UNIVERSITÉ DU QUÉBEC À MONTRÉAL

PRÉVISION DE L'ÉTAT ACTUEL DU PIB QUÉBÉCOIS

MÉMOIRE
PRÉSENTÉ
COMME EXIGENCE PARTIELLE
MAÎTRISE EN ÉCONOMIQUE

PAR
FRANCIS MAYER-CHARTRAND

AVRIL 2025

UNIVERSITÉ DU QUÉBEC À MONTRÉAL
Service des bibliothèques

Avertissement

La diffusion de ce mémoire se fait dans le respect des droits de son auteur, qui a signé le formulaire *Autorisation de reproduire et de diffuser un travail de recherche de cycles supérieurs* (SDU-522 – Rév.12-2023). Cette autorisation stipule que «conformément à l'article 11 du Règlement no 8 des études de cycles supérieurs, [l'auteur] concède à l'Université du Québec à Montréal une licence non exclusive d'utilisation et de publication de la totalité ou d'une partie importante de [son] travail de recherche pour des fins pédagogiques et non commerciales. Plus précisément, [l'auteur] autorise l'Université du Québec à Montréal à reproduire, diffuser, prêter, distribuer ou vendre des copies de [son] travail de recherche à des fins non commerciales sur quelque support que ce soit, y compris l'Internet. Cette licence et cette autorisation n'entraînent pas une renonciation de [la] part [de l'auteur] à [ses] droits moraux ni à [ses] droits de propriété intellectuelle. Sauf entente contraire, [l'auteur] conserve la liberté de diffuser et de commercialiser ou non ce travail dont [il] possède un exemplaire.»

Remerciements

Je tiens à exprimer ma profonde gratitude à mes professeurs Dalibor Stevanovic et Philippe Goulet Coulombe pour leur encadrement et leur soutien tout au long de mes études. Leur expertise et leurs conseils avisés ont été essentiels à la réalisation de ce travail.

Je souhaite également remercier ma copine Bianca Martella et ma fille Billie Chartrand pour leur soutien, leur patience et leurs encouragements constants, qui m'ont permis de surmonter les moments difficiles.

Enfin, je remercie chaleureusement ma famille et mes amis, qui m'ont toujours soutenu avec amour et compréhension tout au long de ce parcours. Leur présence et leur encouragement ont été une source de motivation inestimable.

Table des matières

Liste des tableaux

Table des figures

Résumé

Dans ce mémoire, nous explorons l'utilisation des données de Google Trends pour améliorer les prévisions en temps réel du produit intérieur brut (PIB) du Québec. L'objectif est de déterminer si l'intégration de ces données non traditionnelles dans des modèles économétriques permet d'accroître la précision des prévisions économiques. Nous appliquons plusieurs modèles de prévision, notamment MIDAS (Mixed Data Sampling), UMIDAS (Unrestricted MIDAS), Random Forest (RF), Gradient Boosting Machine (GBM), et LASSO, à des données de fréquences mixtes couvrant la période de 2015 à 2023. Ces modèles sont comparés à un modèle de référence autorégressif (AR(p)) pour évaluer leur performance avec et sans l'inclusion des données de Google Trends. Les résultats montrent que l'intégration des données de Google Trends améliore significativement la précision des prévisions en période de volatilité économique, comme pendant la pandémie de COVID-19. En revanche, ces données apportent moins de valeur ajoutée dans un contexte de stabilité économique, où les modèles traditionnels sans données additionnelles se révèlent tout aussi performants. Cette recherche met en lumière l'importance d'utiliser des sources de données alternatives pour enrichir les modèles de prévision économique, tout en soulignant les limites de ces données dans des environnements économiques stables. Elle contribue ainsi à une meilleure compréhension de l'efficacité des approches innovantes en prévision économique, offrant des perspectives pour les décideurs et les analystes sur l'utilisation de nouvelles sources de données pour affiner leurs modèles.

Mots-clés : Google Trends, prévisions économiques, Produit intérieur brut (PIB), Québec, MIDAS, UMIDAS, Random Forest, GBM, LASSO

Introduction

Dans un monde où les décisions économiques doivent être prises rapidement et avec précision, la capacité de prédire les indicateurs économiques en temps réel est devenue essentielle. La prévision actuelle (nowcasting) permet d'estimer des variables économiques, comme le produit intérieur brut (PIB), avant la publication officielle des données, en utilisant des informations disponibles à des fréquences plus élevées. Cette technique est particulièrement pertinente pour les décideurs politiques et les institutions financières qui doivent réagir rapidement aux évolutions économiques.

Le PIB est un indicateur clé de la performance économique, mais il est souvent publié avec plusieurs mois de décalage, rendant difficile la prise de décision fondée sur des données actualisées. Les banques centrales, par exemple, s'appuient sur des indicateurs économiques actuels pour formuler leurs politiques monétaires. Dans ce contexte, le nowcasting du PIB devient une tâche cruciale pour anticiper les tendances économiques et ajuster les politiques en conséquence. Une estimation précise et rapide du PIB peut également aider les entreprises à planifier leurs investissements et les consommateurs à prendre des décisions financières informées.

Les décisions de politique monétaire reposent sur une compréhension claire de la situation économique actuelle ou imminente, mais elles dépendent de bases de données souvent publiées avec retard. Par exemple, le PIB est publié trimestriellement, tandis que l'indice des prix à la consommation (IPC) est publié mensuellement, créant ainsi un défi pour la synchronisation des données.

Pour surmonter ces contraintes, le nowcasting et la prévision future (forecasting) de certaines variables économiques sont devenues des outils essentiels pour les banques centrales. Ces dernières se concentrent sur des indicateurs clés directement liés à la variable à prévoir. Par exemple, dans un cas où le PIB est la variable d'intérêt, les banques centrales vont souvent utiliser la production industrielle ou l'emploi comme indicateurs. Le principal défi pour les statisticiens est de ne pas perdre l'information contenue dans les données à haute fréquence lorsqu'elles sont agrégées à une fréquence trimestrielle.

Ce mémoire s'inscrit dans cette perspective et a pour objectif de prévoir le PIB québécois en temps réel. L'originalité de ce mémoire réside dans l'utilisation de séries à fréquence mixte incluant des indicateurs québécois, canadiens et certains indicateurs américains. De plus, une attention particulière est portée à l'utilisation des données hebdomadaires provenant de Google Trends, une base de données encore peu exploitée mais prometteuse dans la recherche économique.

Google Trends fournit des données sur le volume des recherches effectuées sur Google, offrant ainsi une mesure de l'intérêt du public pour divers sujets au fil du temps. En analysant les fluctuations de ces recherches, il est possible d'obtenir des indicateurs en temps réel de l'activité économique ou de la demande pour certains biens et services. L'utilisation de Google Trends est devenue plus courante depuis que Google a rendu publiques des données sur le volume relatif de requêtes en 2006.

Ce mémoire cherche à répondre à la question suivante : l'intégration des données Google Trends dans des modèles économétriques permet-elle d'améliorer la précision des prévisions actuelles du PIB québécois ? Nous évaluons notamment l'apport des modèles à fréquence mixte (MIDAS) introduits par Ghysels et al. (2004) ainsi que des modèles sans restriction (UMIDAS) introduits par Foroni et al. (2015). L'intérêt pour ces modèles réside dans leur capacité à exploiter les données de fréquence mixte sans nécessiter l'agrégation, préservant ainsi l'information cruciale des fréquences plus élevées.

En utilisant les données hebdomadaires de Google Trends, ce mémoire explore la corrélation entre les mouvements du PIB et l'attention portée par la population québécoise à ce sujet. Les premiers travaux, tels que ceux de Choi and Varian (2009) ou Askitas and Zimmermann (2009), montrent l'intérêt de ces données pour améliorer les prévisions économiques. Toutefois, il convient de noter que l'utilisation des données de Google Trends présente certains défis, notamment en termes de fiabilité et de représentativité, qui seront abordés dans une section ultérieure.

Ce mémoire est structuré comme suit : le chapitre 1 présente une revue de la littérature ; le chapitre 2 se consacre aux données utilisées ; le chapitre 3 développe la méthodologie ; enfin, le chapitre 4 présente les résultats obtenus.

Chapitre 1 : Revue de la littérature

Une prévision typique des séries chronologiques utilise des données provenant d'une même fréquence comme des données mensuelles ou trimestrielles. Cependant, de nombreux chercheurs ont montré que l'intégration de séries temporelles à fréquence plus élevée peut améliorer la précision des prévisions pour des séries à plus faible fréquence. En effet, les données à haute fréquence permettent de capter des informations importantes entre les différentes observations d'une série à faible fréquence, offrant ainsi une vue plus détaillée des dynamiques sous-jacentes.

Bien que la littérature sur le nowcasting soit encore relativement jeune, les travaux de Ghysels et al. (2004) et Giannone et al. (2008) sont parmi les études les plus citées dans ce domaine. Ils mettent en évidence l'efficacité de l'utilisation de données à fréquence mixte pour améliorer la précision des prévisions économiques.

La littérature utilise principalement quatre modèles pour ce type de prévision : les modèles à facteurs dynamiques (MFD), les modèles MIDAS, les VAR et les modèles bridge (LI). Chacun offre des avantages pour intégrer des données à plusieurs fréquences et améliorer les prévisions.

Cette étude utilise un modèle à facteurs dynamiques pour gérer la complexité des données, ainsi que des modèles MIDAS et UMIDAS pour exploiter les données à fréquence mixte. De plus, nous intégrons des modèles d'apprentissage automatique pour explorer leur potentiel dans ce contexte et un modèle autorégressif d'ordre p (AR(p)) sera utilisé comme modèle de référence pour évaluer la performance comparative de nos approches.

1.1 Modèles à facteurs dynamiques

Les modèles à facteurs dynamiques développés par Stock and Watson (2002) sont considérés parmi les outils les plus performants en macroéconomie pour la prévision de variables économiques à grande échelle. Cette approche est largement adoptée par les banques centrales et les institutions économiques en raison de leur capacité à synthétiser une vaste quantité de données macroéconomiques en un nombre restreint de facteurs sous-jacents, rendant ainsi l'analyse

et la prévision plus efficaces et gérables.

Stock and Watson (2002) montrent que ces modèles permettent d'améliorer la précision des prévisions en réduisant la dimensionnalité des données tout en conservant l'essentiel de l'information économique. Leur étude jette les bases des modèles factoriels dynamiques utilisés aujourd'hui par des institutions comme la Réserve fédérale ou la Banque centrale européenne. Cependant, bien que leur cadre théorique soit robuste, il repose sur des bases de données traditionnelles et ne prend pas en compte des sources alternatives, comme les tendances de recherche en ligne, qui peuvent contenir des signaux économiques anticipateurs.

Cette approche est largement adoptée dans la littérature économique, notamment par Giannone et al. (2008), qui démontrent que les MFD permettent d'améliorer significativement la précision des prévisions en intégrant un large éventail de variables macroéconomiques en temps réel. Leur étude démontre notamment que l'intégration de séries à haute fréquence améliore la réactivité des modèles aux chocs économiques, ce qui constitue un élément central de mon mémoire.

Cependant, bien que ces modèles soient performants pour traiter de grandes quantités de données traditionnelles, leur capacité à exploiter des données alternatives, comme les comportements de recherche en ligne via Google Trends, demeure une question peu explorée. Mon approche s'inspire directement de ces travaux en appliquant un modèle à facteurs dynamiques aux données Google Trends, tout en testant si l'ajout de ces nouvelles sources de données améliore la précision des prévisions.

Contrairement aux modèles MFD classiques qui s'appuient principalement sur des indicateurs macroéconomiques standard tels que le PIB, l'emploi ou l'inflation, cette étude évalue dans quelle mesure des données issues de Google Trends peuvent capturer des tendances économiques en temps réel, en complément des bases de données traditionnelles. Cette intégration permettrait potentiellement de détecter plus tôt des changements dans l'activité économique en captant les préoccupations des consommateurs et des entreprises avant la publication des indicateurs officiels.

Les MFD reposent sur l'hypothèse que les variations observées au sein d'un large ensemble de séries temporelles macroéconomiques s'expliquent par quelques facteurs non observés, appelés facteurs communs. Ces facteurs, qui capturent la majeure partie des dynamiques sous-jacentes, sont extraits à partir des données via des méthodes d'analyse en composantes principales (ACP). Cela permet de réduire la dimensionnalité du problème, rendant le modèle plus robuste et plus simple à estimer.

Cette approche se révèle particulièrement efficace dans le cadre des prévisions économiques

en temps réel. Par exemple, Giannone et al. (2008) appliquent ces modèles à la zone euro et montrent que l'intégration de données macroéconomiques à fréquence mixte améliore considérablement la capacité à anticiper les cycles économiques. Toutefois, leur travail repose uniquement sur des bases de données classiques, et ne considère pas des sources alternatives telles que les moteurs de recherche.

Dans certaines situations, comme celles étudiées par Giannone et al. (2008), les données utilisées pour la prévision ne sont pas toutes observées à la même fréquence. Par exemple, certaines variables économiques sont disponibles mensuellement, tandis que d'autres ne sont disponibles que trimestriellement. Cette disparité rend l'estimation des facteurs communs plus complexe, car elle nécessite de gérer les données manquantes de manière efficace.

Pour aborder cette complexité, ils utilisent un estimateur à deux étapes inspiré des travaux de Doz et al. (2011) et Giannone et al. (2005). La première étape consiste à extraire les composantes principales des données disponibles, même lorsqu'elles sont incomplètes, en appliquant des techniques avancées comme le filtre de Kalman. Ce dernier est un outil puissant pour traiter les séries temporelles avec des observations irrégulières ou manquantes.

Cependant, ces méthodes traditionnelles ne tiennent pas compte de sources de données externes, telles que les tendances de recherche sur Internet, qui peuvent contenir des informations utiles pour affiner les prévisions. Par exemple, durant la crise de la COVID-19, les volumes de recherche pour certains termes économiques (ex. : "assurance emploi", "chômage", "faillite") ont fortement fluctué avant la publication des statistiques officielles, ce qui démontre que ces signaux peuvent être exploités pour améliorer les prévisions économiques.

En intégrant ces données alternatives dans les MFD, mon étude vise à tester si ces nouvelles sources permettent d'affiner les facteurs latents sous-jacents aux dynamiques économiques québécoises. Contrairement aux méthodes d'estimation traditionnelles, l'intégration de Google Trends pourrait potentiellement améliorer la gestion des données manquantes en apportant une information additionnelle en temps réel, ce qui constitue une contribution originale de mon mémoire.

Une gestion efficace des données manquantes est cruciale dans un contexte de fréquence mixte, car elle repose sur l'estimation précise des composantes idiosyncratiques spécifiques à chaque série. Une bonne estimation de ces composantes est essentielle pour maintenir la qualité des prévisions et éviter les biais introduits par les données manquantes.

Traditionnellement, ces biais sont corrigés par des techniques de filtrage comme le lissage exponentiel, le filtre de Kalman ou ACP dynamique. Mon étude explore une autre piste : l'intégration de Google Trends comme source complémentaire de données, permettant potentielle-

ment de pallier certaines lacunes dans les séries économiques classiques.

Les MFD prouvent leur efficacité dans divers contextes économiques, notamment pour la prévision du PIB, de l'inflation et d'autres indicateurs macroéconomiques clés. Leur application dans des contextes de données à fréquence mixte, comme le montrent Giannone et al. (2008) et Doz et al. (2011), améliore de manière significative la précision des prévisions en temps réel.

Toutefois, ces études demeurent ancrées dans des cadres traditionnels où seules les bases de données macroéconomiques classiques sont utilisées. Mon travail innove en proposant d'examiner si les signaux tirés des recherches sur Google Trends peuvent être exploités dans un cadre MFD, permettant potentiellement d'améliorer la réactivité du modèle face aux évolutions rapides du marché.

1.2 Modèle MIDAS

Le concept de MIDAS est introduit par Ghysels et al. (2004) pour résoudre un problème fondamental en macroéconomie : l'intégration de données à différentes fréquences dans un modèle de prévision unique. En effet, dans la plupart des analyses économiques, il est courant de disposer de variables explicatives disponibles à des fréquences supérieures (par exemple, des données mensuelles ou hebdomadaires) par rapport à la variable dépendante souvent disponible à une fréquence inférieure (comme les données trimestrielles ou annuelles). Les modèles MIDAS permettent de combiner ces différentes fréquences de manière parcimonieuse et efficace.

Le modèle MIDAS se distingue par sa capacité à intégrer des variables explicatives à haute fréquence dans un modèle à basse fréquence sans augmenter de manière exponentielle le nombre de paramètres à estimer. Cela est rendu possible par l'utilisation d'un modèle à retards distribués pondérés où les poids des différentes observations à haute fréquence sont estimés par un vecteur de paramètres. Ce vecteur de poids permet de capturer l'effet des variables explicatives tout en évitant la prolifération des paramètres qui pourrait sinon rendre le modèle impraticable.

L'un des principaux défis associés à l'utilisation des modèles MIDAS est la détermination du nombre optimal de retards à inclure dans le modèle. Lorsque les variables explicatives sont à fréquence mensuelle et que la variable dépendante est à fréquence trimestrielle, il devient nécessaire d'inclure plusieurs retards pour capturer toute l'information pertinente. Cependant, cela peut entraîner une prolifération des paramètres à estimer, rendant le modèle complexe et potentiellement instable.

Pour résoudre ce problème, Ghysels et al. (2007) proposent l'utilisation du retard exponentiel d'Almon, une forme fonctionnelle qui permet de réduire le nombre de paramètres tout en

conservant une grande flexibilité. Cette méthode ajuste automatiquement les poids des observations à haute fréquence, offrant ainsi une estimation plus stable et efficace du modèle MIDAS.

