Utilisation de nouveaux indicateurs avancés pour le nowcasting

Comment intégrer de nouveaux indicateurs à nos étalonnages usuels ?

Insee Mesurer pour comprendre

1 CONTEXTE ET OBJECTIFS

1.A DURANT LA CRISE SANITAIRE

Utilisation d'une maquette mensuelle pour estimer le niveau d'activité :

- **Estimation par branche**
  - Jusqu'au niveau A129
  - Puis agrégation
- **Souvent peu d'informations quantitatives disponibles**
  - Quel effet des restrictions sanitaires sur l'activité d'une branche ?
- **Maquette utilisée entre avril 2020 et fin 2021.**
  - Difficilement conciliable avec la reprise de nos méthodes usuelles

Nouveaux indicateurs mobilisés :

- **Déclaration sociale nominative (DSN)**
  - Recours à l'activité partielle par branche
  - Heures rémunérées par branche
- **Dépenses par carte bancaire CB et données de caisse**
- **Consommation d'électricité par branche**
- **Indicateurs de mobilité**
  - Cerema : trafic routier dont poids lourds
  - Google Mobility Reports : fréquentation de différents types de lieux
  - Apple Maps Mobility

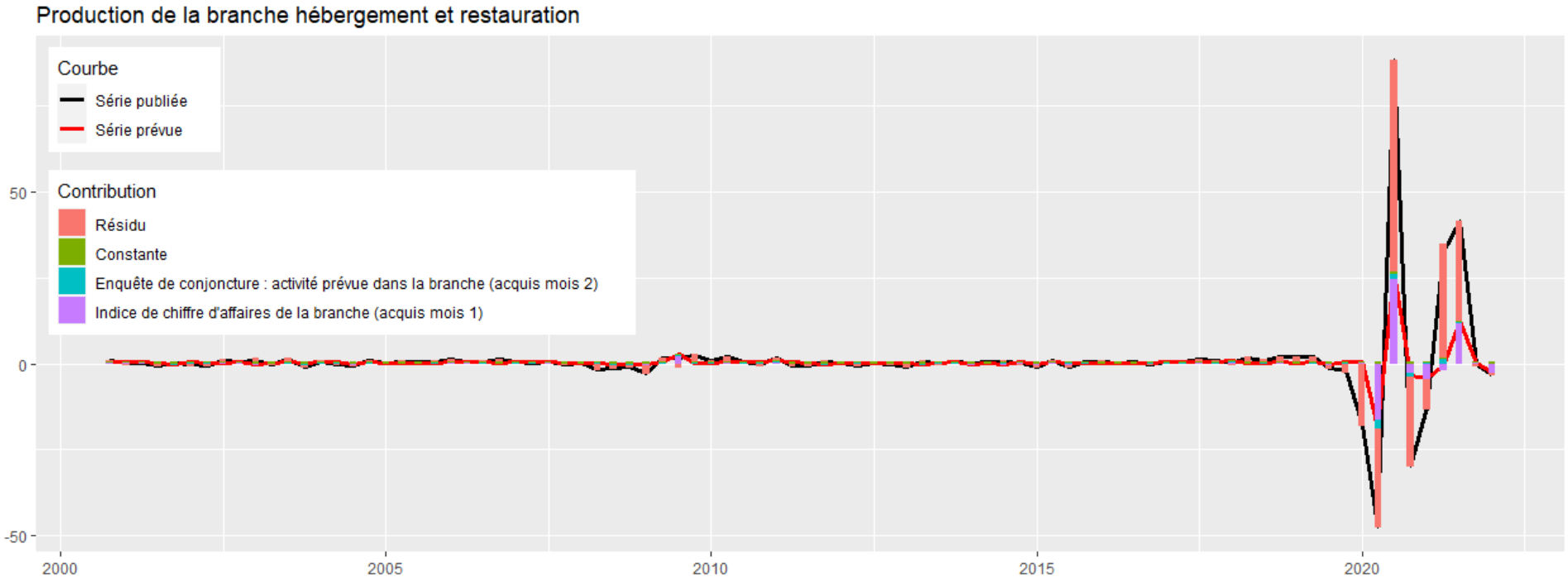
1.B ET APRÈS ?

