ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

GES 824: TRAVAIL DE SYNTHÈSE
ANALYSE D'UNE STRATÉGIE D'INVESTISSEMENT BASÉE SUR L'OUTIL GOOGLE TRENDS

PAR
MANUEL BOLDUC

MONTRÉAL, le 13 mai 2024

TABLES DES MATIÈRES

Page

1. Introduction

Ce travail couvrira l'utilisation de données alternatives dans la conception de méthodes d'investissement algorithmiques. Notre hypothèse principale est que les sources de données alternatives (ou non conventionnelles) puissent contenir de l'information critique à la compréhension de l'évolution du cours de certains actifs financiers.

Nous ciblerons notamment à défricher les opportunités d'utilisation de sources de données alternatives pour des investisseurs individuels. Cela voudra donc dire que nous mettrons l'emphase sur des sources de données accessibles de manière libre, et que nous favoriserons l'extraction d'information à l'aide de méthodes pouvant être facilement implémentées sur un ordinateur personnel.

Au cours de l'article, nous répondrons donc à la question suivante: quelles stratégies d'investissement recommanderait-on à un investisseur individuel qui s'intéresse à l'extraction de signaux depuis des sources de données alternatives? Il est clair que la réponse à cette question pourrait faire l'objet d'une analyse beaucoup plus évoluée que la portée de ce travail de recherche. Nous serons cependant en mesure de présenter les grandes lignes des différentes méthodes existantes dans ce domaine, et de fournir des premières pistes au lecteur intéressé par le sujet.

Afin de justifier notre démarche, nous présenterons dans un premier temps les différentes catégories de données alternatives, ainsi que de l'information générale sur le phénomène récent du *Big data*. Nous présenterons par la suite divers cas d'usages de données alternatives dans des stratégies d'investissement.

Depuis les cas d'usages présentés, nous présenterons une analyse critique des différentes méthodologies proposées pour l'implémentation de stratégies d'investissement algorithmique, en prenant en compte la capacité de réappropriation de ces méthodes par un investisseur individuel. Nous mentionnerons également les principaux enjeux auxquels s'exposent l'investisseur intéressé par des stratégies basées sur des sources de données alternatives.

La contribution principale de notre article sera la mise en œuvre d'une stratégie d'investissement basée sur l'utilisation de données d'intérêt de mots-clés de recherche fournies par l'outil gratuit *Google Trends*. Nous présenterons premièrement une méthode originalement développée par Preis, Moat et Stanley (2013), puis nous évaluerons sa performance sur des périodes boursières plus récentes (backtests). Nous comparerons également le rendement de la méthode avec des méthodes d'investissement plus traditionnelles.

Nous conclurons l'article avec des recommandations pour l'investisseur individuel intéressé par l'extraction de signaux depuis des sources de données alternatives. Nous proposerons quelques pistes supplémentaires intéressantes à explorer pour pousser l'exercice plus loin.

2. Les sources de données alternatives

Le phénomène du *big data* est somme toute un événement récent dans l'histoire moderne. Selon *Wikipédia,* «[l]'explosion quantitative (et souvent redondante) des données numériques permet une nouvelle approche pour analyser le monde ».

Il va donc sans dire que la disponibilité nouvelle et croissante de ce type de données est d'intérêt pour la conception de nouvelles méthodes d'investissement algorithmique, si elles permettent l'accès à de l'information qui ne peut être obtenue depuis des sources de données plus conventionnelles. L'enthousiasme pour ces sources de données se traduit par une croissance constante des dépenses économiques pour y accéder et les analyser: elles sont prévues d'accroître annuellement de 12.8% dans la prochaine décennie (Jansen 2020).

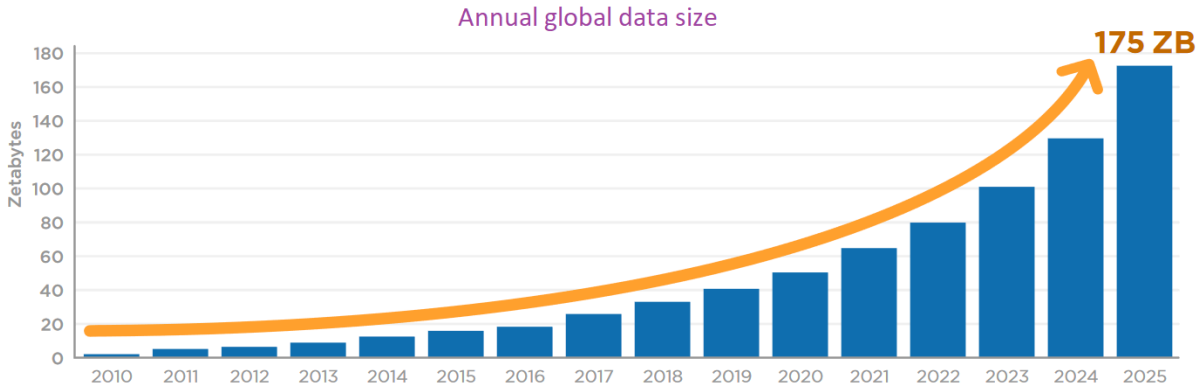

Figure 2.1 Image tirée de Reinsel, Gantz et Rydning (2018)