En plus du modèle MIDAS standard, Foroni et al. (2015) développent également le modèle UMIDAS (Unrestricted MIDAS), qui élimine certaines des restrictions imposées par les formes fonctionnelles de poids dans le modèle MIDAS classique. Le modèle UMIDAS, en ne contraignant pas les coefficients à suivre une forme prédéterminée, offre plus de flexibilité et peut parfois mieux capturer les relations entre les variables à haute fréquence et la variable dépendante.

Kuzin et al. (2011) comparent les performances des modèles MIDAS et VAR (Vecteur Autorégressif) dans le contexte de nowcasting du PIB européen en utilisant des indicateurs mensuels. Leur étude a montré que les deux approches sont plus complémentaires que substitutives. Le modèle MIDAS excelle dans les prévisions à court terme (moins de quatre mois), où il est capable de capter rapidement les fluctuations à haute fréquence, tandis que le modèle VAR se montre plus performant pour les horizons de prévision plus longs (jusqu'à neuf mois), où les dynamiques à basse fréquence prédominent.

Dans ce mémoire, l'horizon de prévision est encore plus court (13 semaines), ce qui justifie pleinement l'utilisation de MIDAS, conçu pour exploiter l'information contenue dans les données à haute fréquence. L'enjeu est donc d'évaluer si l'ajout de Google Trends renforce cette capacité à capter des dynamiques économiques à très court terme ou si leur apport reste limité.

Depuis cette étude, la littérature propose plusieurs améliorations des modèles MIDAS. Par exemple, Foroni et al. (2019) présentent le modèle ARMA-MIDAS, qui associe MIDAS à une composante de moyenne mobile (ARMA). Cette intégration vise à capter les dynamiques récurrentes à court terme qui peuvent être négligées par les approches traditionnelles. Dans une série d'expériences à la Monte Carlo, ils montrent que l'ajout de la composante moyenne mobile améliore significativement la précision des prévisions à court terme et que ces gains sont particulièrement persistants au fil du temps.

Les modèles MIDAS, par leur capacité à intégrer des données à fréquence mixte de manière parcimonieuse et efficace, se sont révélés être des outils précieux pour la prévision économique en temps réel. Leurs applications vont au-delà de la simple prévision du PIB, s'étendant à d'autres domaines comme la finance où la disponibilité de données à haute fréquence est courante. En combinant ces modèles avec des techniques plus récentes comme l'apprentissage automatique ou l'intégration de nouvelles sources de données comme Google Trends, les chercheurs peuvent continuer à affiner et à améliorer les prévisions économiques, en particulier dans des contextes où la rapidité et la précision sont essentielles.

1.3 Modèle UMIDAS

Le modèle UMIDAS ou Unrestricted MIDAS est introduit par Foroni et al. (2015) comme une extension du modèle MIDAS classique. Alors que le modèle MIDAS impose certaines restrictions sur la forme des coefficients associés aux variables explicatives à haute fréquence, le modèle UMIDAS se caractérise par une plus grande flexibilité permettant une estimation moins contrainte des relations entre les variables de différentes fréquences.

Dans un modèle MIDAS traditionnel, la relation entre les variables dépendantes à basse fréquence et les variables explicatives à haute fréquence est souvent modélisée à l'aide d'un polynôme de poids comme le polynôme exponentiel d'Almon. Ce polynôme permet de réduire le nombre de paramètres à estimer en imposant une structure spécifique sur les coefficients. Bien que cette approche soit efficace pour contrôler la prolifération des paramètres, elle peut parfois limiter la capacité du modèle à capturer pleinement la dynamique complexe des données.

Le modèle UMIDAS, en revanche, élimine cette contrainte en ne restreignant pas les coefficients des variables explicatives. Autrement dit, le vecteur de poids appliqué aux retards des variables explicatives est fixé à 1 pour tous les retards, ce qui signifie que chaque observation à haute fréquence est traitée sans pondération prédéfinie. Cette approche non restreinte permet au modèle d'être plus flexible, ce qui peut être particulièrement bénéfique lorsque les relations entre les variables de différentes fréquences ne suivent pas une forme fonctionnelle simple.

L'un des principaux avantages du modèle UMIDAS est sa simplicité. En éliminant les restrictions sur les coefficients, le modèle peut être estimé à l'aide de méthodes standard comme les moindres carrés ordinaires (OLS). Cette simplicité permet de capturer de manière plus directe les relations entre les variables de différentes fréquences, ce qui peut améliorer les performances de prévision dans certains contextes.

Cependant, cette flexibilité accrue a un coût. En augmentant le nombre de paramètres à estimer, le modèle UMIDAS peut souffrir d'une variance plus élevée des estimations, ce qui peut conduire à des prévisions moins précises, surtout lorsque l'échantillon de données est limité. Cette augmentation de la variance est une manifestation du compromis classique entre biais et variance dans les modèles statistiques : bien que le modèle soit moins biaisé en raison de l'absence de restrictions, il est également plus susceptible de s'ajuster de manière excessive aux particularités de l'échantillon de données.

Foroni et al. (2015) testent les performances du modèle UMIDAS à travers une série de simulations Monte-Carlo. Leurs résultats montrent que dans des contextes où l'écart de fréquence entre les variables explicatives et la variable dépendante est faible (par exemple, entre

des données mensuelles et trimestrielles), le modèle UMIDAS peut surpasser le modèle MIDAS traditionnel. Cela s'explique par la capacité du modèle UMIDAS à capturer des dynamiques complexes sans être limité par la forme des coefficients.

Cependant, dans des contextes où l'écart de fréquence est plus large ou lorsque la relation entre les variables est plus linéaire et bien décrite par une forme fonctionnelle simple, le modèle MIDAS avec restrictions peut offrir de meilleures performances en termes de précision de prévision. En pratique, le choix entre UMIDAS et MIDAS dépendra de la nature des données disponibles et des objectifs spécifiques de la prévision.

Le modèle UMIDAS trouve des applications dans divers domaines économiques et financiers où les données à haute fréquence sont disponibles et où une grande flexibilité dans la modélisation est nécessaire. Il est particulièrement utile dans les situations où les relations entre les variables ne peuvent pas être facilement capturées par des modèles paramétriques simples.

Dans le cadre de ce mémoire, cette distinction entre MIDAS et UMIDAS est particulièrement intéressante lorsque l'on travaille avec des données issues de Google Trends. Les tendances de recherche en ligne sont souvent marquées par une forte volatilité, des pics soudains et des évolutions non linéaires, ce qui remet en question l'hypothèse d'une relation régulière entre les fréquences.

Ainsi, il est essentiel d'examiner si un modèle plus souple comme UMIDAS est mieux adapté pour capturer ces signaux à haute fréquence, ou si MIDAS, en imposant une structure plus contrainte, permet d'éviter les fluctuations excessives et d'améliorer la stabilité des prévisions.

Plutôt que d'adopter une approche exclusive, mon étude explore l'intégration des deux modèles afin d'évaluer leur efficacité respective sur les données de Google Trends. Cette démarche permet non seulement de comparer leurs performances en termes de prévision économique à court terme, mais aussi d'identifier dans quelles conditions chaque modèle est le plus pertinent pour exploiter l'information issue des tendances de recherche en ligne.

1.4 Apprentissage automatique

L'apprentissage automatique ou Machine Learning (ML) devient un domaine d'intérêt majeur dans la littérature récente en sciences économiques. Bien que l'intégration du ML en macroéconomie soit relativement nouvelle, ses origines remontent à plusieurs décennies. Par exemple, Lee et al. (1993) sont parmis les premiers à explorer l'application des réseaux de neurones pour tester la non-linéarité dans les séries temporelles économiques. Leur étude compare

la performance des réseaux de neurones avec des tests statistiques traditionnels tels que les tests de Keenan, Tsay, White, et Mcleod-Li, montrant que les réseaux de neurones pouvaient capturer des relations non linéaires complexes mieux que les approches conventionnelles.

De même, Stock and Watson (2002) introduisent l'analyse des facteurs communs dynamiques, une méthode qui est considérée comme une forme de ML non supervisé. Cette approche utilise l'analyse en composantes principales pour réduire la dimensionnalité des données et extraire des facteurs sous-jacents, un processus similaire à ce que font certaines techniques modernes de ML.

Ces dernières années, l'application de l'apprentissage automatique dans la prévision macroéconomique a connu une croissance exponentielle. Cette tendance est en grande partie due à l'augmentation massive de la disponibilité des données ainsi qu'à l'amélioration des algorithmes et de la puissance de calcul. Les méthodes de ML traitent aujourd'hui de vastes volumes de données et détectent des relations complexes souvent négligées par les modèles traditionnels.

L'apprentissage automatique se distingue par sa capacité à apprendre directement des données sans nécessiter de spécifications paramétriques rigides. Cela permet aux économistes de créer des modèles plus flexibles qui peuvent s'adapter aux particularités des données sans les contraintes des hypothèses traditionnelles. En effet, Goulet Coulombe et al. (2022) montrent que l'utilisation de techniques de ML pour la prévision macroéconomique permet de capturer les non-linéarités complexes présentes dans les données économiques, et que ces techniques offrent des gains significatifs en termes de précision de prévision par rapport aux méthodes économétriques traditionnelles. Ce résultat est particulièrement pertinent pour mon étude, car il suggère que les algorithmes de ML pourraient être mieux adaptés que les modèles classiques pour exploiter les signaux issus des tendances de recherche Google, contrairement à leur étude qui applique le ML sur des données macroéconomiques classiques.

De plus, Goulet Coulombe et al. (2021b) démontrent que les techniques de ML peuvent s'adapter rapidement aux chocs économiques, comme celui de la pandémie de COVID-19, en détectant les changements brusques dans les données économiques. Cette capacité d'adaptation est essentielle dans le cadre de ce mémoire, car Google Trends peut capturer des changements de comportement en temps réel, ce qui pourrait rendre les modèles de ML plus réactifs face aux fluctuations économiques à court terme.

Dans le domaine de la prévision avec des données à fréquence mixte, Borup et al. (2021) explorent la combinaison du modèle U-MIDAS avec des algorithmes de ML pour prédire les réclamations d'assurance chômage aux États-Unis à une fréquence hebdomadaire en utilisant des données journalières provenant de Google Trends. Les algorithmes de ML utilisés dans leur

étude incluent le LASSO (Least Absolute Shrinkage and Selection Operator), Elastic Net, et un réseau de neurones artificiel avec trois couches cachées ("hidden layers").

LASSO et Elastic Net sont des techniques de régularisation qui permettent de sélectionner les variables les plus pertinentes parmi un grand nombre de prédicteurs, réduisant ainsi le risque de surapprentissage (overfitting).

Le réseau de neurones artificiel, quant à lui, est capable de capturer des relations non linéaires complexes grâce à ses multiples couches de traitement.

Les résultats de leur étude montrent une amélioration significative de la précision des prévisions par rapport aux modèles traditionnels. En particulier, ils observent une réduction de jusqu'à 63% de la racine des erreurs moyennes quadratiques (RMSE) par rapport au modèle autorégressif d'ordre p (AR(p)) servant de référence.

Leur étude est directement liée à mon mémoire, car elle valide l'idée que Google Trends peut être une source d'information utile dans un cadre de prévision à fréquence mixte.

Une autre étude menée par Lahiri and Yang (2022) applique des techniques de ML pour prédire les revenus fiscaux de l'État de New York en utilisant des données à fréquence mixte. Ils explorent plusieurs algorithmes de ML, y compris des modèles de Gradient Boosting Machine (GBM) avec fréquence mixte, des modèles de GBM avec des facteurs communs dynamiques, ainsi que des variantes du LASSO comme le space group LASSO (sg-LASSO) développé par Simon et al. (2013).

GBM est une méthode d'ensemble qui combine les prédictions de plusieurs modèles faibles pour former un modèle robuste, souvent plus performant qu'un seul modèle.

Le sg-LASSO est une version améliorée du LASSO qui permet de capturer les interactions entre groupes de variables, ce qui est particulièrement utile dans les contextes de données à haute dimension.

Dans leur analyse, Lahiri and Yang (2022) comparent ces modèles avec le modèle ADL-MIDAS traditionnel. Ils trouvent que le modèle de GBM combiné avec deux facteurs communs offrait les meilleures performances en termes de précision de prévision, dépassant les autres modèles étudiés.

L'intégration de l'apprentissage automatique dans la prévision macroéconomique ouvre de nouvelles perspectives pour la modélisation des dynamiques économiques complexes. Les techniques de ML permettent de tirer parti des grandes quantités de données disponibles aujourd'hui, y compris les données non traditionnelles comme celles provenant des moteurs de recherche ou des réseaux sociaux, pour améliorer la précision des prévisions économiques.

Cependant, il est important de noter que la puissance des modèles de ML s'accompagne

également de défis. En particulier, les modèles de ML nécessitent une grande quantité de données pour être efficaces et peuvent être sujets à des problèmes de overfitting s'ils ne sont pas correctement régularisés. De plus, l'interprétabilité des modèles de ML reste un défi majeur, car ces modèles fonctionnent souvent comme des "boîtes noires" difficiles à comprendre pour les utilisateurs finaux.

Malgré ces défis, l'avenir de l'apprentissage automatique en macroéconomie est prometteur. À mesure que les algorithmes deviennent plus sophistiqués et que les techniques pour gérer les données à fréquence mixte s'améliorent, il est probable que l'apprentissage automatique jouera un rôle de plus en plus central dans la prévision économique, fournissant des outils puissants pour aider les décideurs politiques et les analystes économiques à naviguer dans un environnement économique de plus en plus complexe.

1.5 Utilisation de Google Trends dans la prévision macroéconomique

L'intégration des données issues des moteurs de recherche, et plus particulièrement de Google Trends, représente une avancée récente mais significative dans le domaine de la prévision macroéconomique. Google Trends, lancé en 2006, permet de suivre les volumes de recherches effectuées sur Google pour des termes spécifiques, fournissant ainsi des informations en temps réel sur les intérêts et préoccupations des utilisateurs à travers le monde. Cette nouvelle source de données s'est avérée précieuse pour les économistes en leur offrant une manière innovante de capter les tendances économiques avant que les données officielles ne soient publiées.

Choi and Varian (2009), ainsi qu'Askitas and Zimmermann (2009), sont parmis les premiers économistes à montrer l'utilité des données de Google Trends dans les analyses économiques. Leur travail montre que ces données peuvent être utilisées pour prévoir des variables économiques importantes telles que l'évolution de l'emploi, les ventes au détail ou encore la demande immobilière. Ces premières études établient une base solide pour l'utilisation de Google Trends dans la prévision macroéconomique, en prouvant que les recherches en ligne pouvaient refléter les comportements économiques réels.

L'étude de Ferrara and Simoni (2023), menée dans le contexte européen, analyse l'efficacité des données Google Trends pour la prévision du PIB. Ils concluent que ces données ajoutent de la valeur aux prévisions, mais principalement dans les premières semaines suivant la collecte des données. Cette observation est due au fait que les données de Google Trends capturent des

signaux anticipés avant que les données économiques officielles ne soient disponibles. Cependant, une fois que les données officielles sont publiées, le pouvoir prédictif des données Google Trends tend à diminuer. Ferrara et Simoni notent ainsi que l'utilité des données Google Trends est maximale lorsque les informations économiques officielles font défaut.

Götz and Knetsch (2019) approfondit cette analyse en se concentrant sur l'Allemagne. Leur recherche montre que l'intégration des données de Google Trends dans les modèles de prévision pouvait améliorer la précision des prédictions du PIB allemand. Utilisant des modèles de « Leading Indicators » (LI), ils montrent que le choix des mots-clés dans Google Trends variait selon la variable économique d'intérêt, ce qui souligne l'importance d'une sélection judicieuse des termes de recherche pour maximiser la pertinence des prédictions.

Bantis et al. (2021) étendent cette recherche en examinant la valeur ajoutée des données Google Trends dans les prévisions du PIB américain et brésilien, en tenant compte d'un contexte de données massives. Leur étude utilise un modèle à facteurs dynamiques, similaire à celui proposé par Giannone et al. (2008), pour résoudre les problèmes de dimensionnalité souvent rencontrés dans les prévisions économiques. Le modèle à facteurs dynamiques est particulièrement efficace pour extraire les informations pertinentes d'un grand ensemble de données, ce qui en fait un outil puissant pour les prévisions en temps réel. Leurs résultats montrent que les modèles à facteurs dynamiques utilisant les données de Google Trends offraient une précision de prévision supérieure par rapport aux modèles « bridging » utilisés par Götz and Knetsch (2019), ainsi que Ferrara and Simoni (2023). Les auteurs notent également que l'efficacité des données de Google Trends variait selon le contexte économique. Par exemple, les données de Google Trends se révèlent plus utiles pour les prévisions du PIB américain que pour le Brésil, ce qui peut s'expliquer par les différences dans l'accès à Internet et dans les habitudes de recherche en ligne entre ces deux pays.

Dans le contexte canadien, Couture and Stevanovic (2021) analysent l'utilisation des données de Google Trends pour prédire les variables du marché de l'emploi au Canada et au Québec. Leur étude montre que l'intégration des données de requêtes hebdomadaires de Google dans des modèles à fréquence mixte améliorait la précision des prévisions du taux d'emploi, des heures travaillées et du taux de chômage, particulièrement au début du mois, avant la disponibilité des données de l'Enquête sur la population active. Ils soulignent que la haute fréquence des données de Google Trends est cruciale pour la précision des prévisions dans ces contextes.

De plus, Couture (2020) montre l'utilité des données de Google Trends pour prévoir l'activité du marché du travail aux États-Unis. En utilisant des modèles économétriques tels que MIDAS, U-MIDAS et LASSO, celui-ci a montré que les données intramensuelles de Google

Trends améliorent la prévision des mouvements de marché, notamment en permettant des mises à jour en temps réel lorsque de nouvelles données deviennent disponibles.

