# Prédire à l'aide de Google Trends

Dalibor Stevanovic

Université du Québec à Montréal et CIRANO

12 novembre 2021

# Motivations

- Puisque l'Internet est si largement présent dans notre quotidien, la question se pose à savoir si à partir de ces données de navigation nous sommes en mesure de générer des connaissances sur l'activité macroéconomique.
- Ces données sont disponibles très rapidement et permettent d'effectuer de la prévision en temps réel.
- Cette propriété peut être spécialement utile en période de crise. La récente période de la Covid-19 en est un bon exemple.

# Revue de littérature

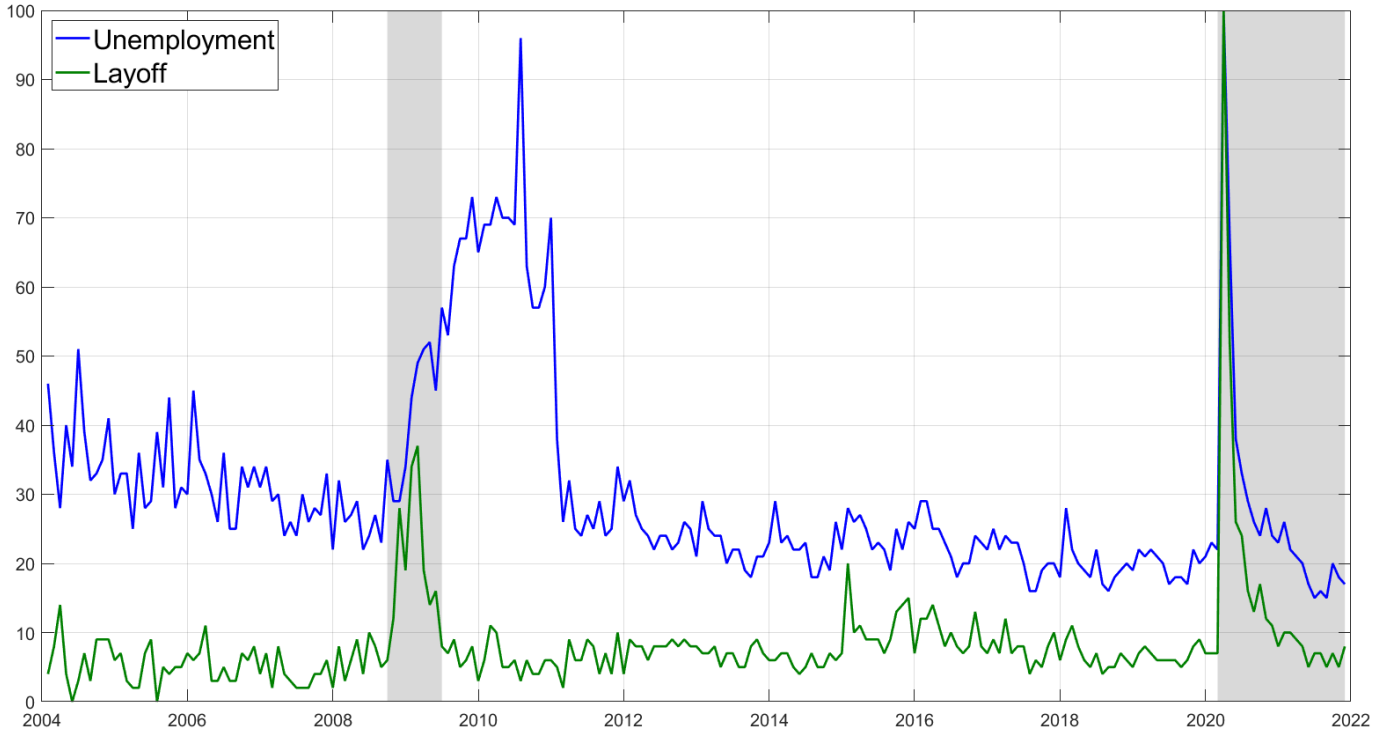
La littérature en ce qui concerne l'utilisation des GT pour la prévision macroéconomique n'est naturellement qu'assez récente. Voici une idée de leur utilisation en matière de prévisions :