Dans le livre *Machine Learning for Algorithmic Trading*, Jansen (2020) dédie un chapitre complet aux types de données qu'il définit comme «alternatives». Il propose notamment de catégoriser ces types de données par les sources qui les produisent. Pour Jansen, les jeux de données alternatives sont principalement produites par:

- Des utilisateurs du web (par exemple, des avis de consommateurs, des données de moteurs de recherches, ou encore des données d'utilisation de réseaux sociaux);
- Des compagnies qui assurent le suivi de transaction commerciales (par exemple, des données de paiement de cartes de crédits, ou des compagnies qui agissent comme intermédiaire dans une chaîne d'approvisionnement);

- Des senseurs connectés (par exemple, des captures d'images par satellite, ou des données de température d'une région agricole)

En plus que les sources de production de ces données alternatives soit nombreuses, elles sont également de plus en plus disponibles sur des sources publiquement accessibles, ce qui facilite notamment leur utilisation pour développer des stratégies d'investissement algorithmique. Nous proposons donc un survol de différentes méthodes d'investissement algorithmiques basées sur des sources de données alternatives dans la prochaine section de ce travail de recherche.

3. Stratégies d'investissement basées sur les sources de données alternatives

Dans le rapport de J.P. Morgan intitulé *Big Data and AI Strategies* (2017), Kolanovic et Krishnamachari présentent des stratégies d'investissement algorithmiques reposant sur différents types de données alternatives. Nous synthétisons en quelques lignes les cas d'usages qu'ils présentent, afin de donner au lecteur intéressé une première idée des sortes de stratégies pouvant être mises en place à l'aide de sources de données alternatives provenant des trois différentes catégories présentées plus tôt (données d'utilisateurs du web, données de compagnies, données de senseurs connectés).

La conception de signaux d'investissement basés sur des données générées par des utilisateurs web passe principalement par des méthodes de *web scraping*, soit l'extraction de données de sites web à l'aide de scripts de programmation. Ce type de stratégie cherche à analyser le sentiment des investisseurs face au cours du marché en temps réel. Cela peut être effectué en évaluant la tendance des sentiments associées à différentes publications sur un

réseau social comme Twitter, par exemple. D'autre part, il est également possible d'analyser des flux de nouvelles en continu, et d'y associer des sentiments positifs ou négatifs par rapport au cours du marché, si certaines de ces nouvelles sont liées à des actifs financiers particuliers.

L'utilisation de données provenant de transactions commerciales, comme l'historique de transactions par cartes de crédits, permet de générer des signaux qui sont plus robustes dans le temps que ceux générés depuis des données d'utilisateurs web. Kolanovic et Krishnamachari (2017) présentent notamment une stratégie d'investissement algorithmique basée sur des données de transactions de certaines compagnies du S&P 500, l'idée étant de prendre une position longue sur les compagnies les plus dépensières et une position courte sur les compagnies les moins dépensières.

Finalement, l'utilisation de données de senseurs connectées permet de générer des signaux d'investissement basée sur l'activité économique perçue dans un certain secteur. Par exemple, il est possible de collecter de l'information sur la géolocalisation de téléphones cellulaires sans avoir à demander le consentement explicite de l'utilisateur. Kolanovic et Krishnamachari (2017) mettent en valeur une stratégie d'investissement basée sur les données de géolocalisation cellulaire afin de déterminer le taux de fréquentation de magasins Lululemon. Une hausse du taux de fréquentation permet alors de prédire une hausse dans le prix de l'action, si elle se traduit en une performance de vente qui n'était pas attendue par les investisseurs.

Il est à noter que plusieurs des exemples fournis dans cette section se basent sur des jeux de données provenant de compagnies qui se spécialisent dans la collecte de données. C'est donc

dire que ce ne sont pas des jeux de données disponibles publiquement, et qu'il y a un coût inhérent associé à la mise en place de ces stratégies. Cela peut déjà être un obstacle non-négligeable pour l'investisseur individuel qui souhaite évaluer les avantages et les inconvénients des différentes stratégies d'investissement proposées.

