

Master 1 Econométrie et Statistique,
parcours Econométrie Appliquée

Évolution et Prévisions mensuelles du prix des denrées alimentaires de 1999 à 2022

Le Roux Noa

Sous la direction de: Olivier Darne

June 2, 2023

Remerciements

Je tiens tout d'abord à remercier mon tuteur de mémoire Monsieur Olivier Darne, pour ses précieux conseils et son soutien constant tout au long de la préparation de cette étude.

Je souhaite également adresser mes remerciements à l'ensemble des enseignants du master ECAP, qui m'ont fourni les outils nécessaires à la réalisation de ce travail, ainsi que pour leur contribution aux travaux futurs.

Je tiens à exprimer ma sincère reconnaissance envers mes camarades de Master, avec lesquels nous nous sommes mutuellement soutenus tout au long de cette année.

Résumé

Dans le contexte d'instabilité croissante des prix des denrées alimentaires, cette étude vise à élaborer divers modèles de prévision, tant économétriques que basés sur le Machine Learning, afin de faire des prévisions sur le prix des denrées alimentaires (*Fao Food Price Index*). Ces modèles seront établis à partir de six variables explicatives, recourant à des données mensuelles couvrant la période allant de janvier 1999 à décembre 2022. Les prédictions sont effectuées en allouant 80% des données à l'entraînement et 20% aux tests, en utilisant un échantillonnage non aléatoire. Les conclusions de notre analyse indiquent que le modèle économétrique ARX Gets surpasse les autres, bien que les modèles XGB Boost et LSTM ne soient pas loin derrière en termes de performances.

Asbtract

In the context of rising instability in food prices, this study aims at developing various predictive models, both econometric and based on Machine Learning, to forecast the price of food commodities (*Fao Food Price Index*). These models will be established using six explanatory variables, relying on monthly data spanning the period from January 1999 to December 2022. The predictions are made by allocating 80% of the data for training and 20% for testing, using a non-random sampling. The conclusions from our analysis indicate that the ARX Gets econometric model outperforms the others, although the XGB Boost and LSTM models are not far behind in terms of performance.

*Time-Series, Forecasting, Machine-Learning, Food Price, R, Python, Finance,
Environment*

Sommaire

1	Introduction	6
2	Présentation des variables	10
3	Analyse exploratoire	28
4	Sélection des variables	42
5	Prévision & évaluation	47
6	Conclusion & Discussion	86
7	Annexe	99
8	Source des données	118

Liste des sigles

AIC : critère d'information d'Akaike

AO : Additive Outliers

ARMAX : Auto Regressive Moving Average with eXternal inputs

ARX : AutoRegressive with Exogenous Variables

BIC : critère d'information bayésien

CFTC : Commodity Futures Trading Commission

CSSED : Cumulative sum of squared error difference (*Somme cumulée de la différence d'erreur quadratique*)

DL : Deep-Learning

FAO : Food and Agriculture Organization of the United Nations

ICE : Intercontinental Exchange

kNN : k-Nearest Neighbors

LOESS : Locally Estimated Scatterplot Smoothing

LOWESS : Locally Weighted Scatterplot Smoothing

LS : Level Shift

LM : Linear Model

LM : Generalized Additive Model (*Modèle Additif Généralisé*)

LSTM : Long Short-Term Memory

LSTM CNN : Long Short-Term Memory and Convolutional Neural Network

MARS : Multivariate Adaptive Regression Splines

ML : Machine-Learning

MLP : Multilayer Perceptron

NOAA : National Oceanic and Atmospheric Administration

OCED : Organisation de coopération et de développement économiques

ONUAA : Organisation des Nations unies pour l'alimentation et l'agriculture

OPEP : Organisation des pays exportateurs de pétrole

R2 OOS : R^2 Out-of-sample (R^2 hors-échantillon)

RMSE : Root Mean Square Error (*Erreur Quadratique Moyenne*)

SVM : Support Vector Machine

TC : Transitory Change

XGB : eXtreme Gradient Boosting

1 Introduction

Dans cette première section, nous édifierons les bases de notre étude. Cela signifie contextualiser correctement notre recherche et définir avec précision la stratégie que nous avons décidée pour mener cette analyse. En somme, nous allons ériger les piliers nécessaires à la compréhension de l'environnement dans lequel notre étude s'inscrit.