- Choi et Varian (2009) - Demande assurance chômage et vente automobile
- Askitas et Zimmermann (2009), D'Amuri (2009), Suhoy (2009), D'Amuri et Marcucci (2017), Nagao *et al.* (2019), Maas (2020) - Taux de chômage
- Vosen et Schmidt (2009), Kholodilin *et al.* (2010) - Consommation privée
- Koop *et al.* (2013) - Production industrielle
- Guzman(2011), Seabold et Coppola (2015) - Inflation
- Kulkarni *et al.* (2009), McLaren et Shanbhogue (2011), Wu et Brynjolfsson (2015) - Prix des maisons
- Borup et Schütte (2020) - Emploi

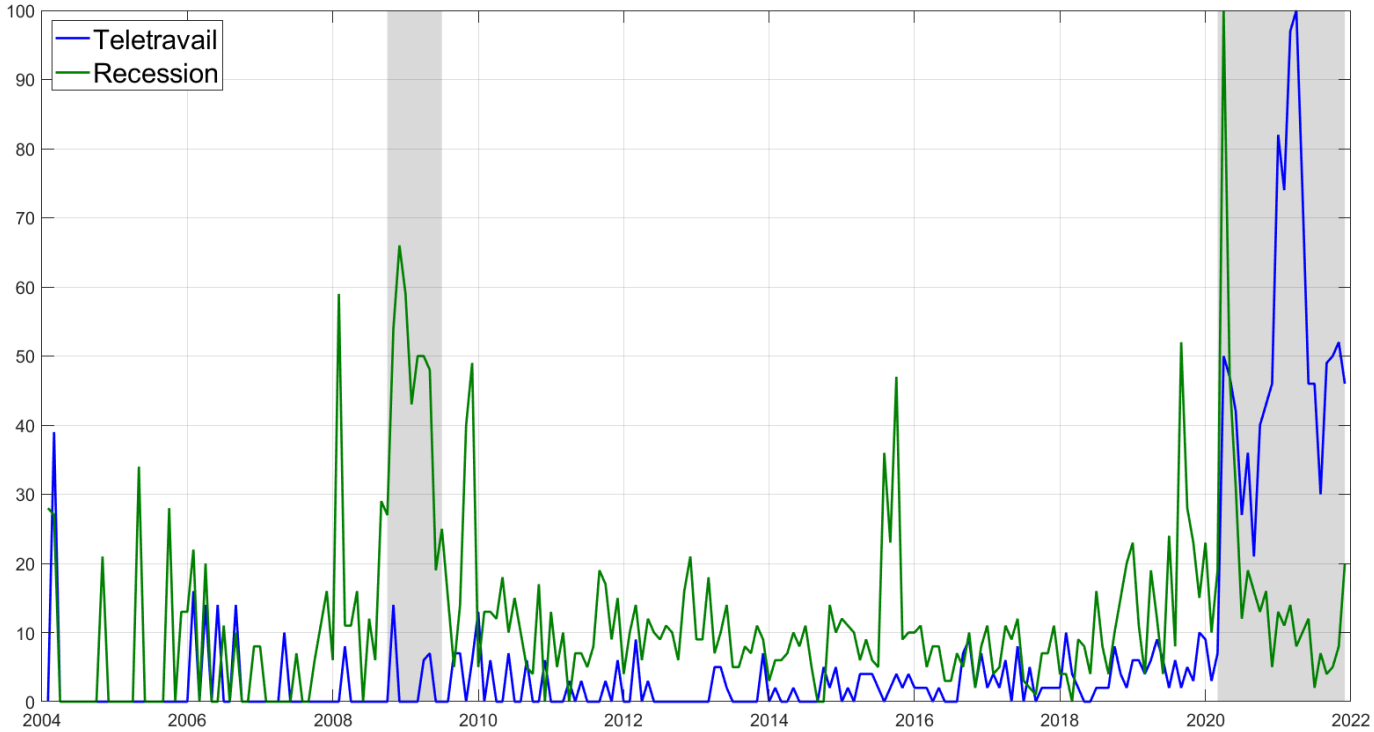
# Google Trends

- Google met à notre disposition un indice de popularité relative selon la période et la région géographique pour laquelle on désire recueillir les données. Par conséquent, ce n'est pas le volume de recherche qui est directement disponible.
- L'indice est normalisé entre 0 et 100. Les valeurs 0 et 100 représentent respectivement les points dans le temps où les recherches pour un mot-clé ont été les moins et plus populaires pour une période et région donnée.
- Les données sont disponibles de janvier 2004 à aujourd'hui.
- Disponible en temps réel en plusieurs fréquences : à la minute (sur une période restreinte) jusqu'à mensuel.

