Bulletin
de la Banque de France
248/5 - SEPTEMBRE-OCTOBRE 2023

BANQUE DE FRANCE
EUROSYSTÈME

Recherche économique

Estimation du commerce mondial en temps réel grâce à l'apprentissage automatique

Un écueil majeur en économie réside dans les longs délais de publication de nombreux indicateurs, ce qui complique l'appréciation du cycle économique en temps réel. Pour y remédier, nous avons construit un « *nowcast* » (une estimation en temps réel) du commerce international. À partir d'une base de données de 600 variables, nous utilisons un nouvel algorithme d'apprentissage automatique, appelé « forêt aléatoire macroéconomique » (*macroeconomic random forest*), qui s'est avéré plus performant que d'autres techniques linéaires et non linéaires. Notre approche comporte trois étapes i) présélection des variables, ii) extraction des facteurs et iii) régression d'apprentissage automatique. Cette approche améliore la précision des prédictions (gain de 15 à 30% par rapport à la méthode en deux étapes de Stock et Watson (2002), et de 30 à 40% par rapport à un modèle autorégressif). Nous donnons des exemples de la performance du modèle pendant la pandémie de Covid-19.

Menzie CHINN
Université du Wisconsin
Baptiste MEUNIER
Banque centrale européenne
Sebastian STUMPNER
Banque de France

Codes JEL
C53, C55,
E37

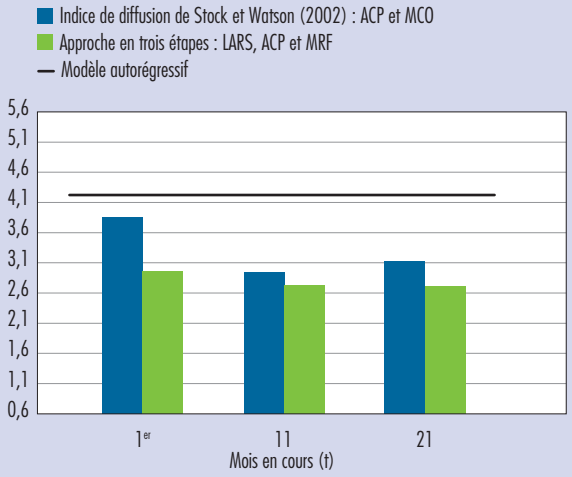
600
le nombre de variables prédictives dans la base de données

26 %
les gains moyens de précision pour notre méthode par rapport à une approche linéaire

40 %
les gains moyens de précision pour notre méthode par rapport à un modèle autorégressif

Évolution de la précision des estimations (RMSE hors échantillon)

(en abscisse : date de la prévision soit le 1er, le 11 et le 21 du mois considéré ; en ordonnée : variation en % sur un an)



Lecture : Une erreur quadratique moyenne (RMSE) faible indique une précision élevée.
Note : L'échantillon couvre la période de janvier 2012 à avril 2022. ACP, analyse des composantes principales ; LARS, régression par moindres angles ; MCO, moindres carrés ordinaires ; MRF, forêt aléatoire macroéconomique.
Source : Auteurs.

1 Les données officielles de commerce mondial sont publiées avec un délai important

L'analyse économique en temps réel est souvent compliquée par le fait que les séries économiques sont publiées avec des délais considérables. Ce constat vaut pour le commerce international : même si certains pays publient rapidement des données **en valeur**, les données **en volume** tardent davantage. Le bureau néerlandais de la planification économique (*Centraal Plan Bureau* – CPB) fournit des estimations utilisées par un grand nombre d'économistes, mais qui sont publiées environ huit semaines après la fin du mois. Autrement dit, les données de mars sont disponibles vers le 25 mai [1]. Cette situation constitue une véritable gageure en matière de politiques, car les décisions doivent s'appuyer sur des informations récentes.

Le problème est d'autant plus préoccupant que nous vivons dans un environnement économique en rapide évolution. En effet, ces dernières années ont été marquées par des crises majeures, comme la pandémie de Covid en 2020 ou la guerre russe en Ukraine depuis 2022. L'objectif de ce projet est d'élaborer un outil permettant de prédire avec exactitude l'évolution du commerce mondial en volume en temps réel ou quasi réel, même en cas de crises de grande ampleur.

S'il est vrai que les données officielles sont publiées avec un délai important, de nombreux indicateurs sont disponibles plus rapidement. Notre récent article (Chinn *et al.*, 2023) exploite ces informations afin de fournir des estimations anticipées des **volumes** d'échanges. Compte tenu de l'ampleur du retard, il ne s'agit pas de prédire les volumes seulement pour le mois t en cours (« *nowcasting* »), mais aussi pour les mois précédents (« *back-casting* » des mois $t-2$ et $t-1$ pour lesquels les données du CPB n'ont pas encore été publiées). Nous faisons également des prévisions pour $t+1$ afin d'évaluer la pertinence de notre méthode pour les évolutions futures.

