Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

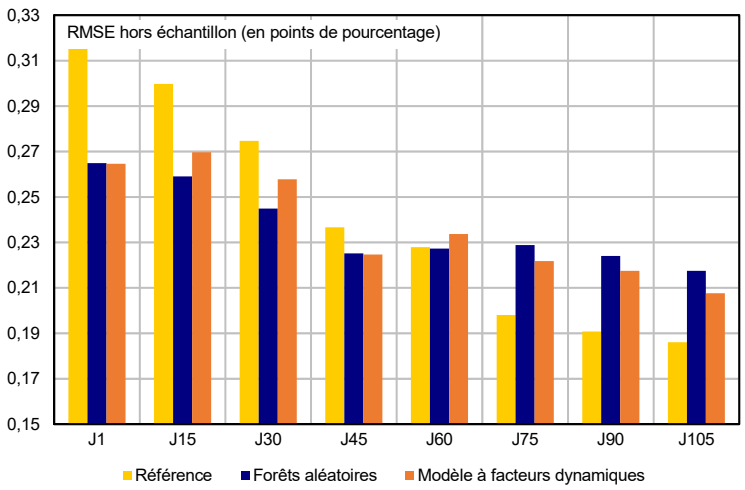
Trésor-éco

N° 254 • Janvier 2020

Améliorer l'estimation du PIB en temps réel grâce aux grands ensembles de données

- Enquêtes de conjoncture, données économiques et séries financières constituent un ensemble conséquent de données mobilisables par le conjoncturiste pour prévoir la croissance économique à très court terme. En particulier, au-delà des indicateurs synthétiques issus des données d'enquêtes de conjoncture, les données sectorielles complètent significativement l'information disponible. L'ensemble des données d'enquêtes contient ainsi plus d'un millier de séries qui peuvent être utilisées pour essayer d'améliorer les prévisions à court terme.
- Les méthodes de prévision traditionnelles ne sont pas adaptées pour traiter toutes ces données. Aujourd'hui la prévision du PIB en temps réel est généralement fondée sur des méthodes de régression linéaire sur un nombre réduit de variables. Or depuis une vingtaine d'années se développent des méthodes statistiques capables de manipuler de bien plus grands ensembles d'informations. Par exemple, les modèles dits à facteurs dynamiques permettent de synthétiser l'information de manière pertinente avec de faibles besoins en ressources de calcul.
- Plus récemment, avec l'augmentation des capacités de calcul, des méthodes fondées sur des techniques d'apprentissage automatique (ou *machine learning*) se sont développées et connaissent une popularité grandissante. Ces méthodes appliquent des moyens nouveaux de tri et de traitement de l'information, tels que les forêts aléatoires (ou *random forests*) ou les réseaux neuronaux.
- Certaines de ces méthodes permettent d'améliorer la performance des prévisions de court terme du PIB en mobilisant de grandes bases de données, incluant en particulier des données sous-sectorielles, sous réserve d'une étape préalable de sélection des variables. Les forêts aléatoires semblent à cet égard constituer une bonne méthode pour sélectionner à différentes dates les variables les plus à même d'apporter de l'information sur le PIB courant.
- En particulier, pour prévoir le PIB, c'est surtout en début de trimestre, avant que les premières données quantitatives ne soient disponibles, que les modèles reposant sur de grandes bases de données sont plus performants que les modèles traditionnels.

Erreur de prévision en fonction de l'horizon de prévision



Sources : Insee, Banque de France, PMI et données financières ; calculs DG Trésor.

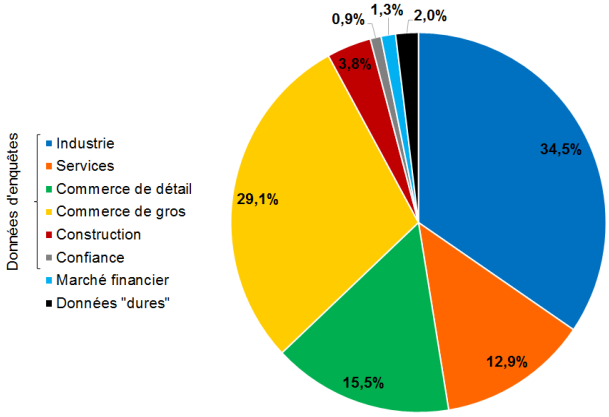
Note de lecture : L'erreur moyenne de prévision est définie ici par la racine carrée de la moyenne des erreurs de prévision au carré. J1 correspond aux prévisions réalisées avec toutes les données disponibles au 1er jour du trimestre considéré, J15 avec les données disponibles au 15e jour, *etc.* ; J105 correspond aux données disponibles juste avant de connaître la première estimation du PIB.

