Journées de méthodologie statistique de l'Insee 2022

Nowcasting PIB : Imputation de variables non encore publiées

Marion CABROL, Kevin FERNANDES

Société Générale CIB - Core Data Science Team

marion.cabrol@sgcib.com, kevin.fernandes@sgcib.com

Mots-clés : PIB français, *machine learning*, imputation, séries temporelles, *nowcasting*

Domaines : *Nowcasting*, apprentissage statistique

Résumé

Afin d'anticiper la position de l'économie dans son cycle, les macro-économistes doivent se doter d'outils robustes, permettant de fournir des prédictions immédiates (*"nowcasting"*) de l'état de l'économie et plus spécifiquement de son agrégat macro-économique de référence : le Produit Intérieur Brut (PIB).

Dans ce contexte, l'article s'intéresse à l'élaboration de méthodes de prévision du PIB français pour le trimestre en cours via l'utilisation de données officielles issues d'institutions publiques (INSEE, Banque de France etc.) et privées (Markit, Bloomberg). En d'autres termes, l'objectif est de prévoir le PIB en se fondant sur un ensemble de données qui inclut à la fois des données d'activité (*"hard-data"*), des données d'enquêtes (*"soft-data"*) ainsi que des données des marchés financiers. Les premières données sont utilisées dans la construction du PIB (production industrielle, ventes au détail, etc.) alors que les secondes sont généralement des enquêtes de confiance. Ces variables à multifréquence sont pour la plupart publiées à différentes dates durant le trimestre qu'elles caractérisent. Elles peuvent également être revues au fur et à mesure du temps, tout comme la valeur du PIB est régulièrement réestimée pendant une période allant jusqu'à trois ans.

Néanmoins, l'objectif final de prédire le PIB de manière quasi-continue implique de pouvoir délivrer des prédictions sans pour autant attendre que l'ensemble des variables ait été publié. Ainsi, au début d'un trimestre, l'ensemble des variables explicatives portant sur ce même trimestre est partiellement vide. Généralement, seulement quelques enquêtes ont été réalisées. A contrario, en fin de trimestre, l'ensemble des variables explicatives est quasi-complet. Ainsi, plus nous avançons dans le trimestre et plus l'information portée par l'ensemble des variables explicatives s'enrichit.

Si les méthodes économétriques comme les modèles à facteurs dynamiques (Banbura et al., 2010 [1] et Bok et al., 2018 [12]) sont largement appliquées aujourd'hui et permettent d'ajuster les prévisions au fur et à mesure du trimestre, l'utilisation des méthodes d'apprentissage statistique (*"machine learning"*) y est plus rare. En effet, l'emploi conventionnel de ces méthodes requiert un ensemble prédéfini de variables observées et bien souvent une profondeur de données importante.

Or, récemment dans de nombreux domaines, les modèles d'apprentissage statitistique, voire d'apprentissage profond, se sont montrés plus probants que les techniques économétriques traditionnelles (Bojer et al., 2021 [11]). Ils permettent, notamment, de gérer plus facilement les données volumineuses. La macro-économie s'orientant davantage vers l'utilisation de diverses données, les modèles d'apprentissage requièrent ainsi une attention particulière, et nécessitent d'être évalués dans l'application du "nowcasting". Cette présente étude analyse donc non seulement l'apport des méthodes linéaires usuellement étudiées dans le domaine du "nowcasting", mais également l'apport des méthodes non linaires issues de l'apprentissage statistique.

Afin d'utiliser le pouvoir prédictif des algorithmes d'apprentissage statistique tout en contournant les contraintes susmentionnées, cet article se propose d'étudier une démarche de modélisation en deux temps.

1. **Imputation des variables non encore publiées :** cette étape consiste à prévoir les données non encore publiées afin d'avoir un ensemble de variables explicatives qui soit complet à tout moment.
2. **Prédiction du PIB :** il s'agit de prédire le PIB à l'aide des variables déjà publiées et des variables explicatives imputées à l'étape précédente.

Cette approche se retrouve dans les travaux de Miller et al. (1996) [29], de Zheng et al. (2006) [46] et de Bouwman et al. (2011) [13].

Cette étude se concentre essentiellement sur la première étape de la modélisation d'"imputation des variables non encore publiées" et s'efforce à déterminer la méthode la plus adaptée. Pour ce faire, l'étude complète les techniques issues de la prévision des séries temporelles, par des méthodes non spécifiques au *"nowcasting"*. À noter que dans nos techniques d'imputation, on distingue les méthodes dîtes de prévisions (qualifiées de méthode d'interpolation par Yoon et al. (2017) [44]) et les méthodes dîtes d'imputation. L'article établit ainsi, une cartographie récente et d'autant plus exhaustive en s'inspirant, entre autres, de recherches liées à l'apprentissage statistique (Bertsimas et al., 2017 [4] ; Buuren et al., 2010 [14] ; Wang et al. 2021 [43]) et de recueils relatifs à l'imputation de données manquantes (Van Buuren, 2018 [42]). C'est d'ailleurs pour cette raison que nous avons fait le choix de nommer cette étape "**imputation de données non encore publiées**".

L'apport de chacune des méthodes est examiné par une procédure d'évaluation reposant sur deux types de mesure :

— Métriques intrinsèques : elles permettent de quantifier la qualité de l'imputation.
— Métriques extrinsèques : elles permettent de mesurer la précision de prédiction du PIB.

On accorde cependant plus d'importance à l'objectif final de prédiction du PIB. On privilégie donc les méthodes qui surpassent sur les mesures extrinsèques tout en contrôlant la cohérence de leurs mesures intrinsèques.

Bien que le papier se concentre sur cette modélisation en deux étapes, il détaille également des solutions concrètes à d'autres défis imposés par la prédiction immédiate. Un de ces défis porte sur les données d'entraînement qui se doivent d'être fidèles à cette notion de "temps réel". En effet, la modélisation *"nowcasting"* se construit sur une chronologie réaliste du flux de données. L'article traite alors des méthodes de reconstruction de calendrier, de corrections de changements de nomenclature ainsi que des méthodes de traitement de séries temporelles adaptées aux révisions macro-économiques. Ces mêmes révisions sont, pour rappel, caractérisées à la fois par une période de valorisation et une date de publication.

Abstract

The authors compare usual econometric methods and machine learning methods including deep learning to build a single "nowcasting" model framework for predicting French current-quarter real gross domestic product (GDP) at any time by tracking real time data flow. The single model framework is implemented as a two-stage procedure :

— The prediction ("imputation") of not yet released explicative variables
— The prediction of GDP based on both available explicative variables and those not yet released but predicted variables.

This work emphasises comparing a vast panel of methods from both prediction and imputation theory in the first stage, i.e. the "imputation" stage. The innovation of the approach is the setting up of this two stage procedure, and combining it with machine learning methods. The authors also propose concrete practical solutions to some challenges imposed by nowcasting predictions, such as variables publishing issues (calendar reconstruction, correction of changes in nomenclature, and time series methods specific to macroeconomic data revisions).

Table des matières

1 Introduction

Prédire de manière immédiate ("nowcasting") en s'appuyant sur une modélisation aussi réaliste que possible peut s'avérer complexe pour tout macro-économiste qui s'y emploie. Cette complexité peut notamment s'expliquer par trois défis propres au "nowcasting".

— En premier lieu, la modélisation recherchée se doit de reposer sur un flux de données réaliste. Cette contrainte se heurte alors à la faible disponibilité d'une exacte chronologie des publications passées de ce flux de données. La modélisation se heurte également à un jeu de données macro-économiques incomplet du fait des différences de date de publication ("ragged edge"). Ces données sont également sujettes à de nombreuses révisions au cours du temps. Ces dernières peuvent être dues à une amélioration de leur précision car davantage de données sont disponibles, ou ont une meilleure qualité. Ces révisions peuvent également être provoquées par un changement dans leur définition ("changement de nomenclature" ou "changement de base"), afin de prendre en compte les transformations de nos économies. Aussi, la prise en compte d'un large choix de données contribuant à améliorer la prédiction du PIB implique une certaine hétérogénéité dans ces données. Cette hétérogénéité se traduit par la différence de nature des variables, ainsi que par la variété de leur fréquence. Enfin, les traitements usuels de séries temporelles deviennent obsolètes pour un flux de données macro-économiques. En effet, ces derniers sont adaptés à la modélisation des observations intuitée par la plupart des prévisionnistes, qui, généralement, ignorent le problème de prévision en temps réel. Ils prédisent en se basant seulement sur les dernières données disponibles. Ainsi, ces prévisionnistes identifient une observation sur la base du couple (période valorisée, valeur), et s'abstiennent d'une identification (période valorisée, date de publication, valeur) nécessaire pour modéliser un flux de données réaliste. De cette manière, les traitements temporels usuels se fondent seulement sur le couple (période valorisée, valeur) utilisé par la plupart des prévisionnistes et non le triplet exploré dans cette étude.

— Une modélisation "nowcasting" appropriée doit également relever un second défi : elle se doit de s'adapter simplement et rapidement, afin de se réajuster quasi-instantanément aux dernières informations disponibles. A minima, la modélisation nécessite de reproduire le comportement humain de l'expert macro-économiste, qui actualise ses prédictions en fonction des dernières informations disponibles.

— Enfin, le bon fonctionnement de la modélisation choisie dépend de son optimisation. Cette dernière doit donc reposer sur une évaluation rigoureuse propre au "nowcasting" (e.g., respect de la temporalité, évaluation sur différents scénarios économiques). De cette façon, le défi final relève de la recherche optimale de modélisation par l'étude de ces multiples paramètres. Les paramètres peuvent être de nature totalement différente. Ainsi, parfois, le paramètre peut être également un modèle dans sa globalité, ou encore une méthode de sélection de variables.

Il existe un grand nombre d'approches pour répondre à ces trois défis pris séparément. Nous proposons d'utiliser un processus de modélisation répondant à l'ensemble de ces défis. Afin de répondre au premier défi, le présent papier fournit des solutions pratiques à l'implémentation et le traitement d'un flux de données réaliste. Le cadre de modélisation unique qui est proposé permet de répondre au second défi. Dans ce cadre, toutes les variables explicatives jugées importantes contribueront à tout moment à la prédiction du PIB. Ainsi, la modélisation n'est pas changée selon la disponibilité des données. Pour ce faire, la modélisation décrite s'articule en deux étapes : une étape d'estimation (ou "imputation") des variables explicatives qui n'auront

pas été encore publiées, et une étape de prédiction du PIB basée sur les précédentes estimations ainsi que les informations disponibles. Le coeur de l'étude porte essentiellement sur cette première étape d'imputation. Ainsi, afin d'assurer une sélection de méthodes pertinente, une large variété de modèles d'"imputation" sont comparés par leur performance lors cette estimation. Nous analyserons également la qualité desdites estimations dans la prédiction du PIB. Pour ce faire, ces estimations obtenues grâce aux différents modèles nourriront alors un modèle de prédiction adapté. Etant donné l'objectif final de l'étude, nous accorderons davantage d'importance aux résultats de cette dernière analyse.

La variété des modèles ici cartographiés émane de différentes familles d'algorithmes, mais également de leur divers domaines usuels d'application. En effet, cette cartographie s'inspire, entre autres, de recherches liées à l'apprentissage statistique (ou "machine learning"), y compris l'apprentissage profond (ou "deep learning"), et de recueils relatifs à l'imputation de données manquantes.

2 Flux de données hétérogènes et son traitement

2.1 Flux de données hétérogènes

2.1.1 La variable d'intérêt : Le PIB Français

Bien que largement remis en cause, le niveau de vie matériel d'un pays reste très largement calculé aujourd'hui par son produit intérieur brut (PIB) et l'évolution de celui-ci. En France, l'INSEE publie le PIB trimestriel environ un mois après la fin du trimestre qu'il caractérise. Il s'agit alors d'une première estimation. Cette estimation est alors revue durant les mois, voire les années qui s'ensuivent. Il n'est pas rare que la valeur du PIB soit corrigée des dizaines de fois durant trois années. Le PIB se calcule en sommant les richesses produites lors du processus de production (valeur ajoutée du secteur public et privé), les taxes associées (TVA) ainsi que d'autres taxes sur des produits particuliers (par exemple : les produits pétroliers et l'alcool) et sur les produits importés (droits de douane). Par analogie, les subventions versées par l'état doivent être retranchées pour obtenir le PIB final.

Afin de mesurer la croissance, il est essentiel d'éliminer l'impact de l'inflation et donc de se baser sur un PIB en volume (ou PIB réel), et non un PIB en valeur. Autrement dit, on se basera sur un PIB à prix constant. En effet, la France n'ayant pas connu de période déflationniste durant les dernières années, le taux de croissance en valeur est nettement plus élevé que celui en volume. Ainsi, l'effet d'une croissance "nette" serait plus difficilement observable sur un taux de croissance en valeur. Par analogie, il est aussi fondamental de se baser sur un PIB désaisonnalisé, afin d'éliminer tout effet saisonnier. La prédiction de cette étude portera donc sur un **PIB en volume et désaisonnalisé** (Figure 1).

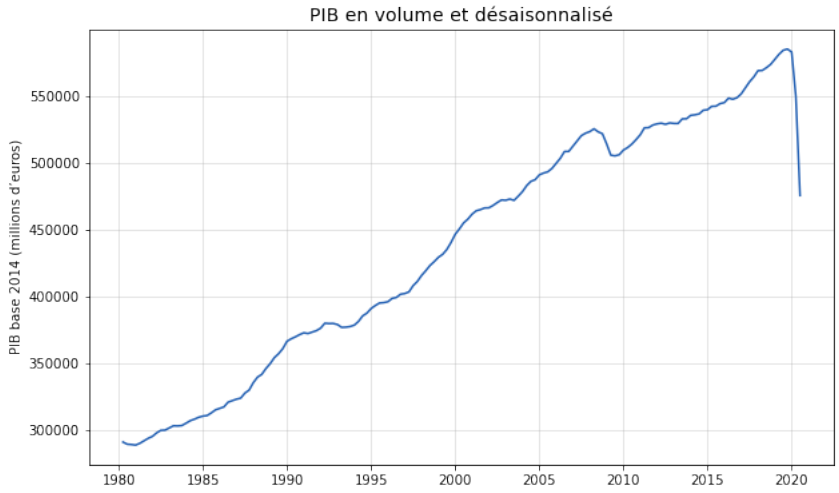

FIGURE 1 – Evolution de la variable d'interêt : le PIB en volume et désaisonnalisé

L'agrégat économique PIB peut être déterminé par différentes approches :

- L'approche par la production, i.e. en additionnant les valeurs ajoutées des agents économiques publics et privés.
- L'approche par le revenu, i.e. en additionnant la rémunération des salariés, les impôts perçus par l'Etat sur la production, les importations (corrigés des subventions reversées) et les excédents d'exploitation dégagés par les entreprises.
- L'approche par la demande, i.e. en additionnant les dépenses réalisées par les ménages, les administrations publiques et les institutions sans but lucratif et les services des ménages

pour acquérir des biens et des services destinés à la satisfaction de leurs besoins.

Ces trois possibilités de calcul reposent sur des variables économiques fondamentalement dissemblables. Ainsi, ces différentes approches de calcul relèvent une idée fondamentale à la démarche retenue pour modéliser le PIB : la détermination de cet agrégat économique peut reposer sur une multitude de variables diverses et variées. Nous nous devons donc d'investiguer sur un large panel de données les variables explicatives les plus pertinentes.

2.1.2 Nature des variables explicatives

Dans ce contexte, l'article s'intéresse à l'élaboration de méthodes de prévision du PIB français pour le trimestre en cours via l'utilisation de données officielles issues d'institutions publiques (INSEE, Banque de France etc.) et privées (Markit, Bloomberg). En d'autres termes, l'objectif est de prévoir le PIB en se fondant sur un vaste ensemble de 634 variables qui inclut à la fois des séries d'activité ("hard-data"), des séries d'enquêtes ("soft-data") ainsi que des séries des marchés financiers.
Certains travaux actuels tendent à démontrer l'utilité des données alternatives pour cette même tâche. On peut notamment penser aux données de tendances de Google, qui d'après Ferrara et al. (2019) [21] sont porteuses d'informations en début de trimestre, ou encore les données textuelles issues de l'actualité économique et financière (Barbaglia et al., 2021 [3]). Néanmoins nous avons fait le choix ici de se concentrer uniquement sur les variables macro-économiques et financières. L'objectif étant d'extraire dans un premier temps le maximum de pouvoir prédictif de ces données plus traditionnelles afin de les compléter, dans de prochains travaux, par des données alternatives.

Les "hard-data" sont des données qui permettent de mesurer des quantités objectives, des volumes tels que le nombre de voitures vendues par exemple. Elles permettent de quantifier des variables à l'aide de calculs robustes. Ainsi, leur valorisation objective du trimestre en cours leur attribue un caractère attrayant dans la prédiction du PIB. À l'inverse, les "soft-data" sont des données qui, par essence, sont subjectives. En effet, elles sont calculées à partir d'enquêtes d'opinions ou de sondages qui donnent initialement une information qualitative. Ces informations qualitatives vont ensuite subir un traitement permettant de leur donner une dimension quantitative. Ainsi, les "soft-data" mesurent généralement une perception des conditions économiques actuelles ou attendues. De ce fait, on leur attribue un décalage dans la prédiction, plus élevé que celui des "hard-data", amplifié par le fait que les agents économiques prennent des décisions économiques en fonction de leur propre perception.

Un exemple de "soft data" est le PMI de IHS Markit. Chaque mois, IHS Markit interroge des chefs d'entreprise via des questionnaires. Ces-derniers doivent répondre aux différentes questions par une augmentation, une diminution ou aucun changement notable. Les réponses obtenues sont ensuite traitées pour obtenir un résultat quantitatif. En effet, pour chaque question, l'indice obtenu est la somme du pourcentage de réponses statuant une augmentation et de la moitié du pourcentage de réponses indiquant qu'il n'y a pas eu de changement. Les indices varient entre 0 et 100. Un indice supérieur à 50 indique une augmentation globale par rapport au mois précédent. De manière analogue, un indice inférieur à 50 indique une diminution globale. Ainsi, à partir de réponses qualitatives, Markit a pu générer un indice reflétant la confiance dans les secteurs étudiés.

Les données financières sont des données de marchés financiers portant sur les 5 familles d'actifs : les taux, les devises, les actions, les matières premières et le crédit. Ainsi, ces données peuvent exprimer, par exemple, le resserrement des conditions financières et de crédit, qui influent sur la consommation des ménages et le développement des entreprises. De manière similaire aux données "soft-data", elles représentent la perception économique d'un groupe spécifique, en l'occurrence, des agents financiers. Il est alors légitime de se poser la question suivante : Est-il

utile de prédire un agrégat économique en se basant, entre autres, sur un ressenti de marché ?

Ferrara et al. (2014) [20] se proposent d'évaluer l'apport des variables financières à un modèle contenant des variables macro-économiques plus classiques. En l'occurrence, des volatilités de différents actifs sont intégrées dans un modèle MIDAS de prédiction du PIB. L'évaluation, qui porte sur la période 2007-2010, incluant ainsi la grande récession, met en exergue l'augmentation de la précision des prévisions grâce à l'ajout de ces variables financières. Ainsi, l'étude de la contribution de ces variables dans la prédiction du PIB semble appropriée.

2.1.3 Cartographie des variables explicatives

Les variables explicatives retenues lors de cette étude sont répertoriées selon 16 catégories.

Catégorie	Nombre de variables
Enquêtes auprès des entreprises	217
Comptes nationaux	208
Commerce	59
Marché du travail	41
Production industrielle	31
Consommation	15
Enquêtes auprès des consommateurs	13
Construction	12
Revenus et investissements	10
Sectoriel	10
Finance	7
Masse monétaire	4
IPC (indice des prix à la consommation)	3
Productivité	2
Comptes relatifs au gouvernement	1
IPP (indice des prix à la production)	1
Total	**634**

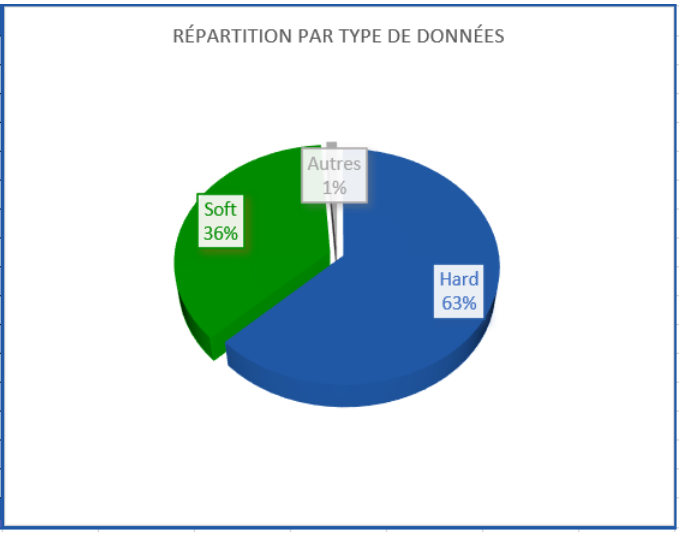

FIGURE 2 – Cartographie des variables explicatives

À noter que certaines intersections des catégories répertoriées ci-dessus (Figure 2) ne sont pas nulles, comme suggère leur définition en Annexe A.1. Cependant, nous avons arbitrairement choisi de classer les variables explicatives dans une seule des catégories, afin d'éviter tout duplicata. Ainsi, les comptes relatifs au gouvernement peuvent également être des comptes nationaux. Cependant, la plupart des variables appartenant aux deux, ont été classées dans les comptes relatifs au gouvernement.

2.1.4 "Ragged-edge" et révisions

En macro-économie, la publication de différents indicateurs introduit deux problèmes majeurs. Le premier est nommé "ragged-edge". Il s'agit à la fois du retard de publication des variables par rapport à la période valorisée et des différences dans les dates de publication de ces dernières. Par exemple, la variable des dépenses de consommation totale [1] (variable trimestrielle) est publiée 30 jours après la fin du trimestre valorisé alors que l'enquête sur les niveaux de production manufacturière [2] (variable mensuelle) est généralement publiée avant la fin du mois valorisé (6 jours avant). La Figure 3 illustre ce flux spécifique de données. Ainsi, au début d'un trimestre, l'ensemble des variables explicatives portant sur ce même trimestre est partiellement

1. FR CONSUMER SPENDING CONA : FRCNPER.D
2. FR SURVEY MANUFACTURING OUTPUT LEVEL - GENERAL OUTLOOK SADJ : FRCNFBUSQ

vide. Généralement, seulement quelques enquêtes ont été réalisées. A contrario, en fin de trimestre, l'ensemble des variables explicatives est quasi-complet. Ainsi, plus nous avançons dans le trimestre et plus l'information portée par l'ensemble des variables explicatives s'enrichit.

Le second problème est relatif aux révisions puisque les "hard-data" peuvent être revues au fur et à mesure du temps, tout comme la valeur du PIB qui est régulièrement réestimée. Etant basées sur des enquêtes d'opinion, les "soft-data"peuvent être obtenues plus rapidement durant le trimestre et ne sont que rarement révisées. Les révisions sont une caractéristique commune des comptes nationaux trimestriels. Afin de répondre aux demandes des utilisateurs, les données trimestrielles préliminaires sont calculées dans de brefs délais dans un premier temps. Puis, elles sont revues ultérieurement lorsque des sources de données de meilleure qualité sont disponibles. Cependant, des révisions majeures peuvent avoir un effet perturbateur. C'est le cas par exemple, si elles sont associées à des changements dans les méthodes statistiques, dans les concepts, les définitions ou les classifications. Cependant, ces révisions, appélées changement de nomenclature, sont plus rares que les révisions issues d'une meilleure qualité de données.