1.1 Contexte

En février 2022, la FAO (*Organisation des Nations unies pour l'alimentation et l'agriculture*) annonçait un nouveau record des prix alimentaires mondiaux, avec une augmentation globale de 20%.^[1] D'après l'économiste de la FAO Upali Gal-
keti *"Les inquiétudes concernant l'état des cultures et les disponibilités d'exportation n'expliquent qu'une partie de la hausse actuelle des prix alimentaires mondiaux. Une poussée beaucoup plus importante de l'inflation des prix alimentaires provient de l'extérieur de la production alimentaire, notamment des secteurs de l'énergie, des engrais et des aliments pour animaux [...] Tous ces facteurs tendent à comprimer les marges bénéficiaires des producteurs alimentaires, les décourageant d'investir et d'accroître la production."* Ce rapport de la FAO n'absorbe en réalité que partiellement les effets du contexte géopolitique du début d'année 2022 avec l'invasion de l'Ukraine par la Russie. De nombreux autres éléments sont à prendre en compte pour tenter d'expliquer l'évolution des prix des denrées alimentaires à l'échelle mondiale.

Cet épisode record très récent illustre un problème important et d'envergure mondiale dans notre société actuelle. La sécurité alimentaire est un enjeu de première

importance pour les gouvernements, consommateurs et producteurs, et les prix des produits alimentaires sont un indicateur important de l'état de cette sécurité. Les prix de la nourriture sont un aspect fondamental de notre vie quotidienne qui affectent nos budgets personnels, notre sécurité alimentaire, ainsi que la stabilité des économies locales et mondiales. Un accès fiable à une alimentation adéquate est essentiel pour la santé physique et mentale de toute la population, il est donc crucial de comprendre les déterminants de ces prix pour anticiper leurs impacts sur notre vie et sur le monde.

Historiquement, les crises alimentaires ont toujours existé. Pour citer les plus récentes, nous pouvons penser à la crise alimentaire de 2007-2008, celle de Malawim ainsi que dans le Corne de l'Afrique en 2011 ou encore la crise alimentaire de 2022. La définition même d'une crise alimentaire peut varier en fonction du pays concerné. Dans les pays moins avancés, elle se caractérise comme "*une situation de pénurie, voire de disette et de famine.*"^[2] Autrement dit, on parle d'une insécurité alimentaire généralisée, où un pourcentage élevé de la population ne dispose pas d'un accès régulier à une alimentation adéquate. En revanche, dans les pays plus avancés où les systèmes agricoles et les infrastructures de distribution alimentaire sont généralement mieux développés, les crises alimentaires peuvent être associée à une augmentation soudaine des prix des denrées alimentaires de base (*inflation alimentaire*), liée par exemple à des ruptures d'approvisionnement temporaires. Cette définition fait écho à la situation actuelle en France où le prix des denrées alimentaires a bondit depuis 2022, grignotant une part toujours plus importante du revenus des plus précaires.

D'après la banque mondiale "*70,6 % des économies à faible revenu, 90,9% des pays à revenu intermédiaire de la tranche inférieure et 87% des économies à revenu intermédiaire supérieur ont enregistré des taux d'inflation supérieurs à 5%, un grand*

nombre d'entre elles affichant même une inflation à deux chiffres. En outre, 84,2% des pays à revenu élevé connaissent une forte inflation alimentaire. Les pays les plus touchés se situent en Afrique, en Amérique du Nord, en Amérique latine, en Asie du Sud, en Europe et en Asie centrale.".[3] Ce rapport présente une synthèse exhaustive de la conjoncture actuelle, caractérisée par une crise alimentaire d'envergure planétaire, dont l'incidence varie selon le niveau de revenu du pays considéré.