1. Des centaines de données peuvent être mobilisées pour prévoir l'évolution du PIB en temps réel

Avant que la première estimation du PIB ne soit publiée – soit une trentaine de jours après la fin du trimestre – le conjoncturiste cherche à prévoir l'évolution du PIB trimestriel en temps réel afin d'appuyer les décisions économiques sur un diagnostic aussi précis que possible. Pour ce faire, un nombre conséquent de données sont disponibles. En plus de données dites « dures »[1] (consommation, production industrielle, *etc.*) et financières (cotation du CAC40, taux d'intérêt, indicateur de volatilité VIX, *etc.*), il existe beaucoup de données d'enquêtes de conjoncture[2] représentatives de secteurs d'activité, à un niveau agrégé (par exemple, l'industrie) ou plus détaillé (par exemple, le secteur automobile). La base de données disponible en France chaque trimestre et utilisée ici contient ainsi plus d'un millier de variables dont la grande majorité correspond à des données d'enquêtes (*cf.* graphique 1).

Ces données conjoncturelles sont publiées tout au long du trimestre, à des dates qui dépendent de chaque indicateur. Ainsi, l'Insee publie ses données d'enquêtes mensuelles dès la fin du mois concerné, et les PMI (*Purchasing Managers Index*)[3] sont également disponibles à cette date, alors qu'il faut attendre une dizaine de jours pour les données d'enquêtes de la Banque de France. Les premières données « dures » sont disponibles plus tardivement : la consommation mensuelle des ménages en biens est disponible avec 30 jours de retard et l'indice de production industrielle (IPI) est connu avec 40 jours de retard. En prenant en compte ces différents délais de publication, il est possible d'élaborer des bases de données par quinzaine, qui retracent l'ensemble d'information à disposition du prévisionniste (*cf.* tableau 1).

Graphique 1 : Composition de la base de données



Sources : Insee, Banque de France, PMI et données financières ; calculs DG Trésor.

Tableau 1 : Date de parution des principaux indicateurs

	Date de parution	Enquêtes	Principales données « dures»
Mois 1	J0	Insee et PMI (mois 0)	
	J15	Banque de France (mois 0)	
	J30	Insee et PMI (mois 1)	Consommation en biens (mois 0)
Mois 2	J45	Banque de France (mois 1)	IPI (mois 0)
	J60	Insee et PMI (mois 2)	Consommation en biens (mois 1)
Mois 3	J75	Banque de France (mois 2)	IPI (mois 1)
	J90	Insee et PMI (mois 3)	Consommation en biens (mois 2)
Mois 4	J105	Banque de France (mois 3)	IPI (mois 2)
	J120	**Publication du PIB**	

Note de lecture : Outre les données déjà présentes dans la base J30, la base J45 contient les enquêtes de la Banque de France correspondant au 1er mois du trimestre ainsi que l'IPI du mois précédent le trimestre considéré.
Note : Ce tableau fait référence aux indicateurs conjoncturels « classiques » ; il ne mentionne pas les variables financières qui sont disponibles quotidiennement, en temps quasi réel.

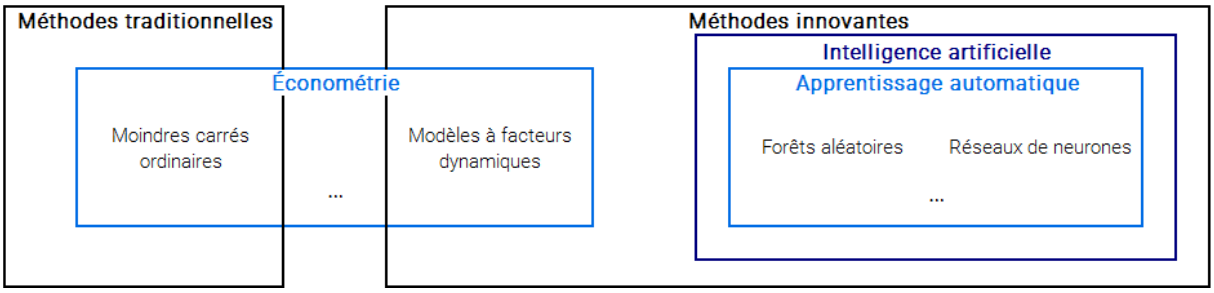
(1) On parle de donnée quantitative ou donnée « dure » par opposition aux données de nature plus qualitative comme les données d'enquêtes. Il peut par exemple s'agir de l'indice de production industrielle (IPI).
(2) Les enquêtes de conjoncture résultent de sondages réalisés le plus souvent mensuellement auprès de panels d'entreprises ou de ménages.
(3) Les PMI sont également des indicateurs conjoncturels. Ils sont produits par l'institut Markit et couvrent de nombreux pays.