Pour envisager des estimations en temps réel du PIB, il faut donc considérer une démarche de modélisation qui repose sur la prise en compte de variables multifréquences, les révisions ainsi que les délais de publications qui diffèrent.

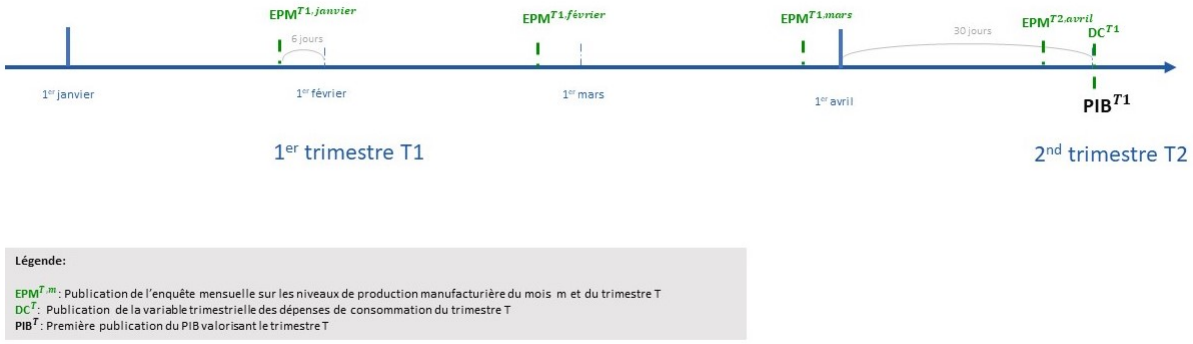

FIGURE 3 – Flux de données multifréquence pour deux variables

2.1.5 Double indéxation de nos séries temporelles

L'un des aspects essentiels et complexes du "nowcasting" consiste à se mettre à la place d'un prévisionniste en temps réel. L'analyse en temps réel repose sur un flux réaliste de données, i.e. en se basant sur les véritables dates de publication de la donnée ainsi que les révisions publiées en temps réel. Autrement dit, Il est essentiel de s'assurer d'utiliser les mêmes données que les prévisionnistes ont vues en temps réel, c'est-à-dire, en prenant en compte la date de publication et le flux des révisions.

Ainsi, à l'inverse des séries temporelles standards, représentées par le couple (*période valorisée*, *valeur*), les séries temporelles d'un flux réaliste de données se caractérisent par le triplet :

(*période valorisée*, *date de publication*, *valeur*).

FIGURE 4 – Illustration de la double indéxation de la variable de production industrielle manufacturière[3]

Le 4^{eme} graphique de la Figure 4 illustre le phénomène de double indéxation appliqué à la production de l'industrie manufacturière [3]. Chaque courbe continue correspond à l'ensemble des révisions pour une période valorisée donnée. Par exemple, les graphiques 2 et 3 montrent les évolutions des révisions pour les périodes valorisées de décembre 2008 et mars 2020.

3. FR INDUSTRIAL PRODUCTION - MANUFACTURING VOLA : FRIPMAN.G

2.2 Traitement du flux de données

2.2.1 Traitements des séries temporelles explicatives

La notion de stationnarité représente un point crucial dans l'économétrie des séries temporelles, où l'estimation des séries non stationnaires peut conduire à des modélisations fallacieuses. Ainsi, lorsqu'il a fallu mettre en place la chaîne de traitement des données, l'une des premières interrogations qui s'est posée concernait la stationnarité des différentes séries temporelles explicatives. Etant donnée la nature majoritairement économique et financière des séries, une étude sur la tendance stationnaire se devait d'être réalisée. En effet, d'après Zivot et al. (2006) [47]), "la plupart des séries temporelles économiques et financières dégagent des tendances ou des non-stationnarités en moyenne". Plus spécifiquement, ces dernières sont "souvent modélisées comme des processus $I(1)$ en présence de tendance déterministe linéaire". À noter ici que la notion de stationnarité fait référence à la stationnarité faible qui suppose des moments d'ordre 1 et 2 indépendants du temps.

À l'inverse des différents travaux qui utilisent les variables explicatives macro-économiques en log-croissance, le choix a ici été fait de les garder en niveau / volume afin d'exhiber précisément la nature de leurs éventuelles tendances sous-jacentes.

Supposons trois variables aléatoires x_t, y_t, z_t telles que :

1. $x_t = a + b.t + u_t$
2. $y_t = c + y_{t-1} + u_t$, où u_t est un bruit blanc gaussien
3. $z_t = a + b.t + z_{t-1} + u_t$

Dans ce cas, x_t possède une tendance déterministe, y_t une tendance stochastique et z_t possède ces deux tendances. Pour retirer une tendance déterministe et linéaire, il faut régresser la série par rapport au temps alors que pour retirer une tendance stochastique (processus $I(1)$), il faut différencier la série. Pour identifier précisément la présence de ces tendances, on peut utiliser la combinaison de différents tests-statistiques : un test unitaire comme celui de Dickey Fuller et un test de stationnarité comme le test de Kwiatkowski-Phillips-Schmidt-Shin (KPPS).

Le traitement détaillé permet d'éviter la sur-différenciation ("overdifferencing" (Zivot et al. (2006) [47])) lorsque nous sommes en présence uniquement d'une série avec une tendance déterministe alors que la transformation log-croissance s'apparentrait, elle, à une différenciation dans le plan logarithmique. Autrement dit, cette dernière transformation est excessive dans le cas d'une tendance deterministe unique. Cependant, la transformation utilisée est plus difficilement interpretable que la transformation log-croissance puisque celle-ci permet de travailler sur un ratio. On portera, néanmoins, plus d'attention au point d'interprétabilité lors de la transformation de la variable d'intêret. Bien heureusement, les résultats obtenus par ces deux types de procédés sont très similaires : la Figure 5 illustre et compare ces procédés.

Ce processus de traitement est largement décrit dans la littérature économétrique. Cependant, qu'en est-il de son application à des séries temporelles caractérisées par le triplet (*période valorisée*, *date de publication*, *valeur*) ?

À notre connaissance, il n'y pas de procédure permettant de stationnariser de telles séries chronologiques. Ainsi, le choix a été fait d'extrapoler le traitement détaillé précédemment à nos séries à double dimension de la manière déroulée dans l'Algorithme 1.

5. FR IMPUTED CONTRIBUTIONS FROM FINANCIAL COMPANIES CURA : FR164713B

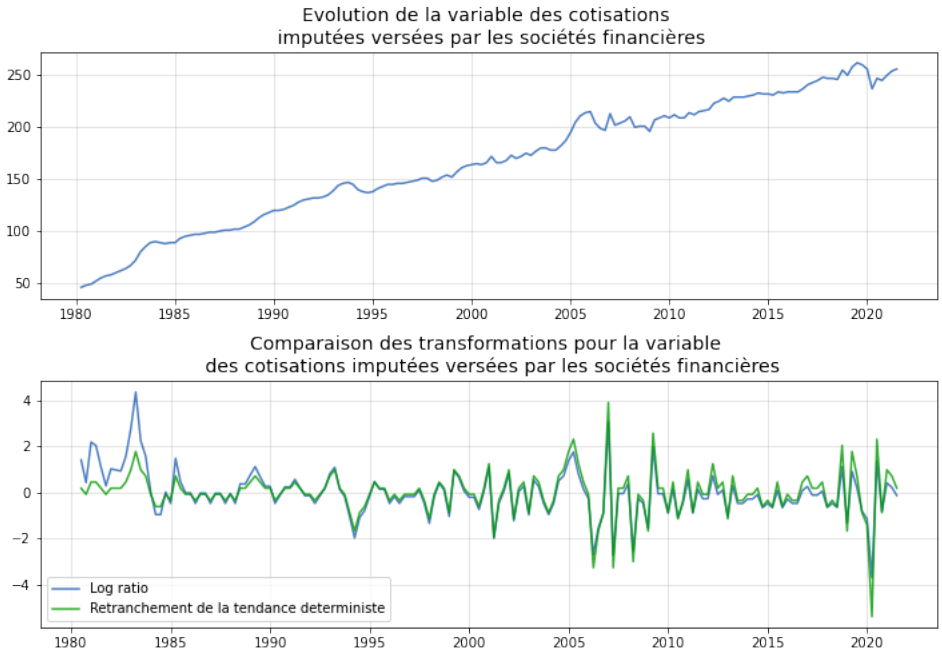

FIGURE 5 – Comparaison de la transformation 'log ratio' et du retranchement de la tendance déterministe dans le cadre de la variable des cotisations imputées versées par les sociétés financières [5] (détectée comme un processus $I(0)$ en présence d'une tendance déterministe).

Algorithm 1 Détection et suppression de tendances dans nos séries double-indéxées

1: Agréger la série à double indéxation par la dernière observation de chaque période.
2: Analyser la présence de différentes tendances sur cette version agrégée et extraire les différents paramètres de tendances
3: Appliquer les traitements de manière séquentielle sur la série initiale :

   a. Tendance déterministe : on retranche la partie linéaire dépendante du temps en retranchant à toutes les estimations d'une même période $(x_t)_i$ la tendance temporelle correspondante : $\alpha + \beta t$

   b. Tendance stochastique : on différencie chaque estimation $x_{t,i}$ avec la dernière estimation disponible à la date i de la période de valorisation précédente $x_{t-1,i}$.

Le traitement de l'Algorithme 1 concerne uniquement les séries économiques à fréquence mensuelle ou trimestrielle. Les séries financières, quant à elle, ont subi un autre traitement adapté à la fois à leur disponibilité quasi-instantanée et à leurs caractéristiques. En effet, elles ne sont ni confrontées au problème du "ragged-edge" ni au problème de révision. Nous avons donc fait le choix d'étudier nos séries financières (hormis pour les indicateurs de volatilité) en croissance logarithmique. Au-delà du fait que certains modèles de la théorie financière reposent sur l'hypothèse de rendements logarithmiques quasi-gaussiens, nous avons également fait ce choix car cette transformation a la bonne propriété de rendre nos séries financières stationnaires.

2.2.2 Transformation de la variable d'intérêt

À l'inverse du traitement des variables explicatives économiques précédemment détaillé, le choix a ici été fait d'étudier le logarithme de la série (Ferrara et al., 2010 [21] ; Banbura et al., 2013 [1]). En effet comme la série du PIB est un processus $I(1)$, la différence des logarithmes, ou encore le logarithme du ratio, permet non seulement de stationnariser la série mais permet

également de garder un certain sens économique, i.e. un taux de croissance continu. Ce sens est d'autant plus important pour la variable à expliquer que pour les variables explicatives. En d'autres termes cette transformation permet d'introduire des unités de pourcentage en se libérant des unités de mesures. La variable cible de l'étude est donc : $y_Q = \Delta log(PIB_Q)$.

Néanmoins, pour un trimestre Q fixé, cette même variable peut être amenée à varier au fur et à mesure des révisions impliquant à la fois PIB_{Q-1} et à la fois PIB_Q. En se plaçant d'un point de vue des marchés financiers, cela semble plus pertinent de prédire la première estimation du PIB_Q qui intervient 30 jours après la fin du trimestre que l'on souhaite valoriser. En effet, le marché peut se contenter d'une valeur approchée, et évaluer ainsi un phénomène de surprise ou non. La différence logarithmique d'intérêt s'effectuera donc entre la première estimation du PIB_Q et la dernière estimation du PIB_{Q-1} disponible à la date de publication de la première :

$$y_{Q,t_1} = log(PIB_{Q,t_1}) - log(PIB_{Q-1,\max(t)_{t\leq t_1}}).$$

2.2.3 Changement de nomenclature

Afin de faciliter l'organisation de l'information économique, les organismes de statistiques doivent respecter certains principes. Ces derniers qualifiés "d'optique" par Rousseau, 1975 [32] représente un découpage fixé d'un ou plusieurs domaines : c'est ce que l'on appelle une nomenclature. Néanmoins, les domaines concernés sont amenés à évoluer et il se peut que la nomenclature précédemment établie ne soit plus aussi adaptée. De ce fait, une nomenclature est difficilement stable dans le temps. Par exemple, les nomenclatures d'activités ont été changées au cours du temps en introduisant des activités nouvelles, comme la construction d'ordinateurs. Ces changements de nomenclature s'appliquent, par exemple, à l'élaboration des comptes nationaux. En pratique, cela se traduit par des sauts inhabituels dans nos séries temporelles, ce qui fausse l'impact de la variable sur le PIB. Afin d'utiliser une série homogène, il est essentiel de détecter ces changements de nomenclature et de rétro propager la nomenclature actuelle. À noter que le jeu de données contient également des séries temporelles macro-économiques disponibles uniquement dans la dernière nomenclature. Elles ont, cependant, le défaut de ne pas s'accompagner de révisions et de dates de publications. Ce sont des séries temporelles usuelles qui nous permettent en revanche d'aider dans la méthode décrite ci-dessous.

On pourrait parcourir l'historique des nomenclatures, relever les différentes dates et corriger manuellement les séries. Il s'avère que l'étude implique plus de 600 séries temporelles et qu'une stratégie automatisée serait plus judicieuse bien que relativement moins fiable. Le choix a donc été fait de mettre en place la procédure de traitement détaillée ci-dessous et illustrée dans la Figure 6.

Cette procédure se déroule deux étapes :

- **Correction externe :** Dans cette étape on agrège la série à double indexation par les dernières publications de chaque période valorisée. Ce qui conduit à l'obtention d'une série temporelle classique que l'on compare ensuite avec la série démunie de toutes révisions. On ajuste ensuite les valeurs pour qu'elles concordent. En effet, on suppose ici que la série dépourvue de révision agit comme référence.
- **Correction interne :** On parcourt ensuite chaque période de valorisation afin de projeter le reste des révisions dans la même nomenclature que l'ultime publication. Pour ce faire, on assimile l'ensemble des révisions d'une période de valorisation à une série temporelle à une dimension que l'on parcourt en appliquant un algorithme de détection d'anomalie adaptée. En supposant que les anomalies détectées sont des changements de nomenclature, on annule les taux d'accroissement suspicieux et on obtient une série de révisions supposée être dans la même nomenclature.

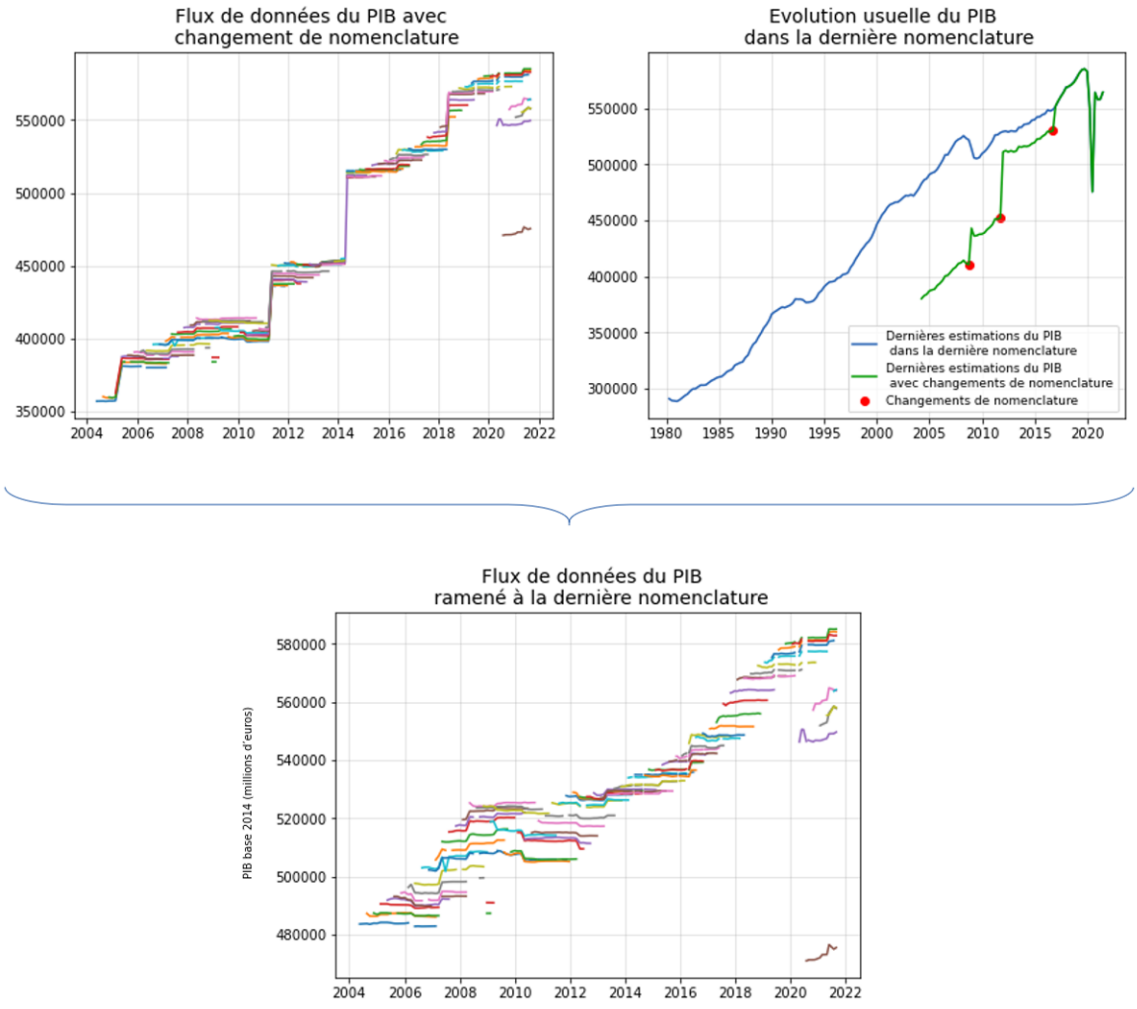

FIGURE 6 – Actualisation du flux de données du PIB à la dernière nomenclature.

En pratique l'algorithme utilisé, inspiré de Tolvi, 2001 [40]), détecte des changements autorégressifs anormaux au sein de nos révisions. En d'autres termes, l'algorithme parcourt séquentiellement la série temporelle étudiée et prédit pas à pas un intervalle de confiance relatif à la prochaine valeur. Si la valeur est à l'extérieur de cet intervalle alors elle est considérée comme étant une anomalie sinon elle est considérée comme normale. En pratique nous avons fixé notre seuil de confiance à 95% et l'ordre p du processus autorégressif à 2.

Bien que notre détection de nomenclature soit automatisée, on apporte une vigililance particulière aux périodes de crise afin de ne pas détecter de faux-positifs, i.e. de réels mouvements liés à la santé de l'économie.

2.2.4 Reconstruction du calendrier

Comme précédemment mentionné, l'analyse en temps réel repose sur un flux réaliste de données, c'est-à-dire, qui se base sur de véritables dates de publication et de révision. Autrement dit, il est essentiel de s'assurer d'utiliser les mêmes données que les prévisionnistes ont vues en temps réel.
Comme précédemment expliqué, les séries temporelles d'un flux de données se caractérisent par le triplet (*période valorisée*, *date de publication*, *valeur*). Cependant, les dates de publication ne sont généralement pas disponibles sur tout l'historique souhaité. Nous nous devons donc de recréer cet historique en nous basant sur les quelques dates de publication disponibles. L'îdée consiste donc à recréer, pour chaque variable, sa dynamique de dates de publication en analysant les spécificités de ses dates disponibles. Puis, on projete cette dynamique apprise sur l'ensemble des historiques manquants afin de reconstruire son calendrier de publication.
En pratique, pour chaque variable, on cherche à trouver les caractéristiques des dates qui régissent la dynamique de publications. On identifie alors 2 familles de caractéristiques : les caractéristiques basées sur des nombres de jours (e.g., publication lors du 3ème jour ouvrable après la fin du trimestre) et les caractéristiques basées sur des jours calendaires (e.g., premier lundi du second mois). Pour chaque variable macro-économique et pour chacune de ses dates de publication disponible, on calcule l'ensemble des caractéristiques descriptives basées sur les 2 familles susmentionnées. On analyse la variabilité de ses caractéristiques sur l'ensemble des dates disponibles afin de sélectionner celles qui affichent le moins de variabilité. On les classe donc par ordre de variabilité croissante afin de rendre explicite les règles de publication retenues par les instituts en charge de leur mise à disposition.

On recrée ensuite l'historique en combinant séquentiellement les caractéristiques des dates gardées. Pour ce faire, on parcourt les caractéristiques ordonnées et on ajuste au fur et à mesure l'inférence de la date de publication. Une étude de performance permet de montrer, qu'en moyenne sur l'ensemble de nos variables mensuelles, il est préférable de garder les 4 caractéristiques les plus prépondérantes.
En prenant le cas de l'indicateur mensuel de production industrielle manufacturière[6] pour la période de valorisation d'avril 2020, on a les 4 caractéristiques prépondérantes de la Table 1.

6. FR INDUSTRIAL PRODUCTION - MANUFACTURING VOLA : FRIPMAN.G

Caractéristique	Valeur	Ordre (variabilité)
date du jour	10^{eme} jour du mois	4
jours à compter de la fin du mois	+41 jours	3
jours ouvrables à compter de la fin du mois	+29 jours	2
nombre de mois	+2 mois	1

TABLE 1 – Caractéristiques régissant la dynamique de publication de la production industrielle manufacturière.

Au fil des 4 ajustements, on infère comme date de publication le 10/06/2020 pour la période valorisée d'avril 2020 de la variable mensuelle de production industrielle manufacturière.

2.2.5 Filtre temporel

Des filtres supplémentaires sont réalisés, afin de s'assurer d'utiliser les mêmes données que celles observées par les prévisionnistes en temps réel.
Ainsi, par exemple, on retire les variables qui sont publiées trop tardivement sur les récentes années. En effet, les variables valorisant le trimestre T qui ont toujours été publiées après l'annonce du PIB français de ce même trimestre T, sur la période 2016 à 2021 (période récente de 5 ans choisie arbitrairement), sont retirées de la base.
De même, les variables qui, dans plus de dix pourcent des trimestres à partir de l'année 2000, n'ont soit pas eu de publication durant le trimestre, soit leur publication est arrivée trop tard par rapport à la publication du PIB, sont supprimées. Aussi, un grand nombre de variables n'étant pas publiées avant les années 2005, nous choisissons de commencer l'historique de données en 2005. En effet, afin de ne pas biaiser les résultats de sélection de variables, il est important de choisir une période où toutes les variables sont publiées.

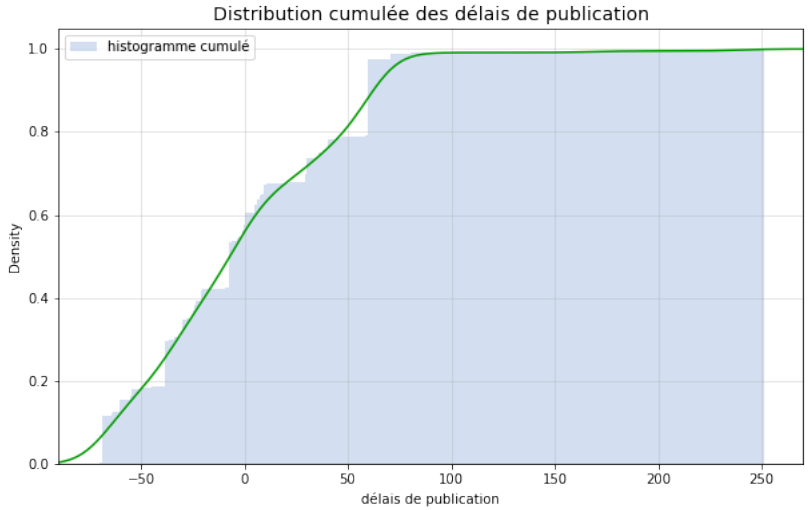

FIGURE 7 – Distribution cumulée des délais de publication de nos variables initiales (0 étant le dernier jour du trimestre).

