et Timmermann (1992). Celui-ci indiquera si les mouvements d'une période à l'autre des nouveaux indices font statistiquement mieux qu'un simple choix pris au hasard. Cette statistique de test n'est pas influencée par la distance entre le vrai indice et la prévision, comme c'est le cas pour l'EQM. Sous l'hypothèse nulle, les signes des prévisions sont indépendants de ceux du vrai indice.

$$S = \frac{\hat{p} - \hat{p}^*}{\sqrt{Var(\hat{p}) - Var(\hat{p}^*)}} \sim N(0,1)$$

$\hat{p}$ représente la proportion des signes bien estimées et $\hat{p}^*$ est l'estimation de son espérance.

Tableau 4.2: Résultats pour le test de signe : Indice construit avec les prévisions des séries

	H = 0			H = 1		
Semaine	Pourcentage même variation	P-Value	Statistique S	Pourcentage même variation	P-Value	Statistique S
1	63,41	0.0029	2.98	62,60	0.0049	2.8165
2	61,79	0.0089	2.6174	58,54	0.0576	1.8985
3	62,60	0.0052	2.7965	63,41	0.0023	3.0480
4	65,85	0.00040890	3.5343	60,16	0.0236	2.2638

Les résultats sont reportés au Tableau 4.2 et indiquent clairement que les indices construits à l'aide des Google Trends performent mieux qu'un simple choix au hasard. De plus, pour les indices prédits avec les données du mois en cours, la proportion des variations qui ont été bien prédites tourne dans les alentours de 61% à 65% selon la semaine du mois. Cela va de pair avec les résultats des prévisions, puisque la proportion des variations pour la période en cours est légèrement supérieure à celle de l'horizon d'un mois qui compte une proportion de 58% à 63%.

Une autre manière d'aborder le problème est de prédire directement l'indice plutôt

que ses composantes. Ainsi, la même méthodologie de prévision a été appliquée au niveau de la prévision directe de l'indice. Afin de garder l'aspect de prévision en temps réel, les nouveaux indices prévus ne sont pas construits directement à l'aide de tout l'échantillon, mais sur le même échantillon qu'utilisent les estimations. C'est-à-dire que l'indice à prévoir est reconstruit chaque mois avec l'ajout d'information. Comme pour les prévisions des séries, on utilise l'erreur quadratique moyenne comme critère de sélection. Ainsi, pour chaque semaine le meilleur modèle sélectionné est celui dont l'EQM est la plus faible.

Tableau 4.3: Ratio des erreurs quadratiques moyennes pour la prévision directe de l'indice

	H = 0				H = 1			
	Semaine 1	**Semaine 2**	**Semaine 3**	**Semaine 4**	**Semaine 1**	**Semaine 2**	**Semaine 3**	**Semaine 4**
MIDAS-AR	0.92724	0.88393	0.92176	0.74987	1.2815	0.71061	1.016	**0.60989***
U-MIDAS-AR	0.7719*	0.82542*	0.78774*	0.81938*	1.0234	1.0468	0.86478	0.86134*
LASSO	**0.69976**	**0.70928**	**0.7202**	**0.70978**	**0.67745**	**0.66091***	**0.64496***	0.6414*

Note : Les nombres représentent les EQM relatives entre les différents modèles présentés précédemment et le modèle de référence $AR(p)$. Les meilleurs modèles sont indiqués en gras. Les étoiles représentent le niveau de significativé auquel le test Diebold-Mariano répond. Ainsi, ***, ** et * correspondent respectivement à un niveau de significativité de 1%, 5% et 10%.

Comme précédemment, les modèles de prévisions directes de l'indice à l'aide des données de Google sont comparés au modèle de référence $AR(p)$. L'utilisation des données du géant du Web offre ici aussi un avantage comparativement au modèle $AR(p)$. Cela dit contrairement aux résultats obtenus antérieurement, le modèle LASSO est celui qui se démarque le plus. Ainsi, l'indice sera construit à l'aide de la méthode LASSO, sauf pour la quatrième semaine de l'horizon d'un mois où ce sera plutôt le modèle MIDAS avec le polynôme exponentiel d'Almon qui sera privilégié.

Les Figures 4.3 et 4.4 montrent respectivement les indices obtenus à l'aide des prévisions directes pour la période en cours et pour l'horizon d'un mois. Dans la même veine que les indices prédits précédemment, ceux-ci semblent bien capter les points de retournements ainsi que les tendances de long terme du marché. Cependant, surtout durant la récession, ils semblent plus loin de la vraie valeur que les indices prédits avec les composantes. Toutefois, ce qui importe davantage ici, c'est la proportion de la variation de court terme commune avec l'indice réel.

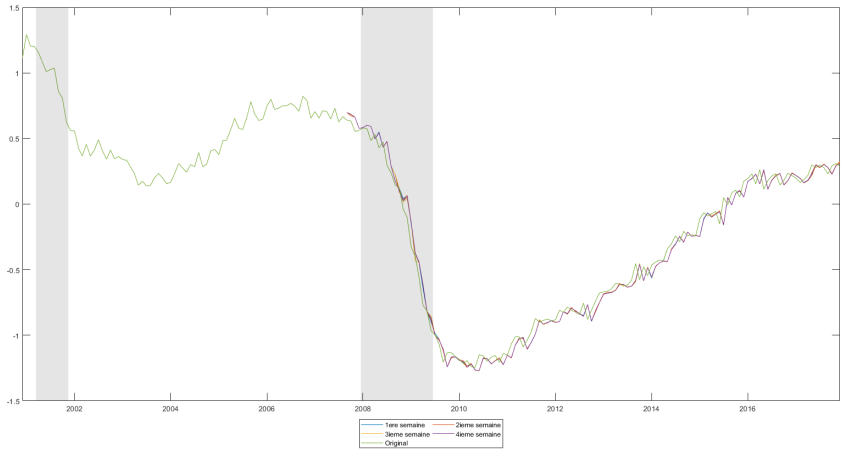


Figure 4.3: Méthode par prévision directe : Indice résultant pour les meilleurs modèles pour H = 0