2. De nouvelles méthodes statistiques permettent de traiter ces grands ensembles de données

Traditionnellement, la prévision économique repose sur des modèles économétriques, qui consistent à postuler une relation le plus souvent linéaire entre, d'un côté, la variable à prévoir (ici le PIB trimestriel) et, de l'autre, des variables explicatives qui sont bien corrélées avec cette dernière.

À ces techniques économétriques se sont ajoutées dernièrement des techniques d'apprentissage automatique, ou *machine learning*, qui relèvent davantage du domaine de l'intelligence artificielle (*cf.* schéma 1). Pour utiliser des techniques économétriques, le prévisionniste doit faire des hypothèses *a priori* sur la forme de la relation entre le PIB et ses variables explicatives, alors que l'apprentissage automatique a une approche heuristique beaucoup plus ouverte, grâce à des algorithmes qui vont apprendre de leurs erreurs.

Schéma 1 : Économétrie et apprentissage automatique



Note : Ne sont présentées dans ce schéma que les méthodes utilisées dans cette étude.

Les modèles linéaires traditionnels utilisés seuls ne suffisent pas à traiter les grands ensembles de données : d'une part, de nombreuses variables sont corrélées ; d'autre part, les procédures de sélection préalable de variables[4] sont mal adaptées aux cas où le nombre de variables est supérieur au nombre d'observations disponibles pour chacune d'elles. Dans ce cas, des méthodes plus récentes, telles que les modèles à facteurs dynamiques (MFD), les forêts aléatoires ou les réseaux de neurones, sont nécessaires pour exploiter ces importants ensembles d'information.

Les MFD font partie du champ de l'analyse dite factorielle[5], dont le principe consiste à résumer en un nombre relativement faible de composantes, appelées facteurs, un ensemble important d'informations[6]. Dans le cadre de la prévision de la croissance trimestrielle du PIB, les facteurs estimés sont ensuite utilisés comme variables explicatives dans le cadre d'un modèle économétrique linéaire classique[7].

La méthode dite des forêts aléatoires[8] est une généralisation de la méthode des arbres de décision, qui vise à prévoir la valeur du PIB à partir d'un ensemble de critères sur les variables explicatives (*cf.* encadré 1). Dans la méthode des forêts aléatoires, on commence par tirer au hasard des échantillons de données parmi l'ensemble des données disponibles, puis, pour chaque échantillon, on optimise un arbre de décision pour prévoir le PIB. L'agrégation des prévisions de chaque arbre permet d'obtenir la prévision finale de la forêt. Cette méthode permet également de classer les variables explicatives par ordre d'importance pour la prévision de la variable d'intérêt[9].

La méthode dite des réseaux de neurones repose quant à elle sur des modèles conçus à l'origine en s'inspirant du fonctionnement des neurones biologiques (l'encadré 2 décrit notamment le principe de l'architecture de ces réseaux, et la façon dont il est estimé).

(4) Voir Krolzig H.-M. et D. Hendry (2000), "Computer automation of general-to-specific model selection procedures", *Oxford department of economics discussion series*.
(5) Pour une présentation détaillée, voir Bessec & Doz (2011), « Prévision de court terme de la croissance du PIB français à l'aide de modèles à facteurs dynamiques », *Document de travail de la DG Trésor* n° 2011/01.
(6) Ces modèles constituent aussi une réponse concrète au besoin de parcimonie. Ils permettent en effet d'éviter le surajustement, c'est-à-dire une situation où l'ensemble de variables exogènes explique très bien le PIB observé dans l'échantillon, mais où l'intégration de nouvelles observations conduit à une très forte dégradation des capacités prédictives du modèle.
(7) La méthode des moindres carrés ordinaires est utilisée pour obtenir la prévision.
(8) Voir Breiman (2001), "*Statistical Modeling: The Two Cultures*".
(9) Il est en effet possible de mesurer l'importance de chaque variable en calculant la hausse de l'erreur de prévision lorsqu'on perturbe les valeurs de ces variables. Si l'erreur de prévision augmente peu, alors cela signifie que cette variable avait peu d'importance pour la prévision. Au contraire, si l'erreur de prévision augmente significativement, alors cette variable est importante pour la prévision.