2.2.6 Agrégation mixte de fréquence

Comme expliqué précédemment, les variables issues des trois types de famille ("hard-data", "soft-data" et financières) ne sont pas forcément échantillonnées à la même fréquence. Les variables financières, comme le VIX, sont majoritairement à fréquence quotidienne et les variables

économiques, comme les enquêtes, sont majoritairement à fréquence mensuelle. Ces différences fréquentielles compliquent notre modélisation. De plus la variable à prédire, le PIB, est une variable trimestrielle.

Afin de répondre à cette problématique de mixte de fréquence, Rousset et & al. (2018) [31] proposent deux solutions dans le cadre de séries explicatives mensuelles et d'une variable à expliquer trimestrielle. La première est d'agréger via la moyenne les variables mensuelles valorisant une même information afin de la trimestrialiser. Si l'ensemble des informations mensuelles d'un même trimestre n'a pas été publié, alors, on moyenne seulement les informations mensuelles disponibles. La seconde option consiste à créer trois variables trimestrielles : une correspondant à la valeur du troisième mois en niveau et deux autres termes de variation : $(x_m, x_m - x_{m-1}, x_{m-1} - x_{m-2})$. Cette dernière solution impose, cependant, des contraintes de signe sur les coefficients des éventuelles régressions a posteriori. Plus globalement, Marsilli (2014) [27] expose différentes techniques d'agrégation de variables. L'objectif est de ramener l'ensemble des variables explicatives à la fréquence de la variable cible. Les méthodes les plus classiques sont celles de "l'agrégation de stock" (utilisée pour des variables en niveau ou en nombre de personnes) et "l'écoulement" qui consiste à sommer partiellement sur des plus petites fréquences. Comme son nom l'indique, cette dernière technique est davantage utilisée pour des variables de flux, comme les dépenses ou l'épargne.

La notion de pondération de variables portant le même type d'information au sein d'une même période est donc une question cruciale de la modélisation. Autrement dit, devons-nous pondérer de la même manière les contributions de la production industrielle du mois de janvier, du mois de février et celle de mars dans la prédiction du PIB du premier trimestre ?

Des études macroéconométriques suggèrent que les contributions de chaque mois peuvent être différentes (Martinez M., 2017 [28]). En effet, les exercices comptables en fin d'année/fin de trimestre, peuvent mener à des contributions différentes sur le PIB trimestriel. Ainsi, moyenner/agréger des variables mensuelles pour obtenir une seule variable trimestrielle n'est pas nécessairement la solution la plus évidente. Nous faisons donc le choix de laisser nos algorithmes choisir la pondération sur chacune des variables, en distinguant chacune des variables mensuelles comme trois variables séparées.

Cependant, ce processus d'agrégation ne peut pas être applicable pour nos variables journalières qui sont uniquement composées de données de marché. En effet, par analogie, il conviendrait alors de créer 92 variables pour représenter chaque jour du trimestre (92 étant le jour maximum d'un trimestre). Ceci supposerait que les jours d'un trimestre renvoient des signaux bien différents quant au PIB. Cette hypothèse ne peut pas s'appliquer ici puisque l'important n'est pas de savoir si le signal potentiellement fort renvoyé par un indicateur, comme le VIX, s'effectue le 3ème jour du trimestre ou le 56ème mais plutôt de savoir sa variation sur le trimestre. De plus, l'algorithme ne serait pas en mesure de detecter le signal renvoyé par le pic de volatilité du 16 mars 2020 (76ème jour du trimestre) en se basant sur le signal renvoyé par le pic de volatilité du 20 novembre 2008 (20ème jour du trimestre). En outre, créer 92 variables pour chacune des variables de marché mènerait à un trop grand nombre de variables qui conduirait à se confronter à la malédiction de la dimension en risquant le sur-apprentissage ("overfitting") et une limitation de l'interprétabilité du modèle.

Toutes les variables explicatives n'influent pas avec les mêmes décalages temporels. Par exemple, la consommation des ménages influence plus rapidement le PIB que la variable du nombre de contrats de construction. Ainsi, les variables doivent également être sélectionnées en optimisant leurs décalages temporels. Etant donné le nombre de variables explicatives, le choix

des décalages temporels se fait automatiquement dans la procédure de sélection de variables. Pour ce faire, on construit pour chaque variable ses homologues décalés temporellement d'un à trois trimestres.

3 Modélisation

3.1 Démarche de modélisation en deux temps

3.1.1 Processus de prédiction

L'objectif final de prédire le PIB de manière quasi-continue implique de pouvoir délivrer des premières prédictions tôt dans le trimestre sans pour autant attendre que l'ensemble des variables ait été publié. On cherche donc à contourner le problème du "ragged-edge". Il faut également s'affranchir du problème de révision puisque les premières publications de certaines variables ne sont que des estimations. Par exemple, la publication de la production industrielle manufacturière [7] pour le mois de janvier 2019 a eu lieu début mars et a été ajustée mensuellement jusqu'à septembre 2020. Etant donné que l'on cherche à estimer la première publication du PIB, qui a lieu 30 jours après la fin du trimestre, seules les premières estimations et révisions avec des délais de publication inférieurs à ces 30 jours doivent être prises en compte.

Les méthodes économétriques comme les modèles à facteurs dynamiques (Banbura et al., 2010 [1] et Bok et al., 2018 [12]) et les régressions MIDAS (Marsilli, 2014 [27]) sont largement appliquées aujourd'hui à la modélisation "nowcasting". Ces méthodes permettent d'ajuster les prévisions au fur et à mesure du trimestre. Cependant, les méthodes d'apprentissage statistique ("machine learning") sont plus rares dans ce domaine. En effet, l'emploi classique de ces méthodes requiert un ensemble prédéfini de variables observées et bien souvent une profondeur de données importante. Quelques travaux s'attaquent à ce problème et abordent le sujet en utilisant, non pas un modèle, mais plusieurs modèles au fur et à mesure de l'avancement du trimestre. Par exemple la modélisation "nowcasting" suggérée par la direction générale du Trésor (Blanchet & al., 2020 [6]) discrétise le trimestre en périodes de 15 jours et crée 8 jeux de données différents constitués des variables alors disponibles. Cette modélisation a l'avantage indéniable de contourner la problématique des données non encore publiées mais peut introduire des obstacles lors de l'interprétabilité du modèle et de sa mise en production. En effet, la gestion de 8 modèles d'apprentissage statistique, basés sur 8 jeux de données différents peut être une tâche ardue à maintenir dans la réalité. De plus, la représentativité de chaque secteur peut également varier tout au long du trimestre menant de ce fait à des biais sectoriels et des problèmes d'interprétabilité.

Afin d'utiliser le pouvoir prédictif des algorithmes d'apprentissage statistique tout en contournant les contraintes susmentionnées, cet article se propose d'étudier une cadre unique de modélisation compatible avec un flux de données en temps réel et applicable à tout moment avant la publication du PIB. Pour ce faire, la démarche de modélisation fonctionne en deux temps (visualisation schématique en Annexe A.2, Figure 22) :

1. **Imputation des variables non encore publiées :** cette étape consiste à prévoir les variables explicatives du PIB non encore publiées afin d'avoir un ensemble de variables explicatives complet à tout moment.
2. **Prédiction du PIB :** il s'agit de prédire le PIB à l'aide des variables déjà publiées et des variables explicatives imputées à l'étape précédente.

Cette méthode contourne bien le problème du "ragged edge" puisque les variables non encore publiées sont imputées et nourrissent le modèle de prédiction. Concernant le problème de révision, nos prédictions s'adaptent aux différentes mises à jour et présupposent que les premières révisions sont de bons indicateurs de leurs valeurs finales correspondantes. Cette modélisation se base ainsi sur un flux de données quasi instantané. De plus, cette approche permet d'obtenir directement, et en une seule structure, une prédiction du PIB. Cette démarche de modélisation se retrouve dans

7. FR INDUSTRIAL PRODUCTION - MANUFACTURING VOLA : FRIPMAN.G

les travaux de Miller et al. (1996) [29], de Zheng et al. (2006) [46] et de Bouwman et al. (2011) [13].

Notre papier se concentre essentiellement sur la première étape de la modélisation d'imputation des variables non encore publiées et cherche à déterminer la méthode la plus adaptée. Zheng et al. (2006) [46] qualifient d'ailleurs ce premier modèle d'imputation de « modèle satellite ». Ainsi, leurs travaux comparent les processus autorégressifs AR, les vecteurs autorégressifs VAR et la marche aléatoire en taux de croissance ("Naïve random wall in growth rate"). Miller et al. (1996) [29] utilisent uniquement un VAR et Bouwman et al. (2011) [13] comparent les AR avec les VAR. Pour définir le modèle d'imputation le plus adapté, notre étude complète ces techniques issues de la prévision des séries temporelles par des méthodes peu répandues dans la littérature "nowcasting".

À noter que dans nos techniques d'imputation, on distingue les méthodes dîtes de prévision (qualifiées de méthode d'interpolation par Yoon et al. (2017)[44]) et les méthodes dîtes d'imputation. En accord avec Yoon et al. (2017) [44], les techniques d'imputation reconstruisent les données manquantes (ici non encore publiées) en tentant de capturer des relations synchrones entre les différentes variables alors que les techniques de prévision (interpolation) reconstruisent les données manquantes en tentant de capturer la relation temporelle et non les relations statiques entre les flux. Autrement dit, la différence porte sur la prise en compte ou non de la dynamique temporelle. L'article établit ainsi une cartographie récente plus complète en s'inspirant, entre autres, de recherche liée à l'apprentissage statistique (Bertsimas et al., 2017 [4] ; Buuren et al., 2010 [14] ; Wang et al. 2021 [43]) et de recueils relatifs à l'imputation de données manquantes (Van Buuren, 2018 [42]). C'est d'ailleurs pour cette raison que nous avons fait le choix de nommer cette étape "imputation de données non encore publiées" incluant à la fois les méthodes de prévision et les méthodes d'imputation stricto sensu.

3.1.2 Sélection de variables

Comme expliqué précédemment, cette étude implique 634 variables explicatives. De plus, les modèles utilisés dans cette étude peuvent s'avérer complexes. De ce fait, un point d'attention doit être porté sur le sur-apprentissage.

En apprentissage statistique, le compromis biais-variance est un point crucial. Le biais permet de mesurer l'erreur liée aux hypothèses erronées de l'algorithme d'apprentissage (composante systématique de l'erreur). La variance permet de mesurer la sensibilité des prédictions en fonction des données d'entraînement (composante aléatoire de l'erreur). Autrement dit, la variance permet d'estimer la capacité du modèle à généraliser à de nouvelles données. Il s'agit donc d'un indicateur de surveillance de sur-apprentissage ("overfitting").

À première vue, on pourrait se dire que plus le modèle est complexe (e.g., nombre de variables explicatives, profondeur de l'algorithme etc.) et plus il est capable d'établir des prédictions performantes. Néanmoins, le compromis biais-variance doit être surveillé : plus on complexifie le modèle et plus son biais diminue au détriment de sa variance. A contrario, plus on simplifie le modèle et plus on diminue sa variance au détriment de son biais. Tout l'enjeu consiste donc à trouver un bon compromis entre une erreur acceptable et une capacité convenable de généralisation.

Dans de nombreux cas de modélisation économétrique et d'apprentissage statistique, la sélection de variables est une étape nécessaire. Cette étape consiste à sélectionner un sous-ensemble de variables $S \subset F$ pertinentes pour maximiser la performance du modèle final à partir d'un ensemble initial de variables $F = (v_1, \cdots, v_d)$. Tout l'enjeu de la sélection consiste donc à trouver le nombre de variables qui respecte au mieux le compromis "biais-variance" évoqué précédemment.

En pratique, la sélection de variables possède bien d'autres avantages. Elle permet de contourner la problématique du fléau de la dimension (Bellman R. en 1961), d'améliorer l'interprétation du modèle, de diminuer le temps d'entraînement du modèle, ou encore d'éviter un sur-apprentissage ("overfitting") du modèle.

On peut diviser les méthodes de sélection de variables en trois catégories : les méthodes filtres, les méthodes enveloppes ("wrapper") et les méthodes intégrées ("embedded").

- Les méthodes filtres sont des méthodes de scoring où la variable est retenue ou non selon un score qu'on lui associe. Ce score est calculé en confrontant la variable à la variable à prédire grâce à une fonction basée sur les valeurs de ces deux dernières. On peut notamment penser aux différents tests statistiques, à la corrélation linéaire ou, pour utiliser des méthodes plus innovantes dans le cadre des séries temporelles, la mesure DTW (Dynamic Time Warpping). Aussi, les méthodes filtres sélectionnent les variables de manière indépendante du modèle de prédiction final.
- Les méthodes enveloppes, quant à elles, sélectionnent les variables en prenant en considération les performances du modèle de prédiction final construit sur ces variables. On peut citer la méthode d'élimination rétrograde des variables (ou "backward feature elimination") et son homologue, la méthode d'élimination des variables par l'avancement (ou "forward feature elimination"). Il s'agit de méthodes itératives. À chaque itération, les performances d'un nouveau modèle basé sur des variables différentes de son prédécesseur sont calculées. En effet, ces méthodes retranchent ou ajoutent successivement une des variables explicatives à un modèle, en comparant aux performances du modèle précédent.
- Les méthodes intégrées sont, comme leurs noms l'indiquent, intégrées au processus d'entraînement des algorithmes d'apprentissage statistiques. On peut citer les méthodes de régularisation comme la régression ridge et lasso (voir partie 3.3.1 pour davantage d'explication) qui sont les plus couramment utilisées.

Pour davantage d'explication et de détail sur ces méthodes, le lecteur peut se référer à la thèse doctorale Dernoncourt, D. (2014) [18].

En pratique

Dans notre cas, nous avons fait le choix d'utiliser deux méthodes de sélection de variables :

- La première, plus traditionnelle aux séries temporelles, consiste à sélectionner les variables en fonction de leur corrélation avec la variable d'intérêt (Brownlee, 2017 [10]). C'est donc une méthode filtres. Appliquée à notre jeu de données, cette méthode suppose que si une variable (ou son homologue décalé dans le temps) est fortement corrélée (en absolue) avec le PIB, alors, elle possède un pouvoir prédictif fort.
- La seconde méthode utilise des outils orientés apprentissage statistique avec une complexité temporelle raisonnable. Cette méthode se base sur l'association de propriétés différentes de deux méthodes intégrées. En effet, on utilise à la fois la régression Lasso (ou "Least Absolute Shrinkage and Selection Operator") et l'importance des attributs de l'algorithme de Forêt aléatoire (voir partie 3.3.2). On distribue alors des rôles différents, néanmoins complémentaires, à ces deux types de sélection de variables afin de capter une grande variété de relations entre les variables explicatives et la variable cible. Ainsi, le Lasso distingue les variables explicatives apportant un pouvoir prédictif linéaire de manière pénalisée. La forêt aléatoire, quant à elle, met en exergue les variables au pouvoir prédictif non linéaire (recherche de partition binaire explicative). Pour ces deux algorithmes, on optimise les paramètres (e.g., paramètre de régularisation pour le Lasso) sur le jeu de données restreint à la dernière observation de chaque trimestre (aucune composante manquante présente sur ces observations). L'idée consiste ensuite à créer un score global qui résume ces deux pouvoirs prédictifs. Pour ce faire, on additionne les coefficients

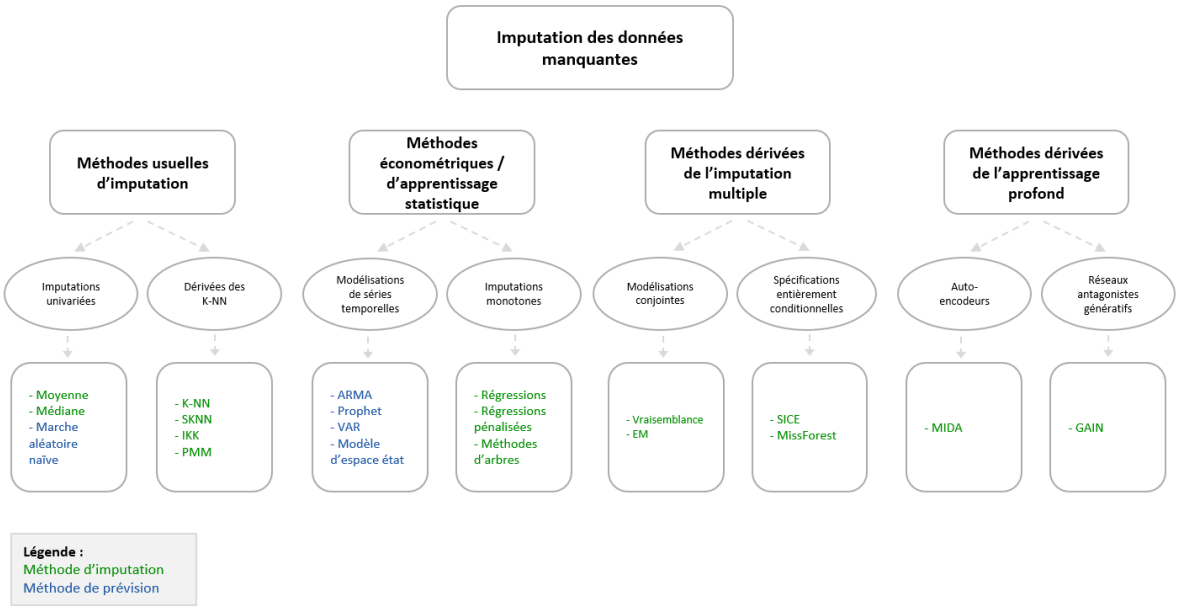

FIGURE 8 – Cartographie des méthodes d'imputation utilisées

de la régression pénalisée avec l'importance des attributs de la forêt aléatoire, préalablement mis à la même échelle. Comme indiqué par Tofallis (2014) [41], cette transformation permet en effet de prioriser les variables déséquilibrées (prédominantes pour un modèle et marginales pour l'autre) au détriment des variables polyvalentes (moyennes pour chacun des modèles).

À noter que la méthode DTW n'a pas été retenue puisque notre modélisation finale incorpore des décalages temporels statiques, et non dynamiques.

3.2 Méthode d'imputation

Comme mentionné, l'objectif principal de cette étude consiste à identifier le modèle "d'imputation" le plus pertinent au regard de notre jeu de données macro-économique et financier. Dans cette partie, l'ensemble des modélisations testées est non seulement présenté mais surtout expliqué dans notre contexte d'utilisation. La cartographie (voir Figure 8) permet de résumer ces méthodes en les insérant dans 4 grandes familles : les méthodes usuelles d'imputation, les méthodes économétriques / d'apprentissage statistique, les méthodes dérivées de l'imputation multiple et les méthodes basées sur l'apprentissage profond. Certains algorithmes peuvent avoir leur place dans plusieurs familles. Dans ce cas, nous avons fait le choix de nous baser sur leurs comparaisons dans la littérature existante sur le sujet. On notera également que nous avons choisi d'indiquer sur la cartographie s'il s'agit d'une méthode initialement de prévision ou d'imputation à proprement parler.

Notation :
Pour les explications de cette partie, on considère X notre jeu de données macro-économiques. Il est composé de n échantillons et d variables explicatives. Un échantillon x_i correspond ainsi à l'information économique disponible pour un trimestre Q à une date donnée t. À cette même date, certaines variables sont publiées alors que d'autres non. On qualifie donc de composantes observées x_{obs} l'ensemble des variables explicatives publiées et de composantes manquantes x_{miss} l'ensemble des variables explicatives non encore publiées. On note alors $x_j = (x_{j,obs}, x_{j,miss})$ et

$\widehat{x_j} = (x_{j,obs}, \widehat{x_{j,miss}})$ sa version imputée, i.e. ses valeurs publiées ainsi que l'estimation de ses valeurs non publiées.

3.2.1 Méthodes traditionnelles

3.2.1.1 Imputation univariée

Moyenne arithmétique, médiane et moyenne équipondérée par trimestre
L'imputation moyenne arithmétique consiste à remplacer les valeurs non encore publiées par la moyenne arithmétique des valeurs publiées. La facilité d'exécution de cette méthode la rend particulièrement attractive. Cependant, cette méthode réduit la variabilité des données proportionnellement au taux de valeurs manquantes. De manière analogue, l'imputation par la médiane est également réalisée afin de diminuer le poids de valeurs extrêmes. Enfin, l'imputation par moyenne équipondérée selon les trimestres est menée dans cette étude. En effet, bien que le nombre d'observations soit relativement stable selon les trimestres, on peut noter de légères différences. Afin de ne pas surpondérer le poids d'un trimestre dans la phase d'imputation, on détermine la valeur moyenne des observations par trimestre, puis cette valeur est moyennée sur l'ensemble des trimestres et est utilisée comme valeur d'imputation.

Marche aléatoire naïve
La marche aléatoire est un des modèles les plus simples mais également fondamental dans la prévision des séries temporelles. Si la variable suit une marche aléatoire, alors, la variable s'éloigne d'un pas aléatoire de sa valeur précédente : $x_t = x_{t-1} + \epsilon_t$. Les pas de la marche sont considérés comme indépendants et identiquement distribués. Le futur de cette variable est sans cesse remis en cause par son état présent, et ne dépend pas de son état passé. Ainsi, la marche aléatoire naïve, comme expliquée par Zheng et al. (2006) [46], consiste à reproduire la dernière valeur (i.e. la dernière croissance résultant de la dernière publication). En effet, $\mathbb{E}(x_t|x_{t-1}) = x_{t-1}$.
Ce procédé implique un traitement différent selon la fréquence de la variable. En effet, le report de la dernière valeur est différent sur nos variables trimestrielles et mensuelles. Dans le cas des mensuelles, lorsqu'une nouvelle publication a lieu, les trois variables représentant les trois mois du trimestre doivent être actualisées avec cette valeur. Dans le cas d'une variable trimestrielle, seulement une variable est actualisée.

3.2.1.2 Imputation dérivée de l'algorithme des K-NN

Dans le cadre d'une prédiction de valeur manquante, les méthodes univariées se base sur une faible part de l'information disponible. En effet, si l'on souhaite prédire la production industrielle, ne pas utiliser les résultats des premières enquêtes disponibles du trimestre mais seulement la dernière valeur ou la moyenne de la production industrielle du trimestre précédent, pourrait s'avérer être une perte considérable d'informations pertinentes. Dans ce cadre macro-économique où certaines variables peuvent avoir un pouvoir de causalité (ou au moins une corrélation forte) sur d'autres, les méthodes univariées ne semblent pas être nécessairement une stratégie appropriée.

Imputation via les K-NN
L'algorithme KNN, ou "k-Nearest Neighbors", permet alors d'incorporer davantage d'informations dans ses prédictions de valeurs manquantes. En effet, l'algorithme prend en compte les variables explicatives pour lesquelles les valeurs ont été publiées. Grâce à ses valeurs publiées, l'algorithme identifie les points voisins où aucune des variables ne possèdent de valeurs manquantes. La notion de 'voisin' se définit par rapport à un paramètre primordial à l'algorithme : une mesure spécifique de la distance. On utilise ici la distance euclidienne. Les valeurs manquantes

peuvent alors être remplacées via la moyenne des valeurs disponibles sur les observations voisines. Les K-NN servent également de base à différents algorithmes d'imputation plus complexes.