Bien que les données de Google Trends montrent leur valeur dans la prévision macroéconomique, elles ne sont pas sans limites. Leur utilité dépend fortement du contexte et de la sélection des mots-clés. De plus, leur pouvoir prédictif tend à diminuer lorsque des données économiques officielles deviennent disponibles. Cependant, avec l'évolution continue des méthodes d'analyse des données et l'amélioration des modèles de prévision, l'intégration de sources de données non traditionnelles comme Google Trends est susceptible de devenir de plus en plus importante.

Les études futures pourraient se concentrer sur l'amélioration de la méthodologie pour sélectionner les mots-clés les plus pertinents ou sur la combinaison des données Google Trends avec d'autres sources de données en temps réel pour améliorer la robustesse des prévisions économiques. En somme, bien que Google Trends ne puisse pas remplacer les données économiques traditionnelles, il s'agit d'un outil complémentaire précieux pour capter les tendances économiques émergentes et pour affiner les prévisions à court terme.

L'ensemble de ces études confirme que les données issues de Google Trends possèdent un potentiel prédictif intéressant en macroéconomie, notamment pour anticiper des variables telles que l'emploi, le PIB et la consommation. Toutefois, la majorité des recherches existantes se concentrent sur des prévisions à grande échelle (États-Unis, Europe, Brésil, Allemagne) et reposent principalement sur des méthodologies économétriques traditionnelles. Cette étude se distingue en appliquant ces concepts à l'économie québécoise et en explorant un horizon de prévision plus court de 13 semaines, ce qui correspond à un besoin accru de réactivité en prévision économique.

De plus, plutôt que de se limiter à une seule approche, cette recherche mobilise plusieurs méthodologies avancées, notamment MIDAS, UMIDAS et des techniques d'apprentissage automatique, afin d'évaluer dans quelles conditions Google Trends apporte la plus grande valeur ajoutée. Alors que certaines études suggèrent que l'efficacité de ces données diminue à mesure que des indicateurs officiels deviennent disponibles, cette étude cherche à déterminer si ces signaux conservent leur pertinence sur un horizon de prévision court et quelles techniques permettent d'en maximiser l'apport. Cette approche contribue ainsi à enrichir la compréhension de l'intégration des données de recherche en ligne dans un cadre économétrique et propose une évaluation plus nuancée de leur rôle dans la prévision macroéconomique.

Chapitre 2 : Données

Les données employées pour cette étude proviennent d'un ensemble diversifié de sources macroéconomiques et financières, englobant des informations du Québec, du Canada, ainsi que des États-Unis. En outre, des données obtenues à partir du moteur de recherche Google ont été intégrées pour compléter l'analyse.

2.1 Données québécoises

Les données québécoises constituent une partie essentielle de cette étude, incluant un large éventail de variables macroéconomiques fournies par Fortin-Gagnon et al. (2022). L'objectif principal est d'analyser le PIB réel québécois, qui est disponible à une fréquence trimestrielle. Pour compléter cette analyse, un ensemble de 11 variables macroéconomiques mensuelles a été inclus, couvrant divers aspects de l'économie québécoise tels que le taux de chômage et l'inflation par exemple.

Il est important de noter que ces séries mensuelles débutent à des périodes différentes, reflétant la disponibilité des données au fil du temps. Pour assurer la comparabilité et l'utilisabilité de ces séries, toutes les données ont été désaisonnalisées lors de leur collecte. De plus, elles ont été rendues stationnaires à l'aide de techniques de transformation qui seront détaillées dans les sections méthodologiques ultérieures de ce mémoire. Cette approche permet d'éliminer les effets saisonniers et de stabiliser les moyennes et variances des séries, conditions essentielles pour une modélisation statistique robuste.

2.2 Données canadiennes

Les données canadiennes utilisées dans cette étude comprennent une sélection de variables macroéconomiques nationales ainsi que des indicateurs financiers clés issus du Toronto Stock Exchange. Au total, 15 variables mensuelles ont été retenues, dont 14 macroéconomiques et une

financière, offrant une vue d'ensemble des dynamiques économiques et financières au Canada. Parmi les données trimestrielles, le PIB canadien joue un rôle crucial, étant disponible un mois avant la publication du PIB québécois pour le même trimestre. Cette différence de calendrier ne signifie toutefois pas que les données provinciales sont immédiatement accessibles à Statistique Canada au moment de la publication du PIB national. En réalité, le PIB canadien repose sur des estimations avancées et des modèles d'agrégation, alors que les données provinciales sont publiées ultérieurement après des processus de validation plus approfondis. Ce décalage méthodologique explique pourquoi, bien que le PIB canadien puisse fournir une première indication des tendances économiques, les données détaillées du PIB québécois ne deviennent disponibles qu'après un temps supplémentaire de collecte et de traitement.

Cela permet néanmoins d'enrichir l'analyse en offrant une perspective nationale pouvant influencer les prévisions économiques au niveau provincial. À l'instar des données québécoises, les séries canadiennes ont été désaisonnalisées pour éliminer les variations saisonnières et ont ensuite été transformées pour les rendre stationnaires. Ce processus est essentiel pour garantir que les séries temporelles soient adaptées aux modèles économétriques utilisés dans cette étude, minimisant ainsi les risques de biais dans les prévisions.

2.3 Données américaines

En ce qui concerne les données américaines, il n'a pas été jugé nécessaire d'intégrer l'ensemble complet de la base de données FRED_MD développée par McCracken and Ng (2016) pour cette étude. Cependant, la sélection de certaines séries macroéconomiques américaines clés s'est avérée pertinente pour compléter l'analyse. Pour ce faire, nous nous sommes inspirés du travail de Goulet Coulombe et al. (2021b), qui ont examiné l'efficacité des modèles d'apprentissage automatique dans la capture des effets de la pandémie de COVID-19 au Royaume-Uni. Leur approche, basée sur les travaux de McCracken and Ng (2016) ainsi que de Fortin-Gagnon et al. (2022), a guidé notre sélection.

Dans le cadre de cette étude, 18 séries macroéconomiques américaines ont été sélectionnées. Ces séries comprennent des indicateurs tels que les taux d'intérêt de la Réserve fédérale, les indices boursiers, les taux de chômage, et d'autres variables cruciales pour comprendre l'impact des dynamiques économiques américaines sur le Canada et, par extension, le Québec. Comme pour les autres ensembles de données, les séries américaines ont été désaisonnalisées et rendues stationnaires pour assurer leur compatibilité avec les modèles économétriques employés dans cette étude.

2.4 Données Google Trends

L'utilisation des données issues de Google Trends est une approche relativement récente dans la littérature économique. En raison de l'accessibilité rapide de ces données et du fait qu'elles ne sont pas soumises à des révisions, elles ont suscité un grand intérêt non seulement parmi les économistes, mais aussi dans divers autres domaines. Google Trends permet d'accéder gratuitement à des données directement extraites du moteur de recherche Google, offrant ainsi une précieuse opportunité de capter les intérêts d'une population cible sur divers sujets. Ces données sont particulièrement riches car elles donnent accès à des informations difficiles à mesurer par des moyens traditionnels en temps réel. Dans le cadre de cette étude, l'accent est mis sur la croissance économique. Les paragraphes suivants détailleront les caractéristiques des données Google Trends et les choix méthodologiques effectués pour leur utilisation.

Google Trends fournit des données disponibles à trois fréquences différentes : mensuelle, hebdomadaire et journalière. Les données mensuelles sont accessibles depuis le 1er janvier 2004 jusqu'à aujourd'hui. Pour les données hebdomadaires, elles ne sont disponibles que pour une période glissante de cinq ans. Par exemple, un utilisateur pourrait obtenir des données allant du 1er janvier 2015 au 1er janvier 2019, mais si la période d'intérêt excède cinq ans, les données seront alors fournies à une fréquence mensuelle. Enfin, les données journalières ne sont disponibles que pour une période de huit mois consécutifs. Par exemple, un extrait de données journalières pourrait couvrir la période du 1er janvier 2023 au 1er août 2023, mais au-delà de cette période, les données seraient agrégées à une fréquence hebdomadaire.

Google distingue deux types de données dans son moteur de recherche : les données en temps réel et les données en temps non réel. Les données en temps réel sont un échantillon de recherches effectuées au cours des sept derniers jours, tandis que les données en temps non réel couvrent la période allant du 1er janvier 2004 à aujourd'hui. Dans le cadre de cette étude, seules les données en temps non réel seront utilisées, car elles offrent une couverture temporelle plus longue et sont mieux adaptées à l'analyse des tendances économiques à long terme.

Les données Google Trends se présentent sous la forme d'un indice de popularité, qui reflète l'intérêt relatif pour un sujet donné dans une région démographique spécifique au cours d'une période définie. Il est essentiel de comprendre que cet indice ne représente pas le volume absolu de recherches, mais plutôt une mesure relative de la popularité d'un sujet par rapport à toutes les autres recherches effectuées dans la même région. L'indice est construit en divisant le nombre de recherches d'un sujet particulier par le nombre total de recherches dans la même région et au cours de la même période. Cette méthode permet de comparer la popularité relative de sujets

dans différentes régions sans être biaisé par les différences de population ou de volume global de recherches.

Les résultats obtenus sont ensuite normalisés sur une échelle de 0 à 100. Une valeur de 0 indique qu'il n'y avait pas suffisamment d'informations sur la requête pour générer un indice significatif, tandis qu'une valeur de 100 représente le sommet de popularité du sujet pour la période et la région sélectionnées. Il est important de noter que deux régions avec le même indice ne reflètent pas nécessairement le même nombre total de recherches, car l'indice est relatif à chaque région. De plus, la valeur de l'indice peut varier selon le moment et l'endroit de l'extraction des données, ce qui ajoute une dimension supplémentaire de complexité à l'analyse.

Google classe les requêtes de recherche en 25 catégories principales, chacune subdivisée en sous-catégories, pour un total de 272 sous-catégories combinées. Certaines de ces sous-catégories contiennent elles-mêmes des sous-divisions, ce qui porte le nombre total de catégories à plus de 1400. Cette classification permet d'organiser et de filtrer les données de manière à capter les tendances spécifiques à différents secteurs d'activité ou sujets d'intérêt. Google filtre également les requêtes pour éliminer les recherches répétées provenant du même utilisateur sur une courte période et les requêtes contenant des caractères spéciaux, telles que des apostrophes, afin d'assurer la qualité et la pertinence des données collectées.

TABLEAU 2.1: Catégories et sous-catégories des données Google Trends

Autos & Vehicles	Photography	Video Games
World Localities	Jobs & Education	Social Networks
Reference	TV & Video	Vehicles
Web Services	Beauty & Fitness	Health
Computers & Electronics	Law & Government	Arts & Entertainment
News	Finance	Real Estate
Food & Drink	Science	Cooking & Recipes
Software	Travel	Financial Planning & Management
Music & Audio	Movies	Anime & Manga
Personal Aircraft	Bicycles & Accessories	Boating
Craft Supplies	Guns & Firearms	Hunting & Shooting
Paintball	Home Appliances	Online Games

Toys & Games	Custom & Performance Vehicles	Office Supplies
Camper & RV	Parenting & Family	DIY & Home Improvement
Automotive Industry	Event Planning	Military
Gun Safety	Nursing	Industrial Goods & Services
Oral & Dental Care	International News	Skincare
Medical Devices & Equipment	Women's Interests	Banking
Women's Health	Biotechnology	Yoga
Business Education	Cosmetics	Business News
Consumer Protection	Cars	Credit Cards
Consulting	Delivery Services	Construction
Diet & Fitness	Consumer Electronics	Education
Credit & Lending	Gardening	E-commerce
Genealogy	Fashion	Health News
Food	Magazines	Food & Grocery Retailers
Real Estate Listings	Home Furnishings	Retailers
Insurance	Search Engines	Jewelry
Sporting Goods	Medical Facilities	TV News
Mobile Devices	TV Shows	Mortgage
Vacation	Nutrition	Veterinarians
Parenting	Weather	Pets & Pet Care
Weight Loss	Pharmaceuticals & Biotech	Wellness
Programming	Wine & Spirits	Religion
Airlines	Retirement	Alcoholic Beverages
Social Media	Anti-Aging	Toys
Camping	Universities	Cancer
Video & Computer Games	Catering	Weddings
Chemical Industry	Comics & Animation	Shoes
Consumer Goods	Corporate Finance	Software Development
Crafts	Solar Energy	Dance
Sports	Discount Stores	Staffing & Recruiting

Distribution	Tattoos	Drug & Alcohol Rehab
Technology News	Environmental News	Telemedicine
Finance News	Textiles	Fitness
Travel Insurance	Food & Drink News	Vacation Rentals
Food Processors	Veganism	Golf
Vegetarian	Healthcare	Video Conferencing
Hiking & Camping	Wedding Planning	Home & Garden
Wine	Home Improvement	Women's Rights
Hospitals	Workwear	Hotels
Writing & Editing	Jewelry & Watches	Zumba
Marketing	Medical Journals	Bicycles
Nursing Homes	Pet Insurance	Nutrition
Roofers	Pharmaceuticals	Women's Interests
Public Safety	Bankruptcy	Publishing
Car Maintenance	Retirement Communities	Car Rentals
Scholarships	Car Sales	Senior Living

Ainsi, dans cette étude, nous avons sélectionné un ensemble de 186 catégories et sous-catégories canadiennes et québécoises, 356 en tout, soigneusement choisies pour leur pertinence économique et leur capacité à refléter les comportements de recherche dans un contexte bilingue. Les catégories sélectionnées se trouvent dans le Tableau 2.1. Ce choix méthodologique permet non seulement d'améliorer la représentativité des données utilisées pour la prévision économique, mais aussi de s'assurer que les particularités linguistiques du Québec sont correctement intégrées dans l'analyse, minimisant ainsi les risques d'interprétation erronée.

Imaginons une analyse des recherches sur Internet au Québec et au Canada. Une catégorie choisie pourrait être « Food ». En tenant compte des particularités linguistiques, l'analyse inclurait des termes en français comme « poutine » et « casse-croûte », qui sont propres à la culture québécoise, ainsi que des termes en anglais comme « burger » ou « fast food », courants dans le reste du Canada. En intégrant ces nuances, cette méthodologie permettrait de mieux représenter les comportements des utilisateurs bilingues ou francophones et d'éviter toute confusion entre les tendances régionales et linguistiques.

Dans le cadre de cette étude, l'objectif est de récolter des données fournissant le plus d'informations possible sur le cycle économique québécois. Pour ce faire, les données Google Trends

seront extraites à une fréquence hebdomadaire. Comme mentionné précédemment, les données hebdomadaires ne couvrent qu'une période de cinq ans, ce qui nécessite de les extraire par échantillons successifs depuis le 1er janvier 2004. Les échantillons ainsi extraits seront ensuite alignés pour former une série chronologique continue de 2004 jusqu'au 31 décembre 2023.

Un défi majeur avec les données Google Trends est la variation des indices en fonction du moment et de l'endroit de l'extraction. Pour surmonter ce problème, cette étude s'inspirera de la méthode développée par Bleher and Dimpfl (2022). Ces auteurs ont mis au point un algorithme en cinq étapes pour aligner les séries Google Trends de manière cohérente.

1. **Extraction des échantillons** : D'abord, des échantillons hebdomadaires de cinq ans sont extraits pour toute la période cible (2004 à aujourd'hui). Chaque échantillon doit se chevaucher d'une année avec le précédent pour permettre une transition en douceur.
2. **Estimation des relations** : Ensuite, on estime une équation linéaire entre les deux premiers échantillons les plus anciens en utilisant la méthode des moindres carrés ordinaires (OLS). L'équation prend la forme $V_{tb} = \alpha + \beta V_{ta} + \varepsilon$, où V_{ta} est l'échantillon le plus ancien et V_{tb} le suivant.
3. **Calcul des nouvelles valeurs** : Les valeurs de l'indice pour V_{tb} sont recalculées en utilisant les coefficients estimés (α et β) pour aligner les deux échantillons.
4. **Combinaison des séries** : Les valeurs ajustées de V_{tb} sont ensuite combinées avec les valeurs originales de V_{ta} pour créer une série alignée.
5. **Répétition du processus** : Ce processus est répété pour l'ensemble des échantillons, jusqu'à ce que toutes les séries soient alignées et combinées en une seule série chronologique cohérente.

Cette méthodologie assure une cohérence temporelle et géographique des données Google Trends utilisées dans cette étude, réduisant ainsi les biais potentiels et permettant une analyse plus précise des tendances économiques.

2.5 Méthodologie des données

2.5.1 Données macroéconomiques

Dans cette étude, les données utilisées proviennent d'une vaste base de données macroéconomiques canadiennes, déjà transformées pour garantir leur pertinence et leur robustesse dans le cadre de l'analyse économétrique. La méthodologie employée pour traiter ces données est

inspirée de l'article intitulé "A Large Canadian Database for Macroeconomic Analysis" par Fortin-Gagnon et al. (2022), qui détaille les étapes nécessaires pour préparer et transformer ces séries temporelles de manière à les rendre stationnaires.

La transformation des séries temporelles macroéconomiques est une étape cruciale avant toute analyse, notamment lorsqu'il s'agit de modèles de prévision. Les séries temporelles économiques, par leur nature, sont souvent non stationnaires, c'est-à-dire que leurs caractéristiques statistiques, telles que la moyenne et la variance, peuvent changer au fil du temps. Cette non-stationnarité peut introduire des biais dans les modèles, rendant les prévisions peu fiables.