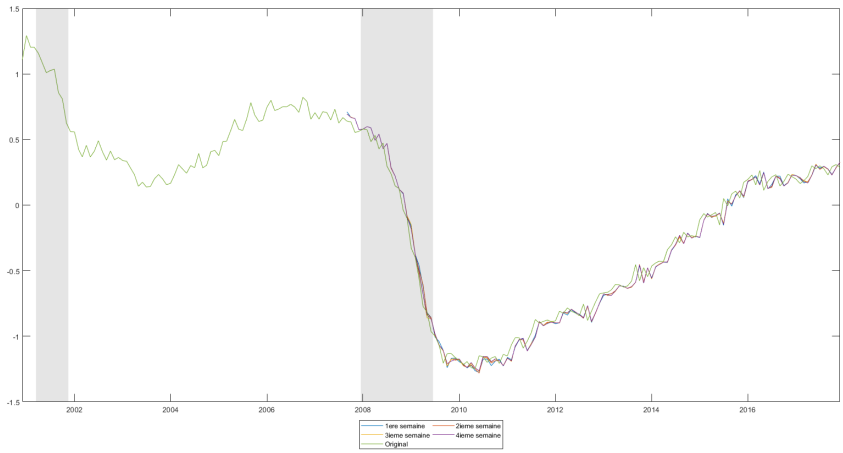


Figure 4.4: Méthode par prévision directe : Indice résultant pour les meilleurs modèles pour H = 1

Tableau 4.4: Résultats pour le test de signe : Indice construit avec la prévision directe

	H = 0			H = 1		
Semaine	**Pourcentage même variation**	**P-Value**	**Statistique S**	**Pourcentage même variation**	**P-Value**	**Statistique S**
1	47,97	0.7157	-0.3643	52,03	0.4553	0.7466
2	46,34	-0.8124	-0.2563	49,59	0.9300	0.0878
3	44,72	0.1965	-1.2916	55,28	0.1023	1.6338
4	43,09	0.0706	-1.8081	55,28	0.2172	1.2341

L'analyse des mouvements de courts termes est faite de la même manière que précédemment. Contrairement aux résultats antérieurs, les variations d'une période à l'autre sont moins bien représentées. Ces résultats sont reportés dans le Tableau 4.4. Aucun des indices ne fait statistiquement mieux que le choix aléatoire à un seuil de 5%. Par conséquent, au niveau des mouvements communs entre les périodes, ces indices performent moins bien que leurs vis-à-vis avec des pourcentages variant de 43% à 48% pour le *nowcasting* et de 49% à 55% pour l'horizon d'un mois. Ce qui montre encore une fois la pertinence de l'utilisation des données de Google Trends afin d'effectuer ce genre de prévision.

CONCLUSION

Le présent mémoire avait pour objectif de créer un indice représentant l'état du marché de travail américain tel que le font Willis *et al.* (2014), mais avec comme particularité l'utilisation des données de recherches Internet de Google, puis, par la même occasion, de vérifier la pertinence de l'utilisation de celles-ci afin de faire des prévisions pour de courts horizons. En plus de l'utilisation de ces données, une des particularités de ces prévisions c'est qu'elles étaient effectuées à l'aide de modèle à fréquences mixtes. Afin de limiter les mots-clés à ceux qui sont potentiellement pertinents pour la prévision des composantes de l'indice, la sélection des mots-clés a été conduite par les données elles-mêmes à l'aide d'une méthode de filtrage puisant ceux-ci d'un bassin qui en contenait plus de 3 400. Ces mêmes prévisions ont été réalisées à l'aide de trois types de modèles. Les modèles MIDAS, d'abord avec polynôme exponentiel d'Almon qui est capable de synthétiser l'information des données à haute fréquence afin de diminuer les paramètres à estimer, puis le U-MIDAS, qui permet une paramétrisation plus souple. Le LASSO a également été utilisé pour sa sélection plus ciblée des mots-clés et des retards qui sont pertinents pour la prévision.

Ainsi, à la lumière des résultats, nous pouvons conclure que l'ajout des données à fréquence hebdomadaire de Google à le potentiel d'améliorer les prévisions pour de très courts horizons. Ceux-ci appuient donc la récente littérature au sujet de la qualité prévisionnelle des Google Trends. En effet, avec leur ajout, de manière générale, les modèles performent mieux que le modèle de référence $AR(p)$ utilisé. Cependant, les semaines importantes au niveau de la qualité des prévisions changent beaucoup selon la variable à prévoir, et l'ajout consécutif des semaines au

cours de l'exercice n'améliore pas nécessairement les résultats. De plus, le pouvoir prévisionnel de ces données semble s'atténuer avec l'augmentation de l'horizon de prévision. En effet, bien que le nombre d'horizons analysés soit faible, on remarque une légère baisse de la performance pour l'horizon d'un mois. Néanmoins, les assez bons résultats se traduisent dans les mouvements à court terme de l'indice construit. Effectivement, cet indice reproduit entre 61% et 65% des mouvements de l'indice réel d'une période à l'autre selon l'horizon et la semaine utilisés.

L'utilisation de ce type de données est de plus en plus présente dans la littérature prévisionnelle en science économique. Les données dites *non conventionnelles* apportent de l'information différente sur des aspects qui sont autrement très difficilement mesurables. Le cas des données de recherche de Google en est un bon exemple, puisque celles-ci ont potentiellement la capacité de quantifier l'attention que porte la population sur un sujet particulier à travers leurs recherches sur Internet. Cependant, un des principaux défauts de ces données est qu'il y a, pour le moment, assez peu d'observation. Il sera donc intéressant de voir comment les modèles de prévision se comportent avec l'ajout d'observations dans le temps, surtout en période d'incertitude économique. Effectivement, la seule période de récession couverte est celle de 2007-2009. Si ces résultats sont vérifiés, l'utilisation des données des recherches faites sur Internet pourrait être considérée comme un ajout intéressant dans un futur non trop lointain.