Imputation via les SKNN

Une des extensions de l'algorithme des K-NN est le SKNN, ou K-NN séquentiel (Kim et al., 2004 [26]). Dans un premier temps, les échantillons disponibles aux différentes dates de publication sont classés en fonction du nombre croissant de leurs valeurs non publiées à ces dates : $(x_1, ..., x_n)$.

$$\forall k, l \in [1, n] \text{ tq } k \leq l, \ \sum_j \mathbb{I}_{(x_{k,j} \text{ est publiée})} \leq \sum_j \mathbb{I}_{(x_{l,j} \text{ est publiée})}$$

Ainsi, x_1 est l'information disponible à une date à laquelle il y a eu le moins d'absence de publication des variables explicatives. Les valeurs non publiées de x_1 sont imputées via l'algorithme K-NN en se basant sur les données totalement observées : $\widehat{x_1} = (x_{1,obs}, \widehat{x_{1,miss}})$. $\widehat{x_1}$ est maintenant considéré comme un échantillon observé et rejoint les données totalement observées. Il peut donc servir de base 'voisine' pour prédire le second échantillon x_2 présentant le moins de valeurs non publiées. Le processus est ensuite répété jusqu'à ce que l'ensemble des données non publiées soit imputé.

Kim et al. (2004) [26] propose donc deux évolutions notables de l'algorithme des K-NN, dans un cadre d'imputation de valeurs manquantes de gènes (un gène représente un échantillon).

- L'estimation séquentielle : les gènes sont imputés par ordre croissant de taux d'absence (taux de valeurs manquantes).
- Les valeurs qui viennent d'être imputées servent d'estimation aux gènes n'ayant pas encore été imputés.

Qu'en est-il dans un cadre où la temporalité doit être respectée ?

Classer les échantillons par ordre de taux d'absence nuit à la chronologie des événements, et de ce fait, devient un risque pour la représentation réaliste du flux de données. Ainsi, lors de la phase de prédiction, l'algorithme SKNN est appliqué non pas en fonction du nombre de composantes de valeurs non publiées mais par ordre chronologique.

Imputation via les IKNN

Une autre extension de l'algorithme des K-NN est le IKNN, ou K-NN itératif. Cet algorithme est décrit par Brás et al. (2007) [7] dans le cadre d'imputation de valeurs manquantes sur la détection de gènes exprimés de manière différentielle, à partir de données de puces à ADN. Comme son nom l'indique, l'estimation des valeurs manquantes est effectuée de manière itérative.

1. Une première phase de l'algorithme consiste à imputer les valeurs manquantes de l'ensemble du jeu de données, avec la méthode de son choix. Par exemple, par une méthode d'imputation en moyenne ou bien en utilisant les K-NN.
2. La seconde phase consiste ensuite à parcourir les échantillons présentant initialement des valeurs manquantes $(x_1, ..., x_n)$. Pour chaque échantillon x_j on impute ses composantes manquantes via l'algorithme K-NN en se basant sur l'ensemble du jeu de données précédemment imputé et publié. Une fois les valeurs de l'échantillon imputées $\widehat{x_j} = (x_{obs}, \widehat{x_{miss}})$, ce dernier est réinjecté dans le jeu de données imputées.
3. On répète cette dernière phase jusqu'à convergence de l'algorithme, i.e. jusqu'à ce que la différence successive des imputations soit inférieure à un certain seuil. Dans notre étude ce seul est fixé à 10^{-3} et l'algorithme arrive à convergence en une dizaine d'itérations.

Qu'en est-il dans un cadre où la temporalité doit être respectée ?

Dans notre cas, l'IKNN est utilisé uniquement durant l'étape d'entraînement. Autrement dit, le parcours de la seconde phase de l'algorithme, s'effectue par ordre chronologique de dates de publication. À l'étape de prédiction, l'imputation des valeurs manquantes se fait grâce à un K-NN usuel où les plus proches voisins sont choisis parmi les échantillons du passé obtenus grâce à

l'IKNN.

Imputation via PMM
L'algorithme adapté au nowcasting de "Predictive Mean Matching" (PMM), ou d'appariement d'après la moyenne prévisionnelle, repose sur l'algorithme décrit par Van Buuren S. (2018) [42].

1. Lors de la **phase d'entraînement**, on calibre un modèle d'imputation K-NN sur le jeu de données d'entraînement ne présentant pas de données manquantes
2. Lors de la **phase d'imputation** du modèle :
   — On prédit dans un premier temps $\widehat{x_i}$ résultant de l'imputation de l'échantillon x_i par le K-NN.
   — Puis, pour une des composantes $x_{i,j}$ initialement manquante, on garde les k valeurs $(x_{1,j}, \cdots, x_{k,j})$ du jeu d'entraînement les plus proches de la valeur prédite $\widehat{x}_{i,j}$.
   — L'imputation finale de cette composante j de l'échantillon x_i correspond alors à la valeur prise par un candidat tiré uniformément parmi les k valeurs retenues.
3. On effectue cette démarche pour l'ensemble des composantes manquantes afin d'obtenir une version imputée de x_i.
4. On itère les étapes 2 et 3 sur l'ensemble des échantillons du jeu de données.

La phase de prédiction pourrait s'effectuer en une étape uniquement en tirant aléatoirement parmi les k voisins du K-NN. Cependant, la séparation en deux étapes permet d'introduire une souplesse quant au modèle d'imputation initiale. Cette méthode aurait pu être appliquée à d'autres algorithmes. Cependant, nous nous sommes restreint à son application à l'algorithme des K-NN.

3.2.2 Méthodes économétriques et utilisant l'apprentissage statistique

3.2.2.1 Modélisations temporelles

ARMA
À l'image de Zheng et al. (2006) [46] qui utilisent un modèle autorégressif (AR) pour imputer les variables explicatives non encore publiées, nous avons fait le choix dans notre étude de reprendre cette idée en l'extrapolant au modèle autorégressif et moyenne-mobile d'ordres (p,q), abrégé en ARMA(p, q) :

$$x_t = \varepsilon_t + \sum_{i=1}^{p} a_i x_{t-i} + \sum_{i=1}^{q} m_i \varepsilon_{t-i}$$

où les a_i et m_i sont les paramètres du modèle et les ϵ_i les termes d'erreur.

Dans notre cas d'imputation, cette méthode univariée s'applique de manière glissante sur chacune de nos variables explicatives tout en s'adaptant à leurs fréquences. La fenêtre glissante est constituée des 24 dernières périodes disponibles. Pour une variable explicative donnée on estime les paramètres p, q ainsi que les coefficients associés à chaque nouvelle publication, grâce notamment à l'optimisation du critère d'information d'Akaike. Ainsi, nous intégrons toujours les informations les plus récentes dans notre modélisation afin de capter les éventuels changements de dynamique.

Le processus d'imputation s'établit de la manière suivante :

1. **Phase d'entraînement :** dans le cadre d'un trimestre Q et d'une variable explicative x_j mensuelle dont aucune des valeurs n'a encore été publiée pour les mois de ce trimestre, on modélise le comportement ARMA sur les deux dernières années : $(x_j^{m-24}, ..., x_j^m)$.
2. **Phase d'imputation :** Puis, on impute séquentiellement $x_j^{Q,1}$ par $x_j^{m+1|m}$, $x_j^{Q,2}$ par $x_j^{m+2|m}$ et $x_j^{Q,3}$ par $x_j^{m+3|m}$.

3. On répète les étapes 1 et 2 à chaque nouvelle publication. Par exemple, lorsque la valeur du premier mois est publiée, on réestime les paramètres du modèle et on impute maintenant $x_j^{Q,2}$ par $x_j^{m+2|m+1}$ et $x_j^{Q,3}$ par $x_j^{m+3|m+1}$.

Prophet

Tout comme la modélisation ARMA, Prophet (Taylor et al., 2018 [39]) est une modélisation permettant la prédiction de séries temporelles. Cette procédure se base sur un modèle additif impliquant des tendances non linéaires $g(t)$, de la saisonnalité annuelle, hebdomadaire et quotidienne $s(t)$, ainsi que des effets de vacances $h(t)$:

$$x(t) = g(t) + s(t) + h(t) + \epsilon_t \text{ avec } \epsilon_t \text{ la composante idiosynchratique.}$$

— Tendance : La procédure fournit deux modèles de tendance possibles : un modèle de croissance saturante et un modèle de tendance linéaire avec points de changement. Dans le cadre de cette étude, le second a été utilisé. Il s'agit d'un ensemble de tendances linéaires par morceaux avec des pentes différentes entre les points de changement. Il se formalise de la manière suivante :

$$g(t) = (k + a(t)^T\delta) + (m + a(t)^T\sigma)$$

Avec k le coefficient de croissance, m le paramètre d'offset', a contient les indications des points de changements, δ contient les ajustements en croissance et σ les ajustements en 'offset'.

— Saisonnalité : La composante saisonnière $s(t)$ permet d'identifier des changements périodiques basés sur des saisonnalités journalières, quotidiennes, hebdomadaires et annuelles. Prophet s'appuie sur les séries de Fourier pour le calcul de ces composantes :

$$s(t) = \sum_{k=1}^{N} a_k \cos(\tfrac{2\pi kt}{P}) + b_k \sin(\tfrac{2\pi kt}{P})$$

Où P désigne la période relative à chaque composante saisonnière (e.g., lors de l'étude d'une série journalière, $P = 7$ correspond à une composante hebdomadaire).

— Vacances et événements : Cette composante permet à Prophet d'ajuster les prévisions lorsqu'un jour férié ou un événement majeur peut modifier les prévisions (e.g., le 'Superbowl'). Pour cela il se base, par défaut, sur un ensemble de dates intégrées mais le module permet l'incorporation de dates ou périodes supplémentaires.

L'utilisation de Prohet pour imputer nos variables explicatives s'est faite exactement de la même manière que l'utilisation des modèles ARMA précédemment expliquée.

VAR

Le modèle vecteur autoregressif d'ordre p, VAR(p), est un modèle statistique permettant de capturer les interdépendances entre plusieurs séries temporelles (contrairement aux méthodes de séries temporelles susmentionnées). Ainsi, chaque variable est expliquée selon ses relations avec ses valeurs passées ainsi que celles des autres variables explicatives :

$$\begin{pmatrix} x_{t,1} \\ x_{t,2} \\ \cdots \\ x_{t,d} \end{pmatrix} = c + A_1 \begin{pmatrix} x_{t-1,1} \\ x_{t-1,2} \\ \cdots \\ x_{t-1,d} \end{pmatrix} + A_2 \begin{pmatrix} x_{t-2,1} \\ x_{t-2,2} \\ \cdots \\ x_{t-2,d} \end{pmatrix} + \cdots + A_p \begin{pmatrix} x_{t-p,1} \\ x_{t-p,2} \\ \cdots \\ x_{t-p,d} \end{pmatrix} + e_t$$

où les A_i sont les paramètres du modèle et ϵ_t le terme d'erreur.

Dans notre cas d'imputation, cette méthode multivariée s'applique de manière glissante sur l'ensemble de nos séries. Cette fenêtre glissante est constituée de 6 ans d'historiques.

1. **Phase d'entraînement :** dans le cadre d'un trimestre Q, on estime les paramètres du modèle via les 24 trimestres précédents $(x^{Q-24}, \cdots, x^{Q-1})$.
2. **Phase d'imputation :** On impute ensuite les données non encore publiées en se basant sur les données du passé déjà publiées ainsi que les paramètres du VAR estimés à l'étape précédente. À noter ici, que les composantes relatives aux trimestres $Q-1$, $Q-2$, $\cdots$, $Q-p$ peuvent varier durant le trimestre Q et ainsi modifier les composantes imputées. En effet, les révisions sont prises en compte via ce procédé.

Plus on incorpore de variables dans le modèle VAR et plus le nombre de paramètres à estimer s'accroît. Etant donnée la profondeur historique limitée de notre jeu macro-économique, cette situation conduit à un problème de sur-paramétrage. La méthode conventionnelle, pour estimer les paramètres dans cette de situation, consiste à se placer dans le cadre bayésien (Vecteur AutoRégressif Bayésien) et donc de fixer des croyances a priori sur les paramètres (Banbura et al., 2010 [2]).

Dans cette étude, lorsque le nombre de variables est trop important (sur-paramétrisation), on choisit d'estimer les paramètres en prenant le problème de minimisation suivant :

$$\arg\min_{c,A_1,\cdots,A_p} \|X_t - c - \sum_{j=1}^{p} A_j \; X_{t-j}\|^2$$

Cependant, à la place d'utiliser des régressions MCO afin d'estimer les différents paramètres, on procède à une descente de gradient à la manière d'un réseau de neurones. En effet, cette méthode permet d'obtenir des 'estimations' dans le cas d'une modélisation sur-paramétrée. En pratique, on utilise une librairie d'apprentissage profond (tensorflow). Pour ce faire, on construit un peuso-réseau composé d'autant d'entrées que de pas d'autorégression, de couches cachées uniques pour chaque entrée et de fonctions d'activation linéaires. Ainsi chaque matrice de paramètre A_j correspond aux poids synaptiques de la couche cachée de l'entrée j.
Pour des raisons de sur-apprentissage, et à l'image du réseau de neurones VAR réalisé par Schubert et al. (2019) [36] une contrainte de régularisation a été imposée sur nos paramètres.

Modélisation espace-état
On modélise ici nos variables explicatives par une modélisation espace-état. Cette modélisation a pour objectif de résumer l'information portée par un grand nombre de variables corrélées entre elles en une représentation plus parcimonieuse. Cette représentation repose sur l'hypothèse que nos variables explicatives ("hard-data", "soft-data" et financières) peuvent être guidées par un nombre restreint de facteurs latents inobservés, censés représenter la santé de l'économie portée par nos variables. Dans le cadre de notre imputation, l'idée sous-jacente consiste donc à estimer les facteurs cachés du trimestre étudié afin de prédire nos variables explicatives non encore publiées.

Plus formellement, on suppose que les facteurs latents suivent un modèle linéaire d'évolution (équation de transition) (1) :

$$F_Q = AF'_{Q-1} + u_Q \quad (1)$$

Avec :
F_Q variable latente à p composants (ou p facteurs) du trimestre Q
A vecteur de taille p qui représente l'évolution du système
u_Q, vecteur de taille p qui est un bruit centré de variance-covariance U

On suppose qu'à tout trimestre Q, les d variables publiées X_Q suivent une transformation linéaire des p facteurs latents F_Q (équation de mesure) (2) :

$$X^Q = CF_Q + v_Q \quad (2)$$

Avec :

X^Q, variables explicatives de notre jeu de données à d composantes
C, matrice (d, n) qui représente le processus de mesure (ou d'observation)
v_Q, vecteur de taille d qui est un bruit centré de variance-covariance R

On suppose en plus que u_t et v_t suivent une loi normale avec pour matrice de variance-covariance :

$$cov\begin{pmatrix} u_Q \\ v_Q \end{pmatrix} = \begin{pmatrix} U & 0 \\ 0 & R \end{pmatrix}$$

L'entraînement d'une telle modélisation consiste à estimer les paramètres A, C, U, R mais aussi l'ensemble des facteurs latents : $(F_Q)_{1 \leq Q \leq T}$. Au vu du grand nombre de variables utilisées dans notre étude, nous pouvons nous placer dans le cadre d'un modèle à facteurs approchés (Bessec et al., 2012 [5]). Dans ce cadre, il existe plusieurs méthodes permettant de calibrer une telle représentation. Les méthodes communément retenues pour l'analyse de données macro-économiques sont l'estimation en deux étapes (Doz et al., 2011 [16]) ainsi que l'estimation par pseudo-maximum de vraisemblance (Doz et al., 2012 [17]).

La calibration de la modélisation espace-état de cette étude se base sur la première méthode, c'est-à-dire, l'estimation en deux étapes. Comme son nom l'indique cette méthode est constituée de deux étapes. Elle est appliquée sur le jeu de données de la manière suivante :

Estimation des paramètres (A, C, U, R) *:* Durant cette étape, on initialise les facteurs (cadre statique) avec une analyse en composantes principales (ACP). Cette ACP est construite sur un jeu de données où seulement la dernière observation de chaque trimestre est prise en compte (i.e. juste avant la première estimation du PIB). De cette façon, le jeu de données ne présente pas de données manquantes. En procédant avec une ACP, on transforme les variables explicatives en p nouvelles variables décorrélées entre elles qui résument l'information économique portée par celles-ci.

— La matrice d'observation C se compose alors l'ensemble des vecteurs propres de l'espace de projection. En effet, la projection des vecteurs propres sur les p facteurs permet de reconstituer les variables explicatives X, ce qui correspond à l'équation d'observation (2).
— R s'obtient grâce à la reconstitution des variables explicatives de l'étape précédente. En effet, R est la matrice de variance-covariance des résidus v de l'équation d'observation (2). Elle peut donc être obtenue en calculant la matrice de covariance de $X - CF$.
— Afin d'estimer les paramètres de l'équation d'évolution (1), on entraîne un modèle vecteur autorégressif (VAR) d'ordre 1, qui permet de capturer les interdépendances temporelles des facteurs obtenues par ACP, comme modéliser dans l'équation (1). La matrice A d'évolution du système est initialisée grâce au coefficient matricielle obtenue dans le modèle VAR.
— De manière analogue à l'obtention des résidus de l'équation d'observation, on calcule les résidus u de l'équation d'évolution (1) en reconstituant les facteurs dans le temps. Ainsi, on obtient U, la matrice de variance-covariance des résidus $F_{Q+1} - AF_Q$.

Estimation des facteurs latents grâce au filtre de Kalman : Le filtre récursif de Kalman permet d'obtenir les meilleurs estimateurs de F_q sachant l'information à l'instant Q et $Q-1$ ($\widehat{F_{Q|Q}}$ et $\widehat{F_{Q|Q-1}}$) ainsi que les meilleurs estimateurs de $\Sigma_{Q|Q}$ et $\Sigma_{Q|Q-1}$, les matrices de variances-covariances associées. En utilisant la théorie bayésienne dans un cadre gaussien, il en découle les 5 étapes du filtre de Kalman à itérer. Elles permettent d'estimer à chaque trimestre les facteurs

latents conditionnellement aux variables observées de la période Q.
Pour tout trimestre $Q = 1, ..., T$:

- $\widehat{F_{Q+1|Q}} = A\widehat{F_{Q|Q}}$ - *équation "moyenne" de prédiction*
- $\Sigma_{Q+1|Q} = A\Sigma_{Q+1|Q}A^T + U$ - *équation "covariance" de prédiction*
- $\widehat{F_{Q|Q}} = \widehat{F_{Q|Q-1}} + K_Q(X_Q - C\widehat{F_{Q|Q-1}})$ - *équation de correction*
- $\Sigma_{Q|Q} = (1 - \mathbb{I} - K_Q C)\Sigma_{Q|Q-1}$ - *équation de correction*
- $K_Q = \Sigma_{Q|Q-1}C^T(C\Sigma_{Q|Q-1}C^T + R)^{-1}$ - *gain de Kalman*

ll ne reste plus qu'à spécifier les conditions initiales. $F_0 = (0)$ est un choix arbitraire mais naturel. Afin d'avoir un a priori qui a peu d'influence sur le modèle nous choisissons une très large matrice diagonale (10^5) de variance-covariance Σ_0.

On se place dans le contexte d'imputation de l'échantillon $x^{Q,t}$ relatif aux informations disponibles à la date t du trimestre Q.

1. Dans la **phase d'entraînement** du modèle, on estime l'ensemble des paramètres et des facteurs grâce à la méthode en deux étapes. Pour ce faire, on prend en compte toute l'information disponible à la date t, y compris les quelques variables publiées du trimestre en cours.
2. La **phase d'imputation** consiste alors à utiliser les équations du filtre de Kalman pour imputer les composantes manquantes de $x^{Q,t}$:
   - Ainsi, au fur et à mesure que les variables sont publiées, on utilise l'équation de correction afin de prédire de F_Q.
   - Puis, on utilise l'équation de mesure afin de prédire nos variables explicatives manquantes.

   À noter qu'en tout début de trimestre, lorsqu'aucune variable n'est encore publiée, on utilise uniquement l'équation moyenne de prédiction afin de prédire F_Q.

La modélisation espace-état nécessite de choisir le nombre de facteurs latents. Pour ce faire, certaines études utilisent des critères d'information (Bai et al., 2002). Dans notre étude, le choix a été fait de se placer dans un paradigme d'apprentissage statistique (*"machine learning"*) et d'appréhender le nombre de facteurs comme un hyper-paramètre de la modélisation. L'enjeu consiste à projeter notre modélisation sur différentes valeurs de r et ainsi obtenir la valeur la plus adéquate.

3.2.2.2 Imputation monotone

Comme décrit par Van Buuren (2018) [42] dans un cadre d'imputation de valeurs manquantes, l'algorithme d'imputation monotone se base sur l'hypothèse de variables explicatives monotones. Un modèle de données manquantes est dit monotone si les variables x_j peuvent être ordonnées telles que si x_j est manquant alors toutes les variables x_k, avec $k > j$, sont également manquantes. Dans le cadre de notre jeu de données macro-économiques, cette hypothèse est généralement vérifiée. En effet, les délais de publication des différents indicateurs restent constants dans le temps et donc leurs ordres d'apparition également.

1. L'algorithme inspiré de Van Buuren (2018) [42], appliqué au "nowcasting", consiste, dans un premier temps, à ordonner de manière croissante les variables en fonction du nombre moyen de valeurs manquantes par trimestre $(x_1, ..., x_d)$. Dans notre cas, cela revient aussi à les ordonner par délais moyens de premières publications au cours des trimestres. On peut donc dire qu'en moyenne x_j est publiée après x_{j-1} et avant x_{j+1}.

2. On parcourt ensuite les variables afin d'**entraîner** un modèle de régression f_j à prédire x_j à partir de ses variables explicatives $(x_1, ..., x_{j-1})$.
3. Puis, chaque valeur manquante de la variable x_j est **imputée** grâce aux valeurs de ses variables explicatives : $\widehat{x_j} = f_j(x_1, \cdots, x_{j-1})$ où f_j représente le modèle de prédiction de x_j. La variable x_j est alors actualisée par de nouvelles valeurs prédites $(x_{j,obs}, \widehat{x_{j,miss}})$ et sert ainsi de variable explicative à la variable suivante x_{j+1}.

À noter que pendant la phase d'entraînement du modèle, on calibre le modèle de la variable à prédire en se basant sur l'ensemble de ses valeurs non manquantes. Il peut arriver qu'une de ses variables explicatives ait une valeur manquante (puisque l'ordre a été établi en moyenne). Dans ce cas, les valeurs manquantes des variables explicatives sont remplacées par leur première valeur publiée au cours du trimestre. Ce problème n'intervient qu'en phase de calibration du modèle. En effet, la phase de prédiction consiste à appliquer strictement l'algorithme d'imputation monotone. Ainsi, pour chacune des variables à prédire, aucune valeur ne peut être manquante dans ses variables explicatives.

L'étude a été menée en utilisant différents modèles de régression : des régressions linéaires simples et pénalisées ainsi que méthodes de boosting utilisant des arbres.

3.2.3 Méthodes dérivées de l'imputation multiple

Il existe deux approches générales à l'imputation multiple de données manquantes : la distribution conjointe ("Joint Modeling" (JM)) et la spécification entièrement conditionnelle ("Fully Conditional Specification" (FCS)).