# Exemples de mots-clés pour le Canada



# Exemples de mots-clés pour le Québec



# Google Trends, suite

- Elles peuvent représenter l'intérêt ou l'attention que porte la population à un sujet précis, une chose habituellement particulièrement difficile à mesurer.
- Les données réagissent rapidement aux évènements d'actualités.
- Cette propriété qu'ont les données de Google Trends à être disponible très rapidement peut alors permettre de corriger les problèmes liés aux délais dans la publication officielle des données.
- Concrètement, cela offre ainsi la possibilité d'effectuer des prévisions pour la période en cours (*nowcasting*).

# Sélection des mots-clés

- Les possibilités de recherches de mots-clés sont quasiment illimitées.
- Une recherche massive de mots-clés effectuée à l'aide de l'outil de planification des mots-clés d'Adwords (Keywords Planner).
  - La plateforme met à disposition une large gamme de mots-clés en lien avec le sujet recherché. Par exemple, pour le sujet « emploi », cet outil propose des mots-clés comme « recherche emploi » ou encore « trouver du travail ». De plus, l'outil de planification propose des mots-clés selon la région géographique d'intérêt.
- Des milliers de mots-clés peuvent être sélectionnés de cette façon.

# Exercice de prévisions

Est-ce que les données de Google Trends sont réellement utiles pour la prévision ?

- Exercice en hors échantillon : 2014M9 - 2019M9 (60 mois)
- Horizons : H = 0 (*Nowcasting*) et 1 mois
- Variables mensuelles à prévoir pour le Canada et le Québec :
  - Emploi (15 ans + et 15 à 24 ans)
  - Emploi à temps plein (15 ans + et 15 à 24 ans)
  - Emploi à temps partiel (15 ans + et 15 à 24 ans)
  - Taux de chômage (15 ans + et 15 à 24 ans)
  - Taux d'emploi (15 ans + et 15 à 24 ans)
  - Heures totales travaillées (15 ans et plus)
- Utilisation de données de hebdomadaires de Google Trends. Les prévisions sont effectuées pour chaque semaine dans le mois.
- Critère d'évaluation : Erreur quadratique moyenne (EQM)

# Méthodes de « filtrage »

1. *Méthode par algorithme statistique*. Pour chaque série à prévoir, l'ensemble des GT seront inclus l'un à la suite de l'autre afin de tester leur pouvoir prédictif. Soit le modèle suivant :

$$y_{t+h} = \alpha + \rho y_{t+h-1} + \beta x_t + \epsilon_{t+h}$$

où $y_t$ est la cible et $x_t$ une série mensuelle de Google Trends. Sous $H_0$, $x_t$ n'a pas de pouvoir prédictif, i.e. $\hat{\beta} = 0$.

2. *Méthode par « intuition »*. Le but de cette méthode est de sélectionner manuellement un nombre restreint de mots-clés qui viendraient englober les grandes lignes du marché du travail. Les mots-clés sont ainsi choisis en fonction de leur utilisation dans la littérature et en raison de leur proximité avec les diverses variables à prévoir.

# Nuages de mots

Mots-clés sélectionnés à l'aide de l'algorithme :


**Canada**


**Québec**

# Modèles de prévisions

- MIDAS-AR

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

$b\left(j;\theta\right)$ est le polynôme exponentiel d'Almon

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

Les prédicteurs à haute fréquence sont contenus dans le vecteur de facteurs estimés $\hat{F}_t$. Les facteurs sont estimés par composante principale (PCA).

# Modèles de prévisions

- U-MIDAS-AR

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

Les prédicteurs à haute fréquence sont contenus dans le vecteur de facteurs estimés $\hat{F}_t$. Les facteurs sont estimés par composante principale (PCA).