Par ailleurs, dans l'article *Big data is a big mess for hedge funds hunting signals*, Kishan (2016) offre un tour d'horizon des dangers potentiels de l'utilisation de données alternatives pour la génération de signaux d'investissement. Il note entre autres choses le danger de se fier à des sources de données de piètre qualité (comme celles obtenues depuis des réseaux sociaux - qui peuvent facilement être corrompues), ou encore l'utilisation de sources de données à la légalité douteuse (comme l'utilisation de la géolocalisation de téléphones cellulaires). Mais, surtout, il prévient que peu de firmes d'investissement ont jusqu'à présent réussi à générer des signaux d'investissement clair depuis l'utilisation de données massives, et que c'est un domaine de la finance qui doit encore atteindre une forme de maturité.

4. Élaboration de signaux d'investissement basée sur l'outil Google Trends

Suite à l'analyse des différents cas d'usages de la section 3, nous sommes en mesure de recommander une stratégie d'investissement algorithmique basée sur l'utilisation de données alternatives pour un investisseur individuel qui a accès à des ressources computationnelles limitées, et qui ne souhaite pas dépenser de l'argent pour acheter des jeux de données alternatifs.

Nous proposons donc à cet investisseur de bâtir une stratégie d'investissement algorithmique basée sur les données disponibles sur l'outil gratuit Google Trends.

4.1 L'outil Google Trends

Selon Berry (2023), «Google Trends est un outil gratuit de Google qui permet d'afficher l'intérêt d'une recherche au fil du temps. Google Trends permet de visualiser les tendances de recherche par lieu, terme de recherche, date, catégorie et type de recherche, et de comparer différentes tendances de recherche.»

Un survol rapide de l'interface de Google Trends permet de constater qu'il est facile d'utilisation, et qu'il permet de filtrer rapidement les données selon des critères spécifiques. Par ailleurs, un bouton nous propose l'extraction de données en format csv, ce qui permet de les traiter directement dans un environnement de développement Python, à l'aide du paquet Pandas.

Ces caractéristiques font de Google Trends un outil qui est utilisé dans plusieurs domaines de recherche (voir l'état de l'art de l'utilisation par Seung-Pyo et al. (2018), par exemple). Notre intérêt particulier pour cet outil est le fait qu'il rend très accessible des données générées par des utilisateurs web, sans devoir utiliser des méthodes de *web scraping*, et que celles-ci nécessitent un traitement minimal, puisqu'elles sont déjà normalisées.