APPENDICE A

DONNÉES MACROÉCONOMIQUES

A.1 Provenance et transformation des séries

Tableau A.1: Variables utilisées pour la construction de l'indice

Variable	Source	Code	Transformation
1.Civilian Unemployment Rate	Federal Reserve Economic Database	UNRATE	D
2.Total unemployed, plus all marginally attached workers plus total employed part time for economic reasons	Federal Reserve Economic Database	U6RATE	D
3.Labor Force Flows Unemployed to Employed : 16 Years and Over	Federal Reserve Economic Database	LNS17100000	DLN
4.Quits : Total Private	Federal Reserve Economic Database	JTS1000QUR	D
5.Civilian Employment-Population Ratio	Federal Reserve Economic Database	EMRATIO	D
6.Employment Level : Part-Time for Economic Reasons, All Industries	Federal Reserve Economic Database	LNS12032194	DLN
7.Job Leavers as a Percent of Total Unemployed	Federal Reserve Economic Database	LNS13023706	D
8.Of Total Unemployed, Percent Unemployed 27 Weeks and over	Federal Reserve Economic Database	LNS13025703	D
9.Job Losers as a Percent of Total Unemployed	Federal Reserve Economic Database	LNS13023622	D
10.Average Hourly Earnings of Production and Nonsupervisory Employees : Total Private	Federal Reserve Economic Database	AHETPI	DD
11.All Employees : Total Private Industries	Federal Reserve Economic Database	USPRIV	DD
12.Indexes of Aggregate Weekly Payrolls of Production and Nonsupervisory Employees : Total Private	Federal Reserve Economic Database	CES0500000035	D
13.All Employees : Professional and Business Services : Temporary Help Services	Federal Reserve Economic Database	TEMPHELPS	DLN
14.Civilian Labor Force Participation Rate	Federal Reserve Economic Database	CIVPART	D
15.Index of Aggregate Weekly Hours : Production and Nonsupervisory Employees : Total Private Industries	Federal Reserve Economic Database	AWHI	DLN
16.Hires : Total Private	Federal Reserve Economic Database	JTS1000HIR	D
17.Monthly average Initial Claims	Federal Reserve Economic Database	MANEMPL	DLN
18.Manufacturing Employment Index	Institute for Supply Management	ICSA	D
19.Percent of firms planning to increase employment	National Federation of Independent Business	PTI	D
20.Expected job availability	University of Michigan	EJA	D
21.Percent of firms with positions not able to fill right now	National Federation of Independent Business	PNTF	D

A.2 Indices du l'activité du marché du travail

Tableau A.2: Résulats de l'analyse par composante principale

Composante Principale	Valeur Propre	% de la Variance	% Cummulatif
1	10.6995	51.1999	51.1999
2	4.5485	21.7656	72.9656
3	3.4357	16.4408	89.4064
4	1.1113	5.3176	94.724
5	0.30658	1.467	96.1911
6	0.20304	0.97162	97.1627
7	0.15078	0.72152	97.8842
8	0.11382	0.54464	98.4288
9	0.082044	0.3926	98.8214
10	0.064638	0.30931	99.1307
11	0.062097	0.29715	99.4279
12	0.040108	0.19193	99.6198
13	0.032435	0.15521	99.775
14	0.024677	0.11809	99.8931
15	0.013734	0.065719	99.9588
16	0.0079819	0.038195	99.997
17	0.00029063	0.0013907	99.9984
18	0.00026252	0.0012562	99.9997
19	6.1716e-05	0.00029532	100
20	4.5649e-06	2.1844e-05	100
21	1.3049e-08	6.2441e-08	100

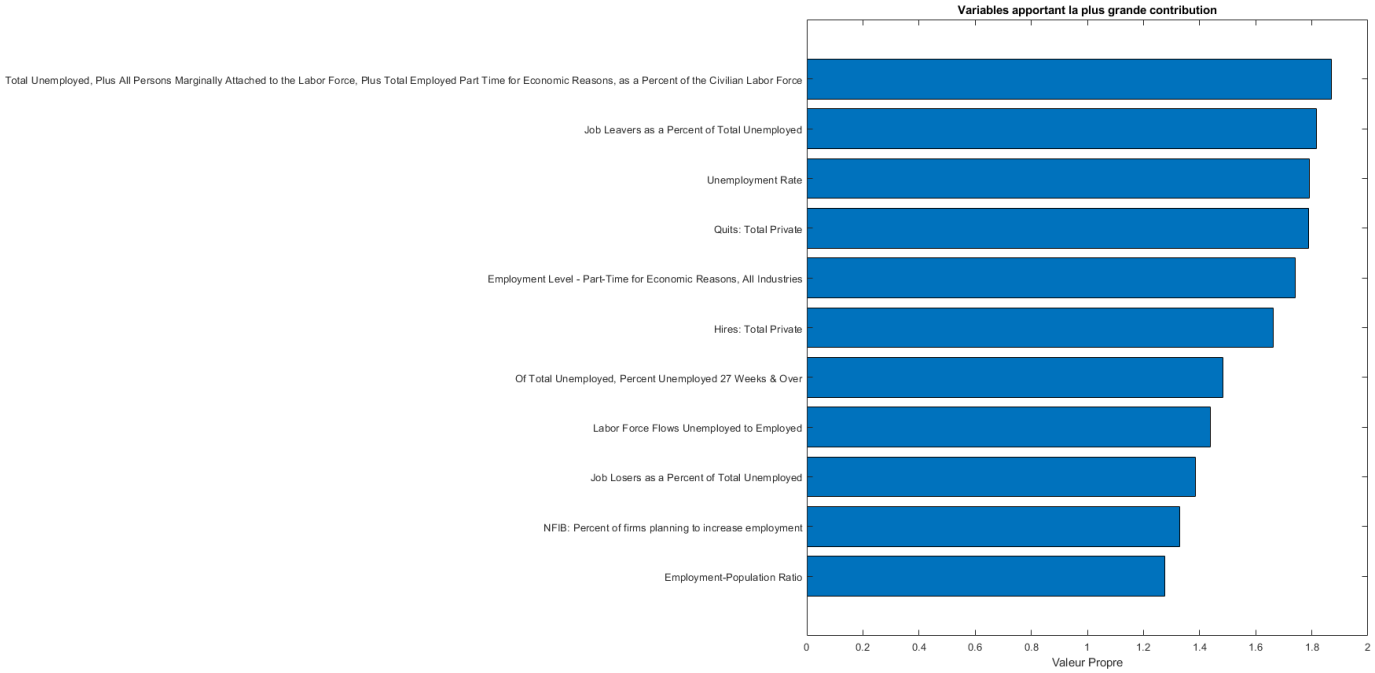


Figure A.1: Top dix des variables qui contribuent le plus à l'indice

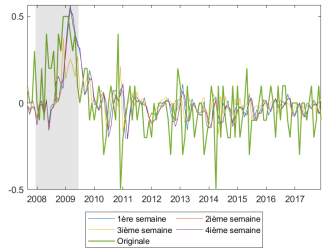
Tableau A.3: Variables sélectionnées avec la méthode de pré-sélection

