Prévoir en continu la croissance française :
Un essai à partir de différents modèles d'apprentissage automatique

Paul-Armand Veillon
Département de la conjoncture, Insee*

L'Insee publie chaque trimestre dans la Note de conjoncture *sa prévision de croissance trimestrielle du PIB pour le trimestre en cours et le ou les deux trimestres suivants. Cette prévision repose sur celles de chacune des composantes du PIB telles que la consommation des ménages ou la production industrielle. Les prévisions de ces composantes sont elle-mêmes établies à partir d'indicateurs conjoncturels comme le climat des affaires ou l'indice de la production industrielle. Si une seule prévision est publiée chaque trimestre, la publication de nouveaux indicateurs est quasi-quotidienne et chacune de ces nouvelles informations est susceptible de faire évoluer l'estimation de croissance économique qui apparaît comme la plus probable à une date donnée. De nouveaux modèles de prévision au jour le jour ou « nowcasting » permettent de tenir compte de ces publications fréquentes de nouveaux indicateurs pour la prévision de la croissance trimestrielle.*

Ces modèles sont élaborés grâce à l'utilisation de méthodes d'apprentissage statistique (dites de « machine learning ») d'une part, et grâce d'autre part à l'accès libre et en temps réel à des centaines d'indicateurs conjoncturels (« open data »). Ainsi, la Réserve fédérale (Fed) d'Atlanta publie depuis 2016 une actualisation de sa prévision de croissance toutes les semaines, en s'appuyant sur un modèle de prévision de ce type.

Ce dossier présente une première proposition de modèles de prévision en continu des variations trimestrielles de la croissance française. Les données utilisées sont notamment les indicateurs conjoncturels publiés par la Banque de France, l'Insee, l'OCDE, Markit et différents services statistiques ministériels (SSM). Plusieurs modèles sont testés, parmi lesquels des modèles d'apprentissage statistique supervisés tels que les forêts aléatoires, et des modèles à facteurs.

Les premiers résultats montrent que la prévision peut varier significativement au cours d'un trimestre (entre +0,2 % et +0,4 % par exemple pour le troisième trimestre 2019), ces variations faisant suite à la publication d'un indicateur en forte hausse ou forte baisse. Les modèles utilisés tendent à converger à la fin du trimestre et ont une erreur, mesurée par la racine de l'erreur quadratique moyenne de prévision, ou Root Mean Squared Forecast Error (RMSFE), d'environ 0,20 point. L'erreur de prévision varie entre 0,28 point au début du trimestre et 0,20 point à la fin du trimestre. L'intervalle de confiance à 80 % pour la prévision de croissance du troisième trimestre 2019 est ainsi passé de [–0,1 ; 0,6] en juillet à [0,0 ; 0,5] fin septembre. ■

* au moment de la rédaction de cette étude. L'auteur remercie Clément Rousset pour son aide.

Prévoir en continu la croissance française

Les prévisions de croissance de l'Insee reposent en très grande partie sur les enquêtes de conjoncture ainsi que sur des indices tels que l'IPI.

La publication de la première estimation du PIB, réalisée par les comptables nationaux, est disponible seulement un mois après la fin de chaque trimestre. Pour autant, la prévision des variations du PIB à court terme constitue un enjeu majeur pour les décideurs économiques. Leurs décisions s'appuient alors sur les prévisions conjoncturelles publiées régulièrement par différents instituts ou entreprises. Par exemple, l'Insee publie dans sa *Note de conjoncture* du mois de décembre et de juin ses prévisions à l'horizon de deux trimestres. Ces premiers chiffres sont révisés lors des exercices de prévision de mars et en octobre. Les prévisions de l'Insee reposent essentiellement sur les enquêtes de conjoncture et sur des indices conjoncturels, tels que les indices de production industrielle (IPI) ou les indices de chiffres d'affaires (CA). Ces prévisions sont ensuite intégrées dans un cadre comptable répliquant celui des comptes nationaux trimestriels et garantissant une cohérence du point de vue des équilibres comptables.

L'essor de nouvelles méthodes statistiques et la multiplication des sources de données permettent de réaliser une prévision en temps réel

Bien qu'une seule prévision soit publiée chaque trimestre, celle-ci peut s'affiner au cours du trimestre de prévision, après chaque nouvelle publication d'indicateurs. La multiplication des sources de données et l'apparition de nouvelles méthodes de prédiction rendent aujourd'hui possible la prévision en continu de l'activité économique à l'aide d'un grand nombre de variables conjoncturelles. Ces méthodes innovantes, dites de « *nowcasting* », proposent un cadre statistique cohérent pour réaliser une prévision quotidienne de la variation du PIB. À titre d'exemple, la réserve Fédérale d'Atlanta, pionnière en la matière, publie presque tous les jours une nouvelle prévision tenant compte de la publication des indicateurs économiques les plus récents. Ceux-ci sont aussi variés que le nombre de permis de construire, les capacités de production, les indicateurs PMI ou bien les enquêtes auprès des directeurs d'achats.

Ces méthodes sont mobilisées ici pour construire un nouvel outil de prévision en continu des variations trimestrielles du PIB français. La base de données utilisée comporte plus d'une centaine de variables temporelles publiées par quatre instituts différents. Une prévision quotidienne est réalisée par ces méthodes capables de synthétiser un très grand nombre de variables en une prévision.

Deux résultats soulignent l'intérêt d'un tel outil. Tout d'abord, l'erreur de prévision décroît continûment au cours du trimestre de prévision, de plus d'un tiers entre le début et la fin du trimestre. Ainsi, à chaque nouvelle prédiction, la qualité, mesurée par l'erreur de prévision empirique, s'améliore sensiblement et la meilleure prévision est celle qui utilise les informations les plus récentes. De surcroît, la prévision, qui varie au cours du trimestre, est fortement sensible à la publication de tout nouvel indicateur. Cette prévision n'est donc pas une donnée figée au cours du trimestre, elle évolue continuellement en fonction de l'information disponible.

La diversité et la fréquence des données disponibles permettent de réaliser une prévision en continu

Le conjoncturiste utilise surtout des données qualitatives sur l'activité économique et des données quantitatives sur la production ou la consommation

Les enquêtes de conjoncture sont les premières données utilisées par les conjoncturistes pour leurs prévisions. Celles-ci sont ensuite enrichies par la publication des premiers indicateurs quantitatifs tels que, entre autres, l'indice de la production industrielle ou les données d'immatriculation. Bien que ces indicateurs fournissent une information plus quantitative que les questions qualitatives des enquêtes de conjoncture, leur délai de publication, supérieur à un mois, peut limiter leur intérêt pour la prévision. À titre d'exemple, l'enquête de conjoncture dans l'industrie manufacturière de l'Insee est publiée 25 jours après le début du mois considéré tandis que l'indice de la production industrielle est publié

40 jours après la fin du mois considéré. Ainsi, à la fin d'un trimestre donné, le conjoncturiste dispose des données d'enquête pour toute la période mais seulement de données quantitatives pour le premier mois. L'essor des données massives (*Big Data*) permet également d'exploiter de nouvelles données telles que les articles de presse, les recherches sur les moteurs de recherche ou encore les données des réseaux sociaux. Cependant, leur apport s'est avéré limité pour la prévision conjoncturelle française (Bortoli et Combes 2015a, Bortoli *et al.* 2017).