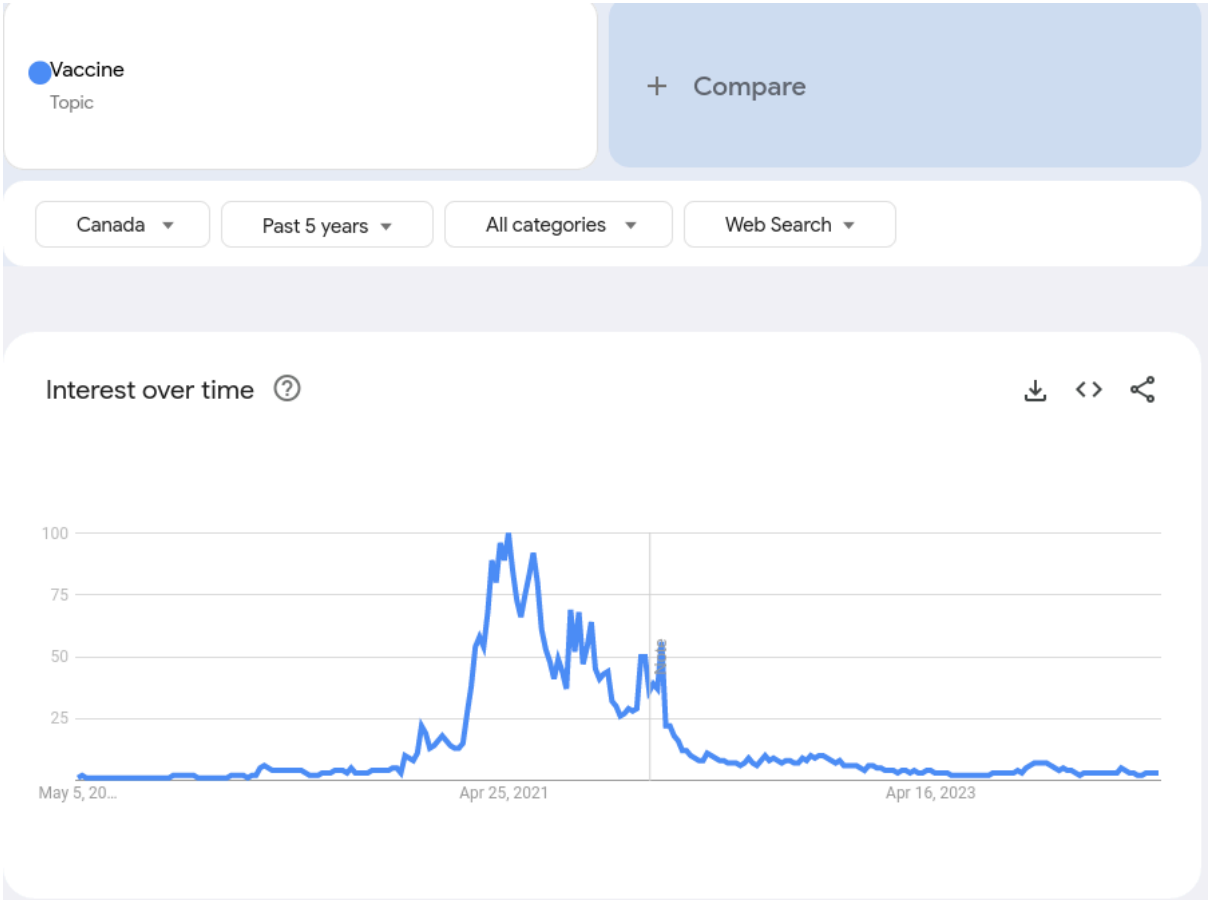

Figure 4.1 Capture d'écran d'une utilisation de l'outil Google Trend. On voit l'intérêt de recherche du mot-clé «vaccine» au Canada au cours des 5 dernières années. Comme on peut l'imaginer, l'intérêt de recherche de ce mot-clé était particulièrement élevé durant la pandémie.

4.2 Méthodologie de Preis, Moat et Stanley (2013)

Dans l'article intitulé *Quantifying Trading Behavior in Financial Markets Using Google Trends*, Preis, Moat et Stanley (2013), décrivent une stratégie d'investissement sur l'indice Dow Jones (DJIA), à partir du volume hebdomadaire de recherches Google sur différents mots-clés associés à l'univers financier.

La prémisse de leur article est la suivante: l'utilisation de données d'utilisateurs web peut offrir une compréhension nouvelle sur le comportement d'un marché affecté par des cours mouvementés, comme celui de la crise de 2008.

Les auteurs se basent sur deux principes importants. Le premier est dérivé de la théorie sur le processus décisionnel d'Herbert Simon (1974), voulant que le processus de décision d'un acteur rationnel débute par une collecte d'information. Le deuxième principe est que le processus de collecte d'information dans le monde contemporain passe majoritairement par la consultation de contenu diffusé sur l'internet.

La stratégie d'investissement proposée dans l'article combine donc ces deux principes fondamentaux. Comme l'outil Google Trends permet d'analyser les tendances de recherche, et plus particulièrement le volume absolu de recherches associés à des mots clés, il est possible de quantifier l'intérêt des utilisateurs web dans le temps face à certains actifs financiers.

Preis, Moat et Stanley (2013) ont donc analysé la performance de signaux d'investissement générés selon le volume de recherche de 98 mots-clés différents, au cours de la période boursière s'étalant du 5 janvier 2004 au 22 février 2011. La sélection des mots-clés s'est basée sur leur fréquence respective d'apparition dans les publications du *Financial Times* entre les mois d'août 2004 jusqu'à Juin 2011.

Les signaux d'investissements sont par la suite conçus en utilisant le taux de changement du volume relatif de recherche pour chaque mot clé $\Delta n(t,\ \Delta t)$, qui est défini mathématiquement

comme:

$$\Delta n(t,\ \Delta t)\ =\ n(t)\ -\ N(t\ -\ 1, \Delta t),$$

où $n(t)$ est défini comme le volume relatif de recherche du mot clé n au temps t (mesuré en semaines). $N(t\ -\ 1, \Delta t)$ est défini récursivement comme:

$$N(t\ -\ 1, \Delta t)\ =\ \frac{n(t-1) + n(t-2) + \ldots + n(t-\Delta t)}{\Delta t},$$

et permet l'amortissement du signal de tendance selon le mouvement du volume de recherche au cours des semaines précédentes.

La stratégie d'investissement algorithmique proposée se décline en deux possibilités. Lorsque le taux de changement du volume de recherche $\Delta n(t\ -\ 1,\ \Delta t)$ est positif, on anticipe une baisse de l'indice DJIA. Inversement, lorsque le taux de changement du volume de recherche $\Delta n(t\ -\ 1,\ \Delta t)$ est négatif, on anticipe une hausse de l'indice DJIA. Les technicalités de la mise en œuvre de la stratégie d'investissement seront décrites dans la prochaine section, avec des exemples concrets pour illustrer son fonctionnement.