Nous avons identifié 600 variables pertinentes pour évaluer l'évolution du commerce mondial. Pour construire notre base de données, nous avons analysé des articles traitant du *nowcasting* du commerce, notamment Keck *et al.* (2010), Guichard et Rusticelli (2011), Jakaitiene et Dees (2012), Bahroumi *et al.* (2016), Martinez-Martin et Rusticelli (2021), Charles et Darné (2022). Nous identifions ainsi des variables qui couvrent différents aspects du commerce mondial (par exemple, données douanières, coûts d'expédition, trafic de marchandises) et, plus généralement, des perspectives macroéconomiques, qu'il s'agisse de l'activité industrielle (par exemple la production d'acier) ou de la consommation des ménages (par exemple, les ventes au détail). Enfin, nous intégrons au modèle le cours des matières premières et des indicateurs financiers.

2 Notre méthodologie repose sur l'apprentissage automatique

La grande nouveauté de notre modèle est qu'il fait appel à des algorithmes d'apprentissage automatique. Après des tests sur différentes catégories d'algorithmes, la technique la plus performante s'est avérée être la « forêt aléatoire macroéconomique » (*macroeconomic random forest,* MRF) de Goulet Coulombe (2020), exposée en détail à l'annexe 1.

Deuxième apport de notre modèle de prédiction : l'approche en trois étapes consiste à présélectionner des variables, puis à extraire des facteurs avant d'utiliser la régression d'apprentissage automatique, approche que nous décrivons à l'annexe 1. Nous avons comparé plusieurs méthodes pour chacune des deux premières étapes : la combinaison la plus performante est celle qui fait intervenir la « *least-angle regression* » (LARS, ou régression par moindres angles) pour la présélection des variables, l'analyse des composantes principales (ACP) pour l'extraction des facteurs, et la forêt aléatoire macroéconomique (MRF) pour la prédiction [2]. La LARS s'inspire de la régression par étapes, utilisée quand il existe un grand nombre de régresseurs potentiels afin d'identifier pas à pas les variables explicatives, mais l'avantage réside dans la similarité en valeur absolue des coefficients de régression lorsque les variables présentent le même niveau de corrélation avec les résidus (cf. annexe 1).

1 Dans certaines économies avancées, les données sont publiées plus rapidement (par exemple, environ un mois pour les États-Unis et à peu près un mois et demi pour la France), mais ces deux pays ne représentent qu'une fraction du commerce international.

2 Les autres méthodes de présélection des variables que nous avons envisagées sont le « *sure independence screening* », la présélection fondée sur les t-statistiques et la moyenne mobile bayésienne itérative. Les autres méthodes envisagées pour l'extraction des facteurs sont l'estimateur à deux étapes, l'estimateur quasi maximal de vraisemblance et l'ACP dynamique. Les autres techniques de régression envisagées sont présentées dans l'annexe 3.

Cette méthode est mise en œuvre de façon séquentielle : (**étape 1**) la LARS sélectionne les 60 prédicteurs les plus pertinents dans notre base de données de 600 variables[3] ; (**étape 2**) nous synthétisons les variables sélectionnées en quelques facteurs en utilisant l'ACP[4] ; (**étape 3**) ces facteurs sont utilisés comme variables explicatives dans la régression du commerce mondial, en se servant de la MRF[5]. La présélection des variables et l'extraction des facteurs ont déjà été utilisées dans d'autres publications (cf. Jardet et Meunier, 2022), mais notre contribution est d'avoir combiné ces deux techniques dans un cadre intégré pour l'apprentissage automatique. L'approche en trois étapes peut être considérée comme une extension de « l'indice de diffusion » de Stock et Watson (2002), utilisé par un grand nombre d'analystes, qui combine ACP et régression des moindres carrés ordinaires (MCO). Nous complétons cette méthode par la présélection des variables et l'apprentissage automatique.

À partir de ces techniques, nous avons établi des prédictions hors échantillon des volumes du commerce mondial de janvier 2012 à avril 2022, représentées dans le graphique 1 *infra*. Nous utilisons une configuration en temps réel en répétant, pour chaque date de la période et sur la base des données auxquelles le prévisionniste aurait eu accès à cette même date, la présélection des variables, l'extraction des facteurs et la régression. Nous gérons le flux de données en temps réel (c'est-à-dire la disponibilité des observations les plus récentes à différentes dates) au moyen du réalignement vertical d'Altissimo *et al.* (2006) (cf. annexe 2).