Sujet de la recherche	**Nombre de mots-clés**
Recherche de travail	79
Chômage	134
Salaire	68
Curriculum vitae et entrevue	23
Mise à pied	36
Total	340

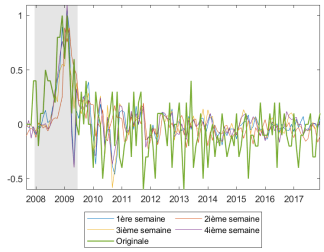
APPENDICE B

RÉSULTATS DES ESTIMATIONS

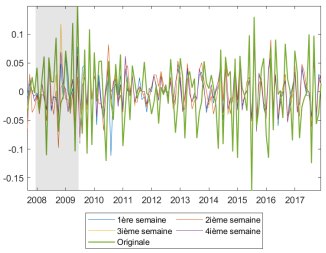
B.1 Graphiques des prévisions



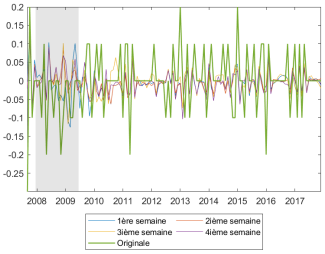
UNRATE



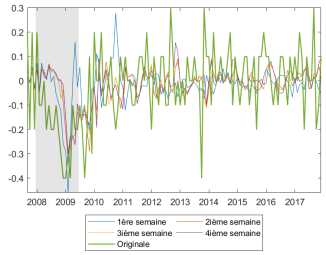
U6RATE



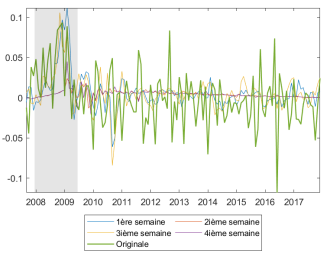
LNS17100000



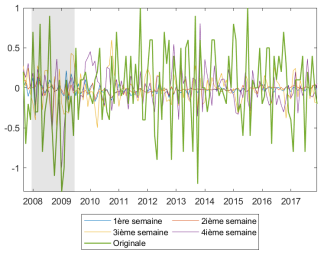
JTS1000QUR



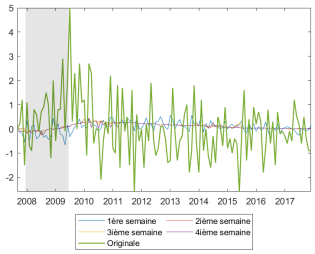
EMRATIO



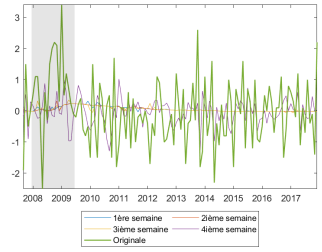
LNS12032194



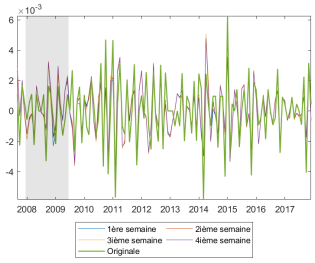
LNS13023706



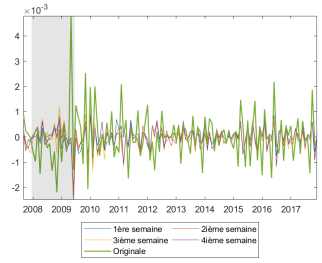
LNS13025703



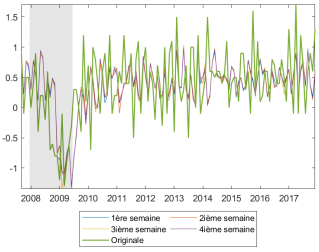
LNS13023622



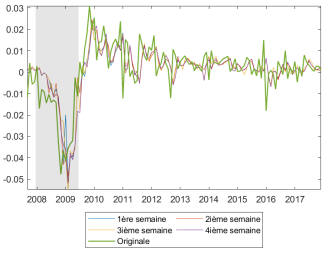
AHETPI



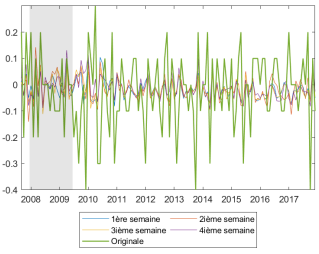
USPRIV



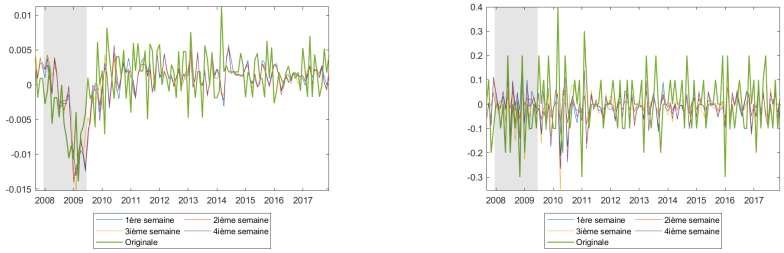
CES0500000035



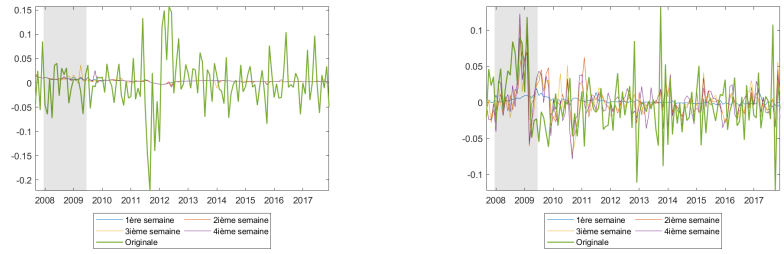
TEMPHELPS



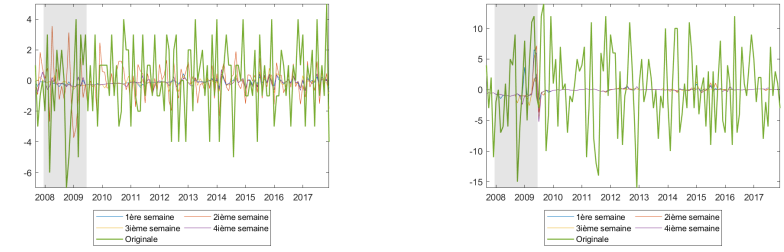
CIVPART



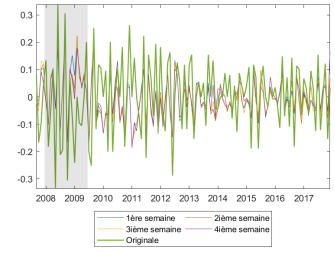
AWHI JTS1000HIR



MANEMPL ICSA

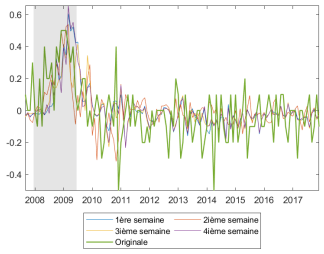


PTI EJA

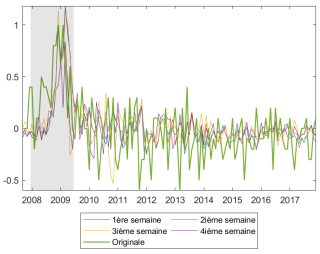


PNTF

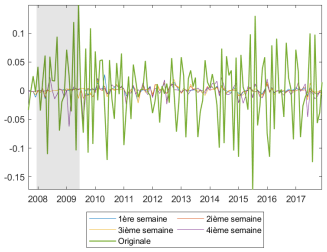
Figure B.1: Prévisions des meilleurs modèles pour chaque semaine (*nowcasting*)