Encadré 1 : Les arbres de décision et les forêts aléatoires

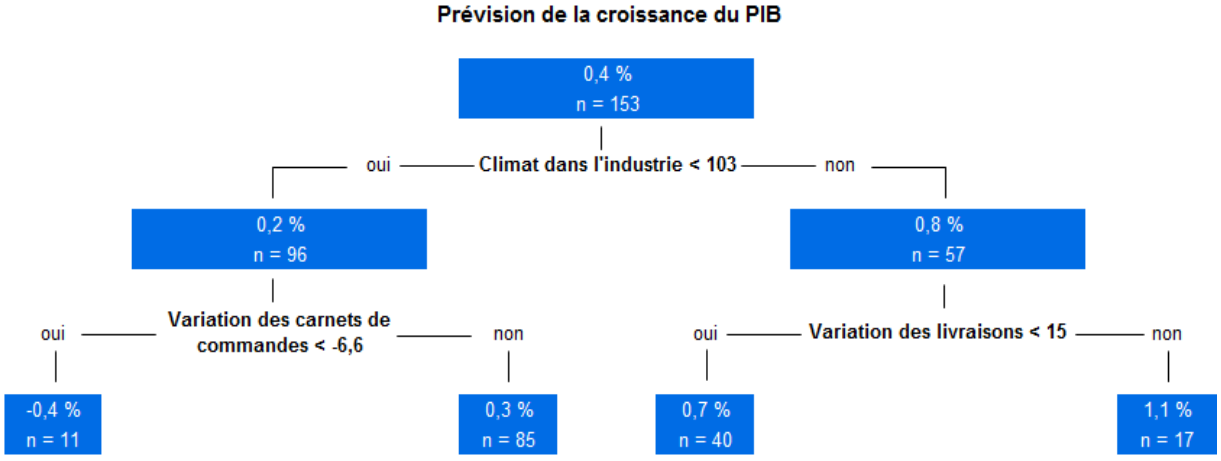
Une forêt aléatoire est un ensemble d'arbres de décision, construits sur des sous-ensembles de données sélectionnées aléatoirement.

Un arbre de décision vise à prévoir une variable d'intérêt à partir d'un ensemble de critères prédéterminé. Sa construction est réalisée *via* un processus itératif, qui vise à séparer, à chaque étape, l'échantillon de départ en deux sous-groupes selon un critère portant sur une variable, de façon à minimiser la variance intra-groupe et maximiser la variance inter-groupes : on définit ainsi un nœud, qui comporte un critère de décision portant sur une des variables de l'échantillon. La procédure s'arrête lorsque le nombre d'observations des différents groupes obtenus est inférieur à un seuil défini *ex ante*.

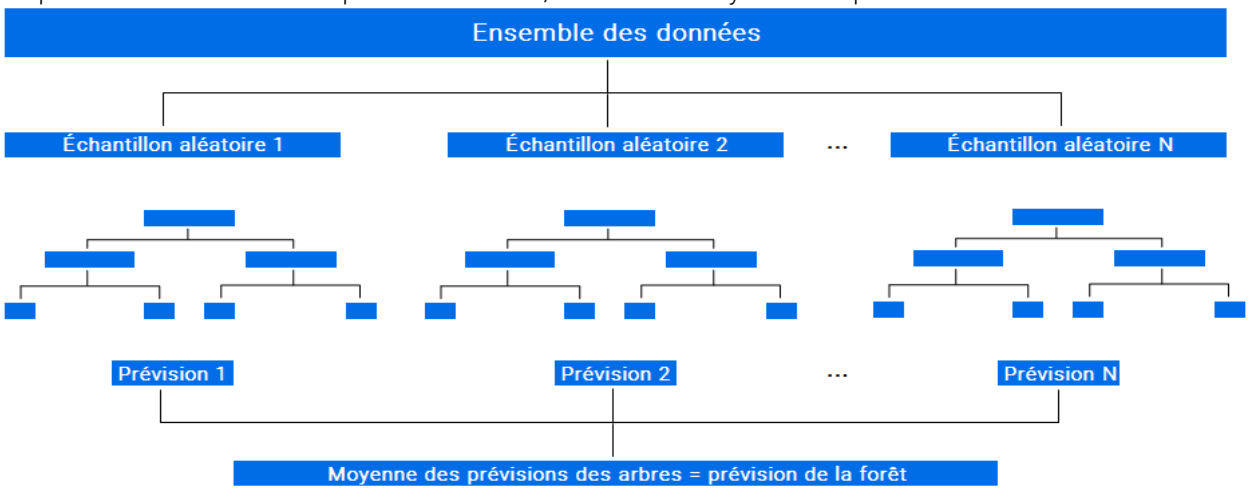
La méthode des forêts aléatoires consiste à générer une multitude d'arbres de décision. Chaque arbre de la forêt est construit sur un sous-ensemble d'observations constitué par tirage avec remise (méthode dite de *bootstrapping*) à partir de la base de données initiale. La constitution par *bootstrap* de l'échantillon permet de rendre le modèle moins sujet au surajustement. De plus, une seconde dimension d'aléa est introduite à chaque scission en différents nœuds, puisque seules certaines variables de la base de données, choisies aléatoirement, sont testées. Ainsi les mêmes variables discriminantes ne sont pas utilisées dans tous les arbres.