Les enquêtes de conjonctures sont les premiers indicateurs de l'activité économique disponibles pour la prévision

L'Insee conduit aujourd'hui une dizaine d'enquêtes de conjoncture aussi bien auprès des ménages que des entreprises des secteurs des services, de l'industrie ou du bâtiment. Leur publication précoce en fait une variable de choix à la disposition du conjoncturiste pour prévoir l'activité économique. Par construction, les enquêtes ont un caractère prospectif : les 20 000 entreprises composant les échantillons des enquêtes de conjoncture sont interrogées sur leur activité, leurs effectifs ou leur production prévue pour les trois prochains mois. Elles sont également interrogées sur la tendance passée de ces variables pour les trois derniers mois. Les modalités des réponses sont « en hausse », « en baisse » et « stable ». Le solde d'opinion, synthétisant les réponses qualitatives, se calcule comme la différence entre les pourcentages de réponses « en hausse » et « en baisse ». Les conjoncturistes déterminent par étalonnage la relation « moyenne » entre ces soldes et l'activité économique pour construire leurs prévisions. D'autres organismes tels que la Banque de France ou l'entreprise Markit réalisent également des enquêtes de conjoncture. Celles-ci apportent une information différente et complémentaire des enquêtes de conjoncture de l'Insee : elles interrogent en effet un échantillon d'entreprises différent sur une période distincte et les questions sont formulées différemment de celles de l'Insee. Les trois indicateurs synthétiques publiés par chacun des organismes, bien que fortement corrélés entre eux, présentent ainsi des fluctuations propres. De plus, il peut être pertinent d'intégrer des données d'enquête portant sur la conjoncture dans la zone euro ou les pays de l'OCDE, telles que celles publiées par cet institut.

Les indicateurs quantitatifs, publiés plus tardivement, apportent, par construction, une information de meilleure qualité sur l'activité économique

Si les enquêtes de conjoncture fournissent un signal utile sur la tendance de l'activité, celui-ci est en partie bruité. En effet, les réponses qualitatives à trois modalités ne peuvent pas donner autant d'information qu'une donnée quantitative. De plus, les questions peuvent être soumises à interprétation (Bortoli *et al.* 2015b). Au contraire, les indicateurs quantitatifs s'appuient sur des données réelles telles que la consommation des ménages ou des données de production. Exceptées les données d'immatriculations, elles sont publiées sous un délai de plus d'un mois, mais donnent une information quantitative au plus proche de la première estimation des comptes trimestriels. Trois indicateurs publiés par l'Insee sont particulièrement déterminants dans la construction de la première estimation du PIB : l'indice de production industrielle, publié sous un délai de 40 jours, est un indicateur avancé de la production industrielle réalisé à partir des données des enquêtes mensuelles de branches. Les séries mensuelles de consommation des ménages en biens, publiées sous un délai de 30 jours, donnent une première estimation des dépenses de consommation finale des ménages. Les indices de chiffres affaires, publiés près de 60 jours après le mois considéré et calculés à partir des déclarations de TVA, renseignent sur les dépenses en services. Pour ces variables, l'acquis de croissance est intégré à l'exercice de prévision. Les variables financières telles que la demande de crédits des particuliers et des entreprises, les taux d'intérêt ou les données boursières ont également une capacité prédictive des variations du PIB. La majeure partie d'entre elles sont publiées mensuellement par la Banque de France et sont intégrées dans la base de données de prévision.

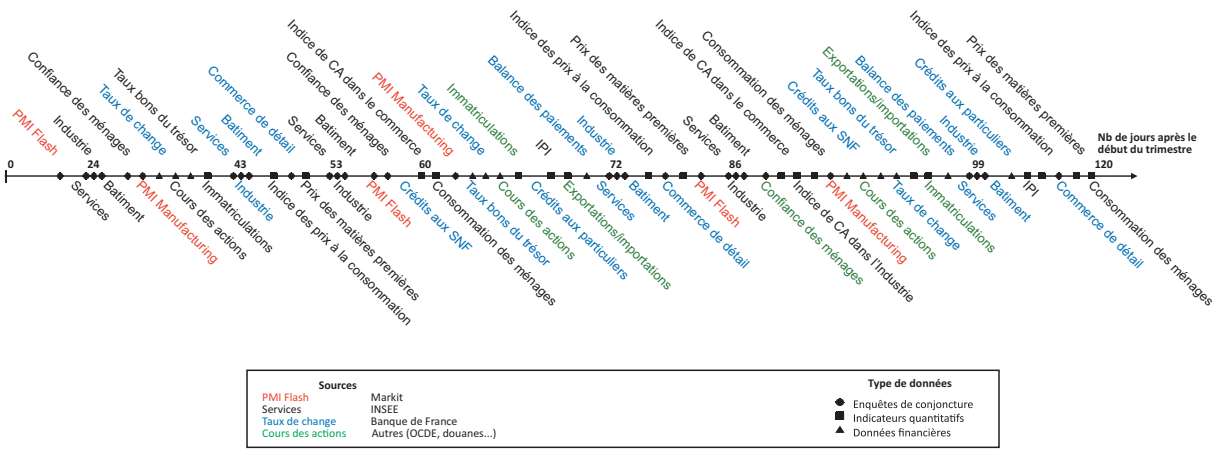
Une nouvelle publication d'indicateurs a lieu à peu près tous les trois jours ouvrés

Sur la *figure* ci-dessous, est représentée la date de publication des principaux indicateurs pertinents pour la prévision conjoncturelle au cours d'un trimestre. Celle-ci commence le premier jour du trimestre considéré et s'achève 30 jours après la fin de ce trimestre, lors de la publication de la première estimation des comptes nationaux trimestriels. Pour un mois précis, les premières données disponibles sont les enquêtes de conjoncture publiées par l'institut Markit et l'Insee respectivement autour de 18 et 24 jours après le début du mois, tandis que l'indice de CA dans l'industrie est publié 89 jours après le début du mois. Au total, sur les quatre mois considérés, de nouvelles données sont publiées au cours de 34 jours sur les 96 jours ouvrés, soit à peu près une nouvelle publication tous les trois jours ouvrés. Ainsi, il est possible de réaliser tous les trois jours une prévision tenant compte d'un nouvel ensemble d'informations. Enfin sur les 64 ensembles de données publiés, 30 sont des enquêtes de conjoncture réalisées par l'Insee, la Banque de France ou Markit, 13 sont des ensembles de données financières publiées par la Banque de France et l'OCDE, 21 sont des ensembles d'indicateurs quantitatifs publiés par la Banque de France, l'Insee et les services statistiques ministériels. Cette diversité d'indicateurs et de sources de données permet de réaliser une prévision à l'aide d'un ensemble d'informations plus large que celui utilisé habituellement par les conjoncturistes mais dont l'exploitation requiert quelques précautions.