3 Notre approche en trois étapes permet d'obtenir des gains importants de précision

Nous mesurons l'efficacité (la précision) du modèle en fonction de l'écart entre les prédictions hors échantillon et les données réelles. Nous nous appuyons sur la racine de l'erreur quadratique moyenne (RMSE), qui mesure l'écart entre les données réelles du taux de croissance sur douze mois du commerce mondial fournies par le CPB (y_t) et la prédiction du modèle ($\hat{y}_t$). Pour des prédictions sur un échantillon allant de 1 a T :

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(\hat{y}_t - y_t)^2}$$

Pour évaluer la précision, nous déterminons d'abord l'ensemble des variables auxquelles auraient eu accès un prévisionniste en temps réel les 1er, 11 et 21 de chaque mois. Puis nous mettons en œuvre l'approche en trois étapes à partir de ces trois ensembles de données distincts.

Les RMSE ainsi obtenues sont affichées dans le graphique 1, en vert pour l'approche en trois étapes, en bleu pour l'approche Stock et Watson (2002) et la ligne noire pour le modèle autorégressif. En allant de la gauche vers la droite, les barres sont de plus en plus petites, ce qui signifie que nous commettons (naturellement) moins d'erreurs à mesure que nous disposons de plus de données. En effet, pour prévoir le commerce au mois *t*, le prévisionniste disposera de davantage d'informations s'il est au mois suivant (*t* + 1, sur la droite du graphique) qu'au mois précédent (*t* – 1, sur la gauche du graphique). Nous atteignons même une RMSE inférieure à 1 % lorsque nous procédons à un *back-casting*. En ce qui concerne les barres vertes, la RMSE diminue d'environ 40% entre les prévisions du 11 du mois *t* – 1 et celles du 11 du mois *t* + 1. Elle diminue de 50% de plus avec les prévisions du 11 du mois *t* + 2.

L'approche en trois étapes affiche systématiquement une meilleure performance que les deux autres méthodes. Bien que le gain de précision varie en fonction de la date et de l'horizon de prédiction, elle permet d'obtenir en moyenne une RMSE inférieure de 26% à celle d'un modèle à la Stock et Watson (2002) et de 40% à celle d'un modèle autorégressif. Nous montrons également que l'approche en trois étapes fondée sur la forêt aléatoire macroéconomique est plus efficace que les autres approches en trois étapes qui reposent sur d'autres techniques de régression linéaire et non linéaire (cf. annexe 3).

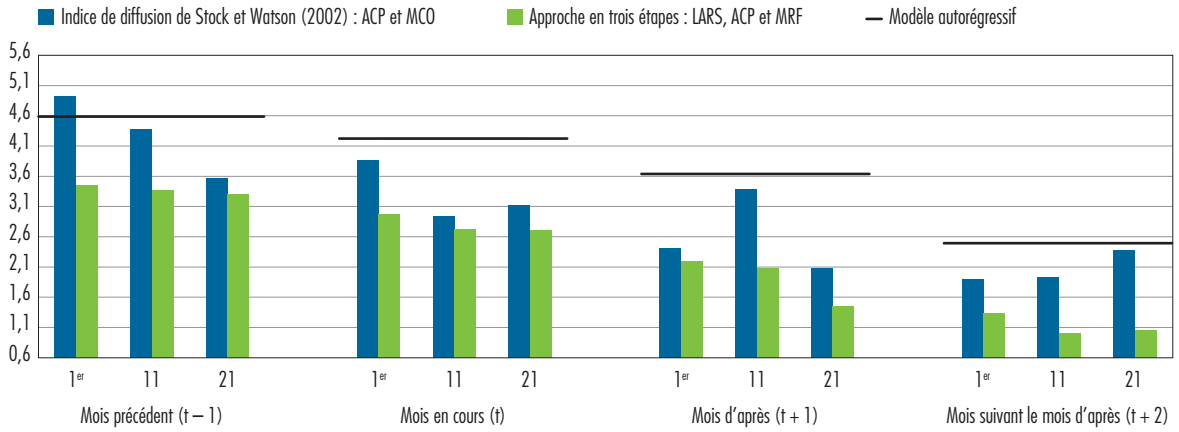
3 Le prévisionniste doit choisir le nombre de variables à inclure dans le modèle. Dans notre cas, nous avons fixé ce nombre à 60 sur la base de tests empiriques.

4 Nous avons déterminé le nombre de facteurs à mettre dans la régression selon le critère d'information de Bai et Ng (2002), comme il est habituel de le faire dans la littérature spécialisée.