La figure ci-dessous représente un arbre de décision, une fois que ses paramètres sont connus. Dans chaque case (ou nœud), la première ligne correspond à la moyenne de la variable d'intérêt (ici la croissance du PIB) des observations se situant dans l'échantillon de ce nœud, et la seconde ligne au nombre d'observations dans cet échantillon. En dessous, on peut lire la condition qui scinde les observations en deux groupes. Cette condition est construite de manière à obtenir des données les plus homogènes (sur la base d'un critère de variance) possibles au sein de ce nœud.

Par exemple, dans cet exemple, pour une observation pour laquelle le climat dans l'industrie vaut 102, la variation des carnets de commande vaut 5 et celle des livraisons vaut 15, la prévision de PIB est de 0,3 %. En effet, le premier nœud (climat = 102 < 103) conduit au sous-arbre de gauche, et le nœud relatif aux carnets de commande (5 > −6,6) conduit à la partie de droite de ce sous arbre. La variable relative aux livraisons n'est pas utilisée pour cette observation.



Lorsque tous les arbres ont été construits, la prévision du PIB correspond à la moyenne des prévisions issues de chaque arbre. Plus précisément, lorsqu'arrive une nouvelle observation, on lui applique un à un chacun des arbres de décision ainsi construits pour en déduire autant de prévisions du PIB, et on fait la moyenne des prévisions obtenues.

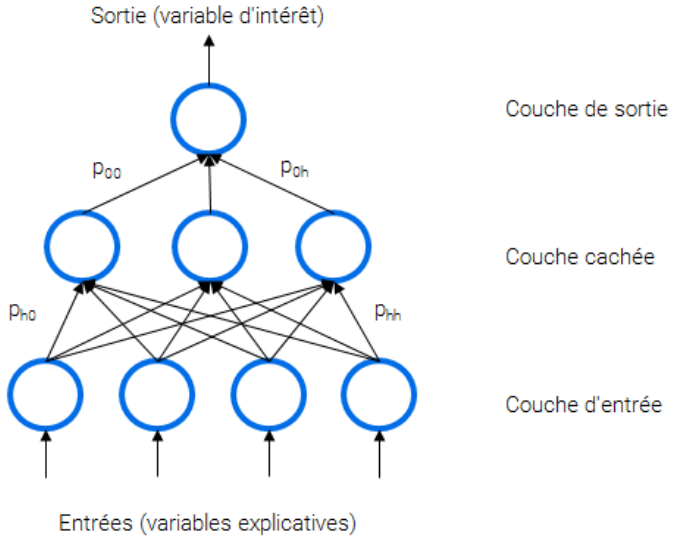


Encadré 2 : Les réseaux de neurones

Les réseaux de neurones sont un ensemble de fonctions qui visent à partir de variables observées à prévoir une variable d'intérêt. La forme de ces fonctions est déterminée à partir d'une base de données, et la prévision est ensuite réalisée à partir d'une nouvelle observation.

Un réseau de neurones est constitué de plusieurs « neurones », dont l'objet est d'agréger des informations, organisés en plusieurs couches : une couche d'entrée, une couche de sortie, et éventuellement une ou plusieurs couches intermédiaires dites « cachées ». Le nombre de neurones d'entrée correspond au nombre de variables explicatives entrant dans le modèle, tandis que le neurone de sortie correspond à la variable d'intérêt et indique donc la prévision de l'évolution du PIB dans le cas d'espèce. Dans une couche, chaque neurone calcule la somme de ses entrées pondérées par les poids affectés (p00, …, phh) et lui applique une fonction de transfert (logistique, sigmoïde…) permettant d'obtenir sa sortie.

La spécification du réseau de neurones dépend donc de l'estimation de ses paramètres (les poids et le biais), une fois choisie son architecture (le nombre de couches, de neurones par couches, les fonctions de transfert, *etc.*). Les poids et les biais sont estimés *via* une méthode numérique qui permet de minimiser l'erreur de prévision[a] ; alors que le nombre de couches et de neurones ainsi que les fonctions de transfert sont le plus souvent déterminés sur des critères empiriques[b].



a. L'algorithme utilisé ici est celui de la descente de gradient.
b. En pratique, la littérature indique que les réseaux de neurones composés d'une seule couche cachée sont performants et sont les plus utilisés pour des finalités similaires à ce travail (Kaastra et Boyd, 1996).