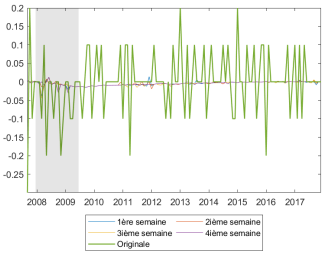
UNRATE



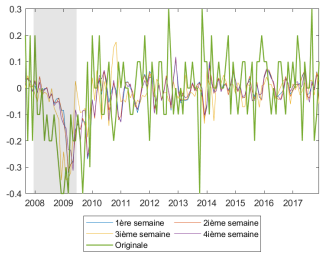
U6RATE



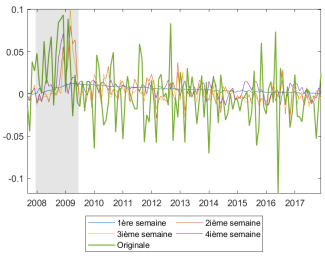
LNS17100000



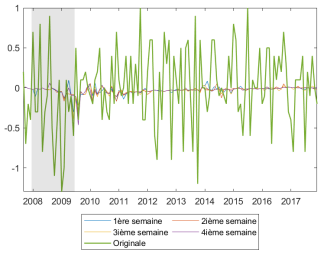
JTS1000QUR



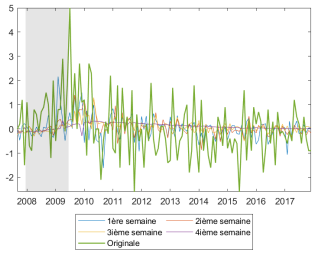
EMRATIO



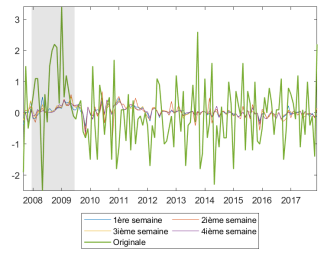
LNS12032194



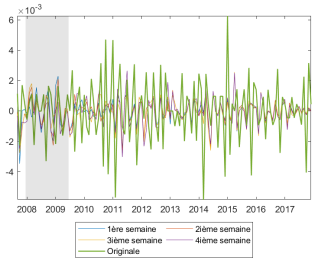
LNS13023706



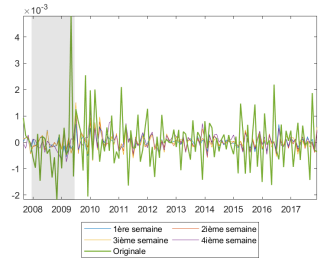
LNS13025703



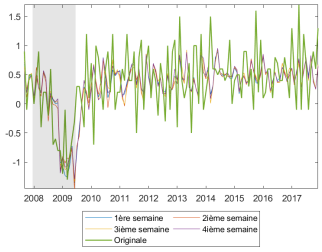
LNS13023622



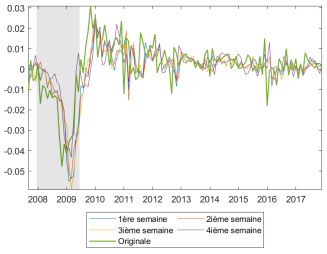
AHETPI



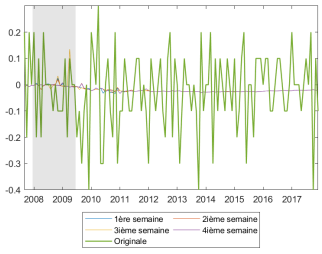
USPRIV



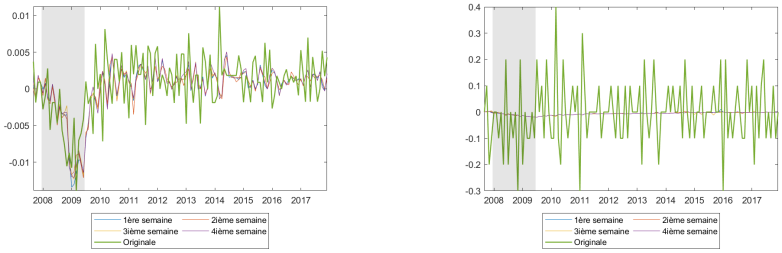
CES0500000035



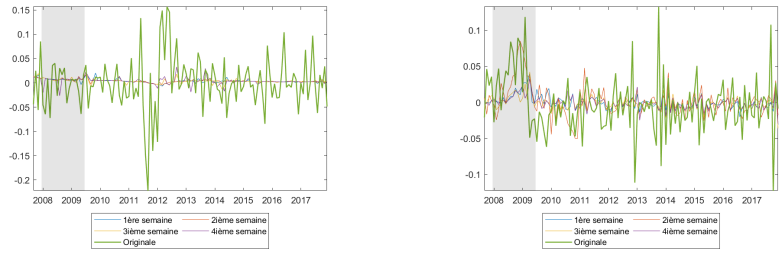
TEMPHELPS



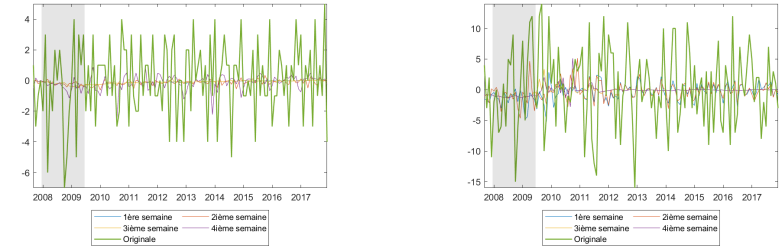
CIVPART



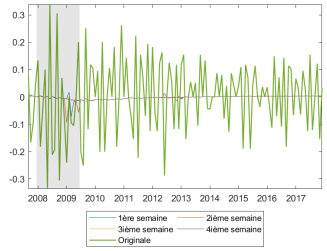
AWHI JTS1000HIR



MANEMPL ICSA



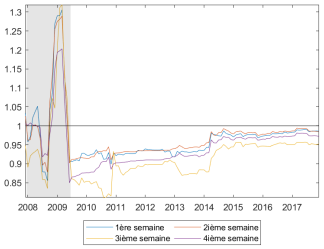
PTI EJA



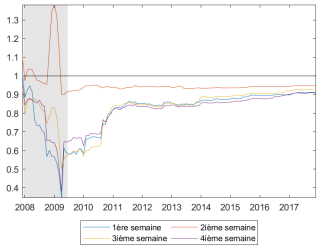
PNTF

Figure B.2: Prévisions des meilleurs modèles pour chaque semaine (H = 1)