5 Dans la configuration de la forêt aléatoire macroéconomique (MRF), la partie linéaire et la partie aléatoire de la forêt aléatoire sont toutes deux composées des facteurs (cf. explication de la MRF en annexe 1). En ce sens, la MRF peut être considérée comme une généralisation de notre modèle linéaire de base des MCO, mais dans laquelle les coefficients de la régression varient en fonction du temps et suivent un algorithme de forêt aléatoire. De plus, les hyperparamètres de la MRF (par exemple, le nombre d'arbres de décision dans la forêt aléatoire) sont fixés sur la base d'une validation croisée.

G1 Évolution de la précision des estimations (RMSE hors échantillon)

(en abscisse : date de la prévision soit le 1er, le 11 et le 21 du mois considéré ; en ordonnée : variation en % sur un an)



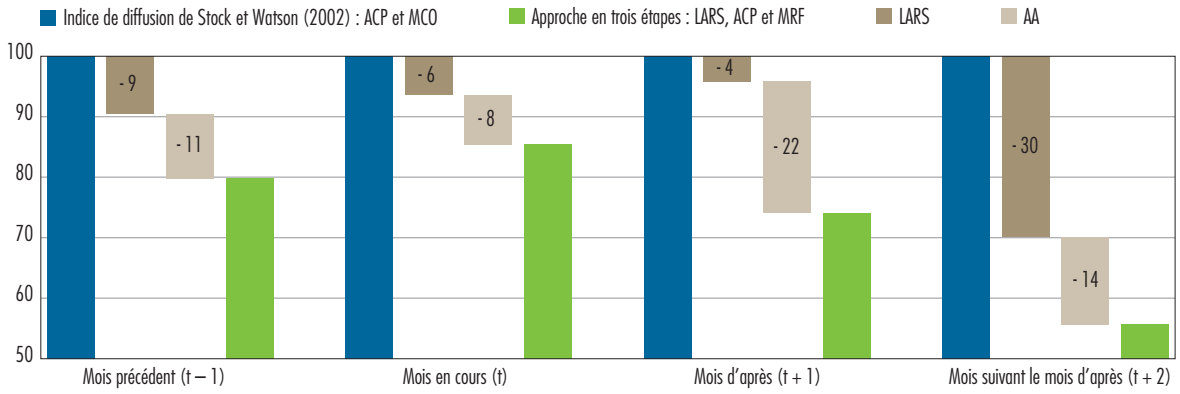
Lecture : Une erreur quadratique moyenne (RMSE) faible indique une précision élevée.
Note : L'échantillon couvre la période de janvier 2012 à avril 2022. La méthode de Stock et Watson (2002) se fonde sur l'ACP et les MCO. L'approche en trois étapes fait appel à la LARS, à l'ACP et à la forêt aléatoire macroéconomique (MRF). L'efficacité du modèle autorégressif est la même sur l'ensemble des données, car le calcul ne dépend pas du jour de la prévision. Cf. glossaire pour la définition des sigles.
Source : Auteurs.

Les gains de précision proviennent de la présélection des variables et de l'apprentissage automatique. Le graphique 2 *infra* affiche les gains obtenus par rapport à l'approche ACP et MCO en deux étapes (Stock et Watson, 2002), en distinguant ceux provenant de la présélection et ceux provenant de la MRF. D'une façon générale, ces deux étapes contribuent substantiellement aux gains de précision. Bien que les contributions dépendent de l'horizon, les gains provenant de l'apprentissage automatique sont plus stables. Ils sont compris entre 10% et 20%.

Enfin, le graphique 3 compare l'évolution du taux de croissance annuel du commerce mondial avec les prévisions de notre *nowcast*. La prédiction en temps réel s'appuie sur les données extraites le 21 du mois, pour un *back-casting* de deux mois (prédiction du mois *t* au mois *t* + 2). Le graphique montre que notre modèle de *nowcast* suit de très près les données réelles sur toute la période considérée et qu'il réussit à prédire avec exactitude les évolutions en période de fortes fluctuations, lorsque les exercices de *nowcasting* sont particulièrement utiles, comme en témoigne l'étroitesse des écarts pendant la

G2 Décomposition des gains de précision (100 = ACP et MCO)

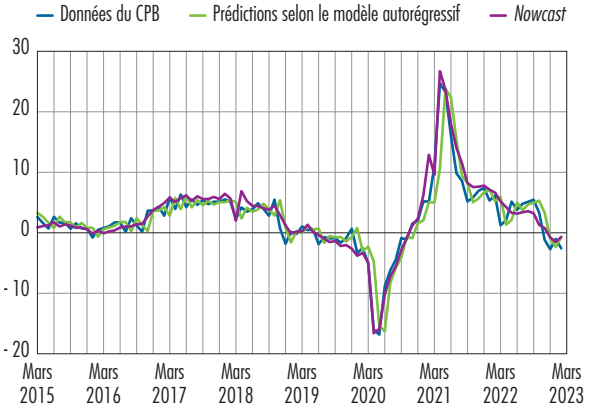
(variation en % sur un an)