Pour remédier à cela, les données macroéconomiques utilisées dans cette étude ont été soigneusement transformées afin de garantir leur stationnarité. Cette transformation implique généralement de prendre la première différence des logarithmes pour la plupart des séries $I(1)$, c'est-à-dire celles qui deviennent stationnaires après avoir été différenciées une fois. Par exemple, les indices de prix et certaines variables comme les taux de chômage et les taux d'intérêt ont été transformés en prenant la première différence de leurs niveaux pour stabiliser leur moyenne et leur variance.

Une étape essentielle du traitement des données consiste à éliminer les effets saisonniers qui pourraient masquer les tendances économiques sous-jacentes. Dans cette étude, toutes les données macroéconomiques utilisées ont été désaisonnalisées avant leur collecte, garantissant ainsi que les modèles de prévision ne soient pas influencés par des variations récurrentes dues à des effets saisonniers prévisibles, comme les fluctuations liées aux cycles économiques ou aux périodes spécifiques de l'année.

L'élimination de la saisonnalité est particulièrement importante car elle permet d'isoler les tendances réelles et les chocs économiques, facilitant ainsi l'interprétation des dynamiques économiques sous-jacentes. Sans ce traitement, les modèles risqueraient d'associer des fluctuations saisonnières aux variations économiques structurelles, ce qui pourrait biaiser les estimations et compromettre la validité des prévisions.

La stationnarité des séries temporelles est un prérequis indispensable pour garantir la validité des modèles de prévision. Les séries non stationnaires peuvent engendrer des résultats trompeurs en raison de fausses corrélations ou de tendances apparentes qui ne sont pas soutenues par des relations économiques réelles. En rendant les séries stationnaires, on stabilise leurs propriétés statistiques, ce qui permet aux modèles économétriques de capturer avec précision les relations économiques sous-jacentes. Cette transformation est donc essentielle pour produire des prévisions fiables et exploitables, tant dans l'analyse économique que dans la prise de décision.

Enfin, pour combler les éventuelles lacunes dans les données et assurer un panel équilibré, l'étude s'inspire de la méthode d'espérance-maximisation basée sur un modèle de facteurs, comme recommandé par Stock and Watson (2002). Cette approche permet de compléter les observations manquantes en les estimant à partir des facteurs communs identifiés dans la base de données, garantissant ainsi une analyse cohérente et robuste.

2.5.2 Données Google Trends

Dans le cadre de l'analyse économique, il est essentiel de travailler avec des séries temporelles qui reflètent fidèlement les tendances sous-jacentes sans être perturbées par des variations saisonnières récurrentes. Les données macroéconomiques, en particulier, sont souvent prétraitées pour retirer ces effets saisonniers, car ces fluctuations peuvent masquer les véritables dynamiques économiques que l'on cherche à analyser. Par exemple, la consommation augmente systématiquement durant la période de Noël, créant ainsi des pics saisonniers récurrents qui ne reflètent pas nécessairement une croissance économique réelle, mais plutôt une habitude de consommation annuelle.

Cependant, toutes les séries temporelles ne sont pas préalablement désaisonnalisées. C'est le cas des données issues de Google Trends, qui ne sont pas ajustées pour les variations saisonnières au moment de leur extraction. Il est donc impératif de retirer la composante saisonnière de ces données afin de pouvoir étudier de manière précise les tendances et les cycles économiques sous-jacents. La présence de saisonnalité dans les données peut conduire à des interprétations erronées, en attribuant à tort des changements périodiques à des tendances économiques réelles alors qu'ils ne sont que des effets saisonniers.

Pour désaisonnaliser les données de Google Trends et extraire les tendances sous-jacentes, nous utilisons l'algorithme Seasonal-Trend decomposition procedure using Loess (STL), introduit par Cleveland et al. (1992). Cet algorithme est particulièrement flexible et puissant pour décomposer une série temporelle en trois composantes distinctes : la tendance (T_t), la saisonnalité (S_t) et le résidu (R_t).

$$Y_t = T_t + S_t + R_t.$$

L'algorithme STL fonctionne en appliquant des techniques de lissage locales (LOESS) à la série temporelle, permettant ainsi une séparation précise entre les différentes composantes. Voici comment la série temporelle Y_t est décomposée :

— T_t : La composante tendancielle, représentant les changements à long terme de la série.

— S_t : La composante saisonnière, qui capte les fluctuations périodiques récurrentes.
— R_t : Le résidu, qui comprend les irrégularités et les variations non expliquées par les deux premières composantes.

L'algorithme STL s'exécute en deux boucles. La boucle intérieure itère entre le lissage de la saisonnalité et de la tendance. D'abord, la composante saisonnière est estimée et retirée de la série, puis la composante tendancielle est calculée sur la série restante. Le résidu est ensuite déterminé en soustrayant les composantes saisonnière et tendancielle de la série originale.

La boucle extérieure, quant à elle, se concentre sur la réduction de l'impact des données aberrantes. Cela permet d'améliorer la robustesse du modèle en minimisant les effets de valeurs extrêmes ou de bruits indésirables qui pourraient fausser l'analyse.

Après la désaisonnalisation, il est également crucial de rendre les séries stationnaires avant de les utiliser dans des modèles de prévision. Une série est dite stationnaire si ses propriétés statistiques, telles que la moyenne et la variance, restent constantes dans le temps. Les séries non stationnaires peuvent conduire à des résultats trompeurs dans les modèles économétriques, car elles peuvent suggérer des relations de long terme qui n'existent pas réellement (phénomène de co-intégration).

Une fois la composante saisonnière retirée à l'aide de l'algorithme STL, les séries résultantes sont transformées afin d'assurer leur stationnarité. Cette transformation peut impliquer la prise de différences successives, une transformation logarithmique ou d'autres ajustements adaptés aux caractéristiques des données.

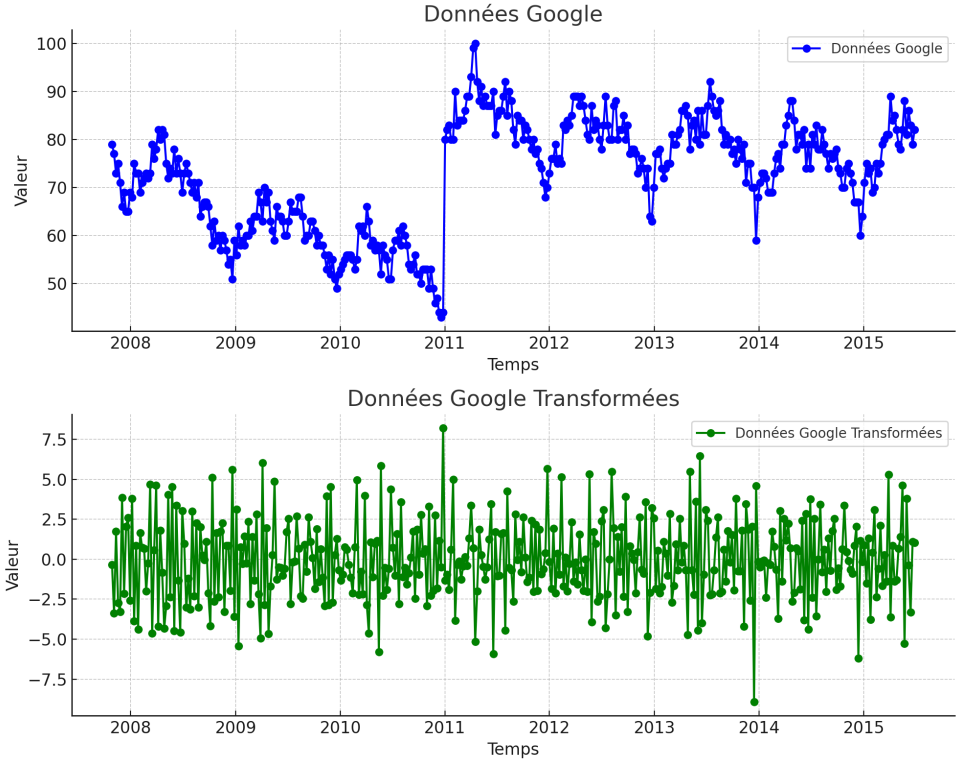
Dans cette étude, la stationnarité des séries issues de Google Trends a été obtenue en appliquant la première différence. Ce choix repose sur plusieurs considérations méthodologiques. Les données de recherche en ligne sont souvent caractérisées par une forte volatilité et des tendances à long terme liées à l'évolution des comportements numériques plutôt qu'à des cycles économiques réels. En appliquant la première différence, on élimine ces tendances et on met en évidence les variations pertinentes à court terme, qui sont les plus utiles pour des prévisions économiques.

De plus, la première différence permet de rendre les séries plus stables en réduisant les fluctuations aléatoires et en limitant les risques de fausses corrélations avec d'autres variables. Cette approche garantit que les modèles économétriques exploitent uniquement les dynamiques de variation d'intérêt et non des niveaux de recherche bruts qui pourraient être biaisés par des effets externes tels que l'évolution du volume global des recherches sur Internet.

Ainsi, cette transformation est essentielle pour tirer pleinement parti des données Google Trends, en permettant aux modèles de capter les véritables signaux économiques et d'améliorer

la fiabilité des prévisions à court terme.

FIGURE 2.1 – Comparaison des données Google avant et après les transformations



La Figure 2.1 ci-dessous illustre le processus de transformation des données de Google Trends dans la catégorie "Auto and Vehicles". L'axe des abscisses représente le temps, avec une période allant du 28 octobre 2007 au 28 juin 2015. Cette plage temporelle a été choisie pour éviter que le graphique ne soit surchargé par une période plus étendue, telle que de 2004 à 2023, ce qui aurait rendu les tendances difficiles à discerner. En se concentrant sur cette fenêtre de temps spécifique, il est possible d'observer de manière plus claire les fluctuations des données de recherche Google Trends pour la catégorie "Auto and Vehicles", permettant ainsi une meilleure interprétation visuelle. L'axe des ordonnées indique la valeur normalisée de l'intérêt de recherche pour cette catégorie sur une échelle de 0 à 100, où 100 correspond au niveau de popularité maximal des termes de recherche observé.

La partie supérieure de la figure présente les données brutes de Google Trends du moment de la récolte, où l'on observe des fluctuations périodiques significatives. La rupture visible souligne l'importance d'utiliser la méthode de Bleher and Dimpfl (2022), comme présenté précédemment. Cette rupture est une manifestation des variations des indices selon le moment et le lieu de l'extraction des données, un problème récurrent avec Google Trends. En appliquant

cette méthodologie, nous pouvons aligner les séries de données de manière cohérente, garantissant ainsi que les variations observées reflètent des tendances réelles plutôt que des artefacts de l'extraction de données.

La partie inférieure de la figure montre les données après l'application de l'algorithme STL, utilisé pour désaisonnaliser et stationnariser la série. On constate que les variations saisonnières ont été efficacement retirées, laissant une série de données plus stable et stationnaire, mieux adaptée à l'analyse économétrique. Les valeurs centrées autour de zéro indiquent que les données ont été normalisées, ce qui réduit les biais potentiels dans les prévisions et permet une meilleure interprétation des résultats.

Ces résultats confirment l'utilité des techniques de désaisonnalisation et de transformation pour améliorer la qualité des données issues de Google Trends avant leur utilisation dans des modèles de prévision économique. En ajustant pour les effets saisonniers et en garantissant la stationnarité, nous pouvons obtenir des prévisions plus précises et mieux comprendre les tendances économiques sous-jacentes.

2.6 La méthode MARX

Dans le cadre de cette étude, comme complément des méthodes de transformation mentionnées plus haut, nous avons utilisé une méthode de transformation des données appelée "Moving Average Rotation of X" (MARX), qui s'est révélée être un outil précieux pour améliorer la précision des prévisions économiques introduit par Goulet Coulombe et al. (2021a). La méthode MARX a été spécifiquement conçue pour traiter les données macroéconomiques à haute fréquence, telles que celles utilisées dans cette recherche, notamment les données canadiennes, québécoises, américaines et celles issues de Google Trends.

La transformation MARX consiste à créer des moyennes mobiles pondérées des variables de la série temporelle, en mettant l'accent sur l'information la plus récente, tout en considérant la continuité entre les différents lags (retards). Cela permet de capter les dynamiques sous-jacentes plus efficacement qu'en utilisant simplement les variables dans leur état brut ou en différenciant les données.

1. **Stabilisation des données** : En appliquant la méthode MARX, les données sont transformées de manière à rendre les relations entre les variables plus stables au fil du temps. Cela réduit le risque que des variations extrêmes ou des changements soudains dans les données affectent négativement les modèles de prévision.
2. **Réduction de la colinéarité** : Les données macroéconomiques sont souvent fortement

corrélées, ce qui peut poser des défis pour les modèles économétriques traditionnels. La méthode MARX atténue ces corrélations en générant des moyennes mobiles qui lissent les fluctuations à court terme, tout en conservant les tendances à long terme.

3. **Meilleure précision des prévisions** : Dans des environnements de prévision complexes, comme ceux impliquant l'apprentissage automatique, la méthode MARX a montré qu'elle pouvait améliorer significativement la performance des modèles en minimisant les erreurs de prévision. Cela est particulièrement vrai lorsque ces transformations sont utilisées en combinaison avec des modèles non linéaires comme les forêts aléatoires (Random Forests) et les arbres de décision boostés (Boosted Trees).
4. **Adaptabilité à divers contextes** : La méthode MARX est également flexible et peut être appliquée à une variété de séries temporelles, ce qui en fait un choix idéal pour traiter les données provenant de diverses sources telles que les marchés financiers, les indicateurs macroéconomiques, et même les données de recherche issues de Google Trends.

Dans cette étude, les données transformées à l'aide de la méthode MARX ont été appliquées aux données canadiennes, québécoises, américaines, et de Google Trends. Chaque ensemble de données a été transformé pour assurer que les modèles de prévision capturent les dynamiques économiques pertinentes sans être perturbés par les fluctuations saisonnières ou les anomalies ponctuelles. Cette approche a permis de maximiser la robustesse des prévisions, en offrant une vision plus claire et plus fiable des tendances économiques à long terme.

L'utilisation de la méthode MARX a permis d'améliorer les performances des modèles en réduisant l'incertitude dans les prévisions, et en permettant une meilleure compréhension des mécanismes économiques sous-jacents. En résumé, la transformation des données via la méthode MARX s'est avérée cruciale pour renforcer la précision et la fiabilité des prévisions économiques dans cette recherche, en optimisant l'utilisation des données disponibles et en minimisant les distorsions liées à la variabilité des séries temporelles.

Chapitre 3 : Méthodologie

Ce chapitre présente la méthodologie adoptée pour effectuer des prévisions en temps réel du PIB québécois. Nous détaillerons les modèles économétriques sélectionnés, ainsi que les méthodes de traitement des données et les techniques d'estimation utilisées pour garantir la précision et la robustesse des prévisions.

Nous commencerons par introduire le modèle à facteur dynamique, essentiel pour gérer la complexité et la dimensionnalité élevée de notre base de données. Ensuite, nous décrirons la spécification des différents modèles de prévision à fréquence mixte, y compris les modèles MIDAS, U-MIDAS, et les modèles d'apprentissage automatique tels que Random Forest (RF), GBM, et LASSO. La section sur l'estimation des modèles expliquera les procédures employées pour obtenir les paramètres des modèles, suivie d'une présentation des critères utilisés pour évaluer la performance des prévisions. Enfin, nous conclurons par une discussion sur l'approche prévisionnelle, qui intégrera l'ensemble des éléments méthodologiques pour aboutir à des prévisions économiques en temps réel précises et pertinentes.

3.1 Modèle à facteurs

Dans cette étude, nous utilisons un modèle à facteur dynamique pour remédier au problème de dimensionnalité élevé, un problème qui survient lorsque le nombre de variables explicatives dépasse le nombre d'observations. Le modèle à facteur permet de condenser l'information contenue dans un grand nombre de variables en un nombre restreint de facteurs communs, tout en capturant l'essentiel des dynamiques économiques sous-jacentes. Ce modèle s'inspire des travaux de Stock and Watson (2002) et peut être formalisé par les équations suivantes.

Soit X_t une matrice $N \times 1$ représentant le vecteur des variables observées à l'instant t, où N est le nombre de variables :

$$X_t = \Lambda F_t + \varepsilon_t.$$

Dans cette équation :

— F_t est un vecteur $r \times 1$ des facteurs communs non observés à l'instant t, avec $r \ll N$.

— Λ est une matrice $N \times r$ des coefficients de chargement des facteurs (factor loadings), qui relie les facteurs F_t aux variables observées X_t.

— ε_t est un vecteur $N \times 1$ des termes d'erreur idiosyncratiques ou spécifiques à chaque variable.

L'objectif du modèle à facteur est de décomposer les variations des variables observées X_t en une partie commune, capturée par F_t, et une partie spécifique à chaque variable, représentée par ε_t. Les facteurs F_t sont supposés capturer les principales dynamiques économiques sous-jacentes, tandis que les termes d'erreur ε_t sont considérés comme du bruit idiosyncratique ou des fluctuations spécifiques à chaque série temporelle.

3.1.1 Modélisation dynamique des facteurs

Les facteurs F_t suivent une dynamique autorégressive, souvent modélisée par un processus VAR (Vecteur Autorégressif) :

$$F_t = \Phi_1 F_{t-1} + \Phi_2 F_{t-2} + \cdots + \Phi_p F_{t-p} + u_t.$$

Dans cette équation :

— $\Phi_1, \Phi_2, \ldots, \Phi_p$ sont des matrices $r \times r$ de coefficients autorégressifs.

— u_t est un vecteur $r \times 1$ des chocs idiosyncratiques non corrélés dans le temps.