C'est d'ailleurs pour cette raison que, dans le cadre de notre étude, nous ne pourrions pas aborder en profondeur les spécificités de chaque régions ou pays. En effet, nous nous intéressons à l'évolution des prix des denrées alimentaires à l'échelle mondiale et donc aux tendances globales, ce qui peut masquer des variations significatives entre les différentes régions et pays. De plus, il convient de souligner que les pays les plus vulnérables aux crises alimentaires sont généralement ceux à faible revenu, caractérisés par une faible disponibilité de données fiables, entravant ainsi une évaluation adéquate de leur situation. Nous retrouvons la même problématique lié au manquement de données dans notre horizon d'analyse. Nous travaillons ici de 1999 à 2022 pour pouvoir prendre en compte certaines variables, bien que cela permette d'étudier les tendances et les événements majeurs qui ont influencé les prix alimentaires au cours de cette période, il est important de reconnaître que les événements historiques antérieurs pourraient également avoir un impact sur les prix actuels. Enfin, nous espérons capter le maximum de l'information avec les variables sélectionnées (*basées sur la littérature scientifique*), certains thèmes seront abordés, mais d'autres aspects pourraient être laissés de côté en raison des contraintes liées aux manquement de données sur la période considérée.

Cette étude tentera donc de répondre à la question suivante : **Dans le contexte de volatilité croissante des prix des denrées alimentaires, quels modèles de prévision, à la fois économétriques et de machine learning, sont capables de fournir les prévisions les plus précises des prix mensuels entre mars 2018 et janvier 2022, en se basant sur un ensemble déterminé de variables explicatives ?**

Pour répondre à cette question, nous commencerons par présenter et justifier les différentes variables utilisées, puis nous réaliserons une analyse exploratoire et appliquerons la méthodologie économétrique sur chacune d'entre elle. Nous établirons ensuite plusieurs modèles de prévisions économétrique ainsi que de Machine-Learning, nous présenterons et comparons les différents modèles dans l'optique d'avoir la prévision la plus optimale possible sur la période donnée. Nous utiliserons un plus grand nombre de variables que les études citées, et nous combinerons l'économétrie au Machine-Learning pour tenter de prédire le plus précisément possible le prix de la nourriture à l'échelle mondiale.

2 Présentation des variables

Dans cette partie, nous passerons en revue chaque variable impliquée. Cette section se concentrera sur la présentation de notre variable dépendante ainsi que de nos variables explicatives, chacune d'elles sera justifiée par la littérature scientifique existante. Nous examinerons également l'évolution de ces variables au fil du temps, en fournissant des explications historiques plausibles pour justifier ces dynamiques et changements observés.