Modélisation conjointe - Algorithme d'espérance-maximisation
Dans le cadre théorique des mécanismes de valeurs manquantes, Schafer et al (2002) [34] reconnaissent le maximum de vraisemblance comme une technique avancée. En effet, sous certains mécanismes propres à la théorie des valeurs manquantes, cette technique permet de donner des estimations de paramètres non biaisées. Néanmoins, dans cette étude, nous ne nous attardons pas sur les propriétés de biais des estimateurs.
Cependant, l'intérêt fort de cette méthode réside dans la détermination de la loi des variables non publiées conditionnellement aux variables observées :

$$f(x_{miss}|x_{obs}, \theta) \text{ avec } x_i = (x_{i,obs}, x_{i,miss}) \text{ et } \theta \text{ ses paramètres}$$

On présuppose alors que les données suivent une modélisation conjointe (JM pour "Joint Modeling") selon une loi spécifique. On suppose généralement que la modélisation conjointe est de type de gaussien :

$$X \sim \mathcal{N}(\mu, \Sigma)$$

Il en découle alors que la loi des variables non publiées conditionnellement aux variables observées suit également une loi gaussienne. La détermination de la loi découle donc simplement de l'identification de ses paramètres de moyenne et de matrice de variance-covariance. Ainsi, en ayant trouvé de manière optimale leur loi conditionnelle, on peut déterminer la valeur des variables non publiées. L'une des manières optimales d'obtenir les paramètres est d'utiliser le maximum de vraisemblance. Autrement dit, on détermine les paramètres de loi μ et Σ qui maximisent la réalisation des échantillons.
La méthode la plus directe consiste à se concentrer uniquement sur les échantillons complets et d'appliquer les estimateurs du maximum de vraisemblance afin d'obtenir $\widehat{\mu}$ et $\widehat{\Sigma}$. Néanmoins cette méthode ne permet pas de prendre en compte l'information potentiellement contenue dans les échantillons ayant des composantes manquantes. L'utilisation de toutes les données disponibles,

même de celles contenant des valeurs manquantes, est alors intuitivement plus attrayante afin d'estimer les paramètres. Cela est rendu possible via l'utilisation de l'algorithme d'espérance-maximisation à des fins d'imputations.

L'algorithme EM, ou espérance-maximisation, est un algorithme d'optimisation itératif identifiant les valeurs les plus probables des paramètres du modèle. Cet algorithme calcule de manière itérative le maximum de vraisemblance du jeu de donnée. Il s'applique également dans l'estimation de valeurs manquantes, et est facilement adaptable à des problèmes complexes de données manquantes. Chaque itération de l'algorithme EM pour l'imputation se déroule en deux étapes : une étape E (étape d'espérance) et une étape M (étape de maximisation par maximum de vraisemblance).

On se place dans le cadre théorique d'imputation suivant (où tout échantillon x peut se mettre sous la forme (x_{obs}, x_{miss})) :

$$\begin{pmatrix} x_{miss} \\ x_{obs} \end{pmatrix} \sim \mathcal{N}(\begin{pmatrix} \mu_{miss} \\ \mu_{obs} \end{pmatrix}, \Sigma) \text{ avec } \Sigma = \begin{pmatrix} P_{miss} & C \\ C & S_{obs} \end{pmatrix}$$

— On initialise le vecteur de la moyenne μ et la matrice de variance-covariance Σ en se basant seulement sur les dates de publication où l'ensemble des variables ont été publiées (estimation par "listwise deletion").

— **Etape E :** cette étape consiste à estimer temporairement les valeurs non publiées en utilisant l'espérance conditionnelle (sachant les valeurs des variables publiées). En réalité, cette estimation temporaire des variables non publiées sert essentiellement au calcul des statistiques (moyenne et matrice de covariance pour la loi gaussienne multivariée) de l'étape M. En pratique on utilise la loi conditionnelle de $x_{miss}|x_{obs}$ pour aboutir aux estimations suivantes :

$$\mathbb{E}(x_{miss}|x_{obs}) = \mu_{miss} + CS_{obs}^{-1}(x_{obs} - \mu_{obs})$$
$$\mathbb{V}(x_{miss}|x_{obs}) = P_{miss} - CS_{obs}^{-1}C^T$$

— **Etape M :** Cette étape consiste à calculer les paramètres de la loi gaussienne multivariée (moyenne et matrice variance-covariance), grâce aux estimateurs de maximum de vraisemblance, en considérant alors le jeu complet de données. On entend par jeu complet de données, les estimations des valeurs non publiées résultant de l'étape E, ainsi que les estimations des valeurs publiées. En pratique on obtient les estimateurs suivants :

$$\widehat{\mu} = \frac{1}{N}\sum_i x_i$$
$$\widehat{\Sigma} = \frac{1}{N}\sum_i (x_i^2 - \frac{\sum_j x_j^2}{N})$$

— On répète alors ces deux étapes, jusqu'à que le vecteur de moyenne et la matrice de variance-covariance aient convergé. Autrement dit, l'algorithme s'arrête lorsque la différence de ces de deux itérations successives est en dessous d'un certain seuil. Dans notre cas, l'algorithme arrive à convergence en une cinquantaine d'itérations (seuil fixé à 10^{-3}).

L'imputation multiple, de l'approche JM, s'effectue directement via la distribution a posteriori des variables manquantes. Cependant, notre étude ne nécessite pas d'analyse utilisant l'imputation multiple. Nous nous restreignons donc à l'utilisation de l'espérance de chacune des distributions a posteriori $\mathbb{E}(f(x_{miss}|x_{obs}, \mu, \Sigma))$.

Spécification entièrement conditionnelle - SICE approche inspirée de l'imputation multivariée par équations chaînées
L'algorithme d'imputation multivariée par équations chaînées (MICE pour "Multiple Imputation by Chained Equations") est une technique multivariée d'imputation multiple. En effet, le premier intérêt de cette technique repose sur les multiples estimations données aux valeurs manquantes d'un jeu de données, créant ainsi plusieurs jeux de données dits "complets". Le second porte sur son imputation multivariée. En effet, les valeurs manquantes sont imputées à partir des valeurs observées pour un échantillon de données, et des relations observées entre variables explicatives.

Comme vu précédemment, les méthodes de modélisations conjointes (JM pour "Joint Modeling") permettent également l'imputation multiple. Néanmoins, elles présupposent que le jeu de données soit induit par une loi multivariée spécifique (e.g., gaussienne). Or, cette hypothèse peut s'avérer trop rigide et inappropriée pour des larges jeux de données, comme cela peut-être le cas dans notre modélisation "nowcasting" (Stuart et al., 2009 [38]). À l'inverse, l'algorithme MICE repose sur l'imputation de type spécification entièrement conditionnelle (FCS pour "Fully Conditional Specification") et, de ce fait, permet davantage de liberté dans ses hypothèses de densités. En effet, chaque variable peut être spécifiée par une densité univariée différente de celles des autres variables. Ces densités sont conditionnelles aux autres variables explicatives. La spécification de la distribution multivariée se fait donc à travers un ensemble de densités conditionnelles, d'où l'expression "Fully Conditional Specification".
L'interêt principal de l'imputation multiple porte sur la mesure de l'incertitude à travers les différentes imputations de données manquantes. Or, ce point n'est pas décisif dans notre étude. On utilise donc un algorithme itératif d'imputation simple, fortement inspiré de l'algorithme MICE. Cet algorithme est qualifié de SICE pour "Single Imputation by Chained Equations" par Khan et al. (2020) [25].

En pratique, l'algorithme appliqué à notre jeu de données macro-économiques se déroule ainsi :

- **Etape 1 :** On spécifie notre modèle d'imputation en indiquant les modèles conditionnels de chaque variable $(f_1, \cdots, f_p)$, où f_1 est le modèle de la variable y_1 conditionnellement aux autres variables : $(x_2, \cdots, x_d)$
- **Etape 2 :** Il s'agit de l'étape d'initialisation. On impute les valeurs non publiées de notre flux de données avec un modèle d'imputation univariée. Par exemple, on peut imputer avec une moyenne ou une marche aléatoire naïve. Cependant, on garde une trace des positions initiales des valeurs manquantes : $(x_{i,miss})_{1 \leq i \leq d}$.
- **Etape 3 :** On choisit une variable x_i parmi nos variables présentant des valeurs non publiées dans notre jeu de données $(x_1, \cdots, x_d)$. Pour cette variable on restreint notre jeu de données aux dates à laquelle la variable est publiée. Puis, on entraîne le modèle de régression f_i, spécifié à l'étape 1, à prédire la variable x_i à partir des autres variables $(x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_d)$, (en prenant en compte leurs dernières imputations).
- **Etape 4 :** On estime ensuite les valeurs manquantes de la variable x_i grâce au modèle de régression f_i entraîné à l'étape précédente pour obtenir $\widehat{x_{i,miss}}$. À noter que les valeurs estimées de x_i serviront par la suite à imputer les autres variables du jeu de données. Elles seront présentes en tant que valeurs explicatives de la variable x_i au même titre que les valeurs $x_{i,obs}$. En d'autres termes, l'estimation des valeurs non publiées d'une autre variable explicative x_j (avec $j \neq i$), à cette même étape 4, se basera notamment sur les estimations $\widehat{x_{i,miss}}$.
- **Etape 5 :** On répète les étapes 2 et 3 sur toutes nos autres variables présentant des valeurs non publiées dans notre reconstruction du flux de données $(x_1, \cdots, x_d)$. Une fois que les d itérations ont été faites, les variables de notre jeu de données présentent alors des premières estimations des données manquantes. On aura alors procédé à un premier

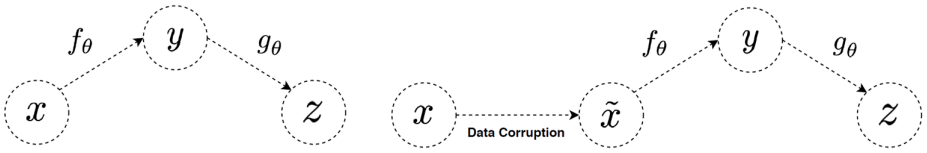

FIGURE 9 – Structures d'un Auto-encodeur simple (AE) et d'un Auto-encodeur de débruitage (DAE). Sources des schémas : Pereira et al., 2020 [30]

"cycle" d'estimation.

— **Etape 6 :** On itère les étapes 2 à 5 sur un certain nombre de cycles, jusqu'à obtenir convergence. On fixe le seuil à 10^{-3}, ce qui, dans notre cas, conduit à un nombre d'térations d'environ 10.

Dans la présente étude, le choix a été fait de toujours utiliser les mêmes méthodes de régression. Ainsi $f_1 \cdots f_p$ appartiennent à la même famille de modèles de régression. Cependant, nous comparons cette technique d'imputation avec plusieurs algorithmes de régression multivariée, de nature variée. Nous utilisons ainsi des modèles linéaires avec ou sans pénalisation, des méthodes d'arbres ou encore des méthodes de boosting.

L'algorithme, ci-dessus, appliqué avec des régresseurs de forêts aléatoires, est également connu sous l'appellation "MissForest" (Stekhoven et al., 2012 [37]).

3.2.4 Méthodes utilisant l'apprentissage profond

Les méthodes dîtes d'apprentissage profond ("deep learning"), par l'élaboration de structures de réseaux de neurones spécifiques, répondent également aux problématiques des données manquantes. Parmi ces structures, deux semblent se détacher du lot : les auto-encodeurs de débruitage (MIDA pour "Multiple Imputation using Denoising Autoencoders") et les réseaux antagonistes génératifs à imputation (GAIN pour "Generative Adversarial Imputation Nets"). Yoon et al. (2018) [45] suggèrent que leur méthode, les GAIN, basée sur la théorie des réseaux antagonistes génératifs (GAN pour "Generative Adversarial Networks"), améliorent de manière significative les méthodes d'imputation plus courantes. De même, Gondara et al. (2018) [24] indiquent que leur structure, appelée MIDA, basée sur la théorie des auto-encodeurs, surpasse l'ensemble des autres méthodes. Au-delà de ces articles initiateurs, des études s'intéressent à leurs évaluations pratiques en les comparant à d'autres méthodes dans plusieurs cadres expérimentales. Ainsi Wang et al. (2021) [43] étudient ces deux méthodes et concluent que bien qu'elles soient moins coûteuses en temps de calcul, elles ne permettent pas de battre les méthodes plus usuelles telles que MICE. Face à ces incertitudes, et afin d'étudier leur apport sur un jeu de données macro-économiques, le choix a été fait de les inclure dans les présents travaux.

Auto-encodeurs pour l'imputation des données manquantes

Les auto-encodeurs (AE) sont des réseaux de neurones composés d'au moins trois couches (couche d'entrée, couche cachée et couche de sortie). Ces réseaux se divisent en deux parties : la partie relative à l'encodeur et la partie relative au décodeur. La première correspond à l'acheminement entre la couche d'entrée et la couche cachée (intermédiaire). La seconde correspond à la partie restante : elle commence à la couche cachée pour terminer à la sortie (voir Figure 9).

Plus formellement, la partie encodeur permet de faire coïncider un vecteur d'entrée x à la représentation cachée y via une transformation non linéaire f (dans le cadre d'un encodeur à une

couche $y = f(x) = s(w_1 x + b_1)$ - où w_1, b_1 représentent respectivement le poids et le biais synaptiques alors que s correspond à la fonction d'activation non linéaire). La représentation y est ensuite accordée à un vecteur z ayant la même taille que x via une fonction non linéaire g (dans le cadre d'un décodeur à une couche $z = g(y) = s(w_2 y + b_2)$). L'entraînement d'un AE consiste donc à mettre à jour l'ensemble de ses paramètres (ici w_1, b_1, w_2, b_2) pour minimiser l'erreur de reconstruction entre x et z via une descente de gradient stochastique.

Les AE peuvent prendre deux types de représentations : sur-complète lorsque la représentation cachée est plus grande que l'entrée et sous-complète lorsque la représentation cachée est plus petite. Dans le cadre des AE classiques, les architectures de type sur-complètes sont peu utilisées. En effet, leur architecture sur-complète n'apprend que l'identité en copiant l'entrée vers la sortie, ce qui crée un scénario de sur-apprentissage et nécessite des ajustements particuliers.

Il y a cependant un domaine où les structures sur-complètes semblent pertinentes : le cas des auto-encodeurs de débruitage (denoising autoencoders : DAE). La principale différence de ces AE avec un AE classique réside dans l'application d'une corruption stochastique aux entrées du modèle pendant la phase d'entraînement. Dans ce cadre, le risque de sur-apprentissage lié à l'éventuelle structure sur-complète est surmonté puisque que le réseau compare maintenant l'entrée d'origine à sa version corrompue. Ainsi, le bruit d'entrée agit comme un terme de régularisation. À ce titre, il existe principalement trois façons de bruiter l'entrée d'un modèle :

- Remplacer par 0 des données d'entrée (analogue au "dropout").
- Ajouter du bruit gaussien aux données d'entrée.
- Ajouter du bruit du type "poivre et sel" aux données (1 ou -1).

L'application d'un DAE à l'imputation des données manquantes de cette étude consiste à entraîner le modèle en se focalisant uniquement sur les observations complètes que l'on démultiplie en observations corrompues via l'utilisation de différents bruits et de différents seuils de corruption. Concrètement, cette méthode consiste à bruiter le jeu de données restreint aux dates où toutes les variables explicatives ont été publiées. Dans notre cas, le bruit gaussien et le bruit qui force plusieurs composantes à zéro ont été utilisés. On entraîne ensuite le DAE à prédire les échantillons complets (toutes les variables ont été publiées) à partir des échantillons précédemment bruités.

Une fois le DAE entraîné et un nouvel échantillon partiellement observé x_j, on remplace ses composantes non publiées par du bruit gaussien afin de nourrir le DAE et d'obtenir ses imputations finales $\widehat{x_j} = (x_{j,obs}, \widehat{x_{j,miss}})$.

Dans ce cadre, il existe une infinité de choix pour construire et entraîner un DAE. En effet, on peut agir sur les différents paramètres de structures, d'entraînements et de corruptions afin d'obtenir une grande variété de DAE. Pour faciliter cette tâche, l'étude réalisée par Pereira et al. (2020) [30] permet de guider notre recherche. Ces derniers analysent 26 publications à ce sujet et dégagent des tendances prédominantes. Ainsi, la structure de DAE de cette étude est très proche de la structure recommandée. Elle a cependant été ajustée à notre problématique. Finalement la structure retenue (que l'on nommera DAE sur-complet) depend d'un paramètre θ qui correspond au nombre de neurones à ajouter à chaque couche de l'encodeur, respectivement à soustraire à chaque couche du décodeur.

En parallèle, nous avons également fait le choix de répliquer le cas d'une structure sur-complète à l'image de celle développée par Gondara et al. (2018) [24], intitulée MIDA. En ajustant l'ensemble des paramètres à notre problématique, on aboutit à la structure de notre second DAE (que l'on nommera DAE sous-complet) qui dépend du même paramètre θ. Dans ce second DAE,

θ correspond au nombre de neurones à soustraire à chaque couche de l'encodeur, respectivement à ajouter à chaque couche du décodeur. Ces structures sont détaillées en Annexe A.2 Figure 23 et Figure 24.

Réseaux antagonistes génératifs pour l'imputation de données manquantes (GAIN) GAIN (Yoon et al., 2018 [45]), pour "Generative Adversarial Imputation Nets", est une méthode d'imputation basée sur la théorie des GAN. Cette théorie fait partie des modèles génératifs et se concentre sur la manière dont les observations ont été obtenues. Pour parvenir à imputer des données manquantes, Yoon et al. (2018) [45] proposent de mettre en concurrence deux réseaux de neurones : un générateur G et un Discriminateur D. La méthode peut être illustrée comme un jeu dans lequel les deux réseaux s'affrontent durant la phase d'entraînement. Le générateur va chercher à imputer les données manquantes de la manière la plus réaliste possible afin de duper le discriminateur, tel un faussaire. Tandis que le discriminateur agit comme un policier en tentant d'identifier les éléments créés par le faussaire. L'idéal recherché est donc la victoire du faussaire, i.e. que les éléments imputés soient si ressemblants des réels que le policier n'ait plus la capacité de distinguer le vrai du faux.

Plus formellement, en partant d'un échantillon $x_i = (x_{i,obs}, x_{i,miss})$, M le masque indiquant les positions des valeurs observées et un bruit aléatoire Z permettant de remplacer initalement les données manquantes, le générateur prédit à la fois les données manquantes et les données observées $G(x_i, Z, M) = (\widehat{x_{i,obs}}, \widehat{x_{i,miss}})$. Puis le discriminateur utilise $\widehat{x_i} = (x_{i,obs}, \widehat{x_{i,miss}})$ et une matrice H appelé Hint afin de prédire le masque $D(x_i, H) = \widehat{M}$. La matrice de Hint permet d'aider le discrimateur en lui donnant une information partielle sur le vrai masque M. Elle peut être échantillonnée à partir d'une loi de Bernoulli (Wang et al., 2021 [43]) ou résulter d'un phénomène plus complexe qui contraint théoriquement le générateur à répliquer les bonnes distributions sous-jacentes (Yoon et al., 2018 [45].) En définisant les fonctions de coût suivantes (cadre d'une observation de dimension d) :

$$\mathcal{L}_D(M, \widehat{M}) = \sum_{j=1}^{d} M_j log(\widehat{M}_j) + (1 - M_j) log(1 - \widehat{M}_j)$$

$$\mathcal{L}_G(Y, M, \widehat{Y}, \widehat{M}) = \sum_{j=1}^{d} M_j log(1 - \widehat{M}_j) + \sum_{j=1}^{d} (1 - M_j)(Y_j - \widehat{Y}_j)^2$$

On entraîne alternativement le discriminateur à minimiser la fonction de coût $\mathcal{L}_D(M, \widehat{M})$ puis le générateur à minimiser la fonction de coût $\mathcal{L}_G(Y, M, \widehat{Y}, \widehat{M})$). En pratique le discriminateur va chercher à minimiser l'entropie croisée entre le masque et sa prédiction alors que le générateur va chercher à maximiser cette même entropie croisée tout en minimisant l'entropie de reconstruction (sortie du générateur).

Concernant l'implémentation de notre GAIN, nous nous sommes appuyés sur l'implémentation des auteurs (Yoon et al., 2018 [45]) auquel nous avons fait des ajustements adaptés à notre problématique. Au cours de nos expériences, nous avons rencontré quelques difficultés lors de l'apprentissage de nos réseaux, notamment leurs instabilités et la non-convergence vers l'équilibre de Nash (Salimans et al., 2016 [33]). En plus de procéder à un réglage de nos hyper-paramètres, nous avons également suivi quelques recommandations de Salimans et al. (2016) [33] afin d'améliorer l'apprentissage de nos réseaux. Parmi ces techniques, la "batch-normalization" a permis de stabiliser l'entraînement tout en améliorant les performances. L'idée de la "batch-normalization" est une normalisation des données sur chacune des couches du réseau et pour chaque batch de données.

3.3 Méthodes de prédiction

Comme précisé précédemment, le coeur de l'étude porte sur la comparaison des méthodes d'imputation et non des méthodes de prédiction. En conséquence, la description des modèles utilisés lors de l'étape de prédiction est plus succincte.

3.3.1 Moindres carrées ordinaires et pénalisation

La méthode des moindres carrés ordinaires (MCO) a ses limites, notamment lorsque le nombre de variables explicatives est supérieur au nombre d'observations. Sans régularisation, l'utilisation des moindres carrés ordinaires peut conduire au risque de sur-apprentissage (problématique de variance) et des problèmes de multi-colinéarité dans les données.

Les méthodes de régression pénalisée tentent de pallier ce problème. La régression Ridge ainsi que la régression Lasso en sont des exemples. Le modèle de Ridge consiste à pénaliser le modèle de régression par une pénalisation $L2$ alors que le modèle de Lasso consiste à pénaliser le modèle de régression par une pénalisation $L1$. L'application d'une pénalisation $L0$ conduit à une méthode de sélection du meilleur sous-ensemble dans le cadre des MCO ("subset selection").

Dans les trois cas, les paramètres de régularisation force les coefficients des régressions à se rapprocher de zéro (égaux dans le cadre de la pénalisation $L0$).

3.3.2 Méthodes ensemblistes

Les méthodes ensemblistes consistent à créer un comité d'experts à partir de prédicteurs faibles. L'enjeu réside dans l'agrégation de ces différents prédicteurs. Pour simplifier, nous distinguerons 3 types d'agrégations : le bagging, les arbres extrêmement aléatoires et le boosting.

Bagging
Dans le cas d'une variable prédite continue, le "bagging", pour "Bootstrap Aggregation", consiste à moyenner les résultats de prédictions de différents modèles indépendants. Cette indépendance des modèles est permise grâce à l'utilisation de jeux de données bootstrapés. Cette technique étudiée par Breiman (1996) [8] permet donc de réduire la variance, et donc de réduire l'erreur de prévision obtenue sur des données n'ayant pas contribuer à la calibration du modèle.

Forêts aléatoires - Bagging amélioré
Les forêts aléatoires (ou "random forest") sont des modèles de bagging d'arbres binaires de décision (CART, acronyme pour "Classification And Regression Trees") où une composante aléatoire est ajoutée sur le choix des variables. Cette méthode développée par Breiman (2001) [9] vise donc à accentuer davantage le caractère indépendant des modèles, et donc à diminuer d'autant plus la variance du modèle.