3.1.2 Estimation du modèle

L'estimation du modèle se fait généralement en deux étapes. Tout d'abord, les facteurs F_t sont estimés à partir des données observées X_t en utilisant l'analyse en composantes principales (ACP), qui minimise la somme des carrés des erreurs idiosyncratiques ε_t. Ensuite, les coefficients de chargement Λ et les paramètres du processus VAR sont estimés à partir des facteurs F_t.

Ce modèle à facteur est particulièrement utile dans le contexte de prévision macroéconomique, où il est fréquent d'avoir accès à un grand nombre de variables économiques, mais où l'objectif est de capturer les tendances globales à travers un petit nombre de composantes principales. En réduisant ainsi la dimensionnalité des données, le modèle à facteur permet d'améliorer la stabilité et la précision des prévisions tout en évitant les problèmes de multicolinéarité qui pourraient survenir dans les modèles standards.

3.2 AR-MIDAS

Le modèle MIDAS (Mixed Data Sampling), développé par Ghysels et al. (2004), permet de régresser une variable à basse fréquence sur une ou plusieurs variables à plus haute fréquence. Pour améliorer la précision de la prévision, un terme autorégressif est ajouté sur la variable d'intérêt, ce qui conduit à un modèle AR-MIDAS. Ce modèle capte à la fois les dynamiques des variables explicatives et l'autocorrélation présente dans la variable cible. La spécification du modèle AR-MIDAS, incluant les variables trimestrielles, mensuelles et hebdomadaires, est donnée par l'équation suivante :

$$Y_{t+h}^{Q} = \mu + \phi Y_t^{Q} + \sum_{k=0}^{K} \beta_k B(k;\theta_1) X_{t-k/f} + \sum_{m=0}^{M} \delta_m B(m;\theta_2) F_{t-m/w} + \alpha Z_t^{Q} + \varepsilon_t.$$

Où :

— Y_{t+h}^{Q} est la valeur prédite du PIB québécois (variable à basse fréquence) à l'instant $t+h$.

— μ est une constante.

— ϕY_t^{Q} est le terme autorégressif, avec Y_t^{Q} représentant la valeur passée du PIB québécois.

— $\sum_{k=0}^{K} \beta_k B(k;\theta_1) X_{t-k/f}$ représente les contributions des variables explicatives à haute fréquence $X_{t-k/f}$ pondérées par le polynôme exponentiel d'Almon $B(k;\theta_1)$, où k est le décalage (lag) et f la fréquence (mensuelle).

— $\sum_{m=0}^{M} \delta_m B(m;\theta_2) F_{t-m/w}$ représente l'effet du facteur Google $F_{t-m/w}$, disponible sur une fréquence hebdomadaire, transformé à l'aide de MIDAS pour s'ajuster à la fréquence trimestrielle. Le polynôme $B(m;\theta_2)$ pondère les contributions des retards m du facteur hebdomadaire.

— αZ_t^{Q} est la contribution de la donnée trimestrielle Z_t^{Q} directement incluse dans le modèle sans transformation MIDAS.

— ε_t est le terme d'erreur, supposé i.i.d (indépendamment et identiquement distribué).

Un élément essentiel dans la spécification du modèle AR-MIDAS est la détermination du nombre optimal de retards pour chaque variable explicative. Cette sélection est cruciale pour capter les effets dynamiques des variables à haute fréquence tout en évitant un modèle trop complexe ou sujet au overfitting.

Dans cette étude, le nombre de retards a été sélectionné en minimisant le critère d'information bayésien (BIC). Cette approche permet de trouver un équilibre entre la qualité de l'ajustement du modèle et sa parcimonie, en pénalisant les modèles trop complexes qui incluent des retards superflus.

Le choix des retards varie selon la fréquence et la nature des variables. Les variables mensuelles, comme les indicateurs macroéconomiques, ont été testées avec un nombre limité de retards (0 à 3 mois) afin de capturer leurs effets récents sur le PIB. Les variables hebdomadaires, notamment celles issues de Google Trends, ont nécessité un horizon plus large (jusqu'à 13 semaines) pour prendre en compte l'effet cumulatif des tendances de recherche.

Cette approche garantit que seuls les retards les plus pertinents sont retenus, permettant ainsi d'améliorer la précision des prévisions tout en évitant la prolifération inutile des paramètres.

3.2.1 Polynôme exponentiel d'Almon

Le polynôme exponentiel d'Almon, introduit par Ghysels et al. (2007), permet de modéliser les poids $B(k;\theta_1)$ et $B(m;\theta_2)$ associés aux retards des variables explicatives à haute fréquence de manière flexible. Ce polynôme peut prendre plusieurs formes avec un nombre limité de paramètres à estimer, ce qui est crucial pour éviter la prolifération des paramètres dans le modèle MIDAS.

Le polynôme exponentiel d'Almon est défini par :

$$B(k;\theta) = \frac{\exp(\theta_1 k + \theta_2 k^2)}{\sum_{j=0}^{N} \exp(\theta_1 j + \theta_2 j^2)}.$$

Où :

— N est le nombre maximal de retards associé à la série de haute fréquence.

— θ_1 et θ_2 sont les paramètres à estimer, qui déterminent la forme du déclin des poids au fur et à mesure que k augmente.

Le polynôme est conçu de manière à ce que les poids $B(k;\theta)$ soient contraints à être positifs et à décliner (rapidement ou lentement) avec les retards, garantissant ainsi que les termes plus anciens contribuent de moins en moins à la prédiction. Ce déclin est assuré tant que $\theta_2 < 0$. La spécification finale du polynôme est déterminée par les données, et le nombre de retards N est effectivement choisi de manière optimale en fonction des données disponibles.

En raison de la nature non linéaire du polynôme $B(k;\theta)$, l'estimation des paramètres ne peut pas être effectuée par les moindres carrés ordinaires. Au lieu de cela, une estimation par moindres carrés non linéaires (NLS) est utilisée pour déterminer les valeurs optimales des paramètres $\theta_1, \theta_2, \beta_k, \delta_m, \alpha$, et ϕ. Cette méthode permet d'obtenir des estimations précises des paramètres, garantissant une modélisation adéquate des dynamiques temporelles présentes dans les données.

3.3 AR-U-MIDAS

Le modèle U-MIDAS (Unrestricted MIDAS), introduit par Foroni et al. (2015), est une version sans restriction du modèle MIDAS, conçu pour traiter les données à fréquence mixte. Contrairement au modèle MIDAS traditionnel, qui utilise un polynôme de poids pour pondérer les retards des variables à haute fréquence, le modèle U-MIDAS n'impose pas de structure prédéfinie sur les poids. Cette absence de restriction confère au modèle U-MIDAS une flexibilité accrue, ce qui peut être particulièrement avantageux lorsque l'écart de fréquence entre la variable d'intérêt et les variables explicatives est relativement faible.

Le modèle U-MIDAS permet d'intégrer ces diverses fréquences sans imposer de structure sur les poids, offrant ainsi une plus grande flexibilité dans la modélisation.

Le modèle U-MIDAS-AR(1), incluant ces différentes fréquences, est formulé de la manière suivante :

$$Y_{t+h}^{Q} = \mu + \phi Y_t^{Q} + \sum_{k=0}^{K} \beta_k X_{t-k/fm} + \sum_{m=0}^{M} \delta_m F_{t-m/w} + \alpha Z_t^{Q} + \varepsilon_t.$$

Où :

— Y_{t+h}^{Q} représente la valeur prédite du PIB québécois (variable à basse fréquence) à l'instant $t+h$.

— μ est une constante.

— ϕY_t^{Q} est le terme autorégressif, avec Y_t^{Q} représentant la valeur passée du PIB québécois.

— $\sum_{k=0}^{K} \beta_k X_{t-k/fm}$ représente les contributions des variables explicatives mensuelles $X_{t-k/fm}$, où fm est la fréquence mensuelle, et k est le décalage (lag).

— $\sum_{m=0}^{M} \delta_m F_{t-m/w}$ représente l'effet du facteur Google Trends $F_{t-m/w}$, disponible sur une fréquence hebdomadaire, transformé pour s'ajuster à la fréquence trimestrielle.

— αZ_t^{Q} est la contribution de la donnée trimestrielle du PIB canadien Z_t^{Q}, directement incluse dans le modèle sans transformation MIDAS.

— ε_t est le terme d'erreur, supposé i.i.d (indépendamment et identiquement distribué).

Le modèle U-MIDAS-AR(1) se distingue par sa simplicité d'estimation. Contrairement au modèle MIDAS traditionnel, qui nécessite des méthodes d'estimation non linéaires en raison du polynôme de poids, le modèle U-MIDAS peut être estimé à l'aide des moindres carrés ordinaires (OLS). Cela rend l'estimation plus accessible et directe, tout en offrant une flexibilité accrue dans la modélisation des relations entre les variables.

Toutefois, cette flexibilité peut entraîner une prolifération des paramètres à estimer, surtout

lorsque l'échantillon de données est limité. Il est donc essentiel de porter une attention particulière à la sélection des variables et des retards inclus dans le modèle pour éviter les problèmes de overfitting et garantir des prévisions robustes. Les variables mensuelles ont été testées avec des retards allant jusqu'à 3 mois, tandis que les variables hebdomadaires, comme Google Trends, ont été évaluées avec un horizon plus large (jusqu'à 13 semaines) pour capter les effets cumulatifs des tendances de recherche. Cette approche empirique permet d'intégrer uniquement les retards les plus pertinents, assurant ainsi un équilibre entre précision prédictive et complexité du modèle.

En résumé, le modèle U-MIDAS-AR(1) offre une approche flexible pour intégrer des données à haute fréquence dans un cadre de prévision économique, tout en tenant compte des spécificités des données mensuelles, trimestrielles, et hebdomadaires. Bien que sa simplicité d'estimation soit un atout, il est important de gérer prudemment le nombre de paramètres pour maintenir la robustesse et la précision des prévisions.

3.4 Apprentissage automatique

3.4.1 Random Forest

Le modèle RF, introduit par Breiman (2001), est un algorithme d'apprentissage automatique qui se base sur la méthode d'agrégation d'arbres de décision pour effectuer des prédictions. Il s'agit d'une méthode robuste et flexible qui a été largement adoptée pour diverses applications en raison de sa capacité à gérer des données complexes et à minimiser le overfitting. Cette approche permet de capter une grande diversité d'informations en intégrant des données à différentes fréquences (trimestrielle, mensuelle ou hebdomadaire) dans un seul modèle de prévision.

Le modèle RF peut être formulé de manière matricielle comme suit :

Soit X un vecteur de K variables explicatives, comprenant les séries temporelles des variables macroéconomiques et les facteurs issus de Google Trends, et Y la variable cible représentant le PIB québécois. Le modèle RF génère une prédiction $\hat{Y}$ en agrégeant les prédictions $\hat{Y}_t$ de T arbres de décision indépendants :

$$\hat{Y} = \frac{1}{T} \sum_{t=1}^{T} f_t(X).$$

Où :

— $\hat{Y}$ est la prédiction moyenne du PIB québécois.

— T est le nombre total d'arbres dans la forêt.
— $f_t(X)$ représente la prédiction de l'arbre t, construit à partir d'un sous-échantillon des données X et en sélectionnant aléatoirement un sous-ensemble de variables à chaque split (division) dans l'arbre.

Chaque fonction de décision $f_t(X)$ correspond à un arbre de décision construit à partir d'un échantillon aléatoire des données et d'un sous-ensemble de variables explicatives sélectionnées aléatoirement. À chaque nœud de l'arbre, une variable est choisie pour optimiser un critère de division, généralement la minimisation de l'erreur quadratique moyenne en régression. Ce processus est répété récursivement jusqu'à ce qu'un critère d'arrêt soit atteint, comme une taille minimale d'échantillon dans les feuilles terminales ou une profondeur maximale de l'arbre.

La prédiction d'un arbre $f_t(X)$ peut être formalisée comme suit :

$$f_t(X) = \sum_{j=1}^{J} w_j \cdot I(X \in R_j).$$

Où J représente le nombre de feuilles terminales, R_j la région définie par les conditions de split, w_j la valeur moyenne des observations dans R_j, et $I(X \in R_j)$ une fonction indicatrice qui prend la valeur 1 si X appartient à R_j, et 0 sinon.

Cette méthode permet de réduire la variance et d'améliorer la robustesse des prévisions. L'implémentation a été réalisée à l'aide de la bibliothèque `randomForest` en R, suivant l'algorithme de Breiman (2001).

Le paramètre de sélection des variables dans le modèle RF détermine le nombre de variables explicatives à prendre en compte à chaque division de l'arbre de décision. Ce paramètre joue un rôle crucial dans la réduction de la corrélation entre les arbres et contribue à l'amélioration de la robustesse du modèle.

En parallèle, le paramètre de complexité de l'arbre contrôle la taille minimale des feuilles terminales dans chaque arbre de décision. Ce paramètre influence directement la profondeur des arbres, où des valeurs plus élevées peuvent conduire à des arbres moins profonds et potentiellement à une réduction du risque de overfitting. Toutefois, cela peut aussi limiter la capacité du modèle à capturer des interactions complexes entre les variables.

L'optimisation du modèle RF consiste à sélectionner les valeurs optimales des hyperparamètres de sélections des variables et de complexité pour minimiser l'erreur quadratique moyenne sur les données d'entraînement. L'algorithme parcourt un espace de grilles d'hyperparamètres pour identifier les combinaisons qui produisent le modèle le plus performant. Une fois les hyperparamètres optimisés, le modèle RF final est utilisé pour générer les prédictions sur les données

de test.

3.4.2 Gradient Boosting Machine

Le modèle de GBM, est une technique d'apprentissage automatique qui combine plusieurs arbres de décision faibles pour créer un modèle robuste capable de réaliser des prévisions précises. Contrairement à un modèle de RF, qui construit des arbres indépendamment, le GBM construit les arbres séquentiellement, en ajustant chaque nouvel arbre pour corriger les erreurs des arbres précédents. Ce processus permet d'améliorer la précision du modèle tout en contrôlant les risques de overfitting.

Le modèle GBM peut être formulé mathématiquement comme une somme pondérée d'arbres de décision $f_m(X)$, où chaque arbre est construit pour minimiser l'erreur de prédiction des arbres précédents. La prédiction finale $\hat{Y}$ du modèle GBM est donnée par :

$$\hat{Y} = \sum_{m=1}^{M} \gamma_m f_m(X).$$

Où M est le nombre total d'arbres construits par le modèle, $f_m(X)$ est l'arbre de décision construit à l'étape m, et γ_m est le taux d'apprentissage (ou "shrinkage"), un paramètre qui contrôle l'impact de chaque arbre sur la prédiction finale. Des valeurs plus faibles de γ_m permettent d'ajuster plus finement le modèle.

Le taux d'apprentissage est un paramètre clé du modèle GBM, qui détermine la contribution de chaque arbre au modèle final. En appliquant un taux d'apprentissage faible, le modèle ajuste plus lentement chaque arbre, ce qui permet d'éviter le overfitting et de s'assurer que le modèle généralise mieux sur de nouveaux ensembles de données. Cela permet au modèle de faire preuve de flexibilité et de précision dans la capture des dynamiques sous-jacentes des données.

La profondeur d'interaction est un autre paramètre crucial dans le modèle GBM, car elle contrôle la complexité des arbres construits. Une profondeur d'interaction élevée permet de modéliser des interactions complexes entre les variables explicatives, en créant des arbres plus profonds. Toutefois, cela peut aussi augmenter le risque de overfitting, car des arbres plus complexes peuvent capturer non seulement les tendances générales mais aussi le bruit spécifique aux données d'entraînement.

Le nombre minimum d'observations dans un nœud détermine combien d'observations sont nécessaires pour qu'un nœud de l'arbre puisse être divisé. Des valeurs plus élevées limitent la complexité des arbres en empêchant la création de nœuds trop petits, ce qui réduit le risque de overfitting. En effet, des nœuds plus grands favorisent des décisions plus robustes en évitant de

modéliser des variations erratiques des données.

Enfin, la fraction de sous-échantillonnage contrôle la proportion d'observations utilisées pour construire chaque arbre. En utilisant une fraction inférieure à 1, le modèle introduit une certaine variance dans le processus de construction des arbres, ce qui aide à réduire la corrélation entre les arbres individuels et, par conséquent, améliore la robustesse et la performance générale du modèle GBM.

3.4.3 Régression pénalisée

Le modèle LASSO (Least Absolute Shrinkage and Selection Operator), introduit par Tibshirani (1996), est une méthode de régression qui ajoute une contrainte de régularisation à la régression linéaire traditionnelle. Cette contrainte vise à réduire la complexité du modèle en imposant une pénalité sur la somme des valeurs absolues des coefficients de régression, ce qui conduit à la réduction de certains coefficients à zéro. Par conséquent, LASSO effectue simultanément la sélection de variables et la régularisation, ce qui le rend particulièrement utile dans des contextes où le nombre de variables explicatives est élevé par rapport au nombre d'observations, ou lorsque les variables sont fortement corrélées.

Mathématiquement, le modèle LASSO peut être formulé comme suit :

$$\hat{\beta} = \arg\min_{\beta_0,\beta_j} \left\{ \frac{1}{2N} \sum_{i=1}^{N} \left(Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}.$$

Où :

— Y_i est la variable cible pour l'observation i.
— X_{ij} sont les variables explicatives.
— β_0 est l'ordonnée à l'origine.
— β_j sont les coefficients de régression pour chaque variable explicative.
— λ est le paramètre de régularisation qui contrôle la force de la pénalisation appliquée aux coefficients.

Dans le modèle LASSO, le paramètre de régularisation λ joue un rôle central. Ce paramètre détermine la quantité de pénalisation appliquée aux coefficients de régression. Un λ plus élevé contraint davantage les coefficients β_j, en réduisant certains à zéro, ce qui équivaut à éliminer ces variables du modèle. Cela permet de simplifier le modèle et d'éviter le overfitting, en ne conservant que les variables les plus pertinentes pour la prédiction du PIB québécois.