5. Analyse et application de la stratégie d'investissement

Afin d'analyser la performance de la stratégie d'investissement présentée dans la section précédente, nous présentons deux exemples possibles de son implémentation. Nous commençons par reproduire les résultats de Preis, Moat et Stanley sur l'utilisation du mot-clé *debt* pour prédire le cours de l'indice DJIA entre le 5 janvier 2004 et le 22 février 2011. Par la

suite, nous évaluerons le rendement de cette même stratégie entre le 5 mai 2019 et le 5 mai 2024, afin de voir si la stratégie se généralise à d'autres périodes boursières.

Afin de présenter des résultats concrets, nous implémentons la stratégie d'investissement autour du fonds négocié en bourse (FNB) *SPDR Dow Jones Industrial Average*, qui se transige sous le sigle DIA. Lorsque le taux de changement du volume de recherche associé au mot-clé *debt* est négatif pour la semaine $t - 1$, on vendra le DIA au début de la semaine t, puis on l'achètera au début de la semaine $t + 1$, puisqu'on anticipe une baisse de l'indice DJIA. Inversement, lorsque le taux de changement du volume de recherche sera positif pour la semaine $t - 1$, on achètera le DIA au début de la semaine t puis on le vendra au début de la semaine $t + 1$.

Le retour cumulatif de notre stratégie sera alors calculé à l'aide de la formule du log retour. Si l'on représente le prix du DIA au début de la semaine t par la fonction $p(t)$, le retour cumulatif de notre méthode changera à chaque semaine soit par $\log(p(t)) - \log(p(t+1))$ lorsqu'on anticipe une baisse ($\Delta n(t-1, \Delta t) > 0$), ou par $\log(p(t+1)) - \log(p(t))$ lorsqu'on anticipe une hausse ($\Delta n(t-1, \Delta t) < 0$). Notons que nous présenterons nos résultats pour un Δt équivalent à trois semaines. Tous les résultats présentés sont par ailleurs accessibles dans un Jupyter Notebook disponible au lien suivant : https://github.com/bolducmanuel/google_trends_algo_trading/blob/main/google_trends_trading.ipynb.

5.1 Reproduction des résultats de Preis, Moat et Stanley

L'exemple le plus convaincant mis de l'avant par Preis, Moat et Stanley (2013) de l'implémentation de leur stratégie est celui de l'utilisation du mot-clé *debt* pour prédire le

cours de l'indice DJIA. Notre implémentation de leur stratégie donne effectivement lieu à des résultats probants, soit un rendement cumulatif de 373 % sur la période boursière s'étalant du 5 Janvier 2004 au 22 février 2011.

Dans la figure 5.1.2, nous comparons la stratégie d'investissement à une stratégie standard d'achat d'une position longue sur le titre DIA le 5 janvier 2004. Cette stratégie aurait en comparaison rapporté un rendement total de 13%, notamment en raison du creux associé à la récession de 2008. Afin de quantifier le risque, nous utilisons la méthode démontrée en classe de la valeur à risque historique. Cela nous permet de constater que le seuil de risque de notre méthode est relativement bas (3% de risque sur une intervalle de confiance de 95%). Cependant, il arrive que notre méthode génère des retours négatifs assez élevés, comme on peut le voir dans la figure 5.1.3.