Pour chaque arbre, on utilise un sous échantillon du jeu de données d'entrée (avec remise). À chaque noeud de l'arbre, on crée un sous échantillon des variables explicatives du jeu de données par tirage aléatoire (par exemple, un échantillon composé de 3 variables explicatives : production industrielle, PMI, consommation des ménages). Une fois le sous échantillon de variables explicatives sélectionné, on cherche pour chaque variable explicative du sous échantillon la séparation optimale des données. Dans le cadre d'une prédiction continue ou "régression" (i.e. quantitative), on cherche la séparation des observations de notre variable explicative qui minimise la variance de $y_{Q,t} = log(PIB_{Q,t}) - log(PIB_{Q-1,t})$ dans les noeuds fils.

Arbres extrêmement aléatoires
Tout comme les forêts aléatoires, les arbres extrêmement aléatoires ou "ExtraTrees" sont des modèles agrégeant des arbres binaires de décision. Cependant, et comme décrit par Geurts et al (2006) [23], dans le cas d'arbres extrêmement aléatoires ou "ExtraTrees", l'ensemble des données d'entrée est utilisé. C'est la première grande différence avec la forêt aléatoire : on n'utilise plus des sous échantillons de données pour construire les arbres individuels. Il ne s'agit donc pas d'un algorithme de bagging. Il est tout de même possible de paramétrer l'ExtraTrees pour qu'il utilise également des sous échantillons.

Comme pour les forêts aléatoires, on utilise un sous échantillon aléatoire des variables explicatives du jeu de données à chaque noeud. Néanmoins, pour chaque variable explicative de l'échantillon, la séparation des données ne se fait plus de manière optimale. La séparation devient aléatoire. Cependant, comme pour les forêts aléatoires, une fois que les séparations de données ont été trouvées pour chaque variable explicative, l'algorithme optimise le choix de la variable explicative, toujours en minimisant la variance de $y_{Q,t}$ dans les noeuds fils (pour un problème de "regression").

Boosting
Tout comme le bagging, le boosting consiste à moyenner (dans le cadre continu) les prédictions des sous modèles. Cependant, le boosting utilise une "moyenne pondérée". La différence principale entre le boosting et le bagging repose sur la construction des sous-modèles.

Dans le cas du bagging, chaque construction de sous-modèle est indépendante de la construction des autres sous-modèles. Dans le cas du boosting, chaque sous modèle est une version améliorée du sous modèle construit précédemment. Cette adaptation repose sur l'attribution d'une pondération plus forte sur les observations que le sous modèle précédent a mal prédites. Ainsi, tout l'objectif du boosting est d'améliorer au fur et à mesure sa prédiction en se concentrant sur les observations les plus complexes à prédire. Le biais de prévision y est amélioré, comparativement à d'autres approches. En effet, dans le cas du bagging, le biais du modèle final (résultant de l'agrégation des sous modèles) est celui d'un seul arbre (l'espérance de la moyenne des arbres est l'espérance d'un arbre). Ce n'est plus le cas avec un algorithme de boosting. En effet, la prévision finale est une combinaison pondérée par les qualités d'ajustement de chaque modèle.

Le **Gradient boosting** proposé par Friedman (1999) [22] est un algorithme dérivé des méthodes de boosting où l'amélioration dans la construction de la séquence de modèles se fait grâce à la direction que prend le gradient de la fonction de perte. Ainsi, la convergence de l'algorithme y est améliorée. Le gradient y est approché par un arbre de régression.

Le **LightGBM** est un algorithme de gradient boosting qui se base sur des histogrammes dans la construction des arbres. L'intérêt principal étant d'accélérer la construction des arbres et de réduire l'utilisation de la mémoire. En effet, l'algorithme regroupe des valeurs des variables explicatives continues dans des bacs discrets, à la manière d'un histogramme, ce qui accélère la procédure de construction de l'arbre. Contrairement à des algorithmes de construction d'arbres plus classiques, la construction de l'arbre ne se fait plus par niveau, mais feuille par feuille. Cette approche permet d'augmenter la précision, mais peut avoir tendance à également augmenter la variance, si les paramètres de cet algorithme ne sont pas eux aussi défiés.

4 Résultats

4.1 Procédure d'évaluation

4.1.1 Métriques

Chaque méthode d'imputation détaillée ci-dessus permet d'estimer des valeurs pour les données initialement non publiées. Ces estimations, combinées aux apports de l'apprentissage statistique, permettent d'obtenir la prédiction finale du PIB. Afin d'être en mesure de comparer les méthodes entre elles, nous avons mis en place une procédure d'évaluation qui repose sur deux types de mesure :

— Métriques intrinsèques : elles permettent de quantifier la qualité de l'imputation, i.e. la première étape de la modélisation.
— Métriques extrinsèques : elles permettent de mesurer la tâche de prédiction du PIB, i.e. la seconde étape de la modélisation.

Métriques intrinsèques
Ces métriques mesurent la qualité de l'imputation en indiquant la ressemblance aux "vraies valeurs". Dans le cadre du nowcasting, les "vraies valeurs" correspondent aux premières estimations des différentes variables macro-économiques pour chacune des périodes valorisées. À l'image des travaux de Bertsimas et al. (2017) [4], on utilise l'erreur absolue moyenne (MAE pour "Mean Absolute Error") et la moyenne des carrés des erreurs (MSE pour "Mean Square Error") pour quantifier les écarts. Pour simplifier la compréhension de la MSE, on utilisera également la racine de l'erreur quadratique moyenne (RMSE pour "Root Mean Square Error") (Schmitt et al., 2015 [35]).

Métriques extrinsèques
Schmitt et al., 2015 [35] suggèrent que peu d'études s'intéressent à l'effet de l'imputation sur une analyse de plus haut niveau comme une tâche de régression ou de classification. Dans notre cas, nous examinons l'apport des méthodes d'imputation pour une tâche de plus haut niveau : la prédiction du PIB. Ainsi, les métriques extrinsèques permettent de mesurer la qualité de nos prédictions. Plus particulièrement, il s'agit de mesurer les performances des algorithmes de prédiction entraînés sur les jeux de données imputées. Dans ce cadre, nous utilisons les métriques de performance les plus communes au domaine du "nowcasting" i.e. le MAE, le MSE, le signe (croissance ou récession) et le RMSE (Banbura et al., 2010 [1] ; Bok et al, 2018 [12]).

Par soucis de rigueur, les mesures de performance sont faites, dans les deux cas, hors-échantillon ("out-of-sample"). Du fait du caractère temporel de nos données, nous attachons davantage d'importance à ce que les modèles et les mesures respectent la chronologie des événements. Ainsi, nos mesures incorporent la notion de "forecast" (Rousset et al., 2018 [31]), et de ce fait, prennent les appellations suivantes, avec "F" pour "forecast" : MAFE (pour "mean absolute forecast error"), MSFE (pour "mean square forecast error"), RMSFE (pour "root mean square forecast error") et F-sign (pour "forecast sign"). À noter que la dernière mesure permet se focaliser sur le sens de la variation (croissance ou décroissance), plutôt que la différence de valeur entre la prédiction et le réalisé.

On accorde cependant plus d'importance à l'objectif final de prédiction du PIB. On privilégie donc les méthodes qui surpassent sur les mesures extrinsèques, tout en contrôlant la cohérence de leur mesure intrinsèque.

4.1.2 Validation croisée temporelle

En apprentissage statistique classique, l'estimation des performances s'effectue généralement par validation croisée. Néanmoins, comme suggèrent Cerqueira et al (2020) [19], cette estimation ne prend pas en compte la structure temporelle.

Ainsi, la validation croisée temporelle avec fenêtre à expansion (ensemble d'entraînement itérativement augmenté) est utilisée dans cette étude. À chaque étape de la validation croisée, l'ensemble d'apprentissage est composé uniquement d'échantillons qui se sont produits avant les échantillons de l'ensemble de données de test. De ce fait, aucune observation future ne peut être utilisée lors de l'entraînement du modèle.

L'entraînement des modèles nécessite une profondeur suffisante d'historique. Ainsi, les premiers trimestres de notre historique de données serviront seulement à l'entraînement des modèles, et ne seront jamais choisis pour les phases de test.

Dans notre cas, l'ensemble d'entraînement débute toujours au Q1 2005 et il est séquentiellement augmenté de deux trimestres au fur et à mesure des itérations de notre validation croisée (voir Figure 10). Ce processus, à l'apparence simple, s'avère plus complexe lors de son application sur notre double indexation temporelle (*période valorisée*, *date de publication*). En effet, il s'agit d'inclure dans l'ensemble d'entraînement uniquement les informations disponibles avant chaque date de l'échantillon de test.

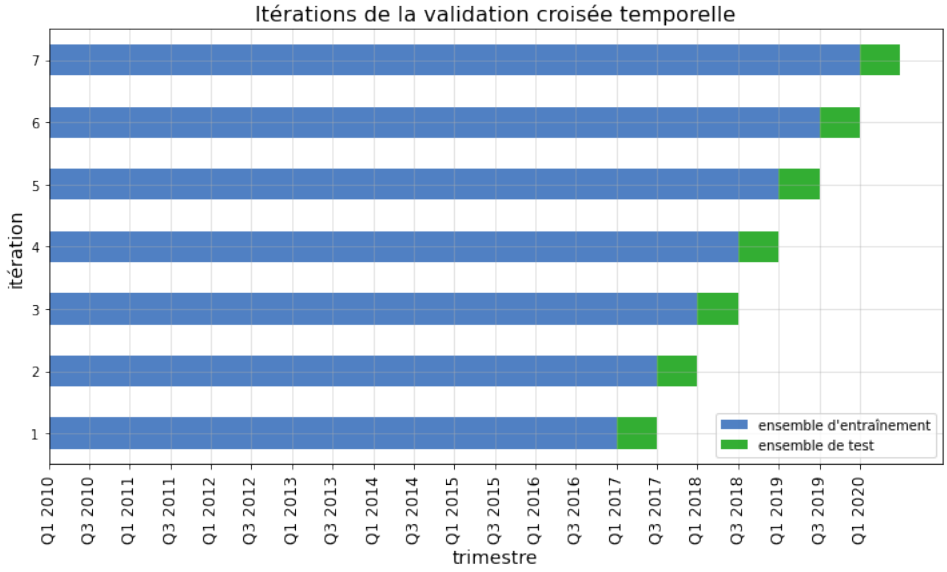

FIGURE 10 – Processus utilisé de validation croisée temporelle

4.2 Expériences

Nous initions notre recherche de meilleures méthodes de modélisation en testant les différentes combinaisons possibles entre méthode d'imputation des variables explicatives et méthode de prédiction du PIB. Pour chaque combinaison, on évalue les performances intrinsèques et extrinsèques grâce à la procédure d'évaluation détaillée en amont. Il peut s'avérer que certaines méthodes d'imputation dépendent d'un ou plusieurs paramètres (et hyper-paramètres). On peut citer, entre autres, le nombre de facteurs latents pour la modélisation espace-état, le nombre de voisins pour les méthodes basées sur les K-NN ou encore le nombre de neurones θ qui s'ajoutent sur chaque couche de l'encodeur du DAE sur-complet. Dans chacun de ces cas, nous testons les

modélisations avec différentes valeurs possibles grâce à une "grille de recherche". À noter également que le nombre de variables sélectionnées et la méthode de sélection de ces variables sont également deux paramètres de modélisation. Dans notre cas, ces expériences ont été coûteuses en temps de calcul. C'est d'ailleurs pour cette raison que nous avons choisi de limiter le nombre de nos variables explicatives.

Nous étudions dans un premier temps les résultats sur un régime économique que l'on pourrait qualifier de "business as usual" avant de s'intéresser tout particulièrement aux résultats relatifs au début de la crise covid-19. Ainsi, la période "business as usual" correspond aux années 2017 à 2019 inclues.

4.2.1 Période hors crise covid-19

Comme évoqué précedemment, la période d'évaluation hors crise covid-19 correspond aux années 2017 à 2019 inclues.

Mesures intrinsèques
La Table 2, ci-dessous, présente les 5 méthodes permettant d'obtenir la plus faible erreur d'imputation des variables explicatives. On notera la prééminence de l'algorithme EM et des méthodes dérivées des K-NN.

Méthode d'imputation	Famille	Méthode de sélection	Nombre de variables	MAFE	RMSFE
EM	JM	score	150	**0.095**	**0.138**
EM	JM	corrélation	150	0.145	0.190
KNN	KNN	corrélation	30	0.210	0.211
IKNN	KNN	corrélation	30	0.224	0.224
SKNN	KNN	corrélation	30	0.249	0.250

TABLE 2 – 5 meilleures méthodes selon les performances intrinsèques - Evaluation sur la période hors crise covid-19

De manière générale, les méthodes de prévisions temporelles (Marche aléatoire Naïve, ARMA, VAR, Prophet) tendent à sous performer les méthodes d'imputation stricto sensu (Figure 11). Les méthodes basées sur les K-NN se démarquent fortement. 3 méthodes basées sur cet algorithme se retrouvent effectivement parmi les 5 meilleures méthodes selon les performances intrinsèques. Les méthodes liées à l'apprentissage profond (DAE et GAIN) réussissent moins bien leur tâche d'imputation que la moyenne et affichent des performances intrinsèques relativement faibles.

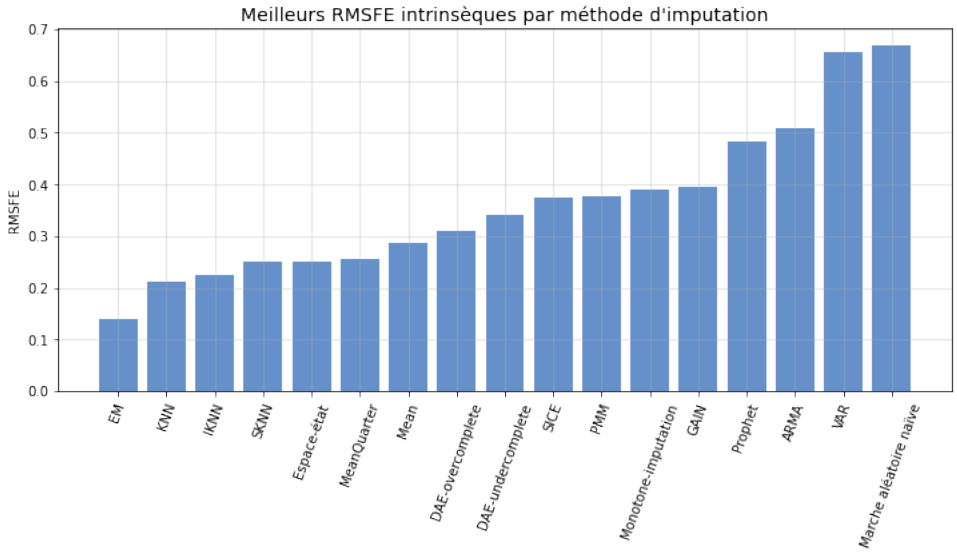

FIGURE 11 – Performances intrinsèques - hors crise covid-19

Mesures extrinsèques

De manière fidèle à notre modélisation en deux étapes, nous évaluons la performance de la tâche de prédiction du PIB, c'est-à-dire, la performance de la seconde étape. Le tableau ci-dessous présente alors les 5 meilleures combinaisons de méthodes d'imputation et de méthodes de prédiction. Les combinaisons optimales sont celles qui minimisent l'erreur de prédiction.

Méthode d'imputation	Famille	Méthode de prédiction	Méthode de sélection	Nombre de variables	MAFE	RMSFE	signe
Espace-état	Séries temporelles	Arbres extrêmement aléatoires	Score	150	**0.00135**	**0.00137**	0.83
IKNN	KNN	Forêt aléatoire	Corrélation	150	0.00139	0.00141	**0.91**
DAE surcomplet	DAE	Forêt aléatoire	Score	150	0.00139	0.00142	0.83
IKNN	KNN	Forêt aléatoire	Score	150	0.00143	0.00143	0.83
SICE	FCS	Forêt aléatoire	Corrélation	150	0.00143	0.00145	0.85

TABLE 3 – 5 meilleures combinaisons selon les performances extrinsèques - hors crise covid-19

On peut d'ores et déjà constater que les méthodes d'arbres et plus spécifiquement les forêts aléatoires sont en haut du classement des méthodes de prédiction. Concernant les méthodes d'imputation, la modélisation espace-état obtient le meilleur RMSFE de prédiction, tout en assurant une bonne prédiction du signe du ratio PIB de 83%. Autrement dit, la modélisation espace-état prédit correctement une croissance ou une récession à 83%. Ainsi, on préférera la méthode d'imputation IKNN qui permet d'obtenir une meilleure certitude du signe à 91%. De plus, elle conserve un RMSFE relativement proche de celui de la modélisation espace-état (0.00139 contre 0.00135). Enfin, on notera qu'une fois de plus, les méthodes de prévision de séries temporelles (hormis pour la modélisation espace-état) tendent à faire moins bien que les méthodes d'imputation (Figure 12).

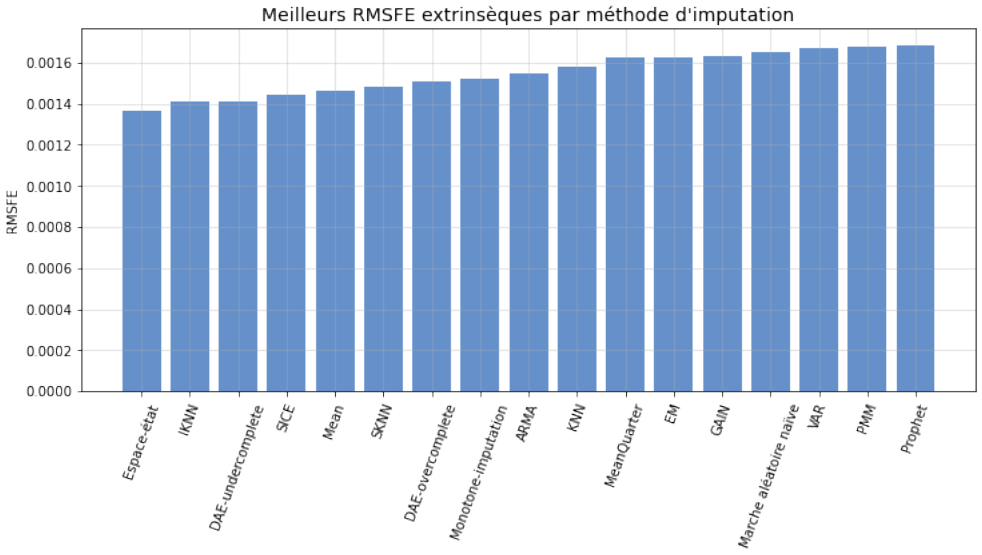

FIGURE 12 – Performances extrinsèques - hors crise covid-19

Evolution des performances entre les 2 étapes du cadre de modélisation
Le classement extrinsèque des méthodes d'imputation diffère nettement du classement intrinsèque. On peut, néanmoins, observer 3 points :

- Modèles à sous-performance constante : Les méthodes sous performantes lors de la tâche d'imputation tendent à persister dans leur imperfection lors de la tâche de prédiction (e.g., Marche aléatoire naïve, VAR, ARMA, Prophet, Gain) quel que soit le modèle de prédiction avec lequel elles sont combinées. La Figure 13 illustre ce point en représentant l'évolution de la performance au cours des 2 étapes du cadre de modélisation.
- Evolution croissante des performances de modèles récents : L'imputation monotone, SICE et les DAE obtiennent de bien meilleures performances extrinsèques qu'intrinsèques. Le rang de l'imputation monotone évolue de la 12ème à la 8ème place, celui de SICE de la 10ème à la 4ème place et celui du DAE sous-complet de la 9ème à la troisième place (Figure 13). Les performances de ces dernières méthodes permettent de souligner l'intérêt des méthodes d'apprentissage statistique et profond dans un contexte de prédiction macro-économique.
- Modèles à considérer : Parmi l'ensemble de ces méthodes, deux semblent se détacher du lot. Il s'agit de la méthode d'imputation IKNN et de la modélisation espace-état. Que ce soit dans leur performance intrinsèque ou extrinsèque, ces deux méthodes donnent des performances à chaque fois parmi les 5 meilleures.

4.2.2 Sensibilité des paramètres

Le paragraphe précédent permet d'ores et déjà de formaliser quelques conclusions. Néanmoins, les performances des modèles d'imputation peuvent être affectées par d'autres paramètres (méthode de sélection de variable, nombre de variables sélectionnées et nature de l'algorithme de prédiction). On décide donc de s'intéresser également aux comportements moyens de ces paramètres d'imputation (Figure 14).
Ainsi la sélection de variable basée sur la corrélation permet d'obtenir, en moyenne, de meilleures imputations de nos variables explicatives que la sélection de variable basée sur notre score. Aussi, plus on augmente le nombre de variables et plus l'erreur d'imputation tend à augmenter.

Cependant, le choix de la méthode de sélection de variables ne semble pas influer sur la précision

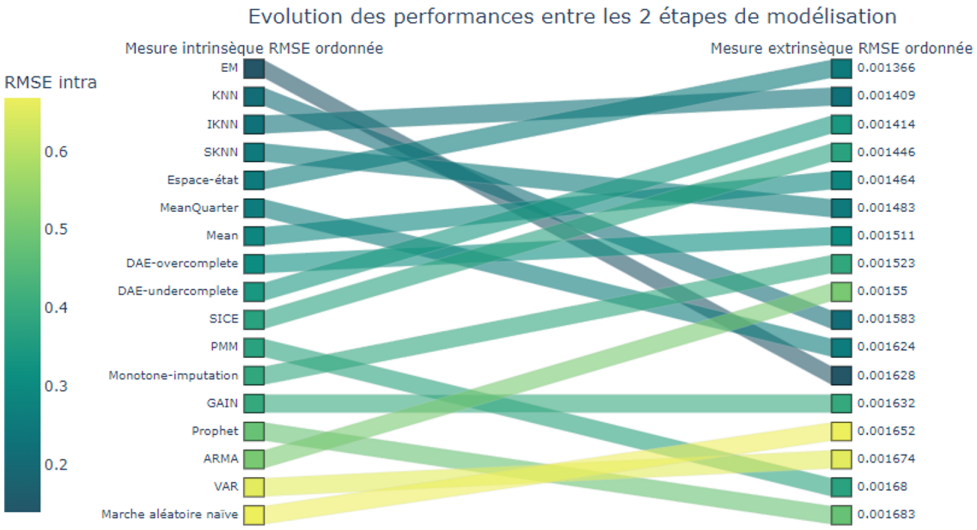

FIGURE 13 – Evolution de la performance intrinsèque à la performance extrinsèque

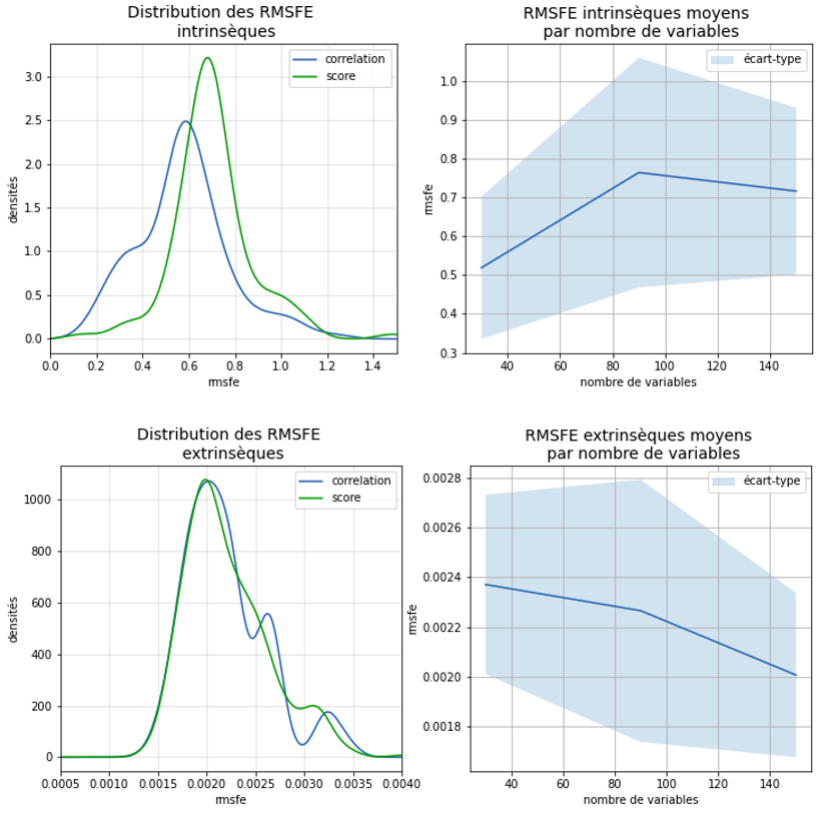

FIGURE 14 – Sensibilité des paramètres

de la prédiction du PIB. À l'inverse, le nombre de variables joue un rôle prépondérant. En effet, plus on l'augmente et plus l'erreur de prédiction tend à diminuer. Comme suggéré par la Figure 15 et la Table 3, les méthodes de Forêt Aléatoire et d'Arbres extrêmement aléatoires tendent à mieux performer que les méthodes boosting et les méthodes linéaires.