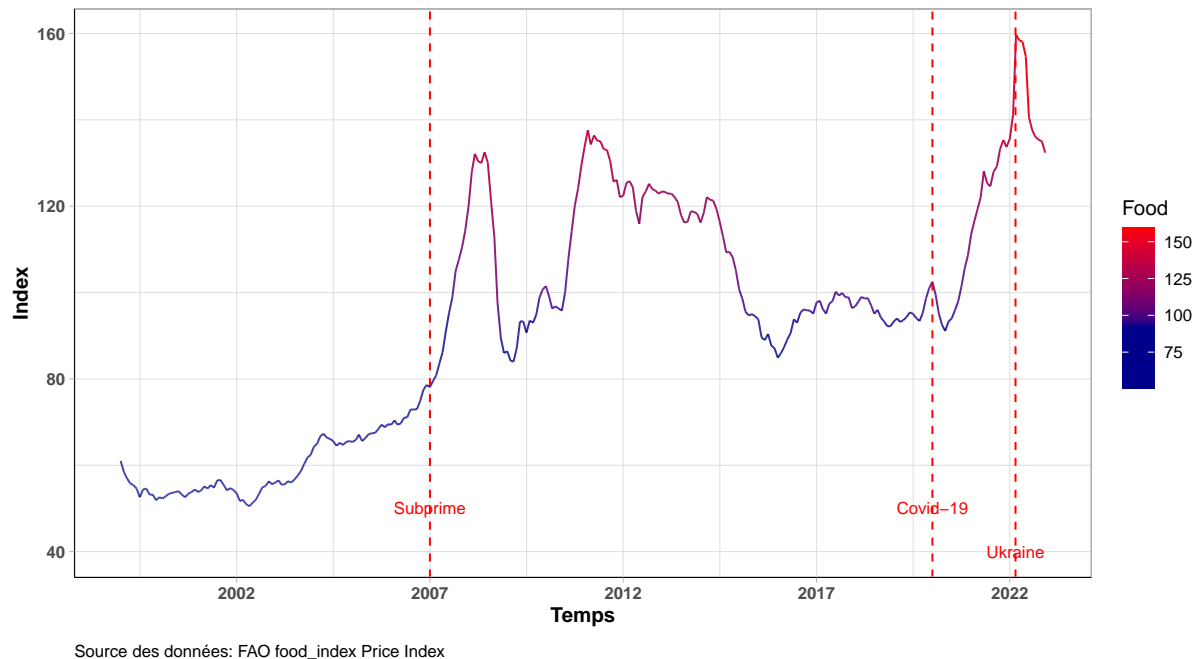
Pour information, la variable dépendante (*ou variable Y*) symbolise la variable que nous souhaitons prévoir, tandis que les variables explicatives (*ou variables X_i*) servent à capter plus d'informations dans le but d'améliorer la prévision de la variable Y . Ce processus est partie intégrante d'une analyse de séries temporelles multivariées. Les séries temporelles se définissent comme des séquences de données, observées à des intervalles de temps successifs, dont l'objectif est souvent de prévoir les valeurs sur un horizon h précisément défini. L'expression "multivariées" renvoie à l'utilisation de plusieurs variables explicatives pour prédire la variable dépendante.

2.1 Variable dépendante : **FAO Food Price Index**

La FAO, connue sous le sigle **ONUAA** pour "*Organisation des Nations unies pour l'alimentation et l'agriculture*", ou plus couramment sous le sigle **FAO**, de l'anglais "*Food and Agriculture Organization of the United Nations*" est une organisation spécialisée du système des Nations unies qui propose le "**FAO Food Price Index**" (*indice mondial des prix des denrées alimentaires*). Ce dernier enregistre l'évolution des prix du marché mondial de 55 produits agricoles et denrées alimentaires en dollars

américains, et est considéré comme un indicateur de l'inflation future et des tendances des coûts dans l'industrie alimentaire. Il est techniquement un index et est calculé selon la formule de Laspeyres.[4] Les poids choisis (*quantité consommées*) étant conditionnel à la période choisie comme référence, à savoir 2014-2016. L'indice global peut se décomposer en plusieurs sous-indices, notamment un indice des prix de la viande, des produits laitiers, des céréales, des huiles et enfin du sucre. L'évolution mondiale des prix de la nourriture nous permet d'observer graphiquement des "chocs" des prix, correspondant à des crises alimentaires d'envergure (*souvent*) mondiale. Une crise alimentaire peut se définir comme une pénurie importante de nourriture, ce qui augmente l'insécurité alimentaire d'une population donnée. Cette pénurie peut être liée à une augmentation brutale des prix, ce qui empêche les ménages de consommer à leur faim. L'origine de l'augmentation des prix est multifactoriel, nous justifierons nos choix dans les prochaines sections.