# Modèles de prévisions

Modèles de régularisation : Les coefficients $\hat{\beta}$ des modèles minimisent les fonctions objectives suivantes :

- LASSO

$$\underset{\beta}{\operatorname{argmin}}\left\{\sum_{t=1}^{T}\left(Y_t - \beta_0 - \sum_{j=1}^{M}\beta_j z_{j,t}\right)^2 + \lambda\sum_{j=1}^{M}\left|\beta_j\right|\right\}$$

- Elastic-Net

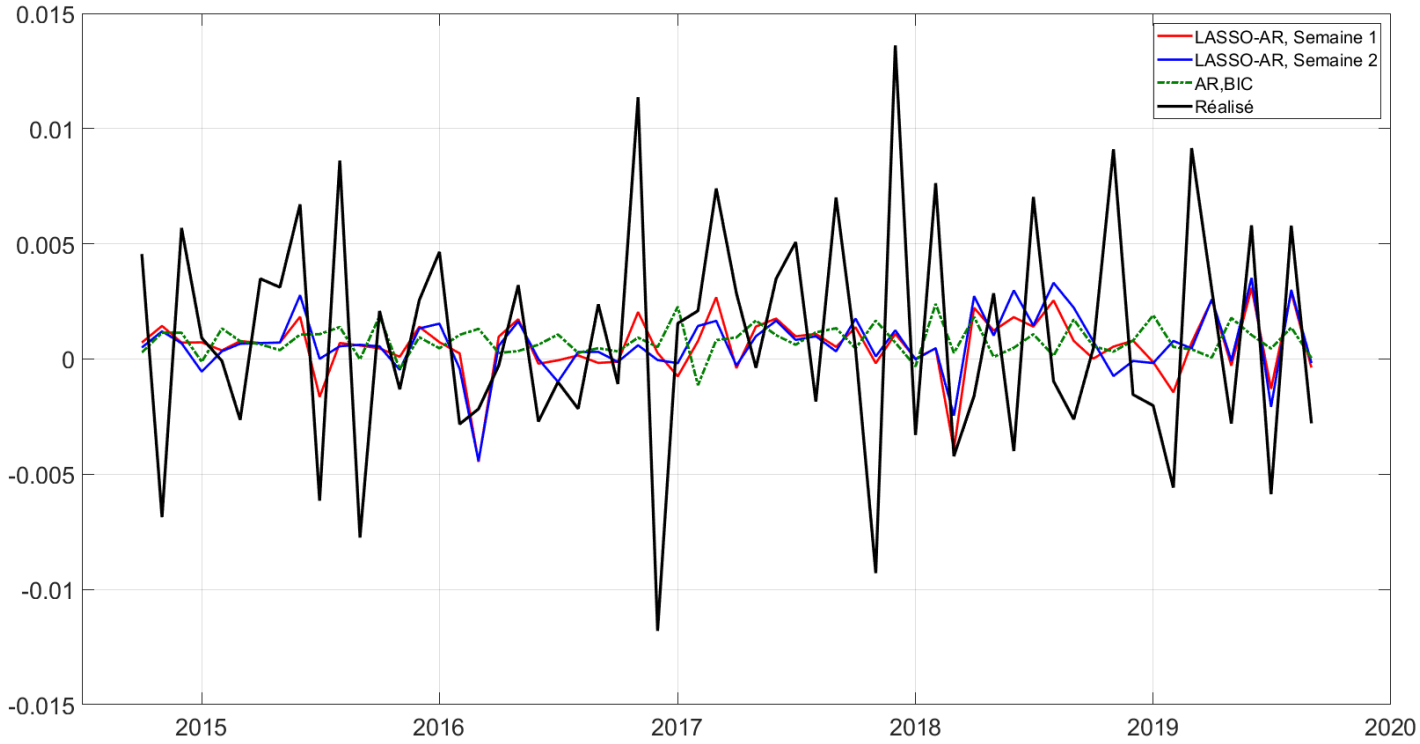
$$\underset{\beta}{\operatorname{argmin}}\left\{\sum_{t=1}^{T}\left(Y_t - \beta_0 - \sum_{j=1}^{M}\beta_j z_{j,t}\right)^2 + \lambda\sum_{j=1}^{M}(1-\alpha)\left|\beta_j\right|_1 + \alpha\left|\beta_j\right|^2\right\}$$

où $z_{j,t}$ représente un GT et $M = p_y + N \times p_x$. Autrement dit, la matrice $Z_t = [z_{1,t}, \ldots, z_{M,t}]$ comprend les $p_y$ retards de la cible et tous les GT et leurs $p_x$ retards.