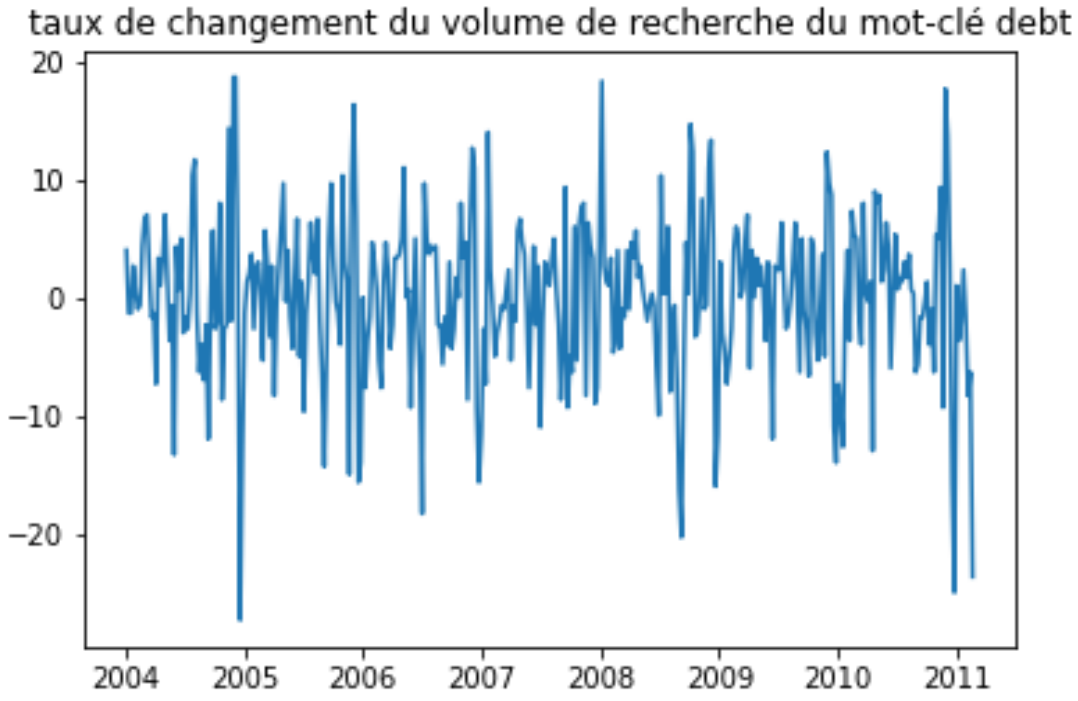

Figure 5.1.1 Taux de changement du volume de recherche du mot-clé *debt* aux États-Unis entre le 5 Janvier 2004 et le 22 février 2011. Le taux de changement est calculé avec un Δt égal à 3 semaines.

Figure 5.1.2 Retours cumulatifs sur la stratégie d'investissement entre le 5 janvier 2004 et le 22 février 2011. Notons que la majorité du rendement de la stratégie provient de la période plus mouvementée de la bourse autour de la récession de 2008.

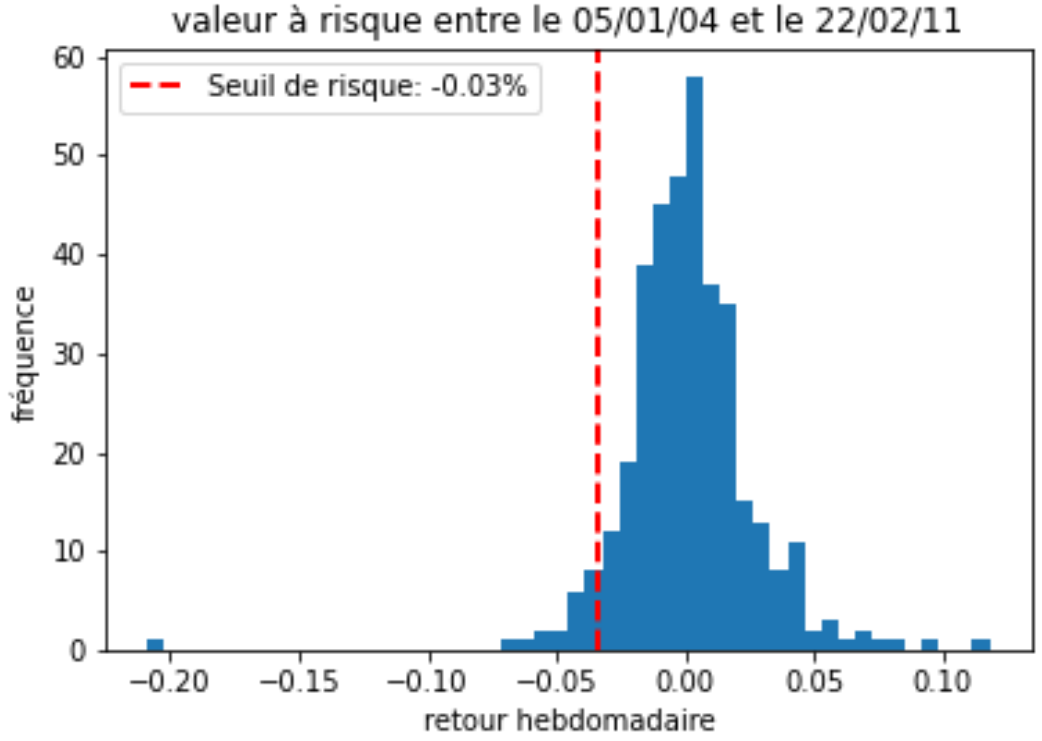

Figure 5.1.3 Histogramme des retours. La valeur à risque est de 3% pour un intervalle de confiance de 95%.

5.2 Résultats pour une période boursière récente

Ayant dorénavant confirmé que notre méthode est fonctionnelle, et que nous sommes en mesure de reproduire les résultats originaux de Preis, Moat et Stanley (2013), nous pouvons maintenant nous intéresser à explorer à l'utilisation de la stratégie d'investissement dans des contextes différents. Nous démontrons que la stratégie produit des résultats encourageants lorsque appliquée au volume de recherche du mot-clé *debt* entre le 5 mai 2019 et le 5 mai 2024.