Le modèle est ajusté en utilisant une procédure de validation croisée pour déterminer la

meilleure valeur de λ, c'est-à-dire celle qui minimise l'erreur de prédiction sur un échantillon de validation. Cette approche garantit que le modèle est à la fois simple et performant, en sélectionnant uniquement les variables explicatives les plus importantes tout en régularisant les coefficients pour éviter les fluctuations erratiques dues à des échantillons spécifiques.

Une fois le meilleur λ déterminé, le modèle final est ajusté sur l'ensemble des données d'entraînement en utilisant cette valeur optimale. Cela permet d'obtenir un modèle de régression robuste qui généralise bien aux nouvelles données.

3.5 Modèle de référence : Modèle Autorégressif AR(p)

Afin d'évaluer la performance des modèles présentés dans cette étude, il est essentiel de les comparer à un modèle de référence. Le choix du modèle de référence permet d'établir une base contre laquelle les performances des autres modèles peuvent être mesurées. Dans ce contexte, le modèle de référence choisi est le modèle autorégressif d'ordre p (AR(p)), en raison de sa simplicité et de son efficacité éprouvée dans la prévision des séries chronologiques.

Le modèle AR(p) est un modèle linéaire qui prévoit la valeur d'une variable en fonction de ses propres valeurs passées. Mathématiquement, il peut être exprimé comme suit :

$$Y_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i Y_{t-i} + \varepsilon_t.$$

Où :

— Y_t est la valeur de la variable d'intérêt (dans ce cas, le PIB québécois) à l'instant t.
— α_0 est une constante (ou l'ordonnée à l'origine).
— α_i sont les coefficients autorégressifs, qui mesurent l'influence des valeurs passées Y_{t-i} sur la valeur actuelle Y_t.
— p est l'ordre du modèle, c'est-à-dire le nombre de retards (lags) pris en compte.
— ε_t est le terme d'erreur ou de bruit, supposé être un bruit blanc, c'est-à-dire une série de valeurs aléatoires indépendantes et identiquement distribuées avec une moyenne nulle et une variance constante.

3.5.1 Importance du modèle AR(p) comme modèle de référence

Le modèle AR(p) est largement utilisé comme modèle de référence en raison de plusieurs avantages. Premièrement, sa simplicité en fait un outil facile à comprendre et à appliquer, ce qui en fait un choix naturel pour comparer la performance des modèles plus complexes. Deuxième-

ment, le modèle AR(p) est capable de capturer efficacement les structures temporelles linéaires dans les séries chronologiques, ce qui lui permet de fournir des prévisions de base raisonnablement bonnes dans de nombreux contextes.

De plus, le modèle AR(p) permet de tenir compte de l'autocorrélation dans les données, en exploitant l'information contenue dans les valeurs passées de la série. Cela le rend particulièrement utile dans les prévisions économiques, où les valeurs passées du PIB, par exemple, peuvent avoir une influence significative sur les valeurs futures.

3.5.2 Sélection de l'ordre p

La sélection de l'ordre p du modèle AR est une étape cruciale. Un ordre trop faible peut conduire à un modèle sous-ajusté, incapable de capturer toute la dynamique de la série temporelle, tandis qu'un ordre trop élevé peut introduire un overfitting, où le modèle devient trop complexe et commence à capter le bruit au lieu du signal réel.

Pour déterminer l'ordre optimal p, le critère d'information bayésien (BIC) a été choisi comme méthode d'évaluation dans cette étude. Le BIC est particulièrement approprié dans le contexte de la prévision du PIB québécois, car il pénalise plus fortement les modèles complexes par rapport au critère d'information d'Akaike (AIC). Cette caractéristique est cruciale dans un cadre de prévision, où l'objectif est de trouver un équilibre entre la précision du modèle et sa simplicité, tout en évitant le overfitting. En privilégiant des modèles plus parsimonieux, le BIC aide à sélectionner un modèle qui généralise mieux aux nouvelles données, réduisant ainsi le risque d'ajuster le modèle au bruit plutôt qu'au signal véritable dans les séries chronologiques.

Suite à cette optimisation, l'ordre p retenu pour le modèle AR-MIDAS est $p = 1$. Ce choix reflète la capacité du modèle à capturer la dépendance temporelle immédiate du PIB tout en maintenant une structure simple et généralisable. En sélectionnant un retard unique, le modèle évite la sur-paramétrisation et garantit une meilleure stabilité des prévisions.

3.6 Approche prévisionnelle

Dans cette étude, l'objectif est d'évaluer la valeur ajoutée des données de Google Trends dans les modèles de prévision du PIB québécois en temps réel. Nous procédons à une analyse de nowcasting sur un horizon de 13 semaines avant la publication du PIB. Chaque semaine, une nouvelle prévision est effectuée en utilisant les données disponibles jusqu'à ce moment-là, et ce processus se poursuit jusqu'à la publication officielle du PIB. Une fois le PIB publié, le cycle

recommence pour la période suivante, et cette démarche est répétée sur une période de cinq ans, couvrant ainsi un total de 260 prévisions en temps réel.

Les modèles sont estimés en utilisant une répartition de 80 % des données disponibles pour l'ensemble d'apprentissage (training set) et de 20 % pour l'ensemble de test (test set). Cette répartition est couramment utilisée en ML et en modélisation prédictive pour assurer un équilibre entre l'entraînement du modèle et son évaluation. Le training set fournit au modèle un volume suffisant de données pour apprendre les relations sous-jacentes entre les variables explicatives et la variable cible, tandis que le test set permet d'évaluer la capacité du modèle à généraliser ses prévisions à de nouvelles données non vues pendant l'entraînement. Les prédictions couvrent la période allant de la première semaine de 2019 à la dernière semaine de 2023.

Comme mentionné, l'objectif principal est de déterminer si l'ajout des données de Google Trends, spécifiquement les catégories et sous-catégories sélectionnées, apporte une valeur ajoutée aux modèles de prévision, et d'identifier les horizons pour lesquels ces données sont les plus pertinentes. En d'autres termes, il s'agit de vérifier à quel point les informations issues des recherches effectuées sur Google, structurées selon ces catégories et sous-catégories, peuvent améliorer la qualité des prévisions du PIB québécois, particulièrement pour des horizons de prévision plus courts.

Pour chaque période de prévision, les modèles sont estimés en utilisant une fenêtre glissante de 13 semaines. Cette approche permet de capturer des dynamiques économiques en temps réel, en s'ajustant continuellement aux nouvelles informations disponibles. Les prévisions sont ensuite comparées trimestre après trimestre, ce qui permet d'évaluer la cohérence et la fiabilité des modèles sur des périodes prolongées.

En choisissant un horizon de cinq ans pour ces prévisions, nous visons à assurer la robustesse des résultats et à observer leur convergence vers une représentation fidèle de la réalité économique. En effet, effectuer des prévisions sur une longue période expose le modèle à diverses conditions économiques, permettant d'évaluer sa capacité à s'adapter à des cycles économiques complets, incluant les variations saisonnières et structurelles. Cette approche permet non seulement de tester la robustesse des modèles, mais aussi d'assurer que les résultats ne sont pas influencés par des événements ponctuels ou des anomalies spécifiques à une période donnée, garantissant ainsi une évaluation rigoureuse de la valeur ajoutée des données Google Trends dans la prévision du PIB québécois.

De plus, pour les modèles MIDAS, les facteurs dérivés des mots-clés de Google Trends sont réestimés à chaque itération pour s'assurer que les modèles capturent les changements structurels au fil du temps. Cette réestimation régulière permet aux modèles de s'adapter aux

nouvelles informations et de mieux refléter les dynamiques économiques actuelles, notamment autour des périodes de fluctuation importante, comme la récession de 2007-2009.

Au total, cette approche produit un échantillon de 260 prévisions, permettant ainsi une analyse exhaustive de la valeur ajoutée des données de Google Trends dans le cadre de la prévision économique en temps réel. Cette méthodologie rigoureuse permet de mieux comprendre l'impact de l'intégration de ces données sur la précision des prévisions et leur utilité dans un contexte économique dynamique.

3.7 Critère d'évaluation des prévisions

Pour évaluer la performance des modèles de prévision, nous utiliserons deux critères principaux : l'erreur quadratique moyenne (EQM) et le test de Diebold and Mariano (1995). Ces outils permettront de comparer les résultats des modèles incluant les données de Google Trends avec ceux n'incluant pas ces données, afin de déterminer leur valeur ajoutée.

3.7.1 Erreur quadratique moyenne

Le EQM est une mesure courante utilisée pour quantifier l'erreur de prévision d'un modèle. Elle se définit comme la moyenne des carrés des écarts entre les valeurs prédites et les valeurs observées. Pour un modèle donné, l'EQM est calculée comme suit :

$$\text{EQM} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2.$$

Où :

— N est le nombre total de prévisions.

— $\hat{y}_i$ est la valeur prédite par le modèle pour l'observation i.

— y_i est la valeur réelle observée pour l'observation i.

Dans le cadre de l'évaluation des performances des modèles de prévision, il est essentiel que l'erreur quadratique moyenne soit la plus faible possible. En effet, une EQM faible indique que les écarts entre les valeurs prédites par le modèle et les valeurs réelles observées sont minimisés. Plus l'EQM est faible, plus le modèle est considéré comme performant, car il reflète une meilleure précision des prévisions. Un EQM réduit signifie que le modèle est capable de capturer les dynamiques sous-jacentes des données avec une grande fidélité, ce qui est crucial pour produire des prévisions fiables et exploitables.

Dans le contexte de cette étude, nous comparons l'EQM des modèles incluant les données de Google Trends avec l'EQM des modèles sans ces données. Le ratio des EQM est donné par :

$$\text{Ratio EQM} = \frac{\text{EQM}_{GT}}{\text{EQM}_{sansGT}}.$$

Où :

— EQM_{GT} représente l'erreur quadratique moyenne des modèles utilisant les données de Google Trends.

— EQM_{sansGT} représente l'erreur quadratique moyenne des modèles n'utilisant pas les données de Google Trends.

Un ratio supérieur à 1 indique que l'utilisation des données de Google Trends améliore la précision des prévisions, car l'EQM est plus faible pour les modèles avec Google Trends.

3.7.2 Test de Diebold-Mariano

Le test de Diebold and Mariano (1995) est utilisé pour comparer la performance de deux modèles de prévision en termes d'erreurs de prévision. Le test examine si les différences entre les EQM des deux modèles sont statistiquement significatives.

Soit ε_{t+h}^{GT} l'erreur de prévision à l'horizon h pour le modèle utilisant Google Trends, et $\varepsilon_{t+h}^{sansGT}$ l'erreur pour le modèle sans Google Trends. La différence des EQM pour chaque période t est définie par :

$$d_t = (\varepsilon_{t+h}^{GT})^2 - (\varepsilon_{t+h}^{sansGT})^2.$$

L'hypothèse nulle H_0 du test est que les deux modèles ont des performances équivalentes, ce qui signifie que la moyenne de d_t est nulle :

$$H_0 : E[d_t] = 0.$$

L'hypothèse alternative H_1 est que l'un des modèles a une performance supérieure, c'est-à-dire que la moyenne de d_t est différente de zéro :

$$H_1 : E[d_t] \neq 0.$$

La statistique de test DM est calculée comme suit :

$$\text{DM} = \frac{\overline{d}}{\sqrt{\hat{V}(\overline{d})}}.$$

Où :

— $\overline{d} = \frac{1}{T}\sum_{t=1}^{T} d_t$ est la moyenne des d_t.

— $\hat{V}(\overline{d}) = \frac{1}{T}(\gamma_0 + 2\sum_{i=1}^{h-1} \gamma_i)$ est une estimation de la variance de $\overline{d}$.

— $\gamma_i = \text{cov}(d_t, d_{t-i})$ est la covariance des différences d'erreurs à des périodes i distinctes.

La statistique DM suit une loi normale centrée réduite $N(0,1)$. Si la statistique DM est significativement différente de zéro (généralement au seuil de 5%), cela suggère que les performances des deux modèles sont statistiquement différentes.

Dans le contexte de cette étude, un ratio Ratio EQM > 1 combiné à un test de Diebold-Mariano rejetant l'hypothèse nulle H_0 indiquerait que les modèles utilisant les données de Google Trends offrent des prévisions plus précises que les modèles sans ces données.

Ensuite, une fois cette première comparaison réalisée, nous sélectionnerons les modèles les plus performants de chaque catégorie, c'est-à-dire ceux qui montrent les meilleurs résultats avec ou sans les données de Google Trends. Ces modèles sélectionnés seront alors comparés au modèle de référence AR(p) mentionnée plus haut, en répétant le processus de prévision et d'évaluation avec une fenêtre glissante. En utilisant l'EQM comme critère de comparaison, nous pourrons déterminer dans quelle mesure les modèles sélectionnés surpassent le modèle AR(p) en termes de performance prédictive. Si l'EQM des modèles complexes reste significativement plus faible que celui du modèle AR(p), cela confirmerait que l'intégration des facteurs complexes, y compris les données de Google Trends, apporte une amélioration substantielle par rapport à un modèle de référence simple. Ce processus itératif sur plusieurs horizons de prévision nous permettra d'identifier les modèles les plus robustes et performants pour la prévision du PIB québécois.

Chapitre 4 : Résultats

Dans ce chapitre, nous présentons les résultats obtenus à partir des modèles économétriques détaillés dans le chapitre précédent. L'analyse débute par l'évaluation de l'apport des données Google Trends aux prévisions de l'état actuel du PIB québécois. Cette évaluation permettra de répondre à la question centrale de ce mémoire.

Les résultats seront ensuite approfondis par la comparaison avec le modèle de référence AR(p). Cette comparaison sera illustrée à l'aide de « glide charts », offrant une visualisation claire des performances relatives des différents modèles.

Enfin, nous réitérerons l'exercice en excluant l'impact du choc économique causé par la pandémie de COVID-19, afin d'évaluer la robustesse de nos modèles dans un environnement moins perturbé. Cette section permettra de vérifier si l'ajout des données de Google Trends reste pertinent même en l'absence d'événements économiques exceptionnels.

4.1 Impact des données Google Trends sur les modèles de prévision

4.1.1 Résultats : comparaison Google Trends

Dans cette étude, l'objectif principal était de déterminer si les données de Google Trends apportaient une valeur ajoutée aux modèles de nowcasting du PIB québécois. Les résultats sont présentés sous forme de ratios des EQM des modèles avec les données de Google Trends par rapport aux modèles sans ces données.

TABLEAU 4.2 – Ratio des EQM

Horizon	MIDAS	UMIDAS	RF	GBM	LASSO
13	1.0044	1.0072	1.0013	0.9659**	0.9470
12	0.9961	1.0859***	1.0016	1.0074	1.0173
11	0.9880	1.0388**	1.0009	1.0015	0.9974
10	0.9817***	1.0379***	0.9996	0.9991	0.9993
9	0.9789***	1.0162***	1.0021	1.0444	1.0092
8	0.9845***	1.0079*	1.0007	0.9815	0.9832
7	0.9429***	1.0804***	1.0018**	0.9449**	1.0477
6	0.9377***	1.0561***	1.0034***	1.0034**	0.9983
5	0.9229***	1.0101	1.0028*	0.9672	1.0388
4	0.9535**	1.0283***	1.0061**	1.0162	1.0836**
3	1.0189	1.0111	1.0095***	0.9825	1.0145
2	1.0133**	0.9948*	1.0052***	1.0430*	0.9839
1	1.0113	1.0193	1.0058**	1.0072	0.9528

Les résultats présentés dans le tableau 4.2 illustrent les ratios des EQM entre les modèles utilisant les données de Google Trends et ceux n'en utilisant pas. Chaque colonne du tableau, par exemple celle du modèle MIDAS, représente le ratio entre la performance du modèle MIDAS avec les données de Google Trends et la performance du même modèle sans ces données. Un ratio supérieur à 1 indique que l'intégration des données de Google Trends améliore la performance du modèle par rapport à son équivalent sans ces données, tandis qu'un ratio inférieur à 1 signifie que les modèles sans les données de Google Trends performent mieux.

L'horizon, indiqué dans la première colonne, représente le nombre de semaines avant la date de publication du PIB québécois. Ainsi, un horizon de 13 signifie que les prévisions sont faites 13 semaines avant la publication du PIB, tandis qu'un horizon de 1 signifie que les prévisions sont faites une semaine avant cette même publication.

Le nombre d'étoiles indiquent le niveau de significativité statistique de ces résultats. Plus précisément, une étoile (*) représente un niveau de significativité de 10 %, deux étoiles (**) indiquent un niveau de 5 %, et trois étoiles, un niveau de 1 %. Par exemple, un ratio marqué de trois étoiles (***), comme pour le modèle UMIDAS à l'horizon de 12 semaines, montre que l'amélioration de la performance avec les données de Google Trends est statistiquement très significative à un niveau de confiance de 99 %. À l'inverse, l'absence d'étoiles indique que l'hypothèse nulle H_0 du Test Diebold-Mariano a été rejetée, suggérant que les différences de

performance ne sont pas statistiquement significatives.

Les résultats montrent que le modèle UMIDAS bénéficie le plus de l'intégration des données de Google Trends, notamment aux horizons plus éloignés de la publication du PIB, où les ratios sont significativement supérieurs à un. Par exemple, entre la 12e et la 6e semaine avant la publication, le modèle UMIDAS affiche des améliorations considérables avec des ratios EQM significatifs, suggérant une forte valeur ajoutée des données de Google. Le modèle MIDAS montre également des améliorations, particulièrement dans les trois semaines avant la publication.