# Résultats hors échantillon : Heures travaillées (Canada)

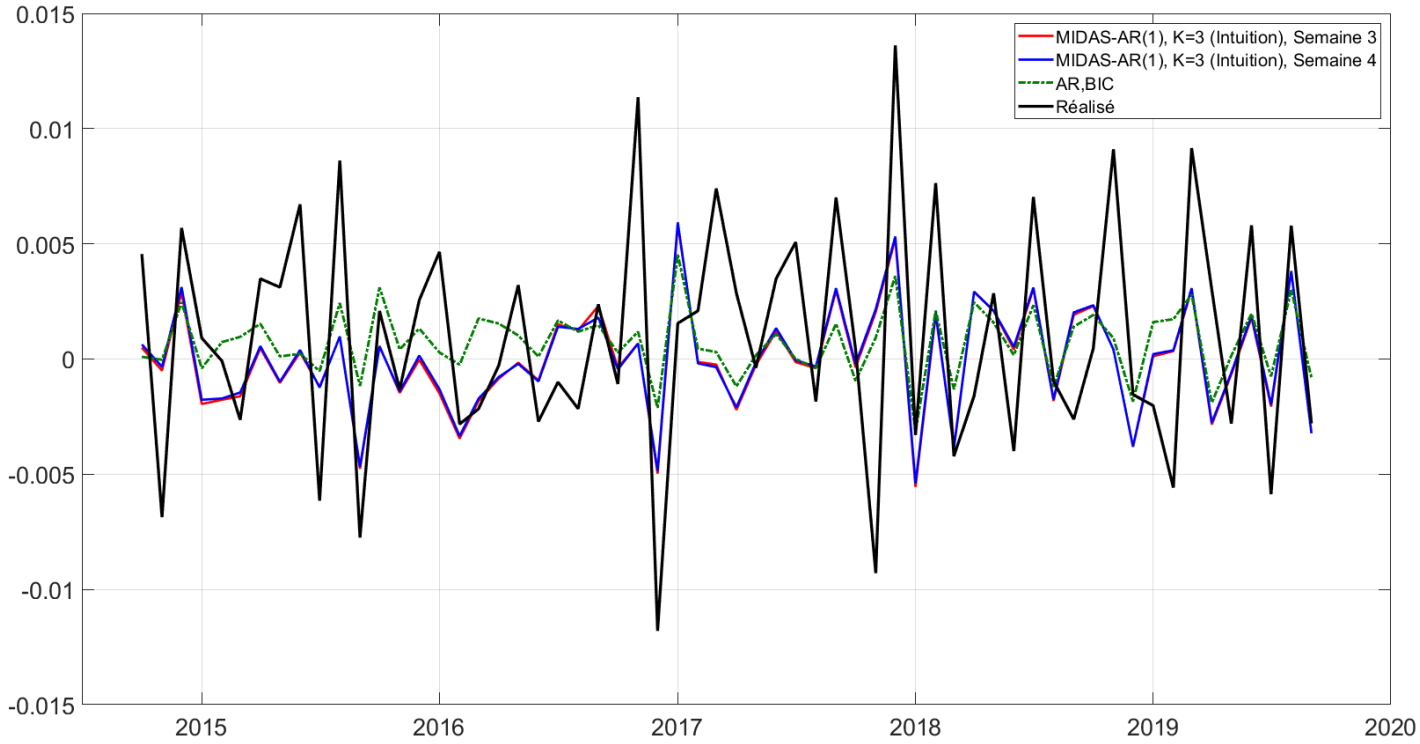
| | | H = 0 | | | | H = 1 | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Séries** | **Modèles** | **Semaine 1** | **Semaine 2** | **Semaine 3** | **Semaine 4** | **Semaine 1** | **Semaine 2** | **Semaine 3** | **Semaine 4** |
| **TOT HRS CAN** | MIDAS-AR(1), K=1 | 1.0176*** | 1.0174*** | 1.0194** | 1.0149*** | 1.0398 | 1.0333 | 1.2665*** | 1.2639*** |
| | MIDAS-AR(1), K=2 | 1.021** | 1.0296** | 1.0329** | 1.0236** | 1.051* | 1.0511 | 1.2834*** | 1.2713*** |
| | MIDAS-AR(1), K=3 | 1.0232** | 1.0368** | 1.0444** | 1.0435** | 1.0397 | 1.0538 | 1.2575*** | 1.2833*** |
| | MIDAS-AR(1), K=BIC | 1.0176*** | 1.0174*** | 1.0194** | 1.0149*** | 1.0398 | 1.0333 | 1.2665*** | 1.2639*** |
| | MIDAS-AR(1) filter, K=1 | 1.0236*** | 1.0229*** | 1.0293*** | 1.036*** | 1.0428 | 1.0453 | 1.2763*** | 1.252*** |
| | MIDAS-AR(1) filter, K=2 | 1.0153 | 1.0333** | 1.0426** | 1.0619*** | 1.0431* | 1.0495* | 1.2901*** | 1.2658*** |
| | MIDAS-AR(1) filter, K=3 | 0.9816 | 1.0148 | 0.9899 | 1.0034 | 1.0631* | 1.0356 | 1.3178*** | 1.2979*** |
| | MIDAS-AR(1) filter, K=BIC | 1.0236*** | 1.0229*** | 1.0293*** | 1.036*** | 1.058** | 1.0497 | 1.2731*** | 1.2552*** |
| | U-MIDAS-AR, K=1 | 1.013*** | 1.0124*** | 1.0179** | 1.0131*** | 1.0655* | 1.0447 | 1.2597*** | 1.2589*** |
| | U-MIDAS-AR, K=2 | 1.0057 | 1.0201* | 1.0302** | 1.0216 | 1.0348 | 1.0263 | 1.2547*** | 1.2582*** |
| | U-MIDAS-AR, K=3 | 1.008 | 1.0227** | 1.0358** | 1.0238 | 1.0363 | 1.0291 | 1.2557*** | 1.2657*** |
| | U-MIDAS-AR, K=BIC | 1.013*** | 1.0124*** | 1.0179** | 1.0131*** | 1.0655* | 1.0447 | 1.2597*** | 1.2589*** |
