IAE Nantes
Économie & Management
Pôle Sociétés
Nantes Université

Évaluation de l'apport des données de recherche en ligne dans la prévision du syndrome grippal en France : application avec Google Trends

Master Économétrie et Statistique
Auteur : ALLAIN Adrien
Encadrant : Darne Olivier
Année académique : 2024 -2025

Cette étude compare des méthodes de prévision du syndrome grippal basées uniquement sur les données d'historique de la grippe à d'autres enrichis des recherches internet, obtenues via Google Trends. Les résultats montrent que les données de Google Trends renforcent la qualité des prévisions, notamment lors des variations atypiques, bien que des limites persistent lors des pics épidémiques les plus intenses.

Mots clés : Prévision, Syndrome grippal, Google Trends, données comportementales, France, R

Problématique et cadre de l'étude

La grippe est une infection respiratoire courante mais qui peut s'avérer grave. Elle est responsable chaque année de 2 à 6 millions de cas en France et de plusieurs centaines de milliers de décès dans le monde. Au-delà de son impact sanitaire, elle engendre des coûts économiques considérables proches du milliard d'euros en France, notamment en raison des consultations médicales, d'hospitalisations et d'absentéisme au travail. Face à ces enjeux, disposer de modèles prédictifs fiables est devenu essentiel pour la surveillance

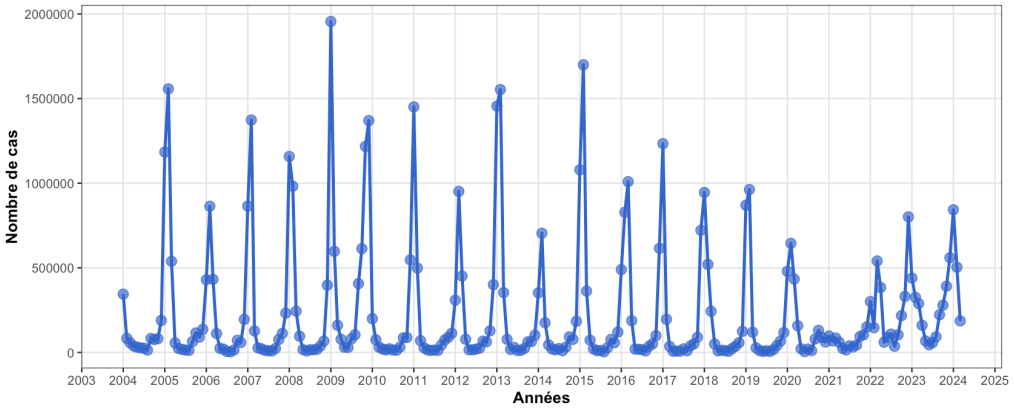
épidémiologique et notamment pour anticiper ce phénomène et permettre de mieux mobiliser les ressources hospitalières, la couverture vaccinale ou encore organiser au mieux les campagnes de prévention.

Les méthodes statistiques de prévision de séries temporelles ont fait leurs preuves dans le contexte de la grippe. Cependant, les approches basées uniquement sur l'incidence grippale peuvent manquer de réactivité face à des changements inattendus.
Dans ce contexte, l'essor du numérique et des données comportementales, comme les recherches internet, offre de nouvelles perspectives en termes de prévisions. La plateforme Google Trends, qui permet d'obtenir en temps réel l'évolution de la popularité des recherches Google, semble être un outil prometteur.

Dès lors, se pose la question de la valeur ajoutée des données issues de Google Trends dans l'amélioration des prévisions du syndrome grippal. Pour y répondre, cette étude propose une comparaison entre des modèles basés uniquement sur les données historiques d'incidence grippale et d'autres intégrant des données comportementales issues des recherches en ligne, en particulier Google Trends.

L'analyse repose sur une série mensuelle du nombre de cas de syndrome grippal en France, issue du Réseau Sentinelles, qui constitue la variable cible à prédire. Pour enrichir les modèles, des séries mensuelles représentant l'évolution de la popularité de recherches sur Google ont été ajoutées. Ces dernières portent sur des symptômes associés à la grippe ainsi que sur des termes génériques liés à l'épidémie. L'ensemble des séries couvre la période de 2004 à 2025.

Graphique 1 : Évolution temporelle de l'incidence de la grippe en France depuis 2004



Méthodologie

Plusieurs approches ont été mobilisées afin de comparer les performances de prévision du syndrome grippal. Des modèles univariés, fondés uniquement sur les données historiques d'incidence (X13-ARIMA-SEATS, SARIMA, Holt-Winters, ADAM, de décomposition et d'ensemble), ont été comparés à des modèles multivariés intégrant des données issues de Google Trends (RLM, SARIMAX…). Les variables exogènes ajoutées aux modèles ont été sélectionnées selon des approches fondées sur des critères d'information. Les performances des modèles ont été évaluées à l'aide de métriques telles que le RMSE, le MAPE et le MAE, ainsi qu'en comparant leur capacité à surpasser un modèle de référence de type naïf saisonnier. Des analyses graphiques ont également permis d'évaluer l'allure des prévisions par rapport à la réalité.