En revanche, les performances du modèle GBM sont plus contrastées. Dans plusieurs cas, le modèle sans les données de Google Trends surperforme celui avec ces données, en particulier à l'approche de la publication du PIB, ce qui pourrait indiquer un overfitting ou une inadéquation des données de Google pour ce type de modèle.

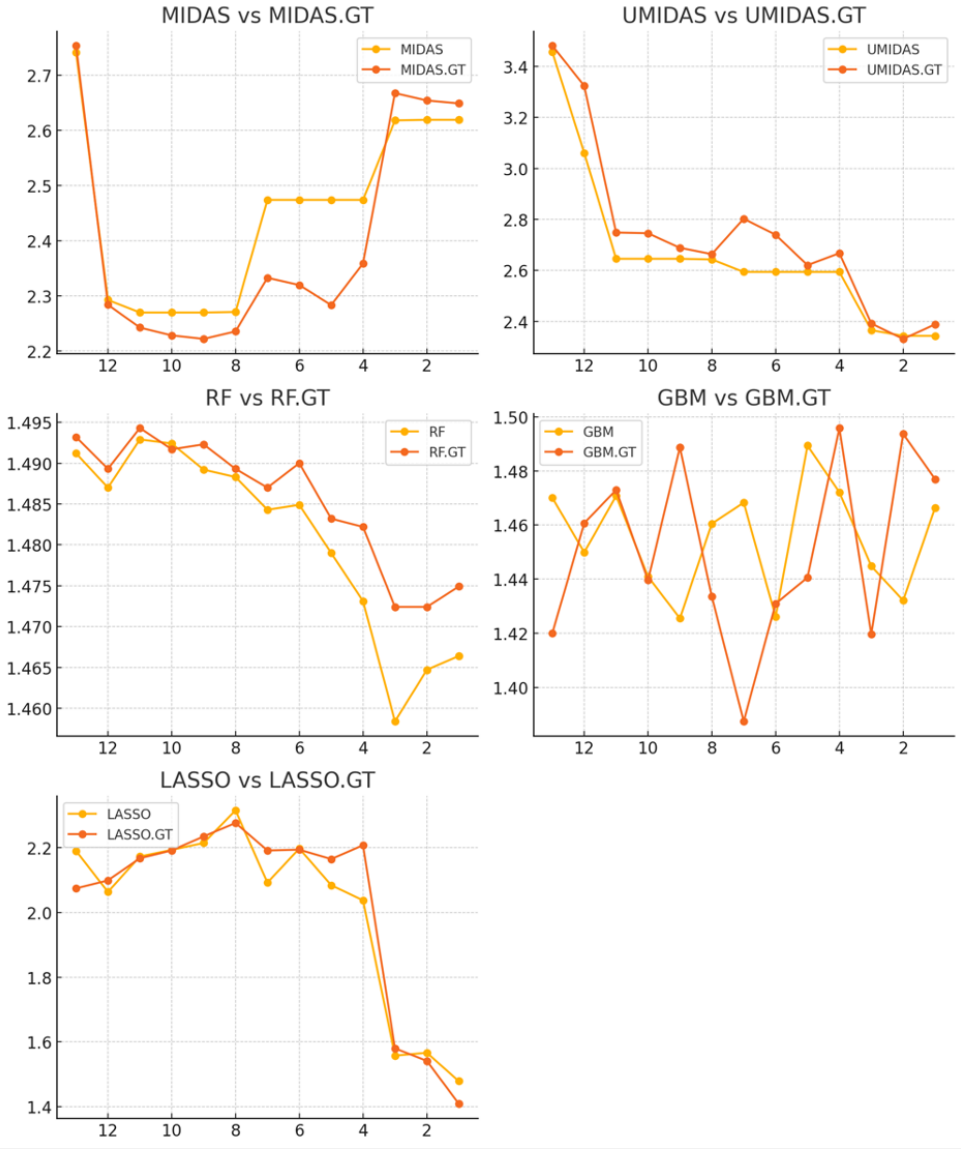
Les résultats présentés pour le modèle RF montrent une amélioration progressive des performances en termes de EQM à mesure que l'on s'approche de la date de publication du PIB, avec l'ajout des données Google Trends.

Pour la plupart de nos modèles, il est intéressant de noter que plus on s'approche de la date de publication du PIB, moins l'impact des données de Google Trends semble prononcé. Cette observation peut s'expliquer par le fait que, à l'approche de la publication, des données économiques plus traditionnelles et plus précises deviennent disponibles, réduisant ainsi l'importance des signaux alternatifs captés par les données de Google Trends.

4.1.2 Résultats : comparaison avec le modèle de référence

Dans cette seconde partie de l'analyse, l'objectif est de comparer la performance des différents modèles de prévision, avec et sans les données de Google Trends, par rapport à un modèle de référence autoregressif AR(p). Cette comparaison vise à déterminer dans quelle mesure nos modèles peuvent améliorer la précision des prévisions par rapport à une approche autoregressive classique.

FIGURE 4.2 – Comparaison des modèles avec le modèle de référence AR(p)



Dans la Figure 4.2, l'axe des abscisses représente le nombre de semaines avant la date de publication du PIB québécois, tandis que l'axe des ordonnées indique le ratio des EQM. Un ratio supérieur à 1 signifie que le modèle de prévision surpasse le modèle de référence AR(p) en termes de précision. Chaque graphique présente les performances du modèle avec et sans les données de Google Trends : la ligne jaune représente le modèle sans les données de Google Trends, tandis que la ligne orange correspond au modèle intégrant ces données.

Les résultats montrent clairement que les modèles UMIDAS, à la fois avec et sans les données de Google Trends, surpassent significativement le modèle AR(p). Cette performance est

particulièrement marquée aux horizons plus éloignés de la publication du PIB, par exemple, entre la 13e et la 10e semaine avant la publication.

Le modèle MIDAS montre également des performances robustes, avec des ratios EQM se situant autour de 2 à 2.7, comparativement au modèle AR(p). Notamment, l'ajout des données de Google Trends améliore légèrement les performances, en particulier à l'approche de la publication du PIB, entre la 3e et la 1ère semaine. Cette amélioration pourrait être due au fait que, bien que les signaux économiques captés par Google Trends soient moins directement liés aux indicateurs traditionnels, ils peuvent offrir des informations précieuses sur les comportements et les attentes des agents économiques, qui ne sont pas immédiatement reflétées dans les données économiques classiques.

Le modèle RF montre des résultats plus modérés, avec des ratios EQM proches de 1, ce qui suggère une performance comparable à celle du modèle AR(p). Cette relative stabilité peut indiquer que le modèle RF, tout en étant robuste, ne bénéficie pas autant des données de Google Trends. Toutefois, il est important de noter que le modèle RF reste performant dans un contexte de prévisions non linéaires, où les relations entre les variables ne sont pas forcément captées par les modèles linéaires comme AR(p).

Le modèle GBM présente des résultats plus contrastés. Il est intéressant de noter que, dans certains cas, le modèle GBM sans les données de Google Trends surpasse celui avec ces données, surtout à l'approche de la publication du PIB. Cela pourrait s'expliquer par le fait que les modèles GBM, qui sont particulièrement sensibles aux surajustements, peuvent capter du bruit lorsque des données additionnelles comme celles de Google Trends sont incluses. De plus, à mesure que la publication du PIB approche, les données économiques traditionnelles deviennent plus abondantes et précises, réduisant ainsi la valeur ajoutée des signaux alternatifs comme ceux de Google Trends.

Enfin, le modèle LASSO montre des performances variées, avec des résultats se rapprochant à ceux du modèle AR(p), en particulier aux horizons plus courts. Cependant, aux horizons plus éloignés, nos modèles surpassent notamment notre modèle AR(p), ce qui pourrait être dû à la capacité du LASSO à sélectionner et à régulariser efficacement les variables pertinentes, y compris les signaux non traditionnels. Une baisse soudaine du ratio EQM est observée à l'horizon de 3 semaines, indiquant une détérioration des performances du LASSO par rapport au modèle AR(p). Cette diminution peut s'expliquer par plusieurs facteurs. Premièrement, il est possible que les variables explicatives sélectionnées par LASSO deviennent moins informatives à ce stade, ce qui réduit la capacité du modèle à capter les dynamiques économiques pertinentes. Deuxièmement, le modèle AR(p) exploite uniquement l'information la plus ré-

cente, ce qui pourrait expliquer pourquoi, à très court terme, les relations autorégressives pures deviennent plus dominantes que les prédicteurs externes. Enfin, il est possible que certains indicateurs économiques de haute fréquence soient trop volatils ou peu pertinents dans cette fenêtre temporelle, entraînant un ajustement sous-optimal du LASSO et, par conséquent, une hausse de l'erreur quadratique moyenne. Ces éléments suggèrent que, dans ce contexte, les gains du LASSO par rapport à un modèle AR simple s'amenuisent fortement à l'approche de la date cible.

L'ensemble de ces résultats doit être interprété dans le contexte économique particulier de la période 2019-2023, qui inclut le choc majeur de la pandémie de COVID-19. Cette période a été caractérisée par une grande volatilité économique et des incertitudes accrues, ce qui a probablement amplifié les avantages des modèles capables de capturer des dynamiques rapides et non linéaires, comme UMIDAS et MIDAS. Les modèles AR(p), qui sont basés sur des dynamiques plus linéaires et des relations historiques stables, ont eu plus de mal à s'adapter aux chocs soudains et aux changements de régime économique induits par la pandémie.

4.2 Robustesse des modèles de prévision en période stable

Après avoir analysé les performances des modèles de prévision économique intégrant les données de Google Trends dans un contexte incluant le choc de la pandémie de COVID-19, il est crucial d'examiner ces mêmes modèles dans un environnement économique plus stable. Cette section se concentre sur les prévisions réalisées entre la première semaine de 2015 et la dernière semaine de 2019, une période marquée par une relative stabilité économique, sans l'impact perturbateur de la pandémie.

L'objectif est de déterminer si les avantages observés grâce à l'intégration des données de Google Trends dans les modèles UMIDAS, MIDAS, Random Forest, GBM et LASSO restent pertinents dans un contexte où les dynamiques économiques sont moins volatiles. En particulier, nous chercherons à savoir si l'amélioration des performances des modèles de prévision, comparés au modèle de référence AR(p), persiste en l'absence d'événements économiques majeurs.

Cette analyse permettra de vérifier la robustesse des résultats obtenus dans la section précédante, en comparant les performances des modèles sur une période plus « normale » sur le plan économique. Elle contribuera également à évaluer si l'ajout de données alternatives comme celles de Google Trends est un outil efficace pour améliorer la précision des prévisions économiques en temps réel, indépendamment des chocs exogènes.

4.2.1 Résultats : comparaison Google Trends sans le choc COVID

Dans cette section, nous analysons la performance des modèles économétriques, avec et sans les données de Google Trends, sur la période de prévisions allant de 2015 à 2019, excluant ainsi le choc économique de la pandémie de COVID-19. Le but est de déterminer si l'intégration des données de Google Trends dans les modèles MIDAS, UMIDAS, RF, GBM et LASSO améliore les prévisions en comparaison avec les modèles sans ces données, dans un contexte économique plus stable.

Dans le tableau 4.3, les résultats sont exprimés sous forme de ratios des EQM. Un ratio inférieur à 1 indique que le modèle sans les données Google performe mieux, tandis qu'un ratio supérieur à 1 suggère que le modèle avec les données Google offre une meilleure performance. La signification des résultats est testée à l'aide du test de Diebold-Mariano, et les niveaux de significativité sont indiqués par des étoiles : *** pour 1%, ** pour 5% et * pour 10%.

TABLEAU 4.3 – Ratio des EQM sans le choc COVID

Horizon	MIDAS	UMIDAS	RF	GBM	LASSO
13	0.9795*	0.6068***	0.9847	0.9880	0.9712**
12	0.9500*	0.6515**	0.9849**	1.0518	0.9916
11	0.9296	0.6871***	0.9900	0.9368**	0.9931
10	0.7779***	0.7970***	0.9954	1.0424	0.9854
9	0.7784***	0.8230**	0.9906	1.0302	1.0024
8	0.8202***	0.8624*	0.9815**	0.9813	0.9730
7	0.9451***	0.8006***	1.0000	0.9352*	1.0270
6	0.9226***	0.7244***	0.9974	0.9865	0.9996
5	1.0085	0.8123***	0.9952	1.0639	0.9670
4	0.9063*	0.7197***	1.0038	1.0992*	1.0128
3	1.0130	0.8138**	1.0141*	0.9961	0.9674
2	1.0082	0.8599**	1.0127**	1.0275	0.9399**
1	0.9380***	0.9139**	1.0102	1.0289	0.9694

Les résultats pour le modèle MIDAS montrent que, dans la plupart des cas, l'intégration des données de Google Trends n'améliore pas la performance du modèle. En effet, les ratios EQM sont généralement inférieurs à 1, indiquant que le modèle MIDAS sans les données de Google Trends est plus performant. Cette tendance est particulièrement marquée aux horizons qui se rapprochent de la publication du PIB (entre 10 et 7 semaines avant la publication), avec des

ratios significativement inférieurs à 1. Cette performance pourrait être attribuée au fait que, dans un contexte économique plus stable, les modèles MIDAS bénéficient moins des signaux non traditionnels comme ceux des recherches Google, ces signaux étant peut-être moins pertinents en l'absence de chocs économiques majeurs.

Le modèle UMIDAS, contrairement à l'analyse précédente, montre une performance significativement meilleure sans l'intégration des données de Google Trends, avec des ratios EQM toujours inférieurs à 1, notamment aux horizons plus éloignés de la publication du PIB. Les résultats indiquent que les données de Google Trends n'apportent pas de valeur ajoutée substantielle dans un environnement économique stable. Ces résultats suggèrent que, bien que les modèles UMIDAS soient puissants pour capter des dynamiques complexes, les signaux additionnels fournis par Google Trends sont peut-être moins essentiels en l'absence de perturbations majeures.

Le modèle RF affiche des résultats mixtes. À certains horizons, l'intégration des données de Google Trends améliore légèrement les prévisions, mais dans d'autres cas, la différence de performance est minime ou inexistante. Avec des ratios très proches de 1 et des niveaux de significativité faibles, il semble que l'apport des données de Google soit moins décisif pour ce modèle dans un contexte économique stable. La nature non linéaire et flexible du RF peut expliquer pourquoi ce modèle ne dépend pas autant des signaux supplémentaires provenant de Google Trends pour améliorer ses prévisions. Cependant, on observe une tendance notable : plus on s'approche de la date de publication du PIB, plus les données de Google Trends ajoutent de la valeur à notre modèle.

Le modèle GBM présente des résultats plus variés. À plusieurs horizons, le modèle avec les données de Google surperforme légèrement. Cependant, ces améliorations ne sont pas systématiques, et il y a même des cas où le modèle sans données de Google performe mieux. Ces résultats pourraient indiquer que, dans un environnement stable, les signaux supplémentaires de Google Trends peuvent parfois ajouter du bruit plutôt que de l'information utile, surtout lorsque le modèle GBM tend à capter des relations complexes qui ne sont pas toujours renforcées par ces données alternatives.

Enfin, le modèle LASSO montre une performance globalement mitigée. À certains horizons, l'ajout des données de Google Trends ne semble pas bénéfique, avec des ratios EQM inférieurs à 1, notamment aux horizons plus éloignés. Cependant, à l'approche de la publication, l'intégration des données Google améliore légèrement les prévisions, comme le montre quatre semaines avant la publication mais pas de façon significative. Cela suggère que, bien que le modèle LASSO soit capable de filtrer les variables moins pertinentes, l'ajout de don-

nées Google Trends n'apporte pas systématiquement de bénéfice en termes de précision des prévisions dans un environnement économique stable.

Les résultats obtenus suggèrent que l'intégration des données de Google Trends dans les modèles de prévision économique, tels que MIDAS, UMIDAS, RF, GBM et LASSO, ne conduit pas nécessairement à une amélioration systématique des performances en période de stabilité économique. En effet, l'absence de chocs économiques majeurs, comme la pandémie de COVID-19, semble réduire la pertinence des signaux non traditionnels captés par Google Trends.