Figure 1 – Évolution de l'indice des prix de la nourriture de janvier 1990 à décembre 2022



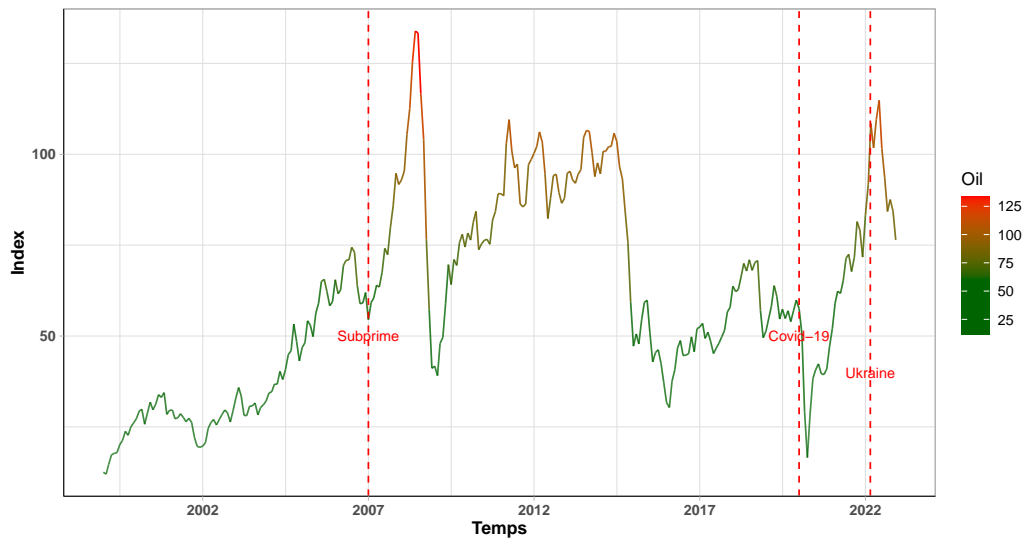
La figure 1 illustre l'évolution de l'indice des prix de la nourriture de 1990 à décembre 2022, en se basant sur des données mensuelles. Globalement, une augmentation de l'indice des prix est observée sur cette période, mais celle-ci est loin d'être régulière et est marquée par plusieurs chocs haussiers importants. Ces chocs incluent, en premier lieu, la crise des subprimes en 2008, suivi de 2010, puis de la pandémie de COVID-19 et du déclenchement de la guerre en Ukraine en 2020 et 2022, respectivement. Nous étudierons de manière plus approfondie ces éléments lors de l'analyse des points atypiques à venir dans la section 3.2.

2.2 Choix des variables explicatives

2.2.1 Cours du pétrole

Comme nous en avons parlé précédemment, d'après la littérature le prix du pétrole semble avoir un rôle à jouer sur le des prix des denrées alimentaires. En effet, il semblerait que lorsque le prix et/ou la volatilité du pétrole augmentent, le prix des aliments augmente.[5] Plus précisément, il semblerait qu'il existe des relations à long terme entre les prix du pétrole brut et de la viande ainsi que des relations de court terme entre le prix du pétrole brut et les céréales.[6].

Figure 2 – Évolution du cours du pétrole WTI



Source des données: F.R.E.D

Graphiquement, on peut observer de “chocs” environ aux mêmes périodes qu’avec l’indice de la FAO. En effet, la premier choc apparent s’observe en 2008, puisque le prix du pétrole atteint son record absolu de 140\$ le baril. Comme pour l’indice des

denrées alimentaires, les marchés spéculatifs semblent être tenus pour responsables.[7]