Note : Méthode de Stock et Watson (2002) : fondée sur l'ACP et les MCO ; Approche en trois étapes : degré de précision final obtenu en utilisant la LARS, l'ACP et la forêt aléatoire macroéconomique (MRF), cf. Chinn *et al.* (2023) ; LARS : présélection selon une régression par moindres angles ; AA : apprentissage automatique avec l'algorithme de forêt aléatoire macroéconomique. Les résultats sont calculés par rapport aux valeurs ACP et MCO normalisées à 100 pour chaque mois. Les résultats sont les gains moyens pour les bases de données du 1er, du 11 et du 21 du mois.
Source : Auteurs.

G3 Estimations (hors échantillon) de l'évolution du commerce mondial en volume pendant la pandémie de Covid-19 en temps réel

(variation en % sur un an)



Note : CPB, *Centraal Plan Bureau* (bureau néerlandais de la planification économique).
Le *nowcast* en trois étapes fait appel à la LARS, à l'ACP et à la forêt aléatoire macroéconomique (MRF). Cf. glossaire pour la définition des sigles.
Source : Auteurs.

pandémie de Covid-19. Les prévisions sont également restées très justes lors de chocs plus récents, comme le déclenchement de la guerre russe en Ukraine début 2022.

⁂

Au total, nous obtenons un modèle de prévision des **volumes** des échanges mondiaux fondé sur une méthode innovante d'apprentissage automatique, la forêt aléatoire macroéconomique. Pour ce faire, nous avons adopté une approche en trois étapes : une présélection des variables et une extraction des facteurs, suivies de l'apprentissage automatique. Plus généralement, cette approche peut être considérée comme un guide pratique pour les prévisionnistes qui souhaitent utiliser l'apprentissage automatique. L'approche est en effet très flexible et peut être appliquée facilement à d'autres variables.

Bibliographie

Altissimo (F.), Cristadoro (R.), Forni (M.), Lippi (M.) et Veronese (G.) (2006)
« New Eurocoin: tracking economic growth in real time », *CEPR Press Discussion Paper*, n° 5633.

Bai (J.) et Ng (S.) (2002)
« Determining the number of factors in approximate factor models », *Econometrica*, vol. 70, n° 1, p. 191-221.

Barhoumi (K.), Darné (O.) et Ferrara (L.) (2016)
« A world trade leading index (WTLI) », *Economic Letters*, vol. 146, p. 111-115.

Charles (A.) et Darné (O.) (2022)
« Backcasting world trade growth using data reduction methods », *The World Economy*, vol. 45, n° 10, p. 3169-3191.

Chinn (M.), Meunier (B.) et Stumpner (S.) (2023)
« Prévoir le commerce mondial en temps réel grâce au *machine learning* : une approche en trois étapes », *Document de travail*, n° 917, Banque de France, juillet.
Télécharger le document

Efron (B.), Hastie (T.), Johnstone (I.) et Tibshirani (R.) (2004)
« Least angle regression », *The Annals of Statistics*, vol. 32, n° 2, p. 407-499.

Goulet Coulombe (P.) (2020)
« The macroeconomy as a random forest », préédition électronique arXiv.

Guichard (S.) et Rusticelli (E.) (2011)
« A dynamic factor model for world trade growth », *Documents de travail du Département des affaires économiques de l'OCDE*, n° 874.

Jakaitiene (A.) et Dees (S.) (2012)
« Forecasting the world economy in the short-term », *The World Economy*, vol. 35, n° 3, p. 331-350.

Jardet (C.) et Meunier (B.) (2022)
« Nowcasting world GDP growth with high-frequency data », *Journal of Forecasting*, vol. 41, n° 6, p. 1181-1200.

Kaiser (H. F.) (1960)
« The application of electronic computers to factor analysis », *Educational and Psychological Measurement*, vol. 20, p. 141-151.

Keck (A.), Raubold (A.) et Truppia (A.) (2010)
« Forecasting international trade: a time series approach », *OCDE Journal: Journal of Business Cycle Measurement and Analysis*, n° 2009/2.

Martínez-Martín (J.) et Rusticelli (E.) (2021)
« Keeping track of global trade in real time », *International Journal of Forecasting*, vol. 37, n° 1, p. 224-236.

Stock (J. H.) et Watson (M. W.) (2002)
« Forecasting using principal components from a large number of predictors », *Journal of the American Statistical Association*, vol. 97, n° 460, p. 1167-1179.

Glossaire

AA : Apprentissage automatique.

ACP : Analyse des composantes principales.