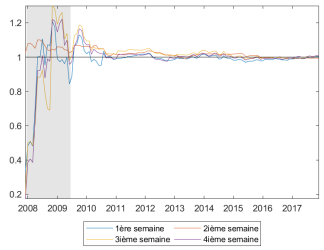
B.2 Graphiques des EQM dans le temps



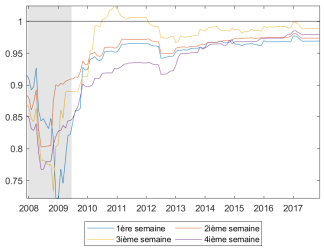
UNRATE



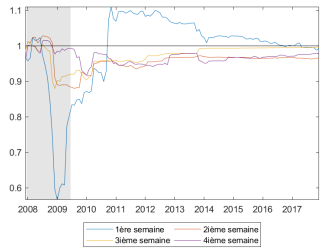
U6RATE



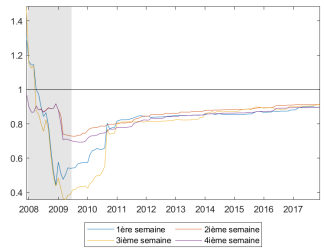
LNS17100000



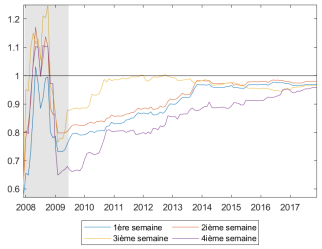
JTS1000QUR



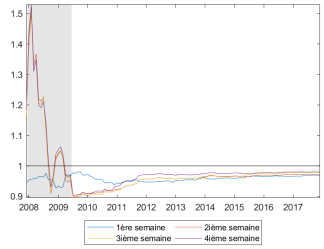
EMRATIO



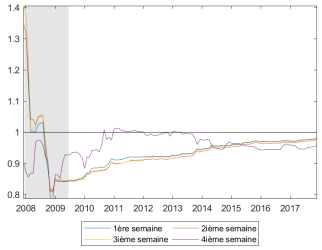
LNS12032194



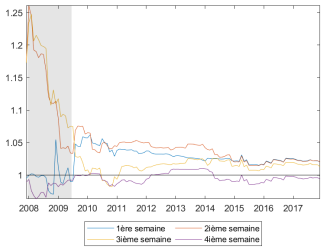
LNS13023706



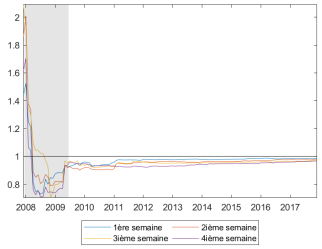
LNS13025703



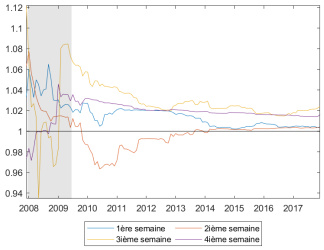
LNS13023622



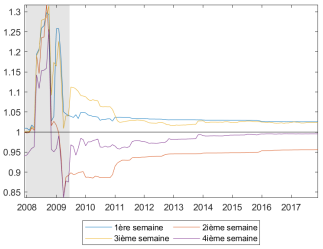
AHETPI



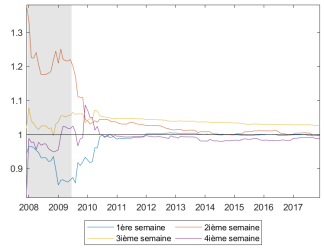
USPRIV



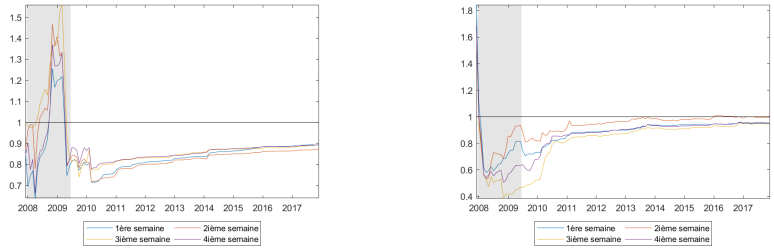
CES0500000035



TEMPHELPS

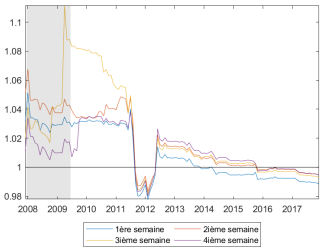


CIVPART

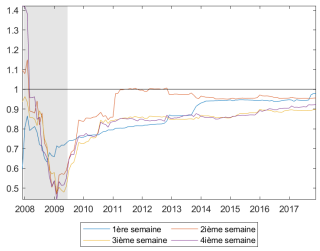


AWHI

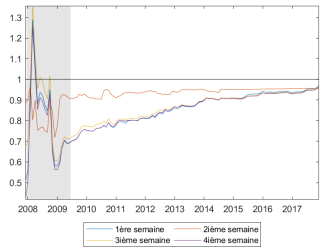
JTS1000HIR



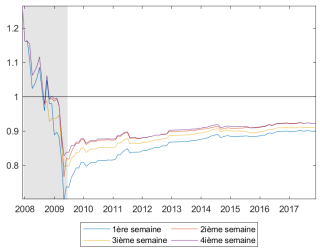
MANEMPL



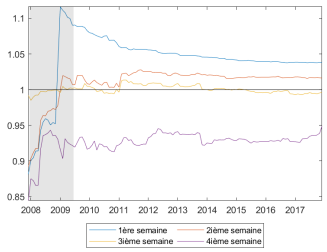
ICSA



PTI

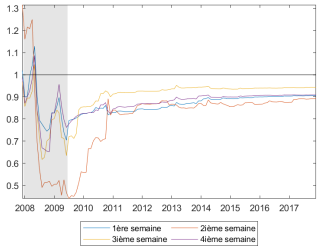


EJA

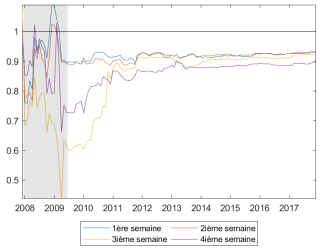


PNTF

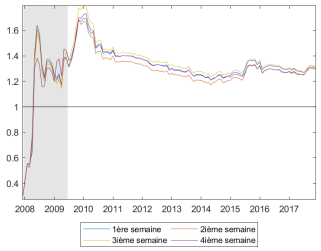
Figure B.3: Moyennes des EQM des meilleurs modèles pour chaque semaine (*nowcasting*)