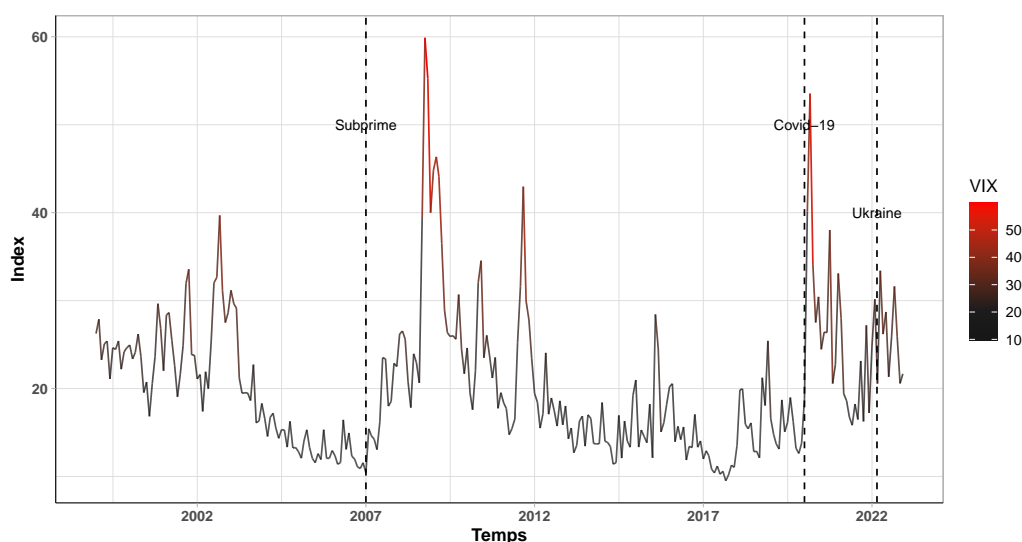
En effet, l'augmentation des prix du pétrole de 50% durant les six premiers mois de 2008 semble être due au manque de régulations sur les contrats futures. C'est en janvier 2006 que la CFTC (*Commodity Futures Trading Commission*) autorise ICE (*Intercontinental Exchange*) à négocier les contrats à terme sur le pétrole brut, ce qui a plus que doublé les prix de références de ces derniers. [8] D'après ce même article, selon une déclaration du ministre qatari du 2 mai 2008, malgré les capacités de production inutilisées, *"l'OPEP n'augmentera pas sa production de pétrole brut, car ce qui se passe actuellement n'est pas une augmentation de la demande de pétrole, mais une forte spéculation sur les contrats à terme. C'est ce qui rend les prix du pétrole si élevés."* La financiarisation du marché pétrolier a donc un rôle majeur dans les fluctuations du cours de ce dernier, notamment via les transactions spéculatives et les contrats à terme. De plus, durant cette période, le mécanisme de formation des prix du marché pétrolier a été perturbé par des incertitudes macro-économiques externes notamment l'excès de liquidité sur les marchés financiers.[9]. Cette forte augmentation des prix perdure jusqu'en 2015, après cette date le ralentissement de l'économie mondiale forme une baisse de la demande de pétrole. Durant cette période on se retrouve dans une offre excédentaire, ce qui fait chuter le prix du baril de 60% entre 2014 et 2015.[10] De plus, la déclaration de l'OPEP en novembre 2014, annonçant que la production de pétrole des pays faisant partie de l'OPEP ne baisserait pas malgré l'augmentation de la production de pétrole des pays ne faisant pas partie de l'OPEP, a également contribué à baisser le prix du baril.[11]

Enfin, la période de 2019-2020, est le dernier choc en date. Les raisons sont les mêmes que pour le prix de la nourriture puisque l'arrêt total de l'économie entraîne différentes mesures, notamment le confinement et l'interdiction de voyager. C'est le 20 avril 2020 que le prix du pétrole subit une chute sans précédent de 300%^[12]. Le COVID ne peut être tenu pour seul responsable puisque la guerre des prix entre la Russie et l'Arabie Saoudite est également à prendre en compte. À cela s'ajoute également le conflit entre la Russie et l'Ukraine.^[13]

2.2.2 VIX : Indice de volatilité

Le VIX aussi appelé "*L'indice de la peur*", est une variable intéressante à prendre en compte dans le cadre de notre analyse, et pour plusieurs raisons. Premièrement, le VIX est un indicateur important de la volatilité du marché boursier, qui est lié aux conditions économiques générales. Il est probable que les fluctuations du marché boursier peuvent avoir un impact direct sur les prix des denrées alimentaires. En effet, il semblerait qu'une variation de l'indice de la peur impacte le prix des denrées alimentaires.^[14]