Au cours de cette période, la stratégie d'investissement nous permet d'atteindre un rendement cumulatif de 137%. Au cours de la même période, une position longue sur le titre DIA nous aurait permis d'atteindre un rendement de 31% (voir figure 5.2.2). On constate que l'écart de performance entre les deux stratégies s'est resserré, même s'il reste significatif. Il est également important de constater que, même si la stratégie s'avère profitable sur toute la période, elle génère des pertes de plus de 100% à un certain point, ce qui voudrait donc dire qu'il aurait fallu injecter de l'argent supplémentaire en cours de route afin de maintenir la méthode à flot.

Notons par ailleurs que, même si le seuil de risque calculé demeure stable (3% à 95% d'intervalle de confiance), il arrive plus d'une fois que les retours hebdomadaires sont négatifs et significatifs (plus de 10% de baisse).

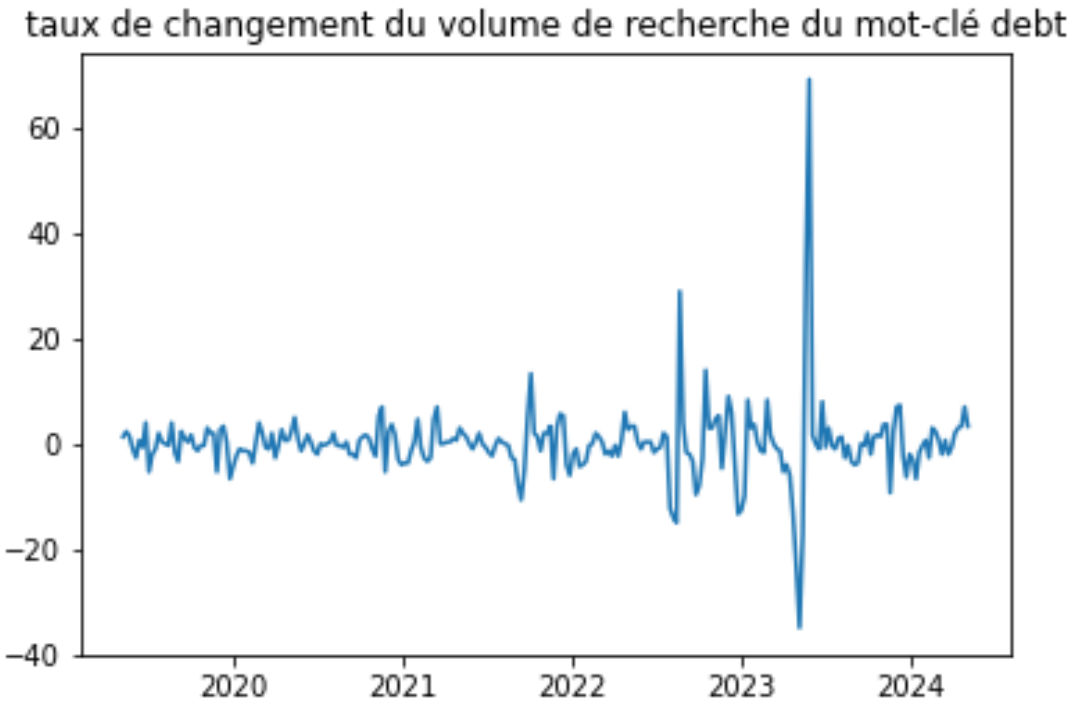

Figure 5.2.1 Taux de changement du volume de recherche du mot-clé *debt* aux États-Unis entre le 5 mai 2019 et le 5 mai 2024. Le taux de changement est calculé avec un Δt égal à 3 semaines.

Figure 5.2.2 Retours cumulatifs sur la stratégie d'investissement entre le 5 mai 2019 et le 5 mai 2024 . Notons que la stratégie d'investissement crée un rendement négatif de plus de 100% au début de la pandémie.