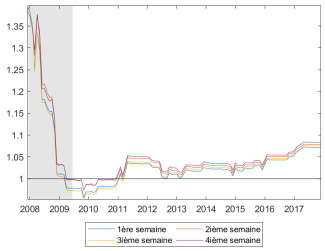
UNRATE



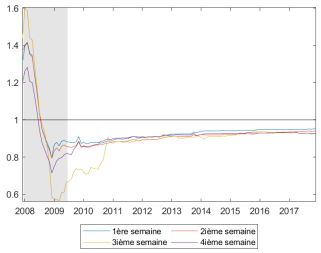
U6RATE



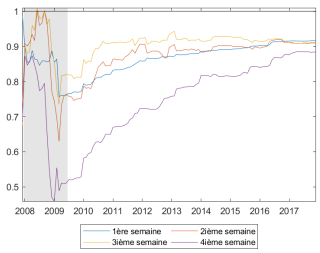
LNS17100000



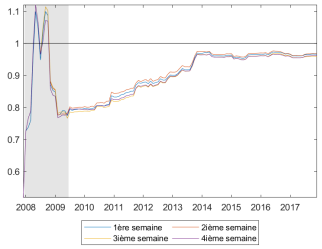
JTS1000QUR



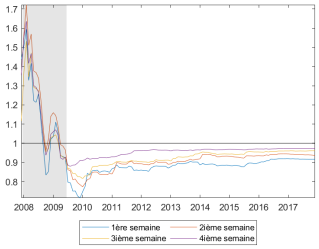
EMRATIO



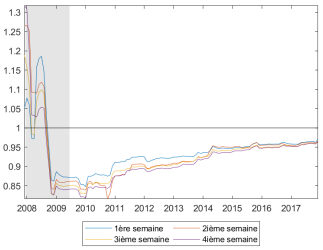
LNS12032194



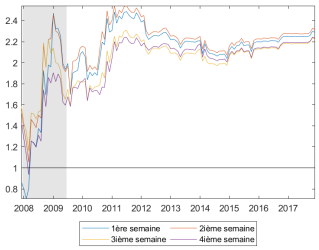
LNS13023706



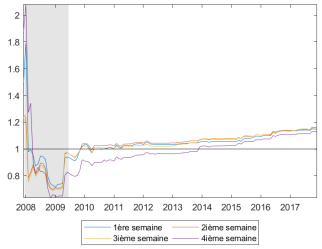
LNS13025703



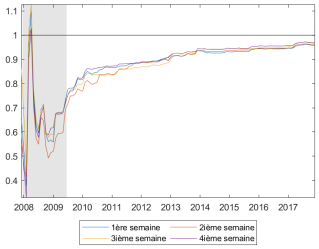
LNS13023622



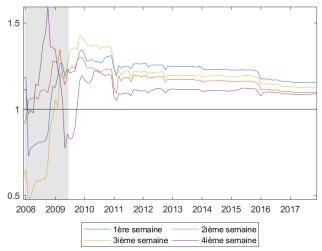
AHETPI



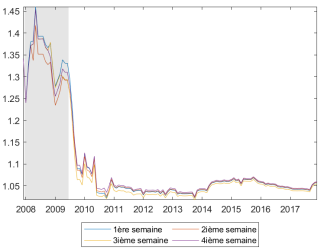
USPRIV



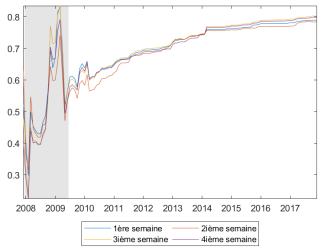
CES0500000035



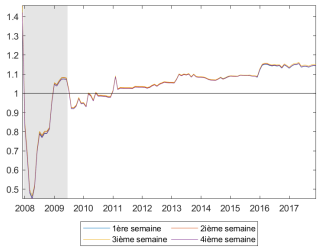
TEMPHELPS



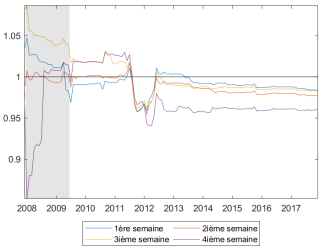
CIVPART



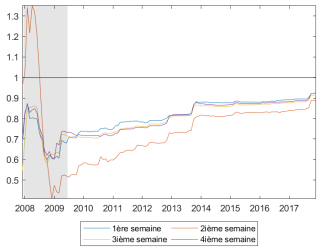
AWHI



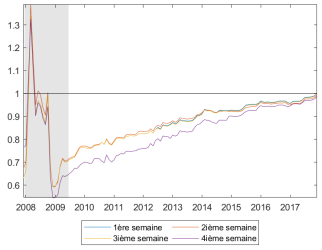
JTS1000HIR



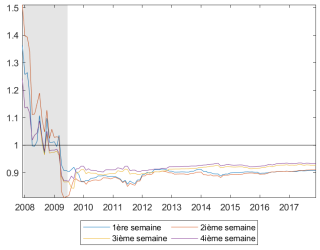
MANEMPL



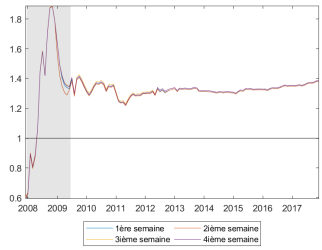
ICSA



PTI



EJA



PNTF

Figure B.4: Moyennes des EQM des meilleurs modèles pour chaque semaine (H = 1)

RÉFÉRENCES

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

Baker, S. et Fradkin, A. (2011). *What Drives Job Search ? Evidence from Google Search Data*. Discussion Papers 10-020, Stanford Institute for Economic Policy Research

Bleher, J. et Dimpfl, T. (2019). *Knitting Multi-Annual High-Frequency Google Trends to Predict Inflation and Consumption*, [Document non publié]. University of Tuebingen.

Campbell, J. Y. et Perron, P. (1991). Pitfalls and opportunities : what macroeconomists should know about unit roots. *NBER macroeconomics annual*, *6*, 141–201.

Chamberlin, G. (2010). Googling the present. *Economic & Labour Market Review*, *4*(12), 59–95.

Chauvet, M., Gabriel, S. et Lutz, C. (2016). Mortgage default risk : New evidence from internet search queries. *Journal of Urban Economics*, *96*(C), 91–111.

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

Choi, H. et Varian, H. (2012). Predicting the present with google trends. *Economic Record*, *88*(s1), 2–9.

Cleveland, R. B., Cleveland, W. S., McRae, J. E. et Terpenning, I. (1990). Stl : a seasonal-trend decomposition. *Journal of official statistics*, *6*(1), 3–73.

Combes, S. et Bortoli, C. (2015). *Contribution from Google Trends for forecasting the short-term economic outlook in France : limited avenues*, Insee – Conjuncture in France.

Croushore, D. (2005). Do consumer-confidence indexes help forecast consumer spending in real time ? *The North American Journal of Economics and Finance*, *16*(3), 435–450.

Da, Z., Engelberg, J. et Gao, P. (2011). In search of attention. *The Journal of Finance*, *66*(5), 1461–1499.

D'Amuri, F. (2009). *Predicting unemployment in short samples with internet job search query data*. MPRA Paper 18403, University Library of Munich, Germany

D'Amuri, F. et Marcucci, J. (2017). The predictive power of google searches in forecasting us unemployment. *International Journal of Forecasting*, *33*(4), 801–816.