Ces nouveaux indicateurs continuent d'apporter de l'information :

	L'indicateur est disponible rapidement après la fin du mois	Le champ mesuré par l'indicateur est très proche de la série à estimer	L'indicateur donne une information quantitative sur l'évolution de la variable d'intérêt	L'indicateur est disponible sur longue période
Enquêtes de conjoncture	X	X		X
Indicateurs traditionnels : IPI, ICA		X	X	X
Nouveaux indicateurs, dont à haute fréquence	X		X	

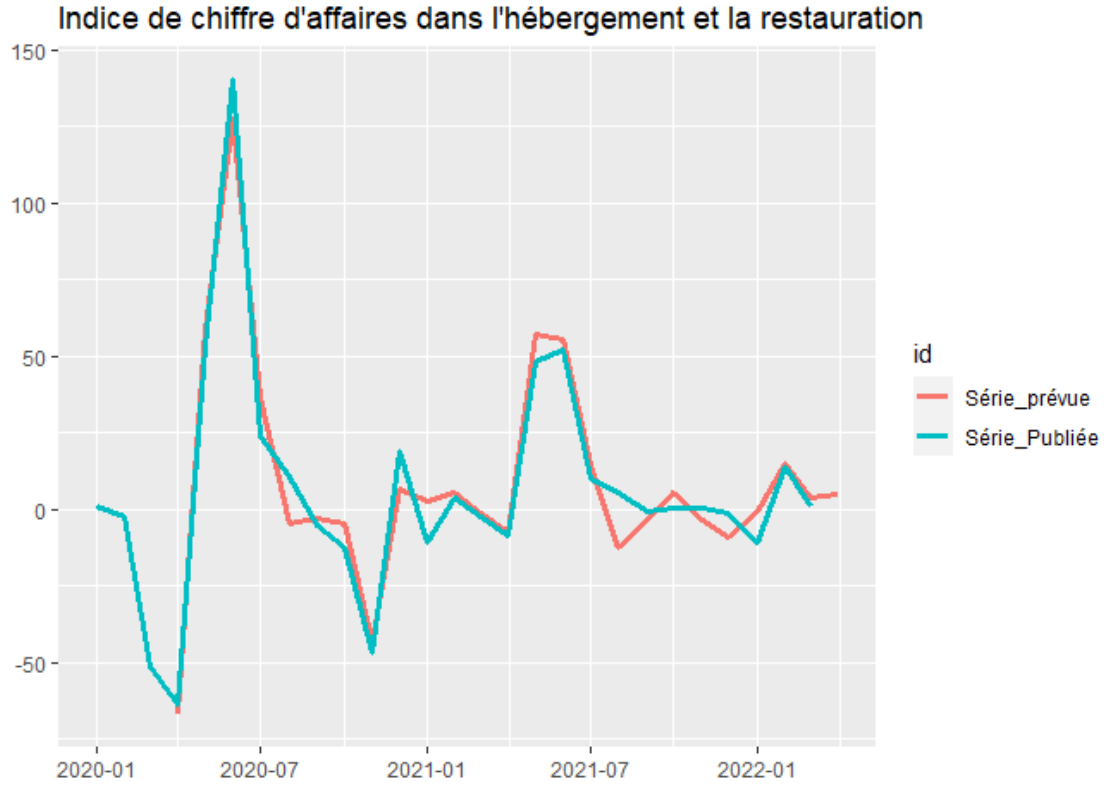
PROLONGER LES INDICATEURS TRADITIONNELS : UN EXEMPLE

Par exemple, pour la production de la branche hébergement-restauration :



Inutilisable en pratique, car l'indice de chiffre d'affaires au mois 1 arrive en fin de mois 3

Mais en prolongeant l'indice de chiffres d'affaires de la branche, grâce à deux indicateurs haute fréquence (*Google Trends*, *Google Mobility Reports*) :



2 DONNÉES ET PRÉ-TRAITEMENTS

Variables prévues et explicatives pour la France :