Grâce aux méthodes d'apprentissage statistique, il est possible de réaliser une prévision à partir d'un nombre d'indicateurs supérieur au nombre d'observations disponibles

Deux problèmes méthodologiques se posent lors de la prévision en temps réel d'un agrégat économique : comment agréger des données de fréquences (mensuelles ou trimestrielles) et de dates de publication différentes ? Comment réaliser une prévision avec un nombre de variables (N) souvent supérieur au nombre d'observations (T), problème que nous noterons « $N > T$ » ? Tandis que l'agrégation de données hétérogènes ou manquantes est une problématique spécifique à la prévision en temps réel, $N > T$ est un problème de prédiction classique, connu sous le nom de problème de la grande dimension. Nous proposons d'appliquer les solutions proposées par la littérature à la prévision de la première estimation par les comptes nationaux trimestriels de la croissance du PIB.

1 - Calendrier de publication des indicateurs conjoncturels



Plusieurs solutions existent pour faire face au problème des données manquantes

Dubois et Michaux (2006) s'interrogeaient déjà sur « le problème des données manquantes » pour la prévision trimestrielle de la production industrielle à partir des enquêtes mensuelles de conjoncture. Leur proposition était alors de créer trois séries trimestrielles correspondant respectivement au premier, deuxième et troisième mois de chaque trimestre. En fonction de la disponibilité des données, ils intégraient une, deux ou trois de ces séries trimestrielles. Cependant, cette méthodologie présente le défaut de multiplier par trois le nombre de variables, ce qui accentue le problème de dimensionnalité. Une variante courante de ces méthodes, dites de « *bridge equation* », consiste à prédire les mois manquants par un modèle auto-régressif. Cependant, la prolongation des données a pour conséquence d'ajouter de l'inertie à la prévision. Ainsi, le choix a été fait ici de calculer une moyenne trimestrielle des données disponibles à la date de prévision pour les enquêtes de conjoncture ou de prendre l'acquis de croissance pour les autres variables. Ces choix ont l'avantage de privilégier la diversité des sources de données à l'ajout de retards d'un nombre restreint de variables et de rendre plus sensible notre prévision à la publication de nouvelles données.

L'ajout d'un grand nombre de variables améliore certainement la qualité d'ajustement du modèle aux données. Cependant, cet ajustement peut se faire au détriment de la prévision. Dans cette situation, qualifiée de « surapprentissage », le modèle estimé est trop proche des données passées utilisées et n'est pas suffisamment généralisable aux évolutions à venir. On préfère alors un modèle parcimonieux, utilisant un nombre restreint de variables (*voir encadré sur le surapprentissage*).

Une solution peut être de sélectionner un nombre restreint de variables. Dubois et Michaux (2006) ont été les premiers à utiliser pour les prévisions du département de la conjoncture une méthode statistique de sélection de variables de type GETS (*General to specific modelling*). Celle-ci consiste à éliminer successivement les variables non significatives en partant du modèle le plus général et en réalisant un certain nombre de tests de spécification à chaque étape. Là où jusqu'alors la sélection se faisait de manière artisanale ou grâce à des algorithmes moins performants tels que les sélections ascendantes et descendantes[1], l'utilisation de GETS a permis, sous certaines conditions, d'obtenir le meilleur modèle de prévision linéaire.

Les modèles à facteurs permettent de résumer un grand nombre de variables en quelques facteurs

Les modèles à facteurs dynamiques proposent de répondre simultanément aux problèmes de données manquantes et de grande dimensionnalité. Sous l'impulsion des travaux de Stock et Watson (2002) et Doz et al. (2011), ces modèles ont connu un essor rapide et sont aujourd'hui utilisés par de nombreux organismes, tels les Fed ou la BCE. De manière générale, les modèles à facteurs permettent d'obtenir une représentation parcimonieuse d'un ensemble de variables, résumé en un nombre restreint de facteurs. Le plus connu d'entre eux est l'analyse en composantes principales. La représentation dynamique de ces facteurs sous la forme d'un modèle espace-état permet quant à elle de tenir compte des valeurs manquantes. Cette méthode, très attrayante conceptuellement, a été appliquée à la prévision de croissance du PIB français par Bessec et Doz (2012) et elle a également été mobilisée dans ce dossier. Un modèle d'analyse en composantes principales (ACP) qui ne tient pas compte de la dynamique des facteurs et qui est communément utilisé dans la littérature est également testé.

1. Les algorithmes de sélections ascendantes ou descendantes de type *stepwise* permettent de tester seulement un nombre réduit de modèles qui s'avèrent généralement ne pas être les plus performants.

Les modèles d'apprentissage statistique apportent de nouvelles solutions au problème de la grande dimension

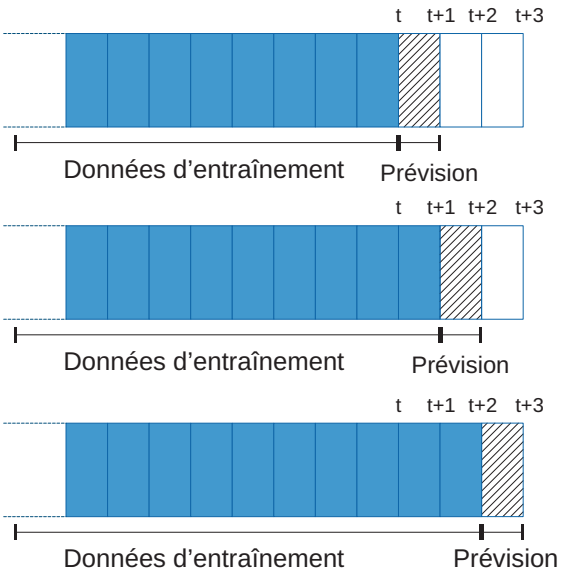
Enfin, les méthodes d'apprentissage automatique (*Machine Learning*, ML) proposent une nouvelle approche pour réaliser les prévisions : la prévision ne repose plus sur une pré-spécification des relations entre la variable endogène et les variables exogènes mais sur l'utilisation d'un algorithme qui trouve le modèle minimisant une fonction objectif. Grâce à leur capacité prédictive, ces algorithmes tels que le LASSO (*Least Absolute Shrinkage and Selection Operator*) ou les forêts aléatoires (Random Forest) ont donné lieu à une littérature émergente portant sur la prévision des agrégats macroéconomiques à l'aide du *Machine Learning*. Biau, Biau et Rouvière (2006) notamment ont appliqué la méthode des forêts aléatoires aux réponses des industriels aux enquêtes de conjoncture de l'Insee pour prévoir la production manufacturière. Cependant, la mise en œuvre de ces méthodes doit respecter un certain nombre de principes élémentaires afin d'éviter l'écueil du surapprentissage. D'autres algorithmes d'apprentissage automatique pourraient être utilisés, tels que les réseaux neuronaux. Cependant, ces modèles dépendent souvent de nombreux paramètres nécessitant une trop grande quantité d'observations pour être optimisés dans le cadre des séries macroéconomiques. Les méthodes retenues sont ici le LASSO[2] et les forêts aléatoires. Le premier modèle permet de construire un modèle linéaire à partir d'un sous-ensemble de variables sélectionnées automatiquement, tandis que le second repose sur la construction d'arbres de décision (*voir encadré*).