CPB : *Centraal Plan Bureau*, bureau néerlandais de la planification économique.

LARS : *Least-angle regression*, régression par moindres angles.

MCO : Moindres carrés ordinaires.

MRF : *Macroeconomic random forest*, forêt aléatoire macroéconomique.

RMSE : *Root mean square error*, racine de l'erreur quadratique moyenne.

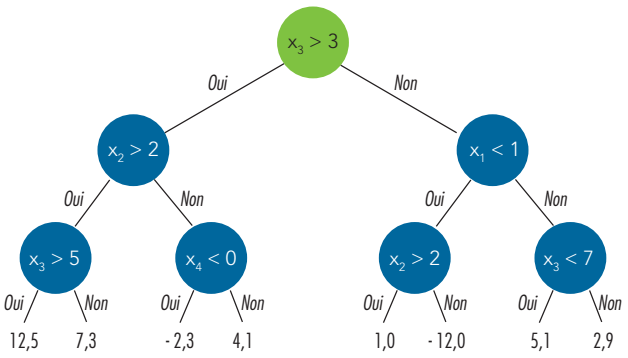
Annexes
Éléments méthodologiques

1 Définitions de certains termes techniques

Comme nous testons différentes techniques de régression (cf. annexe 3), la distinction entre les techniques d'apprentissage automatique fondées sur les **arbres de décision** et celles s'appuyant sur les **régressions** constitue un aspect essentiel de notre travail. La première catégorie (forêt aléatoire, *gradient boosting*), la plus utilisée dans la littérature spécialisée, consiste à agréger plusieurs « arbres de décision » ensemble. La deuxième catégorie, correspondant aux techniques fondées sur la **régression** (forêt aléatoire macroéconomique, *linear gradient boosting*), est beaucoup moins utilisée dans la littérature spécialisée. Il s'agit d'une adaptation de la première catégorie, mais qui utilise des régressions linéaires plutôt que, ou en complément, des arbres de décision.

Un arbre de décision est un algorithme utilisé à des fins de classification ou de régression. Un arbre est composé de nœuds. Chaque nœud est une fourche où s'opère une division en fonction d'un test (une affirmation, qui est vraie ou fausse) évaluant la valeur d'une variable x (pas forcément la même à chaque noeud, par exemple x_3, x_2 ou x_1 dans le schéma ci-dessous). Si x remplit une condition (indiquée dans les bulles dans le schéma ci-dessous), l'algorithme prend un chemin (et dans le même registre, aboutit à une « feuille »), sinon il prend l'autre chemin (une autre « feuille »). Ces « feuilles » mènent à d'autres nœuds, et ainsi de suite. Aux extrémités de tous les chemins possibles se trouvent les dernières « feuilles » qui donnent les prédictions du modèle pour la variable cible (y), illustrées dans le schéma ci-dessous par les chiffres en bas de l'arbre. Le schéma constitue une illustration d'un arbre de décision.

Exemple d'arbre de décision



Source : Auteurs.

Une forêt aléatoire est une « méthode ensembliste » qui utilise un grand nombre d'arbres de décision. L'idée sous-jacente est de construire un grand nombre d'arbres non corrélés entre eux. Ainsi, en calculant la moyenne des prédictions de tous ces arbres, la variance de la prédiction globale est réduite. Un facteur essentiel au succès de la méthode est l'indépendance des arbres. Dans une forêt aléatoire, on crée cette indépendance i) en prélevant un échantillon « *bootstrap* » (méthode à tirer un échantillon aléatoire à partir des données) différent pour chaque arbre et ii) en tenant compte uniquement d'un sous-ensemble de variables pour chaque arbre. Quand les arbres ne sont pas corrélés entre eux, prendre la moyenne des prédictions indépendantes réduit la variance et augmente donc la précision du modèle.

Les forêts aléatoires macroéconomiques (ou MRF, Goulet Coulombe, 2020) constituent un prolongement des forêts aléatoires traditionnelles, qui sont construites en mettant en commun plusieurs arbres décisionnels (d'où le terme « forêt ») pour obtenir une prédiction. Mais ces forêts aléatoires traditionnelles sont souvent trop flexibles pour les séries temporelles macroéconomiques caractérisées par un nombre réduit d'observations. Pour remédier à cette imperfection, la MRF s'appuie sur une partie linéaire $y_t = X_t\beta_t$ comme dans une méthode simple des MCO, où y_t est la variable cible, à savoir le commerce mondial, X_t un vecteur de variables explicatives, et β_t les coefficients associés. Mais, contrairement aux MCO, les coefficients β_t peuvent varier dans le temps en fonction d'une forêt aléatoire. En termes mathématiques, $\beta_t = F(S_t)$, où F désigne une forêt aléatoire basée sur S_t, un ensemble de variables potentiellement différentes de X_t.