3. Les données d'enquêtes détaillées permettent d'améliorer les prévisions de croissance du PIB

Les trois types de modèle – MFD, forêts aléatoires et réseau de neurones – sont appliqués à trois bases de données différentes :

- la base « étroite », constituée de données d'enquêtes agrégées, de données « dures » et de données financières ;
- la base « large », constituée de données d'enquêtes détaillées[10], de données « dures » et de données financières ;
- la base « restreinte », constituées des 100 variables les plus importantes sélectionnées dans la base large par une méthode de forêt aléatoire.

Pour un trimestre, les bases sont construites et les prévisions sont réalisées tous les 15 jours, au gré de la publication des différentes variables. Au total, on dispose donc pour chaque quinzaine, de huit méthodes d'estimation du PIB[11].

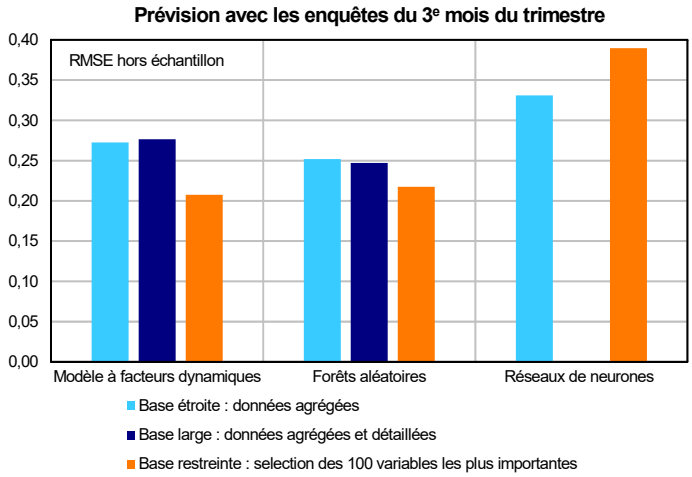
(10) Les données dites détaillées correspondent aux données d'enquêtes représentatives d'un sous-secteur d'activité, par exemple le secteur automobile. En comparaison, les données d'enquêtes qui portent sur un secteur dans son ensemble, comme l'industrie, sont dites agrégées.
(11) Les réseaux de neurones ne sont pas réalisés sur la base « large », le temps de calcul étant trop long.

La comparaison des estimations résultant de ces huit méthodes montre qu'il ne suffit pas d'augmenter la taille des bases de données pour améliorer la prévision, même avec les nouvelles techniques. En effet, les différents modèles donnent des résultats relativement similaires qu'ils soient appliqués à la base étroite ou à la base large (*cf.* graphique 2)[12].

Cependant les résultats s'améliorent sensiblement lorsque les modèles sont appliqués à une base restreinte, c'est-à-dire à l'ensemble de données présélectionnées par les forêts aléatoires dans la base exhaustive.

Les réseaux de neurones semblent en revanche moins précis sur toutes les bases, même la base restreinte. C'est probablement dû au fait que les variables dont on dispose sont trop courtes et trop nombreuses pour permettre l'estimation performante des paramètres nécessaires. Ce problème se pose moins pour la prévision de variables financières, cotées quasi quotidiennement sur les marchés, ce qui fournit une quantité d'observations suffisante pour l'estimation et explique la popularité de ces méthodes dans le secteur financier.

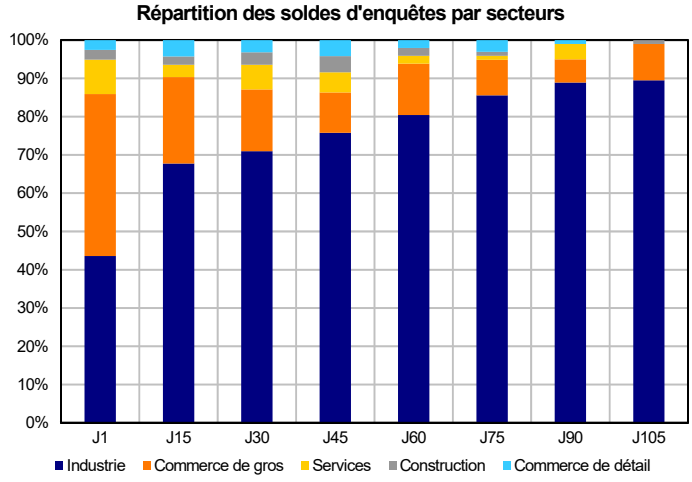
Graphique 2 : Erreur de prévision en fonction du modèle et de la base de données