La performance de ces méthodes est comparée à celle d'un modèle simple utilisant seulement la dynamique de la variable à prévoir (modèle autorégressif à moyenne mobile ou ARMA en anglais) et d'un étalonnage simple utilisant seulement le climat des affaires France.

2. L'hyper-paramètre de régularisation λ a été choisi par *cross-validation* sur les données d'entraînement.

2 - Calcul du RMSFE

Erreur de prévision



$$e_{t+1}=y^{prev}_{t+1}-y^{realisé}_{t+1}$$

$$e_{t+2}=y^{prev}_{t+2}-y^{realisé}_{t+2}$$

$$e_{t+3}=y^{prev}_{t+3}-y^{realisé}_{t+3}$$

$$RMSFE=\sqrt{1/N\sum_{t} e_t^2}$$

Le modèle de prévision donne des résultats qui varient sensiblement au cours du trimestre de prévision, tandis que son erreur diminue d'environ 40 %

La qualité de la prévision est mesurée par la racine des erreurs de prévision au carré, RMSFE (*Root Mean Squared Forecast Error*). Comme présenté sur la *figure 2*, pour une date *t* donnée, le modèle est entraîné avec les données jusqu'à la date t et une prévision est réalisée pour la date *t+1*. L'erreur à la date *t+1* est calculée comme la différence entre la prévision et la valeur effectivement réalisée à la date *t+1*. Le RMSFE se calcule alors comme la racine carrée de la moyenne des erreurs au carré. Les données d'entraînement commencent au quatrième trimestre 2001 et les erreurs de prévision sont calculées sur la période allant du premier trimestre 2011 au premier trimestre 2019. Dans la suite de cette partie, les données de prévision du troisième trimestre 2019 sont présentées à titre d'illustration avec la prévision de la croissance trimestrielle du PIB du troisième trimestre 2019 comme objectif.

Le *tableau 1* présente le RMSFE et la valeur absolue de l'erreur maximale avec les données disponibles 100 jours après le début du trimestre, soit 20 jours avant la publication de la première estimation des comptes trimestriels. Tous les modèles font mieux que ceux retenus en référence à ce moment de la prévision. Le LASSO et les forêts aléatoires sont les modèles retenus avec le plus faible RMSFE.

Si leurs prévisions évoluent de façon assez similaire, les modèles diffèrent par leur volatilité

Au fil de l'arrivée d'informations nouvelles, la prévision évolue sensiblement et différemment selon les modèles. La *figure 3* présente l'évolution de la prévision de la croissance trimestrielle du PIB français du troisième trimestre 2019 fournie par les modèles de type LASSO, forêts aléatoires et ACP. Les trois modèles suivent globalement la même évolution des prévisions, la principale différence étant leur volatilité ou leur sensibilité aux nouvelles publications. Le modèle ACP est le plus

Tableau 1 - Qualité de la prévision des modèles utilisés

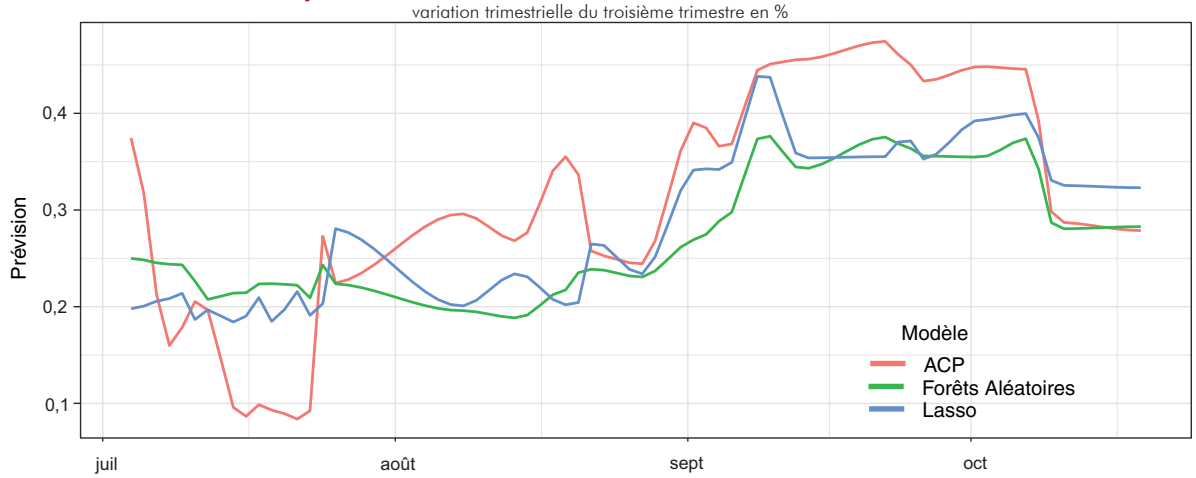
erreurs quadratiques moyennes et erreurs maximales des principaux modèles de prévision entre 2011 et 2019

Modèle	Arima	Climat France	Gets	LASSO	Forêts al.	ACP	Facteur Dynamique
RMSFE	0,33	0,28	0,23	0,20	0,19	0,23	0,22
Erreur Maximale	0,77	0,55	0,65	0,53	0,55	0,62	0,66

Lecture : L'erreur maximale du modèle LASSO à T+100 jours est de 0,53 point (valeur absolue de la différence entre le taux de croissance trimestrielle du PIB prévu et celui réalisé dans l'estimation actuelle) sur la période 2011-2019. Le RMSFE correspondant, de 0,20, est calculé à partir des erreurs de prévisions observées à la date de prévision T+100 jours pour tous les trimestre de la même période.
Sources : Insee, Banque de France, OCDE, Markit, calcul des auteurs

3 - Évolution de la prévision de croissance du PIB au cours du troisième trimestre 2019

variation trimestrielle du troisième trimestre en %



Lecture : le 18 octobre 2019, la prévision à l'aide du modèle des forêts aléatoires est de +0,28 %.
Source : Insee, Banque de France, OCDE, Markit, calcul des auteurs