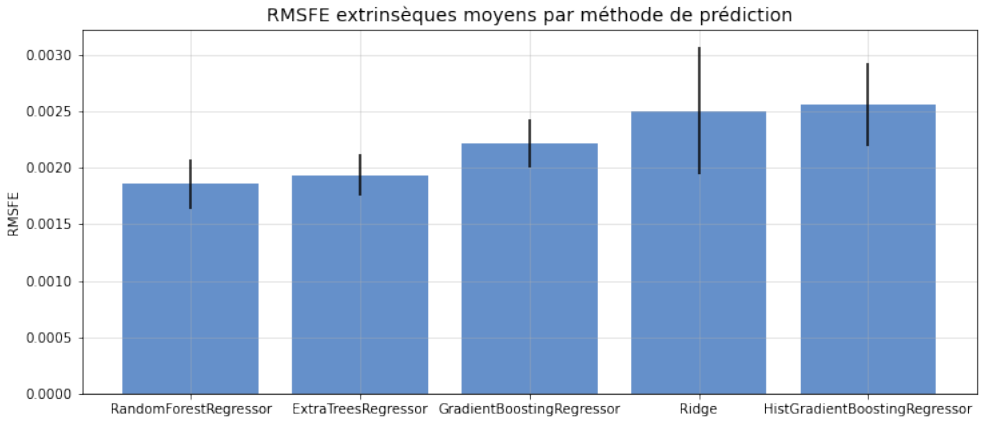

FIGURE 15 – Performances extrinsèques moyennes par famille de prédicteur

À noter que différentes valeurs d'hyperparamètres ont été testées pour chaque famille de modélisation.

4.2.3 Début de crise la covid-19

Contrairement aux évaluations précédentes, cette partie inclut les performances des modèles dans un cadre stressé de l'économie et plus particulièrement le contexte du début de crise covid-19 (Q1 2020 et Q2 2020).

Mesures intrinsèques
Comme indiqué dans la Figure 16, le meilleur RMSFE d'imputation avoisine maintenant les 0.7 alors qu'il avoisinait les 0.14 dans un cadre "business as usual". De manière générale, les RMSFE n'ont plus les mêmes ordres de grandeur et sont multipliés par un facteur de 10 en moyenne.

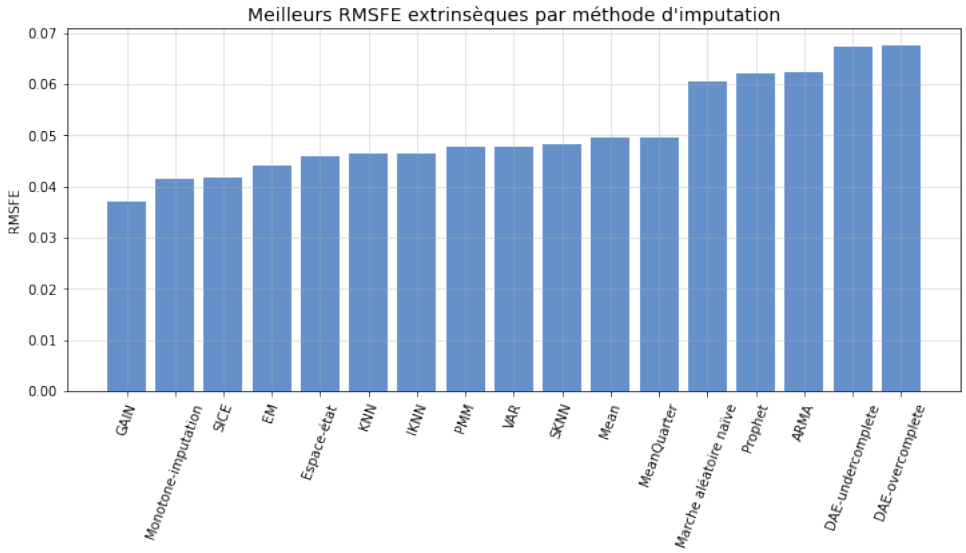

FIGURE 17 – Performances extrinsèques - Crise covid-19

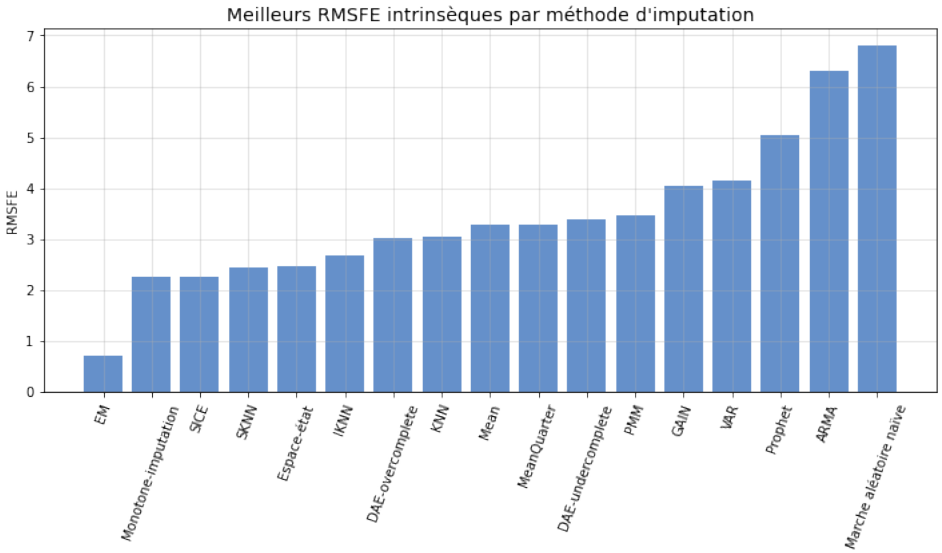

FIGURE 16 – Performances intrinsèques - Crise covid-19

Alors que l'algorithme d'EM conserve sa première place dans les deux contextes économiques, les méthodes basées sur les K-NN se retrouvent, quant à elles, déclassées au profit de méthodes plus complexes (i.e. l'imputation monotone et l'algorithme SICE). Les méthodes de prévisions, quant à elles, occupent toujours la fin du classement.

Mesures extrinsèques

À l'instar des performances intrinsèques, les performances extrinsèques sont grandement impactées par le changement de contexte économique induit par le début de la crise covid-19. En effet, les erreurs de prédictions (RMSFE) n'ont plus du tout les mêmes ordres de grandeur. On observe effectivement un facteur multiplicatif de 10 en moyenne entre les performances des deux contextes économiques.

On assiste également à un changement notable dans le classement des méthodes d'imputation (Figure 17). Seuls les algorithmes SICE et espace-état restent dans le groupe des 5 modèles ayant le plus performés, quel que soit le contexte économique. Les rangs de l'imputation monotone et

de l'imputation GAIN évoluent respectivement de la 8ème place à la 2ème et de la 13ème à la première place. À l'inverse, les DAE se voient déclassés et chutent aux deux dernières places du classement. On retrouve, juste devant les DAE, les méthodes de prévision temporelles (Marche aléatoire Naïve, ARMA, VAR, Prophet) qui, de leurs côtés, continuent de sous-performer.

À noter qu'en période de début de crise, les méthodes permettant de renvoyer les meilleurs résultats se basent sur un ensemble de variables explicatives plus restreint. On soulignera le bouleversement dans le classement des méthodes de prédiction, même si ce point n'est pas un point sur lequel on souhaite s'attarder pour la présente étude. Ainsi, la régression pénalisée Ridge tend, en moyenne, à surpasser l'ensemble des autres méthodes dans le contexte économique stressé durant la période du covid-19.

4.2.4 Prééminence du KNN itératif et de l'imputation monotone

On note une diminution notable des performances entre la période "business as usual" (2017-2019) et la période de début de crise covid-19. Ce constat met en exergue la difficulté de prédire le début de crise covid-19 en se basant seulement sur les données actuelles. De plus, le classement des méthodes selon leurs performances intrinsèques et extrinsèques est très différent d'une période "business as usual" à une période de début de crise covid-19. Aussi, la méthode d'imputation IKNN couplée à l'algorithme de prédiction de forêt aléatoire semble être la modélisation la plus pertinente dans un cadre "business as usual". Cependant, dans un contexte de début de crise comme celui engendré par le début de crise covid-19, la méthode d'imputation monotone (avec pour estimateur interne une régression bayésienne Ridge) couplée à l'algorithme de prédiction Ridge devient un meilleur candidat.

La Figure 18 détaille l'évolution moyenne des RMSFE des deux modélisations susmentionnées, au cours des différents trimestres de la période de test. Autrement dit, ces graphiques représentent l'évolution de l'erreur de prédiction en fonction de l'information macro-économique et financière qui s'enrichit au fur et à mesure du trimestre. Ainsi, les deux graphiques ci-dessous permettent de visualiser des différences notables entre ces deux méthodes.

Dans le contexte économique "business as usual", la modélisation basée sur l'imputation par IKKN voit son erreur diminuer au fur et à mesure du trimestre. A l'inverse, dans ce même contexte économique, l'erreur de la méthode basée sur l'imputation monotone croît au cours du temps. De plus, son erreur globale devient bien supérieure à celle de la méthode précédente.

Dans un cadre de début de crise (en incluant le début de crise covid-19), on remarque que les erreurs moyennes de ces deux modélisations tendent à diminuer au fur et à mesure du trimestre. En revanche, on dénote un écart significatif des erreurs entre ces deux modélisations. En effet, la modélisation basée sur l'imputation monotone surpasse considérablement son homologue adoptant une imputation IKNN.

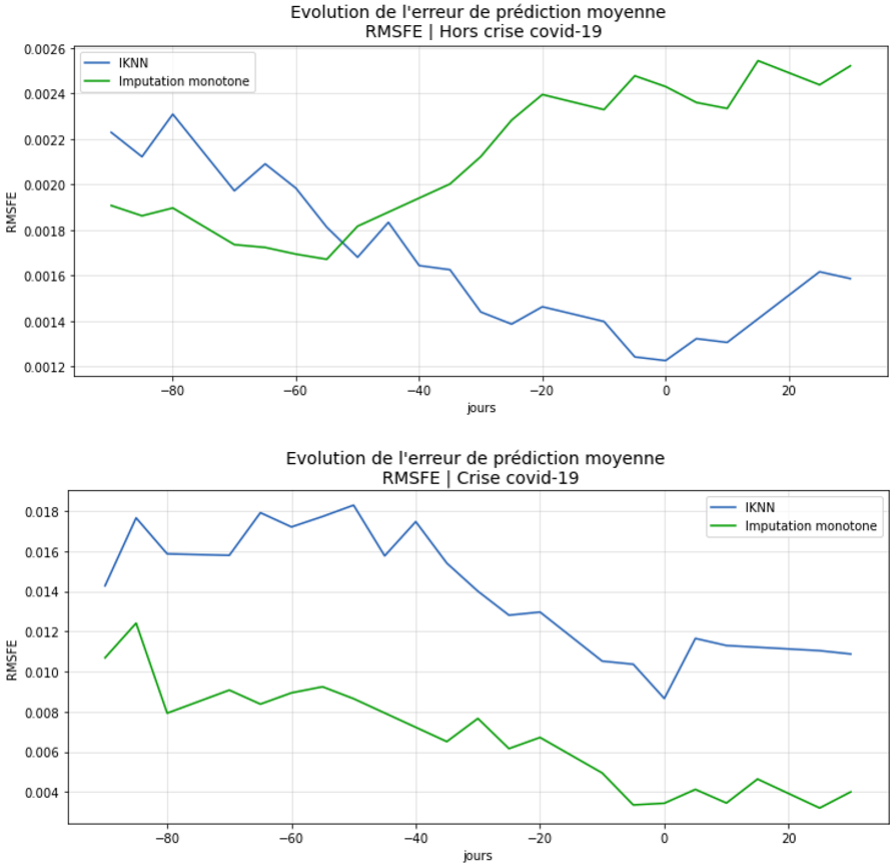

FIGURE 18 – Evolution de l'erreur moyenne de prédiction au fur et à mesure du trimestre

5 Conclusion

Dans ce papier, nous utilisons un cadre de modélisation en deux étapes afin de prédire le PIB français par un procédé "nowcasting". On entend par "nowcasting", l'aptitude d'un modèle à fournir des prédictions immédiates. La modélisation, décrite dans le présent papier, permet d'obtenir des prévisions du PIB français pour un certain trimestre, à tout moment entre le début du trimestre et la première publication du PIB français i.e. 30 jours après la fin du trimestre.

L'étude se concentre essentiellement sur la première étape de cette modélisation, appelée étape d'"imputation". Cette étape consiste à estimer les variables explicatives qui n'auront pas encore été publiées, et qui auront, au préalable, été sélectionnées pour leur pouvoir prédictif. Autrement dit, l'idée est de reproduire un comportement humain d'expert macro-économiste. Ce dernier actualise ses prédictions en fonction des dernières informations disponibles jugées auparavant pertinentes dans leur contribution au PIB. Ainsi, dans le cas où les dernières informations, jugées impactantes pour le PIB, ne seraient pas disponibles, elles sont estimées sur la base d'autres informations, qui elles, auront déjà été publiées. La seconde étape de la modélisation, qui consiste à prédire le PIB français, s'appuie sur l'information publiée et les estimations des variables non disponibles mais estimées grâce à la première étape.

Ainsi, l'étude actuelle propose une adaptation simple : le modèle reste le même quel que soit l'information disponible. De plus, cette adaptation permet de mettre à jour quasi-instantanément la prévision en fonction du flux d'information. La méthodologie présentée permet alors de traiter une grande quantité de données présentées sous forme de flux réalistes et hétérogènes. On entend par flux réaliste, un flux temporel de données indexé selon un calendrier de publication, impliquant de nombreuses révisions mais également des changements de nomenclature. Ainsi, cette construction de flux permet de respecter ce qu'observe le prévisionniste en temps réel. L'hétérogénéité des données, quant à elle, s'explique par leur différente nature ainsi que par le mixte de leur fréquence. En effet, les variables peuvent être des variables macro-économiques "hard", des variables macro-économiques "soft", ou même des variables financières. Leur fréquence peut être trimestrielle, mensuelle ou encore quotidienne.

Afin d'évaluer l'apport des méthodes récentes sur cette modélisation en deux étapes, l'article établit une cartographie récente et d'autant plus exhaustive des méthodes d'estimation contribuant à la première étape, c'est-à-dire, les méthodes d'estimation des variables non publiées. Cette cartographie reprend les méthodes de prévisions temporelles classiques de ce cadre de modélisation (Miller et al., 1996 [29] ; Zheng et al., 2006 [46] et Bouwman et al., 2011 [13]). Elle est alors complétée par différentes méthodes qui s'inspirent, entre autres, de recherches liées à l'apprentissage statistique (ou "machine learning"), y compris l'apprentissage profond (ou "deep learning"), et de recueils relatifs à l'imputation de données manquantes. Bien que répandues dans d'autres domaines comme le médical, les techniques d'imputation de données manquantes sont plus rarement appliquées aux domaines économiques et financiers.

L'évaluation en pseudo temps réel de ces techniques d'estimation de variables non encore publiées, met en exergue la suprématie de techniques innovantes dans ce cadre de modélisation en deux étapes. En effet, l'espérance-maximisation tire son épingle du jeu sur la reconstruction des variables explicatives à la fois dans un cadre économique classique que stressé. Néanmoins, combinée aux modèles finaux de prédiction du PIB, cette méthode affiche relativement moins sa supériorité, au profit d'une méthode dérivée des K-NN, le K-NN itératif.
Les performances des modèles restent néanmoins fragiles lors de la période de crise du covid-19. Dans les circonstances de crise, l'estimation des variables non publiées par l'imputation monotone permet d'amoindrir les imperfections de prévision. Dans ce contexte, les GAIN, méthode d'im-

putation basée sur la théorie des réseaux antagonistes génératifs, permettent toutefois d'obtenir les meilleures performances de prédiction finale de PIB. Cependant, leurs performances relatives à l'imputation des variables non encore publiées en font un moins bon candidat que l'imputation monotone.

Dans notre cadre de prédiction du PIB français, les méthodes d'imputation introduites ici permettent d'obtenir de meilleurs résultats que les méthodes de prévisions temporelles (Marche aléatoire naïve, AR, VAR) suggérées pour la première étape de cette modélisation par Miller et al (1996) [29], Zheng et al. (2006) [46] et Bouwman et al. (2011)[13].

Diverses voies d'amélioration de ces résultats pourraient être explorées. Une des prochaines orientations de cette étude porte sur l'optimisation du nombre de variables explicatives sélectionnées. En effet, les résultats tendent à laisser penser qu'une augmentation du nombre de variables explicatives pourraient continuer d'augmenter la performance du modèle. L'apport de données alternatives est également une des pistes d'amélioration de cette étude, comme suggéré par Ferrara et al. (2019) [21]. Une autre orientation possible consisterait à estimer les variables non publiées en surpondérant davantage les valeurs des variables déjà publiées de la même famille économique. Ainsi, par exemple, l'estimation des variables de construction serait obtenue en surpondérant les valeurs des enquêtes de construction déjà disponibles. La dernière piste consisterait à étudier des modèles qui imputent et régressent simultanément. On pense notamment au modèle BRITS (Cao et al., 2018 [15]) qui, via à un réseau de neurones bidirectionnel, impute des séries temporelles multivariées sans imposer de conditions sur les dynamiques.

Références

[1] Banbura, M., Giannone, D., & Reichlin, L. (2010). Nowcasting.

[2] Banbura, M., Giannone, D., & Reichlin, L. (2010). Large Bayesian vector auto regressions. Journal of applied Econometrics, 25(1), 71-92.

[3] Barbaglia, L., Consoli, S., & Manzan, S. (2021). Forecasting with economic news. Available at SSRN 3698121.

[4] Bertsimas, D., Pawlowski, C., & Zhuo, Y. D. (2017). From predictive methods to missing data imputation : an optimization approach. The Journal of Machine Learning Research, 18(1), 7133-7171.

[5] Bessec, M., & Doz, C. (2012). Prévision à court terme de la croissance du PIB français à l'aide de modèles à facteurs dynamiques. Economie prevision, (1), 1-30.

[6] Blanchet, M. & Coueffe M. (2020). Improved GDP nowcasting using large datasets. Trésor-Economics, No. 254

[7] Brás, L. P., & Menezes, J. C. (2007). Improving cluster-based missing value estimation of DNA microarray data. Biomolecular engineering, 24(2), 273-282.

[8] Breiman, L. (1996). Bagging predictors. Machine learning, 24(2), 123-140.

[9] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

[10] Brownlee, J. (2017). Introduction to time series forecasting with python : how to prepare data and develop models to predict the future. Machine Learning Mastery.

[11] Bojer, C. S., & Meldgaard, J. P. (2021). Kaggle forecasting competitions : An overlooked learning opportunity. International Journal of Forecasting, 37(2), 587-603.

[12] Bok, B., Caratelli, D., Giannone, D., Sbordone, A. M., & Tambalotti, A. (2018). Macroeconomic nowcasting and forecasting with big da-ta. Annual Review of Economics, 10, 615-643.

[13] Bouwman, K. E., & Jacobs, J. P. (2011). Forecasting with real-time macroeconomic data : the ragged-edge problem and revisions. Journal of Macroeconomics, 33(4), 784-792.

[14] Buuren, S. V., & Groothuis-Oudshoorn, K. (2010). mice : Multivariate imputation by chained equations in R.Journal of statistical software, 1-68.

[15] Cao, W., Wang, D., Li, J., Zhou, H., Li, L., & Li, Y. (2018). Brits : Bidirectional recurrent imputation for time series. Advances in neural information processing systems, 31.

[16] Doz, C., Giannone, D., & Reichlin, L. (2011). A two-step estimator for large approximate dynamic factor models based on Kalman filtering. Journal of Econometrics, 164(1), 188-205.

[17] Doz, C., Giannone, D., & Reichlin, L. (2012). A quasi-maximum likelihood approach for large, approximate dynamic factor models. Review of economics and statistics, 94(4), 1014-1024.

[18] Dernoncourt, D. (2014). Stabilité de la sélection de variables sur des données haute dimension : une application à l'expression génique (Doctoral dissertation, Paris 6).

[19] Cerqueira, V., Torgo, L., Smailovic, J., & Mozetic, I. (2017, October). A comparative study of performance estimation methods for time series forecasting. In 2017 IEEE international conference on data science and advanced analytics (DSAA) (pp. 529-538). IEEE.

[20] Ferrara, L., Marsilli, C., & Ortega, J. P. (2014). Forecasting growth during the Great Recession : is financial volatility the missing ingredient ?. Economic Modelling, 36, 44-50.

[21] Ferrara, L., & Simoni, A. (2020). When are Google data useful to nowcast GDP ? An approach via pre-selection and shrinkage. arXiv preprint arXiv :2007.00273.

[22] Friedman, J. H. (2001). Greedy function approximation : a gradient boosting machine. Annals of statistics, 1189-1232.

[23] Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. Machine learning, 63(1), 3-42.

[24] Gondara, L., & Wang, K. (2018, June). Mida : Multiple imputation using denoising autoencoders. In Pacific-Asia conference on knowledge discovery and data mining (pp. 260-272). Springer, Cham.

[25] Khan, S. I., & Hoque, A. S. M. L. (2020). SICE : an improved missing data imputation tech-nique. Journal of big data, 7(1), 1-21.

[26] Kim, K. Y., Kim, B. J., & Yi, G. S. (2004). Reuse of imputed data in microarray analysis increases imputation efficiency. BMC bioinformatics, 5(1), 1-9.

[27] Marsilli, C. (2014). Mixed-frequency modeling and economic forecasting (Doctoral dissertation, Besançon).

[28] Martinez M. (2017), Euro area : caution needed when looking at survey data and growth. Société Générale Cross Asset Research.

[29] Miller, P. J., & Chin, D. M. (1996). Using monthly data to improve quarterly model forecasts. Federal Reserve Bank of Minneapolis Quarterly Review, 20, 16-28.