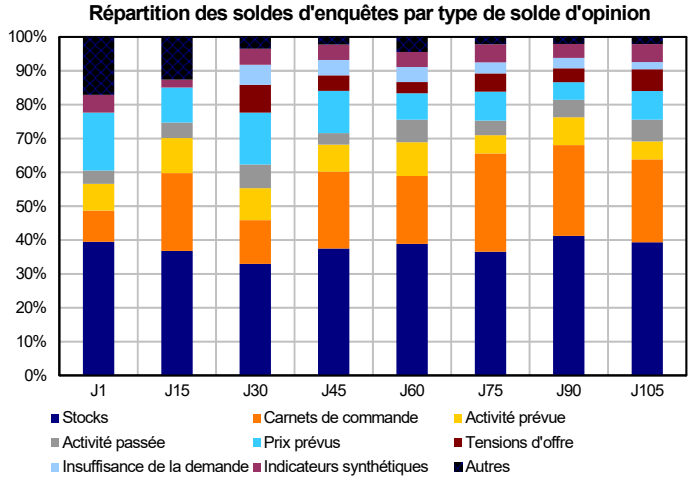


Sources : Insee, Banque de France, PMI et données financières ; calculs DG Trésor.

Note de lecture : Ce graphique présente le RMSE calculé hors échantillon d'estimation en fonction du modèle et de la base de données utilisée. Les réseaux de neurones sur données agrégées et désagrégées n'ont pas pu être estimés en raison d'un temps de calcul trop long.

Graphiques 3 : Répartition des soldes d'enquêtes par secteur et par type de solde d'opinion dans la « base restreinte » de 100 variables





Sources : Insee, Banque de France, PMI et données financières ; calculs DG Trésor.

Note de lecture : Ces graphiques présentent la répartition des soldes d'enquêtes sélectionnés par les forêts aléatoires par secteurs et par types de solde d'opinion, selon l'horizon de prévision. J15 correspond par exemple à la base de données sélectionnée pour une prévision réalisée au 15e jour du trimestre considéré.

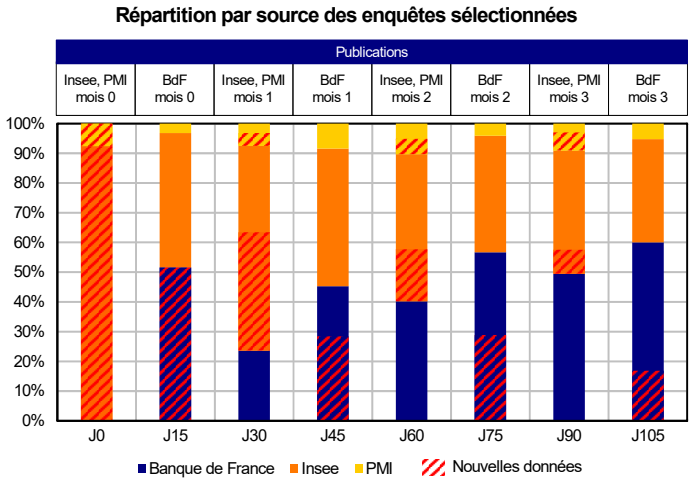
(12) Les modèles sont estimés sur la période 2000-2011 et les prévisions sont réalisées sur la période 2011-2018, avec réestimation des paramètres à chaque point.

La base restreinte contient une majorité de données d'enquêtes détaillées (entre 70 % et 85 % des données selon l'horizon de prévision), montrant ainsi que le détail sous-sectoriel des enquêtes apporte bien une information complémentaire sur l'activité économique.

Plus précisément, ces données sélectionnées par les forêts aléatoires comportent une majorité de données d'enquêtes liées aux secteurs de l'industrie et du commerce de gros, et portent le plus souvent sur l'opinion des entreprises sur leurs stocks de produits finis ou d'intrants, ainsi que sur leurs carnets de commandes. L'industrie renforce par ailleurs sa prépondérance parmi les données sélectionnées dans la base restreinte à mesure que l'on avance dans le trimestre.

La construction de la base par l'algorithme des forêts aléatoires sélectionne généralement les données les plus récentes, en particulier en début trimestre (*cf.* graphique 4), valorisant ainsi les données nouvellement publiées, quelle que soit leur origine.

Graphique 4 : Répartition des soldes d'enquêtes sélectionnés par source



Sources : Insee, Banque de France, PMI et données financières ; calculs DG Trésor.