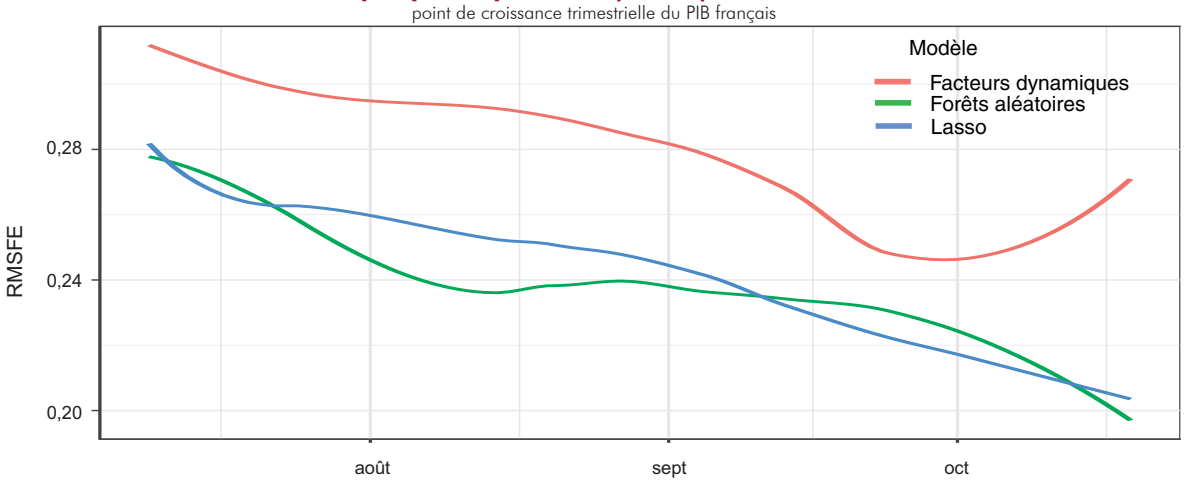
volatil, sa prévision variant entre +0,07 % et +0,47 %, suivi du modèle LASSO dont les résultats varient entre +0,17 % et +0,43 %. Enfin le modèle des forêts aléatoires donne des prévisions variant entre +0,18 % et +0,38 %. Une plus grande volatilité implique également une erreur absolue maximale du modèle plus élevée. Calculée sur l'ensemble des trimestres précédant la dernière estimation, celle-ci varie de 0,53 % à 0,77 % en valeur absolue selon les modèles. Cependant, le LASSO et le modèle des forêts aléatoires, soit les modèles avec les plus faibles RMSFE, se superposent presque parfaitement tout au long du trimestre. Dans la suite de cette partie, nous étudierons plus particulièrement le modèle des forêts aléatoires qui présente le double avantage d'avoir un RMSFE et une erreur maximale relativement faibles.

Au cours du trimestre considéré, l'erreur de prévision diminue de 40 %.

Tout comme la prévision évolue, l'erreur de prévision se réduit au cours du trimestre, au fur et à mesure de la disponibilité des informations sur la situation économique en cours. La *figure 4* présente l'évolution de la prévision par la méthode des forêts aléatoires et la réduction de son erreur de prévision au cours du troisième trimestre 2019. L'erreur de prévision diminue d'environ 40 % entre le début du trimestre et la veille de la publication des comptes nationaux. Autrement dit, l'intervalle de confiance à 80 % de cette prévision est de +/– 0,38 point de pourcentage au début du trimestre de prévision contre +/– 0,25 point à la fin de la période de prévision.

4 - Évolution de l'erreur empirique de prévision (RMSFE) au cours du troisième trimestre 2019

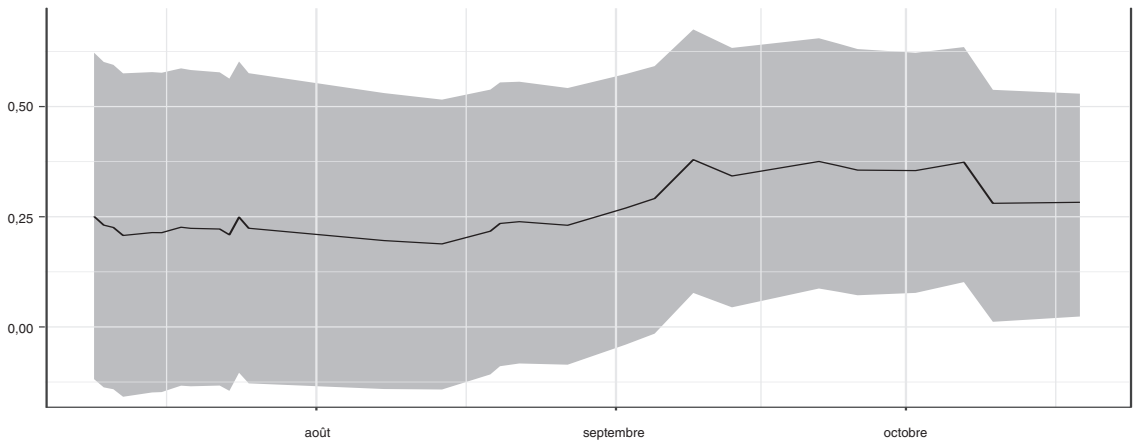
point de croissance trimestrielle du PIB français



Lecture : le 1er septembre 2019, l'erreur de prévision (RMSFE) à l'aide du modèle des forêts aléatoires est de 0,24 point de croissance du PIB.
Source : Insee, Banque de France, OCDE, Markit, calcul des auteurs

5 - Prévision de croissance du troisième trimestre 2019 avec intervalle de confiance à 80 % à l'aide du modèle des forêts aléatoires

variation trimestrielle en %



Lecture : le 18 octobre 2019, la prévision de croissance du PIB français du troisième trimestre à l'aide du modèle des forêts aléatoires est de +0,28 %. L'intervalle de confiance à 80 % est compris entre +0,03 et +0,52.
Source : Insee, Banque de France, OCDE, Markit, calcul des auteurs

Prévoir en continu la croissance française

Les variations de la prévision s'expliquent par la publication d'indicateurs particuliers

La prévision de croissance du troisième trimestre 2019 atteint un point bas mi-août à +0,18 %. Cela coïncide avec la publication de deux indicateurs conjoncturels particulièrement déterminants dans la prévision (cf. ci-après) : l'indice de production industrielle de juin, publié le 9 août, perd 2,2 %, et le solde sur les prévisions de production dans l'industrie manufacturière de l'enquête mensuelle de conjoncture de la Banque de France perd deux points à la même date. Un mois plus tard, la prévision s'élève à +0,39 %, soutenue par la progression de l'indicateur du climat des affaires dans l'industrie, publié par la Banque de France (+3,4 points) et le léger rebond de l'indice de la production industrielle au mois de juillet, publié le 10 septembre (+0,3 %). Début octobre, la chute brutale de la prévision s'explique d'une part par la baisse des indices PMI et par la publication de l'IPI pour le mois d'août, en baisse également.