[30] Pereira, R. C., Santos, M. S., Rodrigues, P. P., & Abreu, P. H. (2020). Reviewing autoencoders for missing data imputation : Technical trends, applications and outcomes. Journal of Artificial Intelligence Research, 69, 1255-1285.

[31] Rousset, C., & Papp, A. (2018). Gets Modelling ou Lasso ? Les différentes méthodes de sélection de variables avec des séries temporelles. L'exemple des étalonnages à l'aide des conjecture.

[32] Rousseau, R. (1975). Pourquoi change-t-on de nomenclature ?. Economie et statistique, 70(1), 63-67.

[33] Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., & Chen, X. (2016). Improved techniques for training gans. Advances in neural information processing systems, 29.

[34] Schafer, J. L., & Graham, J. W. (2002). Missing data : our view of the state of the art. Psychological methods, 7(2), 147.

[35] Schmitt, P., Mandel, J., & Guedj, M. (2015). A comparison of six methods for missing data imputation. j biomet biostat 6

[36] Schubert, M., & Schanze, T. (2019, July). Estimation of Sparse VAR Models with Artificial Neural Networks for the Analysis of Biosignals. In 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 4623-4627). IEEE.

[37] Stekhoven, D. J., & Bühlmann, P. (2012). MissForest non-parametric missing value imputation for mixed-type data. Bioinformatics, 28(1), 112-118.

[38] Stuart, E. A., Azur, M., Frangakis, C., & Leaf, P. (2009). Multiple imputation with large data sets : a case study of the Children's Mental Health Initiative. American journal of epidemiology, 169(9), 1133-1139.

[39] Taylor, S. J., & Letham, B. (2018). Forecasting at scale. The American Statistician, 72(1), 37-45.

[40] Tolvi, J. (2001). Outliers in eleven finnish macroeconomic time series. Finnish Economic Papers, 14(1), 14-32.

[41] Tofallis, C. (2014). Add or multiply ? A tutorial on ranking and choosing with multiple criteria. INFORMS Transactions on education, 14(3), 109-119.

[42] Van Buuren, S. (2018). Flexible imputation of missing data. CRC press.

[43] Wang, Z., Akande, O., Poulos, J., & Li, F. (2021). Are deep learning models superior for missing data imputation in large surveys ? Evidence from an empirical comparison. arXiv preprint arXiv :2103.09316.

[44] Yoon, J., Zame, W. R., & van der Schaar, M. (2017). Multi-directional recurrent neural networks : A novel method for estimating missing data. In Time Series Workshop in International Conference on Machine Learning.

[45] Yoon, J., Jordon, J., & Schaar, M. (2018, July). Gain : Missing data imputation using generative adversarial nets. In International conference on machine learning (pp. 5689-5698). PMLR.

[46] Zheng, I. Y., & Rossiter, J. (2006). Using monthly indicators to predict quarterly GDP.

[47] Zivot, E., & Wang, J. (2006). Unit root tests. Modeling Financial Time Series with s-plus®, 111-139.

A Annexes

A.1 Données et traitements

Catégorie	Type de données	Définition
Enquêtes auprès des entreprises	Soft	Il s'agit d'enquêtes qui permettent de refléter la perception qu'ont les professionnels des secteurs concernés (souvent les chefs d'entreprise) sur la situation conjoncturelle et sur la tendance économique (si on est en phase d'expansion, de contraction ou de stabilité) dans différents secteurs (manufacturiers, services, construction). Les enquêtes peuvent, par exemple, donner des indications sur les prix (s'ils sont à la hausse ou à la baisse), le recrutement dans les secteurs, les délais de livraisons, le taux d'utilisation des capacités de production dans l'industrie. Les données sont obtenues en interrogeant les professionnels sur la situation économique.
Comptes nationaux	Hard	Il s'agit de données quantitatives représentant schématiquement l'activité économique d'un pays, de sa comptabilité nationale. Ils mesurent des flux monétaires représentatifs de l'économie d'un pays pendant une période donnée. La comptabilité nationale est une représentation globale, détaillée et chiffrée de l'activité économique d'un pays dans un cadre comptable équilibré. Elle décrit les ressources et les emplois à un niveau fin pour chaque type de bien ou de service.
Commerce	Hard	Les données « Trade » sont des données portant sur le commerce international (import/export). Cette catégorie peut également contenir des données calculées selon les normes de comptabilité nationale.
Marché du travail	Hard	Comme le nom l'indique, il s'agit de données relatives au marché du travail. Le Bureau International du Travail les définit comme « des statistiques officielles se concentrant sur les activités productives des travailleurs et éventuellement les déficiences du marché du travail, sous de nombreux angles et couvrant de nombreuses dimensions ».
Production industrielle	Hard	L'OCDE définit les données portant sur la production industrielle de la manière suivante : des données faisant « référence à la production des établissements industriels et couvrent des secteurs tels que l'exploitation minière, la fabrication, l'électricité, le gaz et la vapeur et la climatisation ».
Consommation	Hard	Ce sont des données qui quantifient la consommation des ménages par type de biens (e.g. voitures, textile).
Enquêtes auprès des consommateurs	Soft	Tout comme les enquêtes auprès des entreprises, ces données sont issues d'enquêtes, menées cette fois-ci auprès des ménages afin de connaître leurs perspectives concernant l'activité économique. Certaines interrogations d'enquête portent sur la crainte d'être au chômage, sur leur perception du niveau de vie etc.
Construction	Hard	Il s'agit d'indicateurs qui permettent de quantifier les activités de construction. A noter que le PMI construction est classé dans la catégorie « Enquêtes auprès des entreprises ».
Sectoriel	Hard	Ce sont des données qui permettent de mesurer l'activité dans différents secteurs de l'économie Par exemple, certains indicateurs du jeu de données portent sur le commerce de détail et les services.
Revenus et investissements	Hard	C'est une catégorie de données qui est produite selon les normes de comptabilité nationale. Ces données sont relatives au salaire et à l'investissement des ménages, des entreprises ou de l'Etat.
Masse monétaire	Hard	La masse monétaire est l'ensemble des devises et autres instruments liquides de l'économie d'un pays à la date mesurée. La masse monétaire comprend à peu près à la fois les espèces et les dépôts qui peuvent être utilisés presque aussi facilement que les espèces. Ainsi, les indicateurs relatifs à la masse monétaire sont des indicateurs qui vont permettre de quantifier la masse monétaire.
IPC (indice des prix à la consommation)	Hard	L'IPC est une mesure de l'inflation qui repose sur le calcul de l'évolution du prix d'un panier de biens (représentatif de la consommation des ménages) pris entre deux périodes. Ainsi, les données IPC permettent de donner chaque mois la variation du prix de ce panier de mois. Par ailleurs, la composition de ce panier de référence change chaque année.
Productivité	Hard	INSEE définit la productivité de la manière suivante : « En économie, la productivité est définie comme le rapport, en volume, entre une production et les ressources mises en œuvre pour l'obtenir. La production désigne les biens et/ou les services produits. Les ressources mises en œuvre, dénommées aussi facteurs de production, désignent le travail, le capital technique (installations, machines, outillages...), les capitaux engagés, les consommations intermédiaires (matières premières, énergie, transport...), ainsi que des facteurs moins faciles à appréhender bien qu'extrêmement importants, tel le savoir-faire accumulé.
IPP (indice des prix à la production)	Hard	Selon l'INSEE, « les indices de prix à la production (IPP) dans l'industrie pour le marché français mesurent l'évolution des prix de transaction, hors TVA, de biens issus des activités de l'industrie et vendus sur le marché français. Les indices de prix à la production dans l'industrie pour les marchés extérieurs traduisent l'évolution des prix de transaction (convertis en euros, donc incluant les effets de change), FAB, de biens issus des activités de l'industrie française et vendus sur les marchés extérieurs. » Le prix FAB, Franco à bord, est le prix d'un bien à la frontière du pays exportateur ou prix d'un service fourni à un non-résident. Il comprend la valeur des biens ou des services au prix de base, des services de transport et de distribution jusqu'à la frontière, les impôts moins les subventions.
Comptes relatifs au gouvernement	Hard	Ces données quantitatives des comptes relatifs au gouvernement.

A noter que les définitions des catégories ci-dessus s'appliquent au jeu de données de l'étude, et peuvent de ce fait diverger quelque peu des définitions traditionnelles. Ainsi, il n'y a pas de données de crédit dans la catégorie 'Masse monétaire', il n'y a pas non plus de données de services dans la catégorie 'Consommation'.

FIGURE 19 – Catégories de variables et définitions

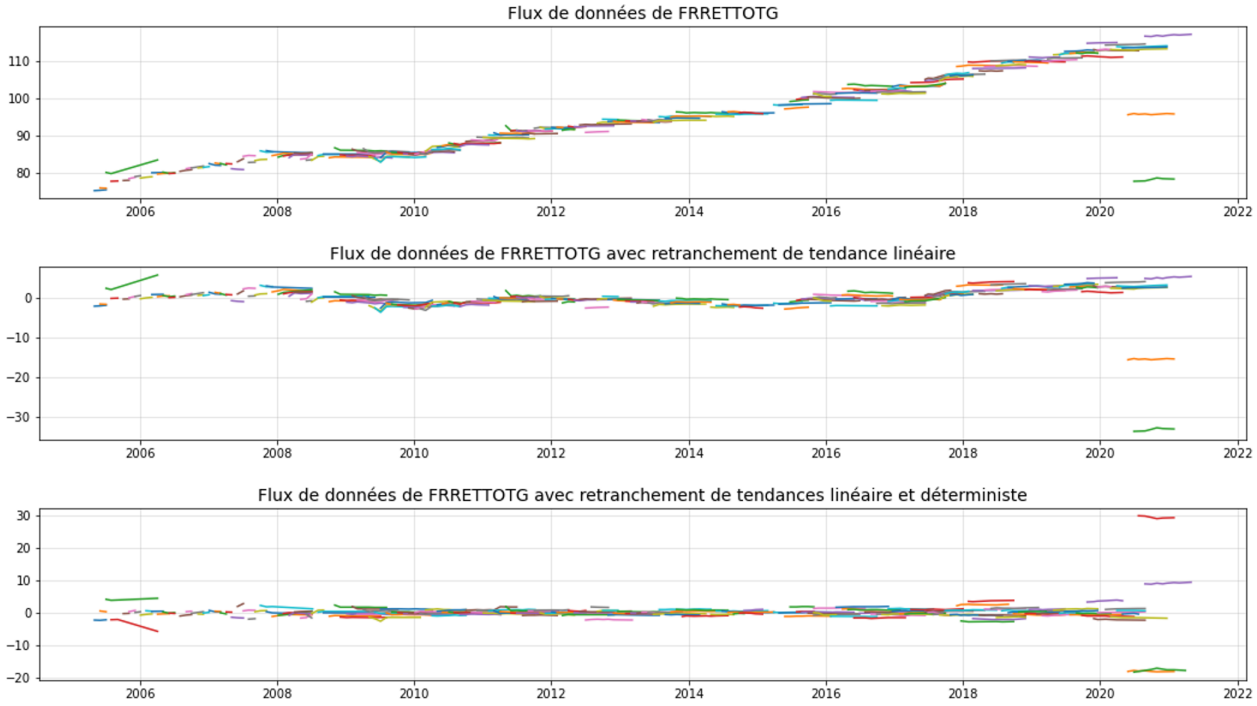

FIGURE 20 – Application à FRRETTOTG du processus décrit dans l'Algorithme 1 pour la suppression de tendances.

Trimestre	date de disponibilité	FRPMGINDQ_0	FRSURTETQ_1	FRPMMEXQ_1	FRSURTDMQ_0	FRSURPRPQ_2	FRSUROTIQ_2	FRSURTWPQ_2	FRPMANDQ_0	FRPMIS_Q_1	FRPMIEXCQ_1_lag_1
Q3 2019	2019-10-23	-0.2398281	-0.2862869	-0.14208073	0.194745065	0.054068937	0.90582275	-0.03057576	-0.19312888	-0.18289273	-0.45353877
	2019-10-25	-0.2398281	-0.2862869	-0.14208073	0.194745065	0.054068937	0.90582275	-0.03057576	-0.19312888	-0.18289273	-0.45353877
	2019-10-29	-0.2398281	-0.2862869	-0.14208073	0.194745065	0.054068937	0.90582275	-0.03057576	-0.19312888	-0.18289273	-0.45353877
	2019-10-30	-0.2398281	-0.2862869	-0.14208073	0.194745065	0.054068937	0.90582275	-0.03057576	-0.19312888	-0.18289273	-0.45353877
Q4 2019	2019-10-01										-0.86395244
	2019-10-04										-0.86395244
	2019-10-07										-0.86395244
	2019-10-08										-0.86395244
	2019-10-09										-0.86395244
	2019-10-15										-0.86395244
	2019-10-22										-0.86395244
	2019-10-23								-0.19412384		-0.86395244
	2019-10-24				0.065493102				-0.19412384		-0.86395244
	2019-10-25				0.065493102				-0.19412384		-0.86395244
	2019-10-29				0.065493102				-0.19412384		-0.86395244
	2019-10-30				0.065493102				-0.19412384		-0.86395244
	2019-10-31				0.065493102				-0.19412384		-0.86395244
	2019-11-01	-0.73300539			0.065493102				-0.19412384		-0.86395244
	2019-11-04	-0.73300539			0.065493102				-0.19412384		-0.86395244
	2019-11-05	-0.73300539			0.065493102				-0.19412384		-0.86395244
	2019-11-07	-0.73300539			0.065493102				-0.19412384		-0.86395244
	2019-11-08	-0.73300539			0.065493102				-0.19412384		-0.86395244
	2019-11-12	-0.73300539			0.065493102				-0.19412384		-0.86395244
	2019-11-13	-0.73300539			0.065493102				-0.19412384		-0.86395244
	2019-11-14	-0.73300539			0.065493102				-0.19412384		-0.86395244
	2019-11-20	-0.73300539	-0.01447138	-0.08857963	0.065493102				-0.19412384	-0.39649764	-0.86395244
	2019-11-21	-0.73300539	-0.01447138	-0.08857963	0.065493102				-0.19412384	-0.39649764	-0.86395244
	2019-11-22	-0.73300539	-0.01447138	-0.08857963	0.065493102				-0.19412384	-0.39649764	-0.86395244
	2019-11-26	-0.73300539	-0.01447138	-0.08857963	0.065493102				-0.19412384	-0.39649764	-0.86395244
	2019-11-27	-0.73300539	-0.01447138	-0.08857963	0.065493102				-0.19412384	-0.39649764	-0.86395244
	2019-11-28	-0.73300539	-0.01447138	-0.08857963	0.065493102				-0.19412384	-0.39649764	-0.86395244
	2019-11-29	-0.73300539	-0.01447138	-0.08857963	0.065493102				-0.19412384	-0.39649764	-0.86395244
	2019-12-02	-0.73300539	-0.01447138	-0.08857963	0.065493102				-0.19412384	-0.39649764	-0.86395244
	2019-12-03	-0.73300539	-0.01447138	-0.08857963	0.065493102				-0.19412384	-0.39649764	-0.86395244
	2019-12-04	-0.73300539	-0.01447138	-0.08857963	0.065493102				-0.19412384	-0.39649764	-0.86395244
	2019-12-05	-0.73300539	-0.01447138	-0.08857963	0.065493102				-0.19412384	-0.39649764	-0.86395244
	2019-12-06	-0.73300539	-0.01447138	-0.08857963	0.065493102				-0.19412384	-0.39649764	-0.86395244
	2019-12-10	-0.73300539	-0.01447138	-0.08857963	0.065493102				-0.19412384	-0.39649764	-0.86395244
	2019-12-11	-0.73300539	-0.01447138	-0.08857963	0.065493102				-0.19412384	-0.39649764	-0.86395244
	2019-12-12	-0.73300539	-0.01447138	-0.08857963	0.065493102				-0.19412384	-0.39649764	-0.86395244
	2019-12-19	-0.73300539	-0.01447138	-0.08857963	0.065493102				-0.19412384	-0.39649764	-0.86395244
	2019-12-20	-0.73300539	-0.01447138	-0.08857963	0.065493102				-0.19412384	-0.39649764	-0.86395244
	2019-12-23	-0.73300539	-0.01447138	-0.08857963	0.065493102	-0.72426521		0.23401242	-0.19412384	-0.39649764	-0.86395244
	2019-12-24	-0.73300539	-0.01447138	-0.08857963	0.065493102	-0.72426521		0.23401242	-0.19412384	-0.39649764	-0.86395244
	2019-12-25	-0.73300539	-0.01447138	-0.08857963	0.065493102	-0.72426521		0.23401242	-0.19412384	-0.39649764	-0.86395244
	2019-12-27	-0.73300539	-0.01447138	-0.08857963	0.065493102	-0.72426521		0.23401242	-0.19412384	-0.39649764	-0.86395244
	2019-12-30	-0.73300539	-0.01447138	-0.08857963	0.065493102	-0.72426521		0.23401242	-0.19412384	-0.39649764	-0.86395244

FIGURE 21 – Agrégation trimestrielle des données

A.2 Modélisation

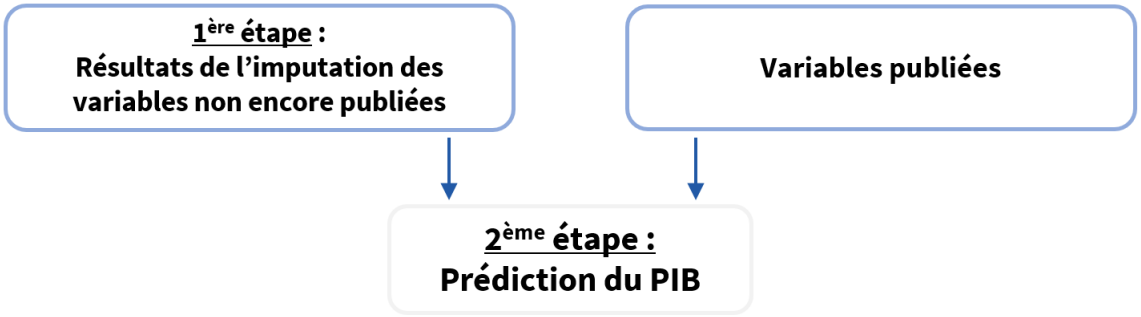

FIGURE 22 – Modélisation en deux temps

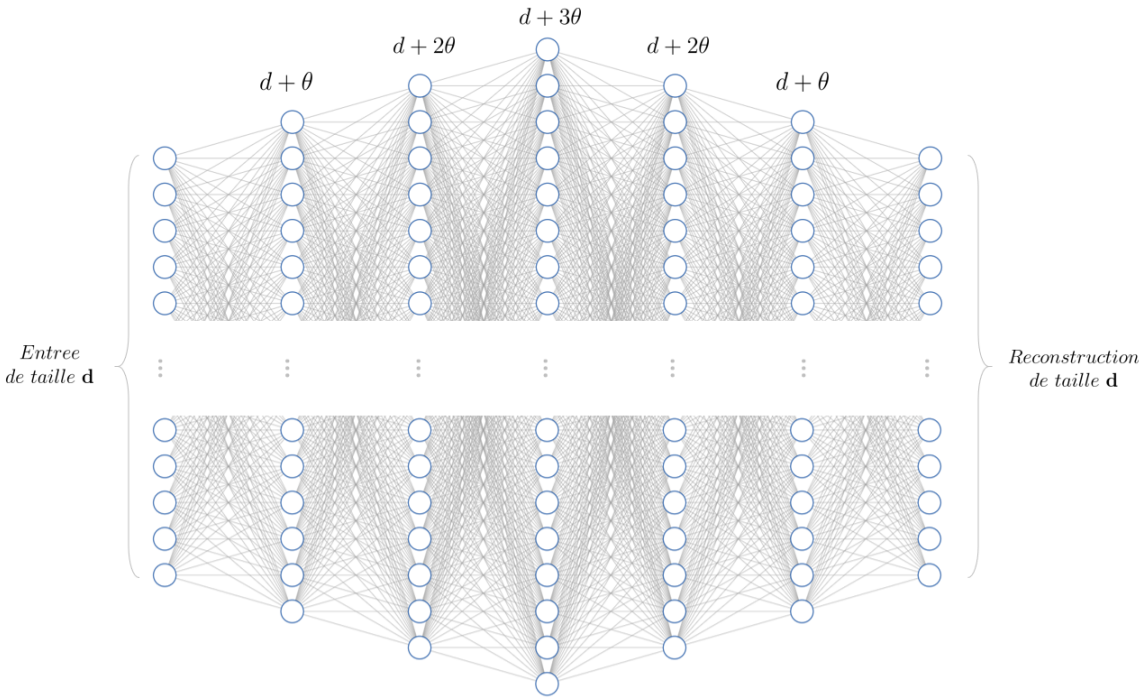

FIGURE 23 – Schéma d'un DAE sur-complet

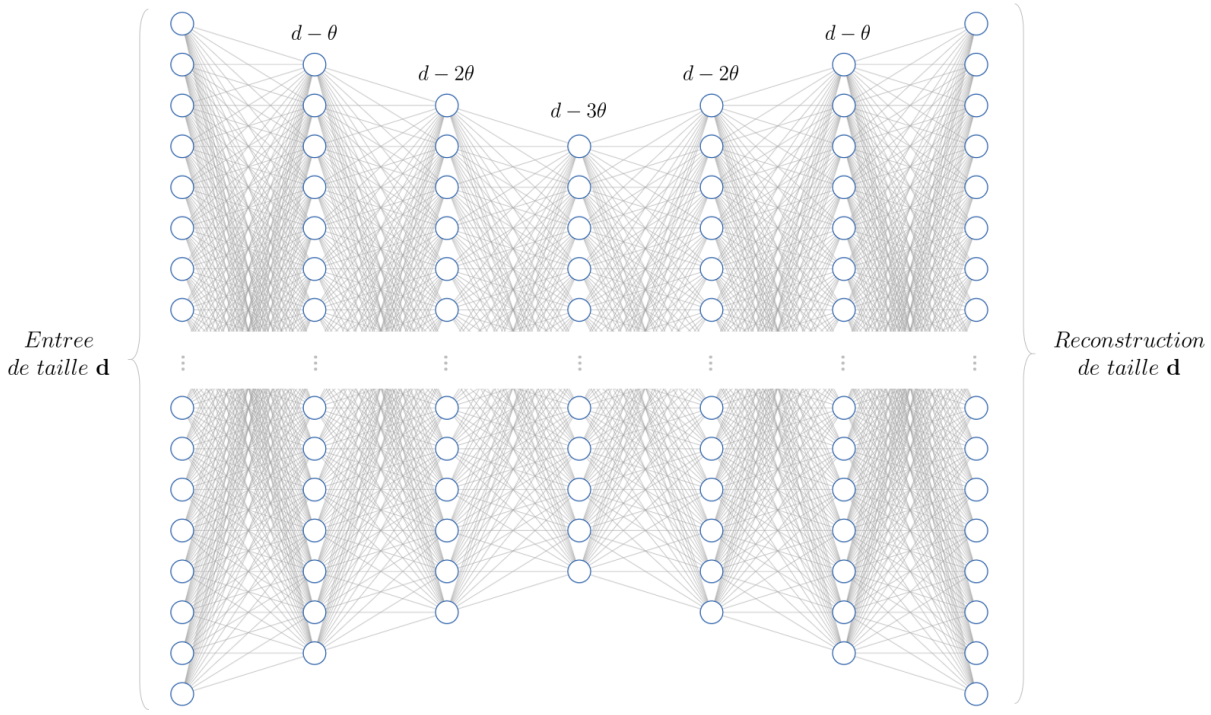

FIGURE 24 – Schéma d'un DAE sous-complet