Résultats

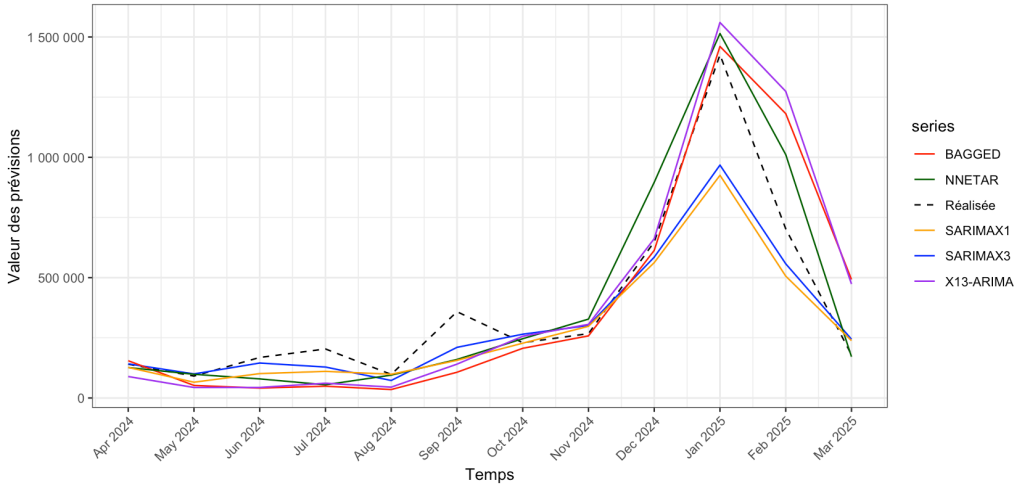
Tableau 1 : Indicateurs de performance des meilleurs modèles de chaque catégorie

Modèles	Type	RMSE	MAE	MAPE
BAGGED	Univarié	191 288	128 656	50.23
X13-ARIMA	Univarié	208 913	142 851	52.48
SARIMAX1	Multivarié (1 variable)	172 262	106 402	25.14
SARIMAX3	Multivarié (4 variables)	150 177	90 232	21.24
NNETAR	Multivarié (5 variables)	140 744	99 124	27.06

Le modèle BAGGED se distingue comme étant le meilleur modèle univarié, mais reste nettement dépassé par les modèles incluant des variables exogènes. Ce dernier présente un MAPE supérieur de 25 points à celui du modèle SARIMAX1, qui intègre seulement une variable (« Grippe »). Ce résultat souligne la valeur ajoutée des données comportementales en termes de performances. L'ajout de plusieurs variables améliore les résultats, comme le montre le modèle SARIMAX3, qui s'impose comme le meilleur modèle. Cependant, malgré ses performances, ce dernier a du mal à prédire le pic épidémique de janvier, contrairement aux modèles univariés et à NNETAR, qui parviennent à mieux l'anticiper. Par ailleurs, NNETAR offre également de très bonnes performances avec le RMSE le plus faible. À l'inverse, les variations atypiques en début de période sont mieux

captées par SARIMAX3 que par les autres modèles, illustrant la complémentarité entre modèles univariés et multivariés.

Graphique 11 : Comparaison des prévisions des meilleurs modèles et les valeurs observées



Conclusion

Les résultats montrent que l'ajout de données issues de Google Trends améliore nettement les performances prédictives. Toutefois, cette amélioration peut se faire au détriment de la précision lors de pics épidémiques importants, un élément critique en santé publique. Ces observations soulignent la complémentarité des approches. Les modèles univariés, comme X13-ARIMA-SEATS et BAGGED, offrent une modélisation robuste de la saisonnalité, tandis que les modèles intégrant des données comportementales, comme SARIMAX3, permettent de capter des signaux faibles et des variations atypiques. Ainsi, une combinaison de ces approches apparaît comme une voie prometteuse pour affiner les prévisions et renforcer leur fiabilité. De plus, les performances intéressantes et la capacité à anticiper le pic épidémique du réseau de neurones (NNETAR) relèvent de l'intérêt à explorer ce type d'approche plus complexe pour la prévision du syndrome grippal.

Bibliographie

Kandula, S., & Shaman, J. (2019). Near-term forecasts of influenza-like illness: An evaluation of autoregressive time series approaches. Epidemics, 27, 41–51. https://doi.org/10.1016/j.epidem.2019.01.002