L'algorithme des forêts aléatoires permet d'identifier les variables déterminantes dans la prévision de la croissance trimestrielle du PIB

Pour les forêts aléatoires, il est possible de mesurer l'importance de chacune des variables dans la prévision (*voir encadré Prévision à l'aide de forêts aléatoires*). Celle-ci est calculée comme le gain prédictif associé à chaque variable. Par exemple, le solde sur la production future dans l'industrie manufacturière permet de réduire de 13,5 % le RMSFE pour une prévision réalisée mi-juillet. Les *tableaux 2 et 3* affichent les dix variables les plus déterminantes pour les prévisions réalisées respectivement aux mois d'octobre et juillet, soit un mois après la fin du troisième trimestre 2019 et le premier mois de celui-ci. La grande majorité des indicateurs déterminants sont relatifs à l'industrie manufacturière. En effet, la production industrielle a une très forte contribution aux variations trimestrielles du PIB, contribution en proportion plus importante que sa part dans la valeur ajoutée de l'ensemble des secteurs. Par ailleurs, les variables les plus déterminantes sont issues d'un grand nombre de sources différentes : OCDE, Insee, Banque de France, Markit. La multiplicité des sources permet donc d'améliorer significativement la prévision. Enfin, les indicateurs qui s'apparentent à des signaux faibles, tels que le cours des actions, font partie des variables déterminantes en juillet mais sont remplacées en octobre par des indicateurs quantitatifs tels que l'indice de la production industrielle.

Tableau 2 - Importance des variables dans la prévision à l'aide du modèle à forêt aléatoire à mi-octobre (*T*+100)

Variables	Importance
Autres produits industriels (C5), évolution des commandes reçues – Banque de France, septembre	12,7
Industrie manufacturière, évolution passée de la production – Banque de France, septembre	12,4
Industrie manufacturière, prévision de production – Banque de France, septembre	10,8
Indice de la production industrielle, industrie manufacturière – Insee, août	10,0
Indice de la production industrielle, biens intermédiaires – Insee, août	9,9
Indice de la production industrielle, biens d'investissement – Insee, août	9,1
PMI manufacturier – Markit, septembre	8,5
Autres produits industriels (C5), prévisions de la production– Banque de France, septembre	7,0
Climat des affaires dans l'industrie du bâtiment – Insee, septembre	6,9
Biens d'équipements (C3), prévisions de la production– Banque de France, septembre	6,7
PMI manufacturier, nouvelle commandes – Markit, septembre	6,6
Consommation mensuelle des ménages, biens manufacturés – Insee, août	5,6

Ces nouveaux outils permettent donc de suivre en temps réel l'évolution de la prévision en fonction de la publication des indicateurs. Ils permettent aux conjoncturistes, en complément de ces outils habituels, de répondre à de nouvelles questions telles que : quelle a été l'évolution de la prévision au cours du trimestre ? quels indicateurs ont particulièrement influencé la prévision ? ou encore quelle est la précision de notre prévision a un instant donné ? Cependant, ce premier prototype présente quelques limites et nécessitera des recherches complémentaires. En premier lieu, l'apprentissage automatique ou *Machine Learning* est un domaine de recherche qui a effectué une profonde mutation ces dix dernières années et qui ne cesse de se développer. Ainsi, les modèles d'apprentissage automatique présentés dans ce dossier peuvent eux-mêmes devoir évoluer en fonction des progrès réalisés dans ce domaine. Par ailleurs, la prévision en temps réel de la croissance trimestrielle repose sur une analyse statistique et ne peut pas se substituer à une analyse économique. Elle ne permet pas d'établir de relation causale entre l'évolution d'un indicateur et la croissance du PIB mais traduit une corrélation entre certains indicateurs et l'évolution du PIB, établie sur des données passées. Enfin, sa performance pendant un trimestre ne permet pas d'attester de sa robustesse. Il est donc impossible d'anticiper son comportement en temps de crise, période pendant laquelle les indicateurs évoluent, par définition, très différemment de leur tendance passée. ■

Tableau 3 - Importance des variables dans la prévision à l'aide du modèle à forêt aléatoire à mi-juillet (*T*+15)

Variables	Importance
Industrie manufacturière, prévision de production – Banque de France, juin	13,5
Composite Index, business survey, OCDE – OCDE, juin	9,5
Cours des actions, France – OCDE, juin	8,5
Autres produits industriels (C5), prévision de production – Banque de France, août	8,0
Cours des actions, USA – OCDE, juin	7,8
Indicateur de retournement conjoncturelle dans les services – Insee, juin	6,9
Matériels de transport (C4), évolution des commandes reçues – Banque de France, juin	6,7
Climat des affaires, industrie manufacturière – Insee, juin	6,6
Matériels de transport (C4), évolution des commandes étrangères reçues – Banque de France, juin	6,6
PMI manufacturier, nouvelles commandes– Markit, juin	6,1
Autres produits industriels(C5), évolution des commandes reçues – Banque de France, juin	5,9

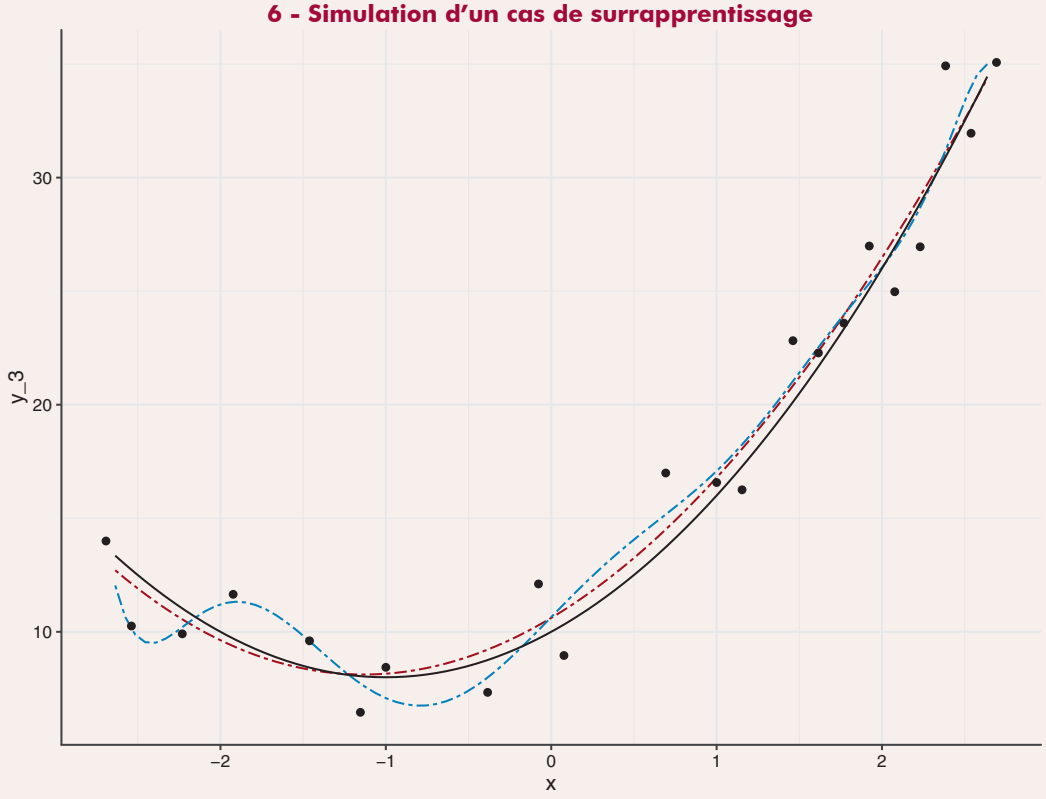
Lecture : mi-juillet, la variable Composite Index, Business survey de l'OCDE améliore de 13,5 % le RMSFE
Source : Insee, Banque de France, OCDE, Markit, calcul des auteurs