Source	Nombre de séries	Type de données
IPI	6	Indicateur traditionnel
ICA	6	Indicateur traditionnel
Enquêtes de conjoncture Insee	120	Soldes d'opinion, climats
Apple Maps Mobility	6	Fréquentation
Cerema	2	Relevés des sites de comptage
Google Mobility	3	Fréquentation
Google Trends	6	Part du volume de recherches
DSN	110	Nombre d'heures rémunérées ou d'activité partielle
Dépenses par carte bancaire CB + données de caisse	19	Montants en euros
Box office	1	Chiffre d'affaires
RTE	31	Consommation électricité par branche (gros sites industriels directement raccordés par RTE)

Mais un nombre moins important d'indicateurs pour les autres pays (Allemagne, Italie, Espagne, Royaume-Uni)

Traitement des séries *Google Trends* et *dépenses CB* :

- **Méthode de désaisonnalisation :**
  - Glissement : on compare chaque jour à son *jour comparable* d'avant la crise sanitaire
  - Avant-crise :
    - **Google Trends : moyenne 2017-2019**
    - **CB : 2019**
- **Autres traitements :**
  - Google Trends : l'import de données quotidiennes sur plusieurs années est délicat
  - CB :
    - **somme des données cartes bancaires *CB* et des données de caisse ;**
    - **les données de carte bancaire sont corrigées du taux d'utilisation de ce moyen de paiement et de l'évolution des prix**

CERTAINES SÉRIES N'ONT PAS PU ÊTRE DÉSAISONNALISÉES

Les autres séries :

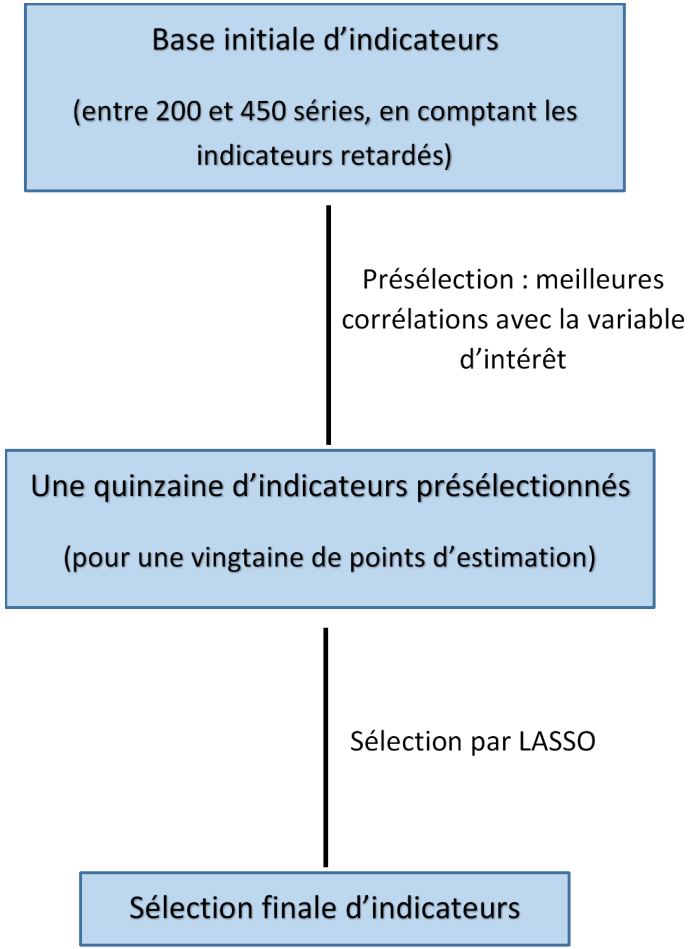
- **Sont parfois déjà désaisonnalisées ou traitées**
- **D'autres sont utilisées brutes**
  - Indicateurs de mobilité disponibles depuis février 2020, avec pour référence janvier/février 2020
    - **Google Mobility Reports**
    - **Apple Maps Mobility**
    - **Trafic routier Cerema**
  - Ces séries sont seulement mensualisées, puis passées en variation

3 LA SÉLECTION DE VARIABLES

3.A MÉTHODE INITIALE

La désaisonnalisation des indicateurs n'est pas le seul problème posé par l'apparition de nouveaux indicateurs :