| | U-MIDAS-AR filter, K=1 | 1.0088** | 1.0105*** | 1.0241** | 1.0158* | 1.0334 | 1.0331 | 1.2532*** | 1.2523*** |
| | U-MIDAS-AR filter, K=2 | 0.9998 | 1.0224** | 1.0382** | 1.0356** | 1.0336 | 1.0398 | 1.256*** | 1.2555*** |
| | U-MIDAS-AR filter, K=3 | 0.9764 | 1.0264** | 1.0407*** | 1.0297** | 1.024 | 1.0384 | 1.2535*** | 1.2616*** |
| | U-MIDAS-AR filter, K=BIC | 1.0122*** | 1.0489 | 1.0617 | 1.0577* | 1.0248 | 1.0354 | 1.274*** | 1.2693*** |
| | LASSO-AR | **0.856*** ** | **0.9269** | 1.1952** | 1.1457* | 1.059* | 1.1385*** | 1.3826*** | 1.3739*** |
| | ELASTIC-NET-AR | 0.9222** | 0.9545 | 1.1441* | 1.1132 | 1.0745** | 1.1301*** | 1.3562*** | 1.4119*** |
| | | | | | | | | | |
| | **INTUITION** | | | | | | | | |
| | MIDAS-AR(1), K=1 | 1.0183*** | 1.0154*** | 1.0177*** | 1.0148** | 1.0495* | 1.0402 | 1.2713*** | 1.2645*** |
| | MIDAS-AR(1), K=2 | 1.0285 | 1.0183 | 1.008 | 1.0067 | 1.0528* | 1.0387 | 1.2659*** | 1.2701*** |
| | MIDAS-AR(1), K=3 | 1.0019 | 1.0038 | **0.9371*** | **0.9339**** | 1.0372 | **1.006** | 1.2663*** | 1.2606*** |
| | MIDAS-AR(1), K=BIC | 1.0073 | 1.01 | 0.9371* | 0.9339** | 1.0372 | 1.0099 | 1.2724*** | 1.2679*** |
| | U-MIDAS-AR, K=1 | 1.0122*** | 1.0129*** | 1.0145** | 1.0118*** | 1.0355 | 1.0338 | 1.2604*** | 1.257*** |
| | U-MIDAS-AR, K=2 | 1.0285 | 1.0183** | 1.008 | 1.0067* | 1.0528 | 1.0387 | 1.2659*** | 1.2701*** |
| | U-MIDAS-AR, K=3 | 1.0019 | 1.0038* | 0.9371 | 0.9339** | 1.0372 | 1.006 | 1.2663*** | 1.2606*** |
| | U-MIDAS-AR, K=BIC | 1.0088** | 1.0105*** | 1.0241** | 1.0158* | 1.0334 | 1.0331 | 1.2532*** | 1.2523*** |
| | LASSO-AR | 0.9994 | 1.0445 | 1.2114** | 1.2462*** | **1.0165** | 1.0224 | 1.3284*** | 1.2822** |
| | ELASTIC-NET-AR | 1.0099 | 0.9844 | 1.2293** | 1.2489** | 1.0327 | 1.0243 | **1.2443***** | **1.2127**** |