La LARS (Efron *et al.*, 2004) est un algorithme itératif de sélection des variables selon un processus ascendant. La sélection de variables est vide au départ. L'algorithme y ajoute le prédicteur x_i le plus corrélé à la variable cible y,

puis il augmente la valeur (absolue) du coefficient β_i de façon à ce que la corrélation de x_i avec le résiduel $(y - \beta_i x_i)$ diminue, jusqu'à ce qu'un autre prédicteur x_j présente une corrélation similaire avec $y - \beta_i x_i$. x_j est alors ajouté à la sélection d'indicateurs et la procédure se poursuit en déplaçant les deux coefficients β_i et β_j dans la même proportion jusqu'à ce qu'un autre prédicteur x_k présente une corrélation aussi élevée avec le résiduel (maintenant $y - \beta_i x_i - \beta_j x_j$). L'algorithme fournit ainsi un classement de toutes les variables en fonction de l'ordre dans lequel elles sont ajoutées dans l'algorithme. Comme il prend en compte les variables déjà sélectionnées, il garantit la complémentarité entre toutes les variables retenues.

2 Comment gérons-nous le flux de données en temps réel ?

Dans la réalité, les dates de publication asynchrones des différentes variables conduisent à une asymétrie. Chaque variable a un nombre différent d'observations manquantes à la fin de la base de données, en fonction des retards de publication. Pour remédier à cette problématique, nous appliquons le « réalignement vertical » d'Altissimo *et al.* (2006). Pour chaque variable, nous considérons la dernière valeur disponible comme la valeur actuelle et toute la série est réalignée en conséquence. Par exemple, si un prévisionniste souhaite établir un *nowcast* en mars 2023 avec une variable dont la dernière observation remonte à décembre 2022, la série est « réalignée » en prenant la valeur de décembre 2022 comme valeur pour mars 2023. En termes mathématiques, si la dernière observation de x_t à la date T remonte à T - k, la série réalignée $\tilde{x}_t$ est $\tilde{x}_t = x_{t-k}$ pour tout t entre 0 et T.

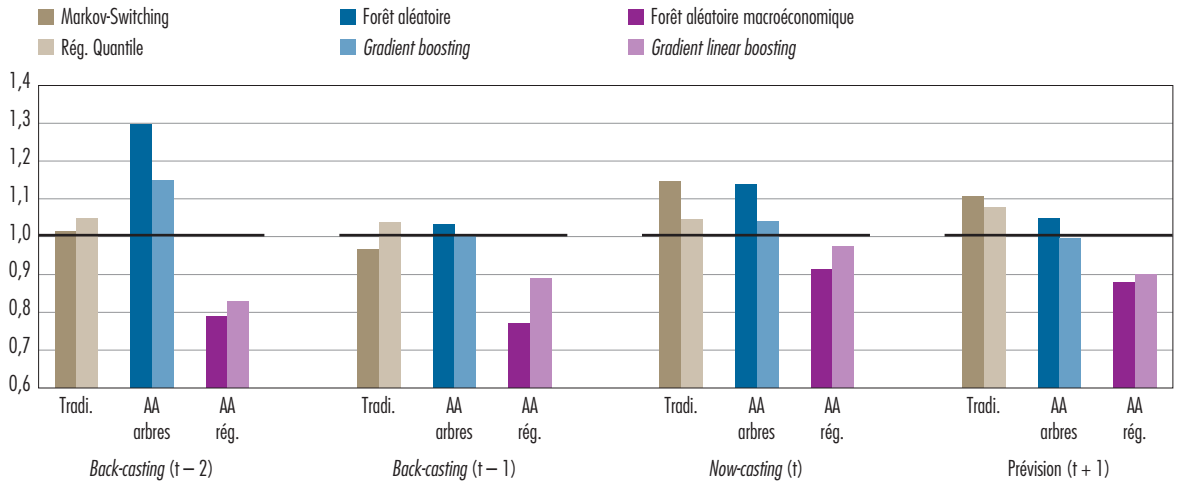
Par ailleurs, la valeur de certaines séries après la date de la prévision est parfois connue. Par exemple, un prévisionniste en mars 2023 pourrait souhaiter faire un *back-casting* des échanges mondiaux en janvier 2023 étant donné le décalage temporel considérable avec lequel les données sont publiées. Par exemple, cela peut être le cas si le prix de pétrole (connu chaque jour) est une variable explicative : le prévisionniste aura, dès mars 2023, à sa disposition les prix du pétrole en février et mars 2023. Or, il est possible qu'il dispose déjà des observations de février 2023 pour les séries les plus rapidement publiées. Pour en tenir compte et ne pas perdre ces observations « excédentaires », nous étendons le réalignement vertical. La série est réalignée dans la direction opposée, comme dans Altissimo *et al.* (2006), en prenant $\tilde{x}_t = x_{t+k}$ au lieu de $\tilde{x}_t = x_{t-k}$. Mais, contrairement à Altissimo *et al.* (2006) où la série réalignée $\tilde{x}_t$ remplace la série originale x_t, la série réalignée est considérée comme une nouvelle variable. Dans notre exemple *supra*, notre procédure aboutira à la création de deux nouvelles séries : une série réalignée en t – 1, c'est-à-dire utilisant les prix du pétrole en t + 1 (février 2023) pour estimer le commerce en t (janvier 2023), et une autre série réalignée en t – 2, c'est-à-dire utilisant les prix du pétrole en t + 2 (mars 2023) pour estimer le commerce en t (janvier 2023).