- **Beaucoup d'indicateurs disponibles, et leur nombre varie :**
  - En fonction de la date où la prévision est effectuée
  - Les indicateurs mis à disposition par des entreprises privées peuvent être retirés ou modifiés à tout moment
- **Les variables d'intérêt sont différentes selon les périodes**
  - Activité mensuelle en période de forte crise
- **Une sélection de variables manuelle :**
  - serait trop chronophage dans ce contexte
  - risque d'occulter des relations intéressantes


Base initiale d'indicateurs
(entre 200 et 450 séries, en comptant les indicateurs retardés)
Présélection : meilleures corrélations avec la variable d'intérêt
Une quinzaine d'indicateurs présélectionnés
(pour une vingtaine de points d'estimation)
Sélection par LASSO
Sélection finale d'indicateurs

3.B ET SI ON AJOUTE DES DONNÉES ?

INCOHÉRENCE DES RÉSULTATS LORS DE L'AJOUT DE DONNÉES

Parfois, lors de l'ajout de nouvelles séries ou un élargissement de la présélection :

- **Un étalonnage auparavant intéressant peut passer à du sur-apprentissage**
  - **Trop de variables conservées par rapport à notre faible nombre de points**
- **D'autres étalonnages restent intéressants, mais voient leurs performances (RMSE, R^2…) s'amoindrir**
- **Huang *et al.*, 2008 : en situation non asymptotique, le Lasso a des comportements non désirables. Ils conseillent de plutôt utiliser la méthode Adaptive Lasso (Zou, 2006)**

Différence entre un Lasso et un Adaptive Lasso

- **Un Lasso est une régression pénalisée via la norme 1 des coefficients :**
  - On minimise $$\hat{\boldsymbol{\beta}}(\text{lasso}) = \arg\min_{\boldsymbol{\beta}} \left\| \mathbf{y} - \sum_{j=1}^{p} \mathbf{x}_j \beta_j \right\|^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$
- **Un Adaptive Lasso minimise plutôt** $$\arg\min_{\boldsymbol{\beta}} \left\| \mathbf{y} - \sum_{j=1}^{p} \mathbf{x}_j \beta_j \right\|^2 + \lambda \sum_{j=1}^{p} w_j |\beta_j|$$
  - Les pondérations w_j sont obtenues via un premier estimateur
  - Zou, 2006 : en cas de colinéarité entre certaines variables, Ridge est conseillé comme premier estimateur.