# Avant publication




Prévision des heures travaillées pour le Canada (Semaine 1 et 2)

# Après publication

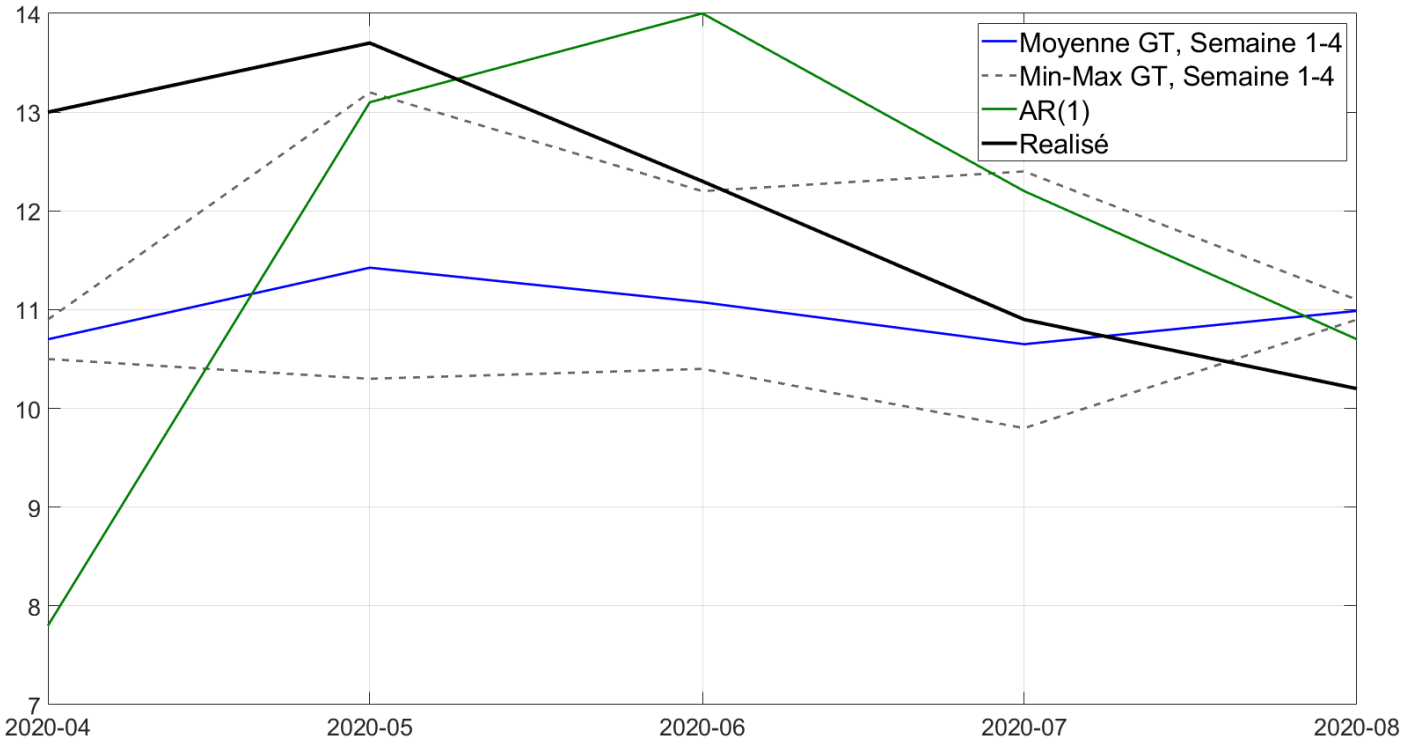



Prévision des heures travaillées pour le Canada (Semaine 3 et 4)