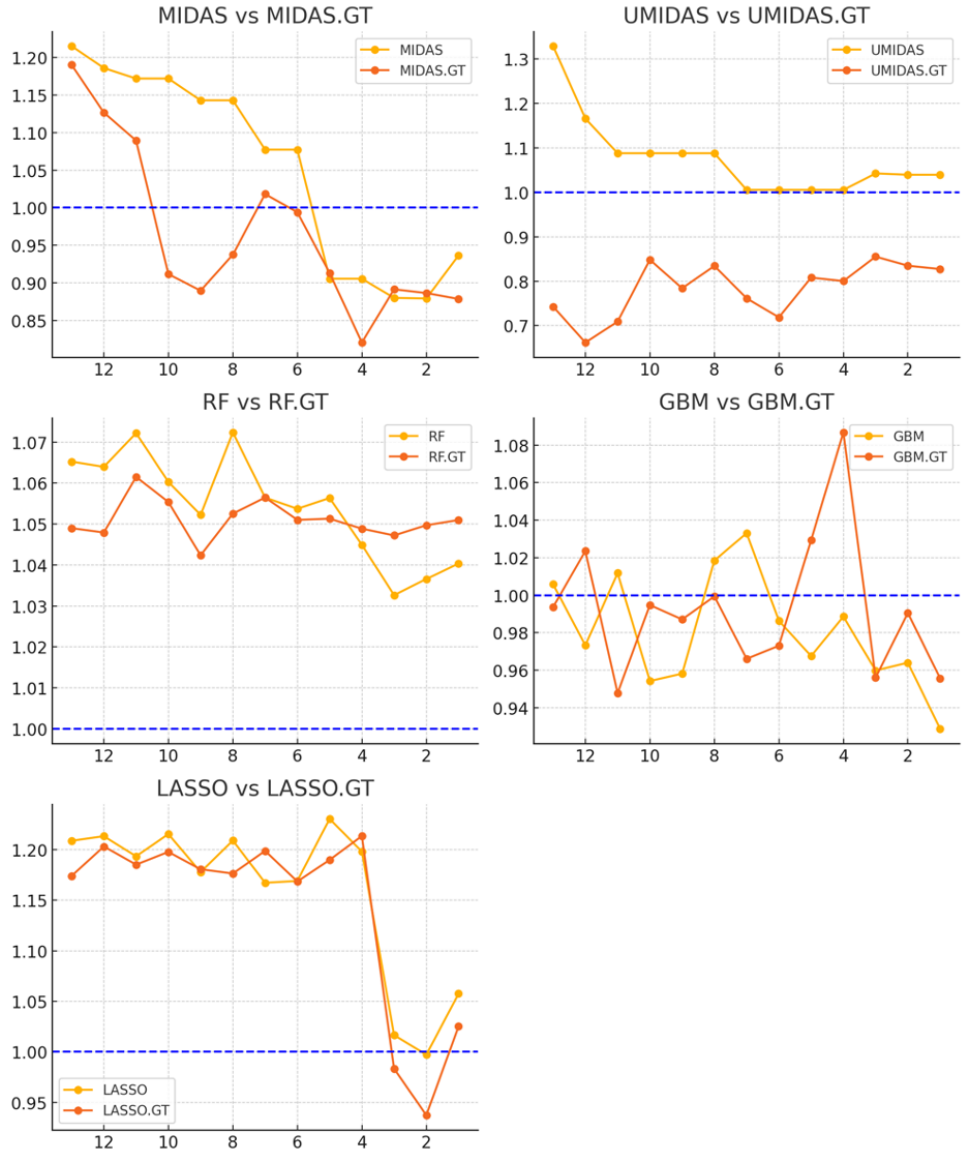
Dans un environnement stable, les modèles traditionnels sans données supplémentaires semblent suffisants pour capter les dynamiques économiques sous-jacentes. Cependant, l'apport des données de Google Trends pourrait être plus pertinent en période d'incertitude accrue ou de volatilité économique, où les comportements des agents économiques sont plus difficiles à prévoir avec les seuls indicateurs traditionnels.

Ces résultats soulignent l'importance de considérer le contexte économique dans lequel les prévisions sont réalisées. Si les données de Google Trends peuvent offrir une valeur ajoutée dans des périodes perturbées, leur pertinence semble moins marquée dans des environnements économiques stables, où les modèles traditionnels sans données supplémentaires restent efficaces.

4.2.2 Résultats : comparaison avec modèle de référence sans le choc COVID

Dans cette analyse, nous comparons la performance des différents modèles économétriques – avec et sans les données de Google Trends – par rapport au modèle de référence AR(p), en excluant l'impact de la pandémie de COVID-19. L'objectif est de comprendre dans quelle mesure ces modèles peuvent surpasser le modèle autoregressif AR(p), en particulier dans un environnement économique plus stable, tel qu'observé entre 2015 et 2019. Cette comparaison est essentielle pour évaluer l'efficacité des modèles dans des contextes économiques normaux, en l'absence de chocs exogènes majeurs.

FIGURE 4.3 – Comparaison des modèles sans le choc COVID avec le modèle de référence AR(p)



Dans la Figure 4.3, l'axe des abscisses représente le nombre de semaines avant la date de publication du PIB québécois, tandis que l'axe des ordonnées indique le ratio des EQM. Un ratio supérieur à 1 signifie que le modèle de prévision surpasse le modèle de référence AR(p) en termes de précision. Chaque graphique présente les performances du modèle avec et sans les données de Google Trends : la ligne jaune représente le modèle sans les données de Google Trends, tandis que la ligne orange correspond au modèle intégrant ces données.

Les résultats pour le modèle MIDAS montrent une performance généralement supérieure à

celle du modèle AR(p) sur presque tous les horizons éloignés étudiés. En particulier, aux horizons de 13 à 6 semaines avant la publication du PIB, le modèle MIDAS surpasse de manière significative pour la plupart des horizons le modèle AR(p) avec des ratios des EQM systématiquement supérieurs à 1. Cela suggère que le modèle MIDAS est capable de capter des dynamiques économiques pertinentes à des horizons éloignés, ce qui n'est pas toujours possible avec un modèle autoregressif classique comme AR(p). Cette performance pourrait s'expliquer par la capacité du MIDAS à intégrer des données à haute fréquence, offrant ainsi une meilleure réactivité aux changements économiques qui surviennent bien avant que les effets ne soient visibles dans les indicateurs plus traditionnels utilisés par le modèle AR(p).

Toutefois, à mesure que l'on se rapproche de la publication du PIB, la supériorité du modèle MIDAS sur le modèle AR(p) semble s'atténuer légèrement, bien que le modèle MIDAS continue de performer adéquatement. Cela peut être attribué au fait que, à des horizons plus courts, les données traditionnelles deviennent plus prédominantes, et l'avantage des données à haute fréquence s'amenuise.

Le modèle UMIDAS montre également des performances nettement supérieures au modèle AR(p), en particulier à des horizons plus éloignés. Les résultats indiquent que le modèle UMIDAS est particulièrement efficace pour capter des dynamiques économiques complexes qui pourraient être négligées par des approches plus linéaires. Le fait que le modèle UMIDAS surpasse de manière significative le modèle AR(p) aux horizons de 13 à 9 semaines suggère que ce modèle est mieux équipé pour anticiper les variations économiques sur une plus longue période.

Cependant, peu importe l'horizon, le modèle UMIDAS avec les données Google Trends (UMIDAS.GT) montre une performance inférieure au modèle AR(p), avec des ratios EQM tombant en dessous de 1. Cela pourrait s'expliquer par une potentielle surspécification ou une dépendance excessive aux données à haute fréquence.

Le modèle RF montre des résultats intéressants. Dans plusieurs cas, ce modèle présente des performances comparables, voire légèrement meilleures que celles du modèle AR(p). Les avantages du modèle RF par rapport au modèle AR(p) résident probablement dans sa capacité à capturer des relations non linéaires et des interactions complexes entre les variables, ce qui peut être particulièrement utile dans un environnement économique stable mais avec des fluctuations subtilement influencées par des variables exogènes.

Néanmoins, aux horizons très courts (proches de la publication), le modèle RF ne montre pas de supériorité significative par rapport au modèle AR(p), ce qui pourrait indiquer que, dans des contextes où les relations linéaires deviennent plus dominantes, le modèle RF perd une

partie de son avantage.

Le modèle GBM affiche des résultats mixtes. Bien qu'il montre parfois une performance supérieure au modèle AR(p), en particulier 12 semaines, 5 semaines et 4 semaines avant la publication du PIB, il y a aussi des cas où il sous-performe. Cela pourrait être attribué à la sensibilité du modèle GBM au surajustement, surtout dans des périodes où les données économiques sont relativement stables. Le modèle GBM est conçu pour maximiser l'ajustement des données, ce qui peut conduire à une meilleure performance dans des contextes volatils, mais il peut également capturer du bruit lorsqu'il est appliqué dans des environnements plus prévisibles.

Enfin, le modèle LASSO montre des performances variées par rapport au modèle AR(p). Aux horizons éloignés, le modèle LASSO (en particulier avec les données Google Trends) dépasse généralement le modèle AR(p), ce qui indique que ce modèle est capable de sélectionner les variables les plus pertinentes pour la prévision à long terme. Toutefois, aux horizons plus courts, l'avantage du modèle LASSO s'atténue, et dans certains cas, il sous-performe par rapport au modèle AR(p). Cela pourrait être dû à la nature parcimonieuse du LASSO, qui peut éliminer des variables potentiellement importantes lorsque les données deviennent plus riches à mesure que l'on se rapproche de la publication.

Il est important de noter que l'absence du choc de la COVID-19 dans cette analyse modifie significativement le contexte économique. Dans un environnement plus stable, comme celui de 2015 à 2019, les avantages des modèles capables de capturer des dynamiques complexes et non linéaires (comme le MIDAS, UMIDAS, et RF) sont amplifiés, car ils peuvent exploiter efficacement les variations subtiles des indicateurs économiques. En revanche, les modèles plus simples comme AR(p) pourraient ne pas capturer ces dynamiques aussi efficacement, mais leur robustesse face à des fluctuations économiques plus modérées leur permet de rester compétitifs, voire supérieurs dans certains cas.

En conclusion, cette analyse montre que, même en l'absence de chocs économiques majeurs, les modèles MIDAS et UMIDAS continuent de surpasser le modèle AR(p), en particulier aux horizons éloignés, grâce à leur capacité à intégrer des données à haute fréquence et à capturer des dynamiques économiques complexes. Toutefois, à mesure que l'on se rapproche de la publication du PIB, les avantages de ces modèles tendent à diminuer, laissant place à une plus grande pertinence des modèles plus traditionnels et linéaires comme AR(p).

Conclusion

L'objectif principal de ce mémoire est d'évaluer l'apport des données Google Trends dans la prévision du PIB québécois en temps réel et de déterminer si ces données peuvent améliorer la précision des modèles économétriques dans différents contextes économiques. Nous avons comparé plusieurs modèles, incluant MIDAS, UMIDAS, RF, GBM et LASSO, avec et sans l'intégration des données de Google Trends, tout en les confrontant à un modèle de référence autoregressif AR(p).

Les résultats de notre étude montrent que l'intégration des données de Google Trends apporte une valeur ajoutée significative aux modèles de prévision en période de volatilité économique accrue, comme celle causée par la pandémie de COVID-19. En particulier, les modèles UMIDAS et MIDAS montrent des améliorations notables aux horizons éloignés de la publication du PIB, suggérant que les données de Google Trends captent des signaux anticipateurs pertinents qui ne sont pas immédiatement visibles dans les données économiques traditionnelles. Les données de Google Trends se révèlent utiles pour améliorer les performances des modèles, notamment lorsque les décisions économiques doivent être prises rapidement dans un environnement incertain.

Cependant, nos résultats indiquent également que cette valeur ajoutée n'est pas uniforme et dépend du contexte économique et de l'horizon de prévision. Dans un environnement plus stable, tel que celui observé entre 2015 et 2019, l'intégration des données de Google Trends dans les modèles économétriques n'a pas systématiquement amélioré la précision des prévisions. En effet, dans plusieurs cas, notamment avec les modèles UMIDAS, l'ajout de ces données a même pu dégrader les performances, suggérant une potentielle surspécification ou une dépendance excessive à des signaux qui ne sont pas toujours pertinents en l'absence de perturbations économiques majeures.

Les modèles MIDAS et UMIDAS se sont particulièrement distingués par leur capacité à exploiter des données à haute fréquence et à capturer des dynamiques économiques complexes, surpassant le modèle de référence AR(p) dans de nombreux cas, en particulier aux horizons

éloignés. Ces résultats montrent que ces modèles sont bien adaptés aux prévisions dans des contextes où les informations économiques disponibles sont limitées et où il est crucial de réagir aux changements économiques le plus rapidement possible. Toutefois, lorsque l'on se rapproche de la date de publication du PIB, les avantages de ces modèles diminuent, laissant une plus grande place aux modèles plus traditionnels, tels que le modèle AR(p), pour lesquels les données à basse fréquence plus fiables deviennent prédominantes.

En conclusion, l'étude suggère que l'intégration des données de Google Trends dans des modèles économétriques peut être un outil puissant pour le nowcasting, mais cette pertinence est fortement contextuelle. Les résultats suggèrent que l'utilisation de ces données est particulièrement bénéfique dans des périodes de volatilité économique élevée, où les comportements des agents économiques peuvent changer rapidement et ne pas être entièrement captés par les indicateurs traditionnels. Cependant, dans des environnements économiques stables, les modèles traditionnels sans données supplémentaires semblent suffire pour capter les dynamiques sous-jacentes. Ainsi, l'application des données de Google Trends dans les modèles économétriques doit être adaptée aux conditions économiques spécifiques et aux horizons de prévision considérés.

Cette recherche ouvre la voie à de futures études qui pourraient approfondir l'intégration de données non traditionnelles, comme celles de Google Trends, dans des modèles économétriques avancés. Les études futures pourraient explorer des méthodes pour optimiser la sélection des catégories de recherche ou combiner ces données avec d'autres sources en temps réel pour affiner davantage les prévisions économiques. En somme, bien que Google Trends ne remplace pas les données économiques traditionnelles, il offre une source complémentaire précieuse pour capturer les tendances émergentes et affiner les prévisions à court terme, surtout dans des environnements incertains.

Bibliographie

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

Bantis, E., Clements, M. P., and Urquhart, A. (2021). Forecasting GDP Growth Rates Using Google Trends in the United States and Brazil. *SSRN Electronic Journal*.

Bleher, J. and Dimpfl, T. (2022). Knitting Multi-Annual High-Frequency Google Trends to Predict Inflation and Consumption. *Econometrics and Statistics*, 24 :1–26.

Borup, D., Rapach, D. E., and Schutte, E. C. M. (2021). Mixed-Frequency Machine Learning : Now- and Backcasting Weekly Initial Claims with Daily Internet Search-Volume Data.

Breiman, L. (2001). Random Forests. *Kluwer Academic Publishers*, 45 :5–32.

Choi, H. and Varian, H. (2009). Predicting the Present with Google Trends.

Cleveland, W. S., Grosse, E., and Shyu, W. M. (1992). Local Regression Models. In *Statistical Models in S*, page 68. 1st edition.

Couture, H. (2020). Prévision de l'activité du marché du travail aux états-Unis à l'aide des données de Google trends. Master's thesis, Université du Québec à Montréal.

Couture, H. and Stevanovic, D. (2021). Analyse du marché du travail à l'aide des données de Google Trends.

Diebold, F. X. and Mariano, R. S. (1995). Comparing Predictive Accuracy. *Journal of Business Economic Statistics*, Vol. 20, No. 1, Twentieth Anniversary Commemorative Issue (Jan., 2002) :pp. 134–144.

Doz, C., Giannone, D., and Reichlin, L. (2011). A two-step estimator for large approximate dynamic factor models based on Kalman filtering. *Journal of Econometrics*, 164(1) :188–205.

Ferrara, L. and Simoni, A. (2023). When are Google data useful to nowcast GDP ? An approach via pre-selection and shrinkage. *Journal of Business & Economic Statistics*, 41(4) :1188–1202. arXiv :2007.00273 [econ].

Foroni, C., Marcellino, M., and Schumacher, C. (2015). Unrestricted Mixed Data Sampling (MIDAS) : MIDAS Regressions with Unrestricted Lag Polynomials. *Journal of the Royal Statistical Society Series A : Statistics in Society*, 178(1) :57–82.

Foroni, C., Marcellino, M., and Stevanovic, D. (2019). Mixed-frequency models with moving-average components. *Journal of Applied Econometrics*, 34(5) :688–706.

Fortin-Gagnon, O., Leroux, M., Stevanovic, D., and Surprenant, S. (2022). A large Canadian database for macroeconomic analysis. *Canadian Journal of Economics/Revue canadienne d'économique*, 55(4) :1799–1833.

Ghysels, E., Santa-Clara, P., and Valkanov, R. (2004). The MIDAS Touch : Mixed Data Sampling Regression Models.

Ghysels, E., Sinko, A., and Valkanov, R. (2007). MIDAS Regressions : Further Results and New Directions. *Econometric Reviews*, 26(1) :53–90.

Giannone, D., Reichlin, L., and Sala, L. (2005). Monetary Policy in Real Time. *MIT Press*, 19.

Giannone, D., Reichlin, L., and Small, D. (2008). Nowcasting : The real-time informational content of macroeconomic data. *Journal of Monetary Economics*, 55(4) :665–676.

Götz, T. B. and Knetsch, T. A. (2019). Google data in bridge equation models for German GDP. *International Journal of Forecasting*, 35(1) :45–66.

Goulet Coulombe, P., Leroux, M., Stevanovic, D., and Surprenant, S. (2021a). Macroeconomic data transformations matter. *International Journal of Forecasting*, 37(4) :1338–1354.

Goulet Coulombe, P., Leroux, M., Stevanovic, D., and Surprenant, S. (2022). How is machine learning useful for macroeconomic forecasting ? *Journal of Applied Econometrics*, 37(5) :920–964.

Goulet Coulombe, P., Marcellino, M., and Stevanović, D. (2021b). CAN MACHINE LEARNING CATCH THE COVID-19 RECESSION ? *National Institute Economic Review*, 256 :71–109.

Kuzin, V., Marcellino, M., and Schumacher, C. (2011). MIDAS vs. mixed-frequency VAR : Nowcasting GDP in the euro area. *International Journal of Forecasting*, 27(2) :529–542.

Lahiri, K. and Yang, C. (2022). Boosting tax revenues with mixed-frequency data in the aftermath of COVID-19 : The case of New York. *International Journal of Forecasting*, 38(2) :545–566.

Lee, T.-H., White, H., and Granger, C. W. (1993). Testing for neglected nonlinearity in time series models. *Journal of Econometrics*, 56(3) :269–290.

McCracken, M. W. and Ng, S. (2016). FRED-MD : A Monthly Database for Macroeconomic Research.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). A Sparse-Group Lasso. *Journal of Computational and Graphical Statistics*, 22(2) :231–245.

Stock, J. H. and Watson, M. W. (2002). Forecasting Using Principal Components From a Large Number of Predictors. *Journal of the American Statistical Association*, 97(460) :1167–1179.

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society Series B : Statistical Methodology*, 58(1) :267–288.