Encadré 1 : Le surapprentissage

L'objectif d'un modèle prédictif est de formuler une prévision la plus exacte possible d'une variable inobservée à partir d'observations auxiliaires. À cet effet, la priorité n'est pas de maximiser la qualité de l'ajustement aux données utilisées pour l'estimer : l'objectif est d'avoir un modèle suffisamment généralisable pour obtenir une bonne prévision sur de nouvelles observations. La qualité d'un modèle de prévision est donc évaluée sur un ensemble de données différent de celui-ci utilisé pour sa construction. Pour ce faire, l'ensemble de données initial est scindé en un échantillon d'apprentissage, destiné à estimer les caractéristiques du modèle et un échantillon de validation, destiné à évaluer ses performances sur un échantillon inconnu.

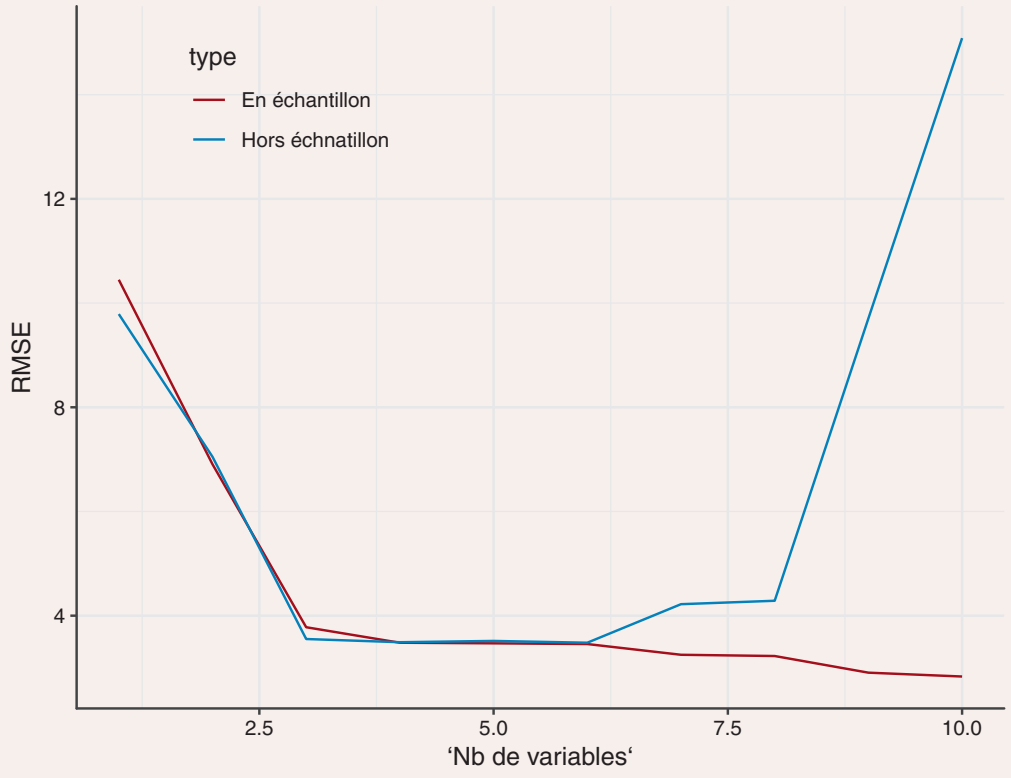
La capacité d'un modèle à être généralisable est intrinsèquement et inversement liée à sa complexité, comme énoncé par le principe d'Ockham. Plus un modèle est simple, moins ses performances empiriques sont dépendantes des particularités des données utilisées pour l'estimer. Pour illustrer ce propos, supposons qu'un ensemble de données ait été généré par une fonction *f* auquel on ajoute un bruit epsilon, tel que *y(x)=f(x)+epsilon*. On observe seulement *y(x) et x*. L'objectif du prévisionniste est de trouver la fonction *g* qui approxime au mieux *f*. On peut approximer cette fonction par un polynôme de degré *p*, le modèle sera d'autant plus complexe que *p* sera élevé. Sur la *figure 6* est représentée à titre d'exemple la fonction *f* que l'on cherche à estimer (en noir), une estimation à l'aide d'un polynôme de degré 2 (en rouge) et un polynôme de degré 11 (en bleu). Alors que le polynôme de degré 11 s'ajuste mieux aux données, le polynôme de degré 2 propose une meilleure estimation de la fonction *f*. Le polynôme de degré 11 capte à tort une partie de l'aléa introduit dans le processus de génération des données. Nous sommes dans une situation de suraprentissage.

6 - Simulation d'un cas de surrapprentissage



La *figure 7* représente l'erreur, mesurée par le RMSE (Root Mean Square Error, ou la racine de la moyenne des écarts au carré entre les valeurs prévues et réalisées) en échantillon et hors échantillon lorsque nous augmentons le nombre de variables. L'erreur en échantillon décroît avec le nombre de variables utilisées. En effet, plus il y a de variables plus le modèle peut s'adapter aux données entraînement. Cependant, pour un nombre de variables supérieur à 4, l'erreur hors échantillon augmente. Nous sommes alors dans une situation de surapprentissage. Le modèle n'est pas suffisamment généralisable et capte à tort de l'aléa. ■

7 - Prévision de croissance du troisième trimestre 2019 avec intervalle de confiance, à l'aide du modèle des forêts aléatoires



Encadré 2 : Prévision à l'aide de forêts aléatoires

Les « forêts aléatoires » sont une technique d'apprentissage automatique, proposée par Leo Breiman en 2001. Cet algorithme repose sur la construction de multiples arbres de décision, construits à partir d'échantillons de données légèrement différents.