# Exercice de prévisions en temps réel (*Nowcasting*)

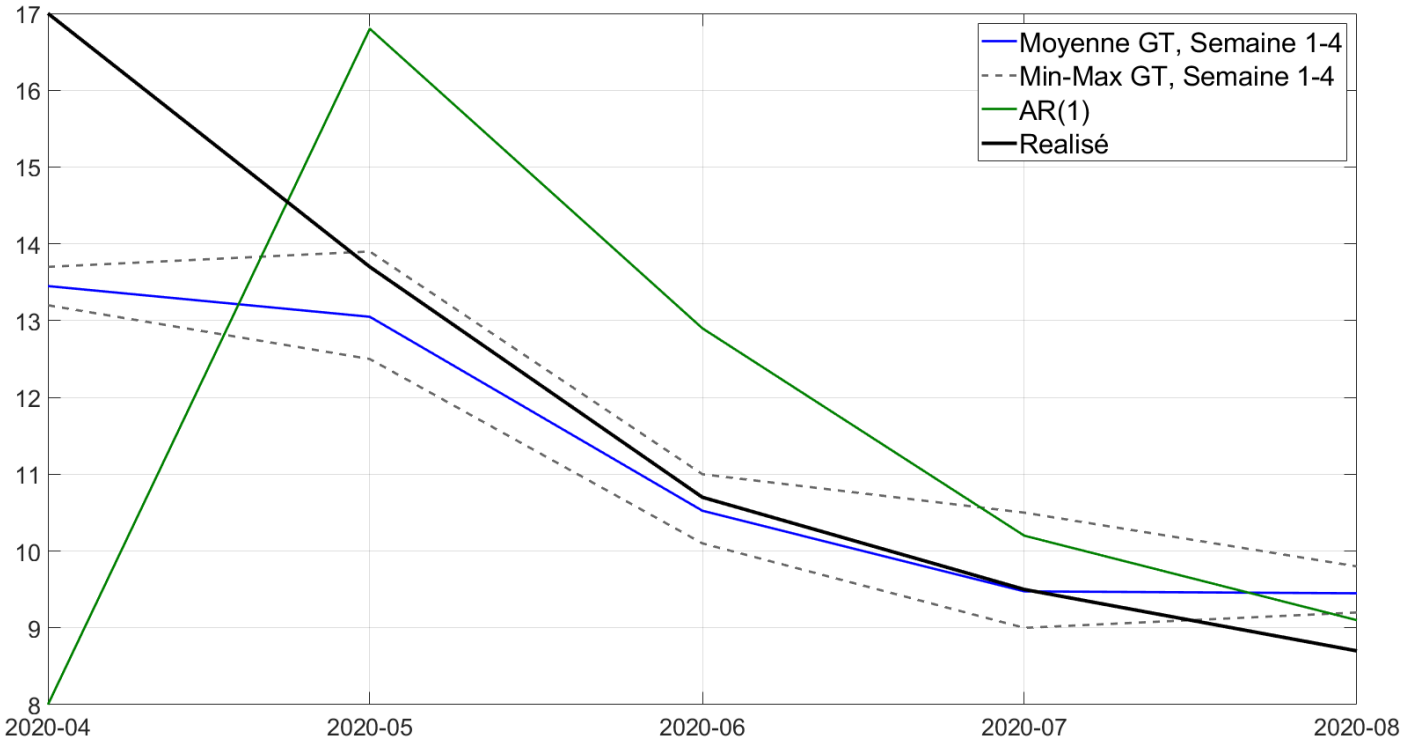
- Au début de la pandémie de la Covid-19, les données hebdomaires de Google Trends ont été utilisées afin de prévoir en temps réel l'emploi, le taux de chômage et les heures travaillées mensuelles pour le Canada et le Québec.
- Les données hebdomaires de Google Trends du mois en cours étaient utilisées afin de prévoir celui-ci (*Nowcasting*).
- Les prévisions étaient mises à jour hebdomadairement à l'aide de modèles à fréquences mixtes.

# Taux de chômage (Canada)




Prévision du taux de chômage pour le Canada (avec ARDI-GT, Ridge)

# Taux de chômage (Québec)




Prévision du taux de chômage pour le Québec (avec ARDI-GT, Ridge)

# Conclusions

- Les Google Trends peuvent contenir de l'information pertinente plus rapidement que les indicateurs standards.
- Elles peuvent également être particulièrement utiles lors des périodes de crise.
- Aussi, elles peuvent approximer l'information autrement difficilement mesurable par les indicateurs standards :
  - Sentiments
  - Anticipations