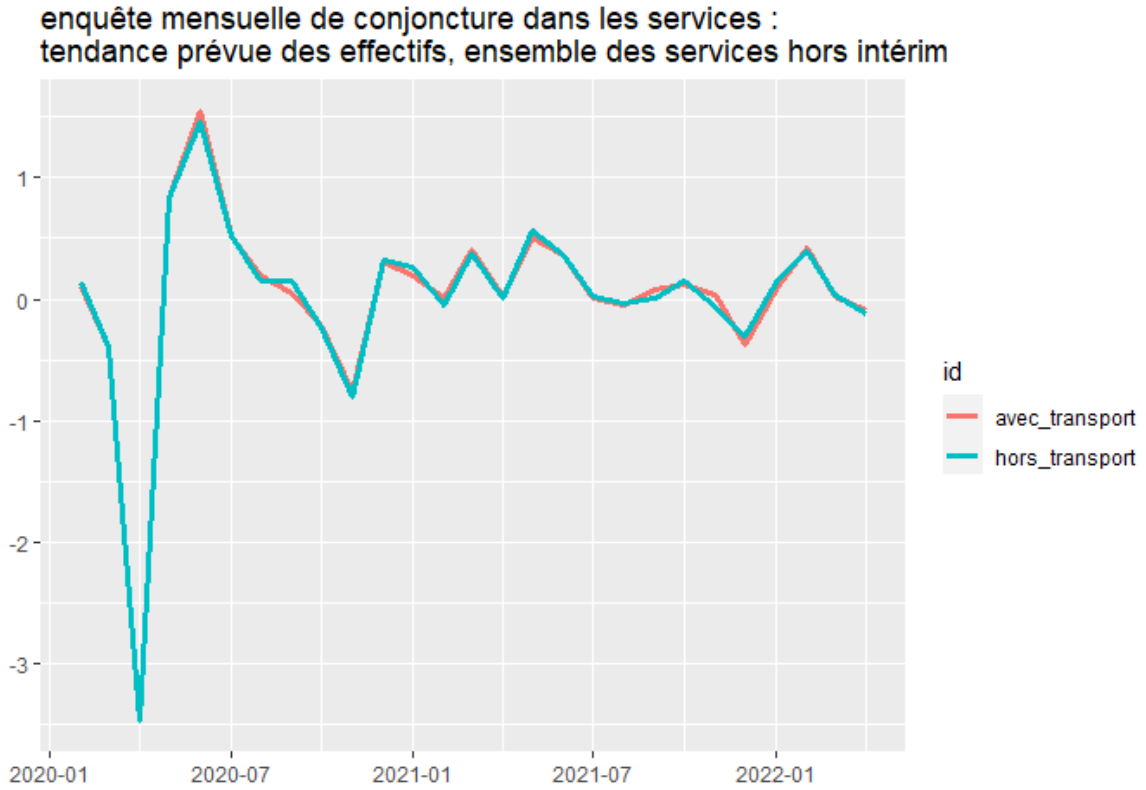
3.C AMÉLIORER L'ÉTAPE DE PRÉSÉLECTION

Le bon comportement de la méthode est soumise à hypothèse :

- **Zou *et al.*, 2008 : dans un contexte**
  - **où les comportements ne sont pas asymptotiques (nombre de points limités) ;**
  - **où le nombre de variables est environ égal ou supérieur au nombre de points.**
- **Alors le bon comportement du Adaptive Lasso repose sur l'hypothèse que les variables à éliminer sont peu corrélées aux variables à conserver**
  - **Mais on ne connaît pas les variables à conserver**
  - **Dans les faits : toutes les variables doivent être peu corrélées entre elles**

UNE HYPOTHÈSE CONTRAIRE À UNE PRÉSÉLECTION PAR CORRÉLATION

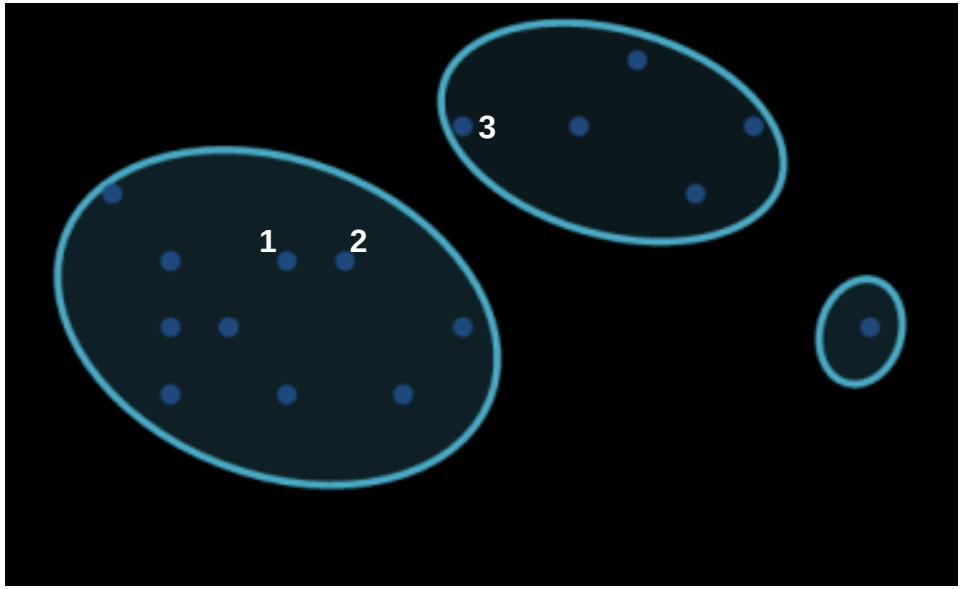
De nombreux indicateurs sont fortement corrélés :



On modifie la présélection des variables, pour respecter au mieux l'hypothèse de corrélation faible :

- Les variables sont regroupées par cluster

- Présélection : on prend un indicateur par cluster



4 CONCLUSION

Autres questions méthodologiques étudiées :

- Diminuer l'importance des points du premier confinement
- Faire une estimation robuste aux révisions et à l'instabilité de certains indicateurs

Et maintenant ?