3 Qu'en est-il des autres techniques de régression non linéaire ?

L'un des intérêts de l'approche en trois étapes est sa souplesse. Elle permet en effet d'intégrer différentes techniques de régression. Avant de retenir l'algorithme de la forêt macroéconomique aléatoire, nous avons testé d'autres approches non linéaires. Les régressions non linéaires « traditionnelles » – la régression de Markov et la régression par quantile – forment un premier groupe. Le second groupe est lié aux techniques d'apprentissage automatique. Dans ce second groupe, nous distinguons les techniques fondées sur les **arbres de décision** et celles s'appuyant sur la **régression**. Une des techniques fondées sur les **arbres de décision** est la forêt aléatoire, qui agrège les prédictions tirées d'un grand nombre d'arbres de décision. Le *gradient boosting* en est une autre. Mais cette technique diffère de la forêt aléatoire en ce sens qu'au lieu de regrouper plusieurs arbres indépendants, elle construit un modèle en ajoutant des arbres de manière itérative, chaque nouvel arbre dépendant des résultats des arbres précédents. La forêt aléatoire macroéconomique appartient à la catégorie des techniques fondées sur la **régression**. Comme l'explique l'annexe 1, il s'agit d'une extension de la forêt aléatoire qui recourt à une régression (non linéaire). La deuxième catégorie comprend également le *linear gradient boosting*, une technique qui fonctionne de la même manière que le *gradient boosting* traditionnel expliqué précédemment, mais qui utilise des **régressions linéaires** au lieu des **arbres de décision**.

Degré de précision de l'erreur quadratique moyenne hors échantillon par rapport aux valeurs de référence linéaires (MCO)

(variation en % sur un an)



Lecture : Une erreur quadratique moyenne (RMSE) faible indique une précision élevée.
Note : Le degré de précision est mesuré par la RMSE hors échantillon entre janvier 2012 et avril 2022. L'efficacité est calculée par rapport aux valeurs de référence des MCO (ligne droite noire à 1,0). Les résultats sont obtenus pour la moyenne des bases de données reflétant les données disponibles pour un prévisionniste le 1er, le 11 et le 21 du mois, en utilisant une LARS pour présélectionner les 60 régresseurs les plus explicatifs, les facteurs étant extraits par ACP sur l'ensemble présélectionné. AA arbres = techniques d'apprentissage automatique fondées sur des **arbres de décision** ; AA rég. = techniques d'apprentissage automatique fondées sur des **régressions linéaires**.
Source : Auteurs.

Comme nous le voyons sur le graphique *infra*, qui montre le degré de précision par rapport aux MCO (ligne noire), la forêt aléatoire macroéconomique surpasse toutes les autres techniques utilisées. Tous les résultats sont calculés selon l'approche en trois étapes avec la LARS et l'ACP. Par conséquent, tout écart éventuel ne peut provenir que de la méthode de régression utilisée. Même dans ce cas, la forêt aléatoire macroéconomique (violet foncé) s'avère nettement plus efficace que les MCO. Elle surpasse également les autres méthodes non linéaires, notamment l'apprentissage automatique (AA) traditionnel fondé sur des **arbres de décision**. Le gain de précision est substantiel et constant sur les différents horizons. La seule méthode dont le niveau de précision est proche de la MFR est le *gradient linear boosting* (violet clair), qui est une autre technique d'apprentissage automatique fondée sur les régressions linéaires. Il semble donc que cette catégorie de techniques soit plus efficace pour la prévision non linéaire.

Éditeur
Banque de France

Directeur de la publication
Claude Piot

Rédaction en chef
Olivier de Bandt

Secrétaire de rédaction
Nelly Noulin

Réalisation
Studio Création
Direction de la Communication

ISSN 1952-4382

Pour vous abonner aux publications de la Banque de France
https://publications.banque-france.fr/
Rubrique « Abonnement »