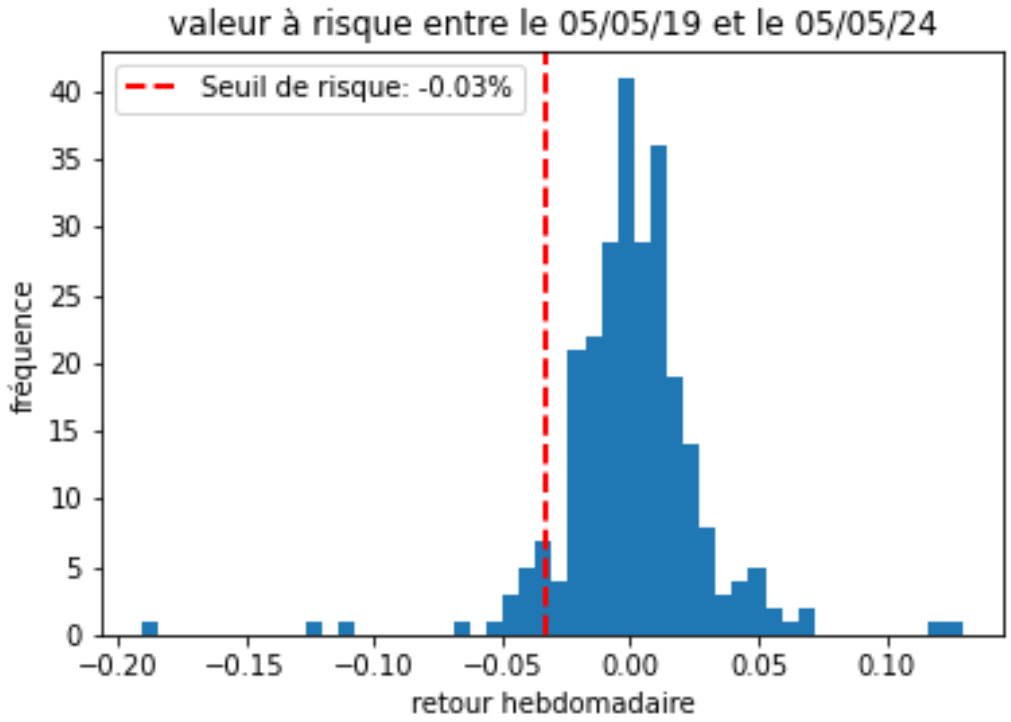

Figure 5.2.3 Histogramme des retours. La valeur à risque est de 3% pour un intervalle de confiance de 95%.

6. Conclusion

Nous avons démontré, au cours de cet article, qu'il est possible pour un investisseur individuel d'établir une stratégie d'investissement à l'aide de l'outil Google Trends. L'analyse de performance de la stratégie nous indique cependant que les signaux d'investissements générés peuvent être particulièrement erronés, même s'ils sont généralement peu risqués (valeur à risque de 3% pour un intervalle de confiance de 95%). Cela étant dit, nous avons délibérément concentré nos efforts sur la génération de signal, plutôt que sur une évaluation de rendement plus poussée de notre stratégie d'investissement. Nous considérons que cela permet une meilleure réappropriation de notre travail de recherche par un investisseur intéressé par l'utilisation de sources de données alternatives. Nous rappelons par ailleurs que tous nos résultats et notre méthodologie de programmation est

disponible au lien suivant: https://github.com/bolducmanuel/google_trends_algo_trading/blob/main/google_trends_trading.ipynb

Références bibliographiques

Berry, Sarah. “Qu’est-ce que Google Trends ?” SEO.com, https://www.seo.com/. Accédé le 13 mai 2024.

“Big data.” Wikipédia, 5 Apr. 2024. Wikipedia. Accédé le 13 mai 2024.

Google Trends. https://trends.google.com/trends/. Accédé le 13 mai 2024.

Jansen, Stefan. Machine Learning for Algorithmic Trading: Predictive Models to Extract Signals from Market and Alternative Data for Systematic Trading Strategies with Python. Second edition, Packt, 2020.

Jun, Seung-Pyo, et al. “Ten Years of Research Change Using Google Trends: From the Perspective of Big Data Utilizations and Applications.” Technological Forecasting and Social Change, vol. 130, May 2018, pp. 69–87. ScienceDirect, https://doi.org/10.1016/j.techfore.2017.11.009.

Kishan, Saijel. “Big Data Is a Big Mess for Hedge Funds Hunting Signals.” Bloomberg Professional Services, 13 Dec. 2016, https://www.bloomberg.com/professional/blog/big-data-big-mess-hedge-funds-hunting-signals/.

Kolanovic, Marko, and Rajesh Krishnamachari. Big Data and AI Strategies. JP Morgan, https://cpb.us/faculty.sites.uci.edu/dist/2/51/files/2018/05/JPM-2017-MachineLearningInvestments.pdf. Accédé le 13 mai 2024.

Preis, Tobias, et al. “Quantifying Trading Behavior in Financial Markets Using Google Trends.” Scientific Reports, vol. 3, no. 1, Apr. 2013, p. 1684. www.nature.com, https://doi.org/10.1038/srep01684.

Rydning, D. R. J. G. J., Reinsel, J., & Gantz, J. (2018). The digitization of the world from edge to core. *Framingham: International Data Corporation*, *16*, 1-28.

Simon, Herbert A. Administrative Behavior: A Study of Decision-Making Processes in Administrative Organization. 3d ed, Free Press, 1976