- Généraliser et éprouver la méthode
  - Paramètres à optimiser
  - Propagation des erreurs entre prévision de l'indicateur et son utilisation dans un étalonnage traditionnel
- Élargir le champ des étalonnages fonctionnels

Retrouvez-nous sur

insee.fr

Robin Navarro & Jérémy Marquis
Département de la conjoncture
INSEE

Insee Mesurer pour comprendre

ANNEXES

On modifie la présélection des variables, pour respecter au mieux l'hypothèse de corrélation faible :

- Les variables sont regroupées par cluster
  - Graphe complet où
    - les indicateurs sont les sommets ;
    - la distance entre deux sommets A et B est égale à 1 – |corr(A,B)|
  - Clustering hiérarchique, avec 12 à 15 clusters.
  - Deux indicateurs du même cluster sont bien corrélés entre eux, et faiblement corrélés s'ils sont dans des clusters différents
- Présélection : on prend un indicateur par cluster
  - Cet indicateur est choisi par corrélation avec la série à prévoir.

LES SÉRIES GOOGLE TRENDS SONT INSTABLES

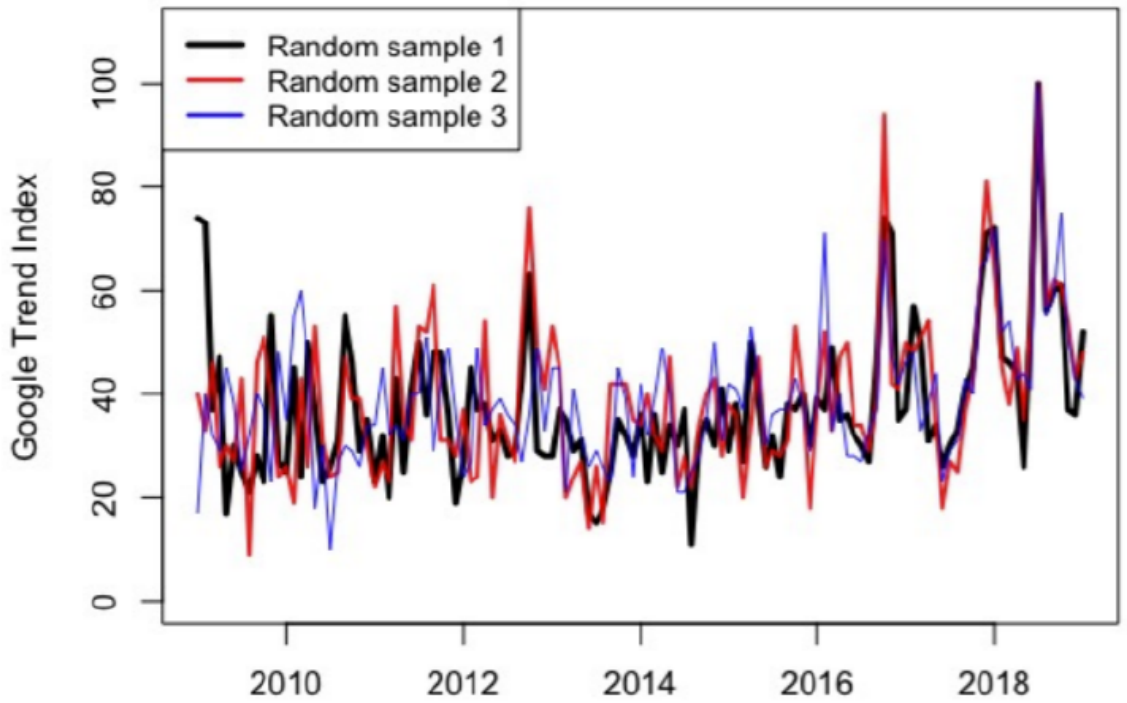


Figure - Google Trend correspondant à la thématique "croissance du PIB" aux États-Unis téléchargée plusieurs fois (Medeiros et Pires 2021).