Dickey, D. A. et Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, *74*(366), 427–431.

Diebold, F. et Mariano, R. (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics*, *13*(3), 253–63.

Foroni, C., Marcellino, M. et Schumacher, C. (2015). Unrestricted mixed data sampling (midas) : Midas regressions with unrestricted lag polynomials. *Journal of the Royal Statistical Society : Series A*, *178*(1), 57–82.

Ghysels, E., Santa-Clara, P. et Valkanov, R. (2004). The midas touch : Mixed data sampling regression models. *Centre interuniversitaire de recherche en analyse des organisations (CIRANO)*.

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

Guzman, G. (2011). Internet search behavior as an economic forecasting tool : The case of inflation expectations. *Journal of economic and social measurement*, *36*(3), 119–167.

Kholodilin, K. A., Podstawski, M. et Siliverstovs, B. (2010). *Do Google Searches Help in Nowcasting Private Consumption ? : A Real-Time Evidence for the US*. Discussion Papers of DIW Berlin 997, DIW Berlin, German Institute for Economic Research

Kulkarni, R., Haynes, K. E., Stough, R. R. et Paelinck, J. H. (2009). *Forecasting housing prices with Google econometrics* George Mason University School of Public Policy, Paper No.2009-10.

Ludvigson, S. C. (2004). Consumer confidence and consumer spending. *Journal of Economic Perspectives*, *18*(2), 29–50.

McLaren, N. et Shanbhogue, R. (2011). Using internet search data as economic indicators. *Bank of England Quarterly Bulletin*, *51*(2), 134–140.

Niesert, R. F., Oorschot, J. A., Veldhuisen, C. P., Brons, K. et Lange, R.-J. (2019). Can google search data help predict macroeconomic series ? *International Journal of Forecasting*.

Pesaran, M. H. et Timmermann, A. (1992). A simple nonparametric test of predictive performance. *Journal of Business & Economic Statistics*, *10*(4), 461–465.

Ross, A. (2013). Nowcasting with google trends : a keyword selection method. *Fraser of Allander Economic Commentary*, *37*(2), 54–64.

Seabold, S. et Coppola, A. (2015). *Nowcasting prices using Google trends : an application to Central America*. Policy Research Working Paper Series 7398, The World Bank

Stock, J. H. et Watson, M. W. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American statistical association*, *97*(460), 1167–1179.

Suhoy, T. (2009). *Query Indices and a 2008 Downturn : Israeli Data*. Discussion paper, Bank of Israel.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

Vosen, S. et Schmidt, T. (2009). Forecasting private consumption : survey-based indicators vs. google trends. *Journal of Forecasting*, *30*(6), 565–578.

Willis, J. L., Hakkio, C. S. *et al.* (2014). *Kansas City Fed's Labor Market Conditions Indicators (LMCI)*, Federal Reserve Bank of Kansas City.

Wu, L. et Brynjolfsson, E. (2015). *The Future of prediction : How Google searches foreshadow housing prices and sales.Economic Analysis of the Digital Economy*, (p. 89–118). Chicago : University of Chicago Press