Les arbres de décision permettent de diviser un ensemble d'observations en groupes homogènes selon un ensemble de variables discriminantes (variables prédictives) et une variable de sortie (variable prédite). Ils présentent l'avantage d'être faciles à mettre en œuvre et de proposer une représentation graphique interprétable. Ces arbres sont construits à partir de l'algorithme CART[1] (Breiman, 1984). Le principe général est de partitionner récursivement l'ensemble de données. À chaque division, les deux sous-ensembles construits sont les plus homogènes pour la variable prédite[2]. La dernière étape, dite d'élagage, consiste à construire le sous-arbre optimal à partir de l'arbre final construit à l'étape précédente. L'idée sous-jacente est que l'arbre final est constitué d'un très grand nombre de branches. Cet arbre possède une très grande variance et un biais faible ; nous sommes dans une situation de surapprentissage. Une solution est alors de construire une famille de sous-arbres à partir de l'arbre final élagué, et de choisir parmi cette famille l'arbre minimisant l'erreur de prédiction.

La *figure 8* représente un arbre de décision pour la prévision des variations trimestrielles du PIB. Cet arbre a été réalisé avec l'ensemble des indicateurs disponibles 20 jours avant la publication des comptes nationaux trimestriels. Il se lit de la manière suivante : si un trimestre, l'acquis de croissance de l'IPI au mois 2 est supérieur à –1,5 %, le climat des affaires dans le bâtiment normalisé est supérieur à 1,9 et l'acquis de croissance des exportations est supérieur à –1,7 %, alors la prévision de croissance est de +0,97 %. Le pourcentage au-dessous de la prévision indique la part des observations de l'échantillon faisant partie de cette classe. On peut également noter que l'algorithme a aussi bien sélectionné des indicateurs quantitatifs que des variables issues des enquêtes de conjoncture.

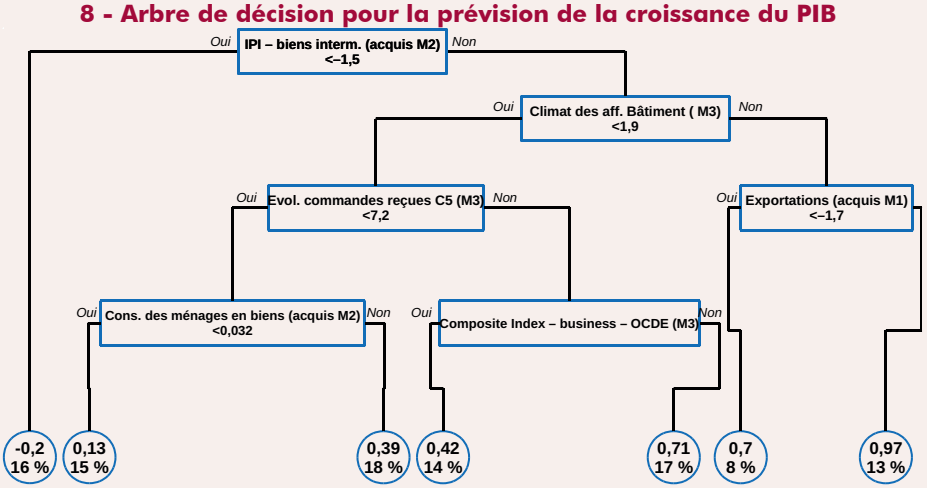
Cependant, cet algorithme souffre d'un défaut majeur d'instabilité. Autrement dit, une légère modification de l'échantillon peut conduire à un arbre de décision et à des prédictions très différentes. L'idée proposée par Breiman est alors d'agréger les prédictions d'un ensemble d'arbres, générés avec une part d'aléa. L'algorithme est alors le suivant :

- Tirage avec remise d'un nombre *N* d'observations pour constituer un échantillon d'entraînement
- Sélections aléatoires de *p*/3 variables parmi l'ensemble des *p* variables prédictives disponibles
- Construction d'un arbre de décision à partir de ces variables et de l'échantillon tiré grâce à l'algorithme CART
- On répète 1 000 fois cette opération pour générer 1 000 arbres de décision différents. La prédiction finale est alors donnée par la moyenne des prédictions données par chaque arbre.

Ainsi, chaque arbre a été généré par un ensemble d'apprentissage différent et leurs prévisions sont faiblement corrélées. Un des critères importants pour le conjoncturiste est l'interprétabilité du modèle construit. Celle-ci est rendue possible pour les forêts aléatoires par la quantification de l'importance des variables, calculée comme le gain prédictif associé à chaque variable. Les *tableaux 2 et 3* présentent l'importance respective des variables dans la prévision des variations trimestrielles à deux dates différentes. ■

1. Classification and Regression Trees
2. Plus précisément, à chaque partition, les deux sous-ensembles minimisent les variances à l'intérieur des sous-groupes.

8 - Arbre de décision pour la prévision de la croissance du PIB



Lecture : si l'acquis d'IPI est inférieur à –1,5 %, la prévision du modèle est de –0,1 %. 16 % des données de l'échantillon d'entraînement ont un acquis d'IPI inférieur à –1,5 %.

Bibliographie

Bessec M. et Doz C. (2012), « Prévision de court terme de la croissance du PIB français à l'aide de modèles à facteurs dynamiques », *Économie et prévision*, 2012, n°199

Breiman L. (2001), « Random forests ». *Machine learning*, n°45, p.5-32

Breiman L., Friedman J., Olshen R. and Stone C. (1984), « Classification and regression trees », Wadsworth & Brooks

Bortoli C. et Combes S. (2015), « Apports de Google Trends pour prévoir la conjoncture française : des pistes limitées », *Note de conjoncture*, Insee, mars, p.43-56

Bortoli C., Combes S. et Renault T. (2017), « Comment prévoir l'emploi en lisant le journal », *Note de conjoncture*, Insee, mars, p.35-43

Bortoli C., Gorin Y., Olive P.-D. et Renne C. (2015), « De nouvelles avancées dans l'utilisation des enquêtes de conjoncture de l'Insee pour le diagnostic conjoncturel », *Note de conjoncture*, mars, p.25-41

Doz C., Giannone D. et Reichlin L. (2011), « A two-step estimator for large approximate dynamic factor models based on Kalman filtering » , *Journal of Econometrics*, 2011, n°164

Dubois E. et Michaux E. (2006), « Étalonnages à l'aide d'enquêtes de conjoncture : de nouveaux résultats », *Économie & Prévision*, n°172

Stock J. et Watson M. (2002), « Forecasting using principal components from a large number of predictors », *Journal of the American Statistical Association*, 2002, n°460 ■