Note de lecture : Au 45e jour du trimestre considéré, lors de la publication de l'enquête de la Banque de France pour le 1er mois du trimestre, l'algorithme des forêts aléatoires sélectionne comme données les plus importantes 43 % de données provenant des enquêtes Banque de France, dont 28 % de données « nouvelles » (*i.e.* de données provenant de l'enquête du 1er mois du trimestre), 44 % de données provenant des enquêtes Insee et 8 % de données provenant des enquêtes PMI.

4. Les forêts aléatoires et les modèles à facteurs dynamiques sont surtout performants en début de trimestre

Par rapport aux méthodes traditionnelles, les forêts aléatoires et les MFD améliorent la performance de la prévision de croissance du PIB au début du trimestre. En revanche, les réseaux de neurones fournissent des résultats moins satisfaisants, quel que soit le moment du trimestre considéré.

Plus précisément, par rapport à un modèle linéaire avec présélection plus classique des variables[13], les MFD et les forêts aléatoires donnent de meilleurs résultats pendant les deux premiers mois du trimestre, c'est-à-dire sur la première moitié de la période durant laquelle des prévisions sont réalisées. Cette prédominance est particulièrement forte durant les deux premières quinzaines du trimestre (*cf.* graphique en première page).

Dans la mesure où les premières données disponibles sont les données d'enquêtes, alors que les premières données « dures » ne sont disponibles qu'à la fin du deuxième mois du trimestre, les nouvelles méthodes permettent de choisir plus efficacement les enquêtes à utiliser que les algorithmes de sélection de variables plus classiques. En revanche, sur le dernier mois du trimestre et le mois suivant, lors duquel l'estimation de la croissance n'est pas encore disponible, la méthode traditionnelle tire plus pleinement profit de la publication de l'indice de production industrielle du premier mois du trimestre[14].

Les nouvelles méthodes sont par ailleurs mobilisables dans des temps de calcul très raisonnables, si bien que le gain en précision n'est pas acquis au prix d'un rallongement de la durée d'estimation.

(13) Les modèles sont ici comparés avec un modèle linéaire appliqué à un ensemble de variables sélectionnées avec la méthode *General to specific* (GETS) (voir Krolzig et Hendry (2000), "*Computer automation of general-to-specific model selection procedures*"), composé uniquement des indicateurs synthétiques des enquêtes et des principales données « dures ».

(14) Dans le modèle de référence, un poids important est donné à l'indice de production industrielle. Dans la méthode des forêts aléatoires, seule une partie des variables explicatives est testée à chaque scission, pour éviter le surajustement au modèle. Ainsi, mécaniquement, l'indice de production industrielle ne va pas intervenir sur chaque arbre, et son poids sera beaucoup plus faible dans la forêt que dans le modèle de référence.

En revanche, contrairement aux modèles plus traditionnels, dans ces méthodes il n'est pas possible d'analyser directement les contributions des différentes variables explicatives à la prévision du PIB[15]. Les méthodes nouvelles fournissent une estimation rapide de la croissance qui ne se substitue donc pas à un diagnostic sectoriel détaillé sur l'économie.

Maël BLANCHET, Mélanie COUEFFE

(15) Les contributions des variables à la prévision ne s'obtiennent pas immédiatement pour le modèle à facteurs dynamiques, mais peuvent être obtenues après quelques calculs *ad hoc*. En revanche, les contributions à la prévision pour le *random forest* ne sont pas calculables.

Éditeur :

Ministère de l'Économie
et des Finances
Direction générale du Trésor
139, rue de Bercy
75575 Paris CEDEX 12

Directeur de la Publication :

Bertrand Dumont

Rédacteur en chef :

Jean-Luc Schneider
(01 44 87 18 51)
tresor-eco@dgtresor.gouv.fr

Mise en page :

Maryse Dos Santos
ISSN 1777-8050
eISSN 2417-9620

Derniers numéros parus

Décembre 2019

N° 253 Les exportations françaises de biens vers l'Union européenne
Orhan Chiali

N° 252 Le recours à la modélisation macroéconomique dans l'évaluation des politiques publiques
Cyril De Williencourt, Florian Jacquetin

Novembre 2019

N° 251 Enjeux du *gender budgeting* en France
Axel Brunetto, Colette Debever, Mounira Nakaa, Louise Rabier

N° 250 Plateformes numériques et concurrence
Marion Panfili

https://www.tresor.economie.gouv.fr/Articles/tags/Tresor-Eco

Direction générale du Trésor

@DGTrésor

Pour s'abonner à la *Lettre Trésor-Éco* : tresor-eco@dgtresor.gouv.fr