Figure 3 – Évolution du VIX



Avec cet indice nous serions en mesure de capturer l'impact d'une baisse de confiance des consommateurs ainsi que des investisseurs, ce qui entraîne bien souvent une baisse du niveau général de la demande. De plus l'incorporation du VIX dans notre analyse nous permet d'estimer l'impact de '*l'effet de propagation/contagion*'[15] des marchés financiers, sur le prix des denrées alimentaires. Il s'agit ici d'appréhender l'impact de l'incertitude économique mais également l'interconnexion des économies et marchés financiers mondiaux.

2.2.3 OVX : La volatilité du pétrole

Nous avons précédemment souligné, en nous appuyant sur la littérature scientifique, l'influence du prix du pétrole sur le coût des produits alimentaires. Il serait instructif d'explorer l'incertitude liée à ce paramètre, qui joue un rôle essentiel dans

notre économie actuelle. L'intégration de l'OVX, qui mesure la volatilité future du prix du pétrole, comme variable explicative additionnelle dans une analyse temporelle de l'indice des prix des denrées alimentaires de la FAO, pourrait fournir des renseignements supplémentaires et améliorer la précision des prédictions. En effet, l'OVX pourrait permettre de comprendre les effets de la volatilité future du prix du pétrole sur le coût des denrées alimentaires, qui pourraient différer de l'impact du prix actuel du pétrole. De plus, cet indicateur donne un aperçu de la perception des investisseurs quant à la stabilité des marchés financiers. Nous avons déjà le VIX à notre disposition, mais il serait intéressant de déterminer si l'incertitude ou la volatilité du prix du pétrole a un impact plus significatif que l'incertitude globale de l'économie (*mesurée par le VIX*) sur le prix des denrées alimentaires.

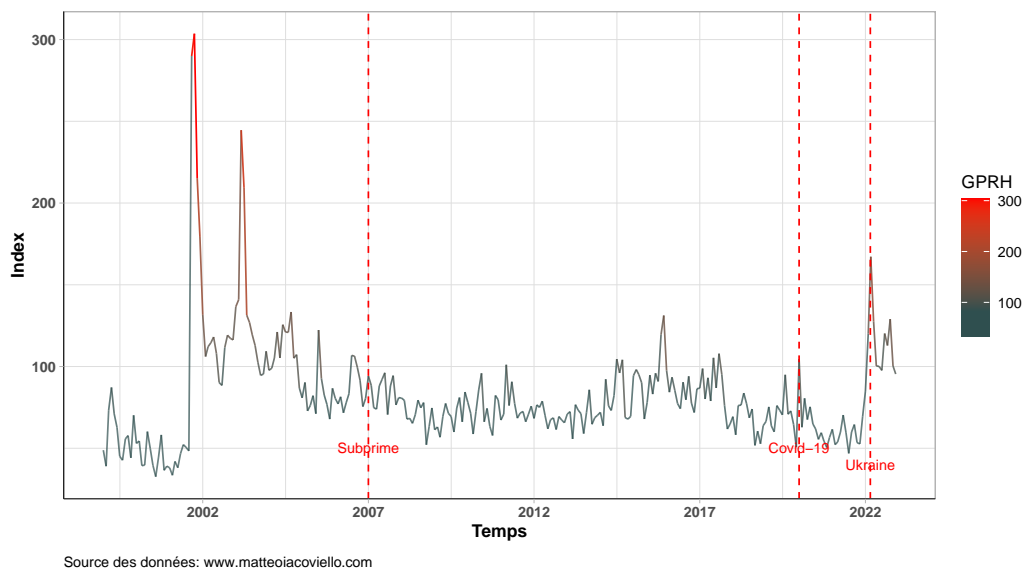
2.2.4 GPR : Indice de risque géopolitique

Le 24 février 2022, la Russie envahit l'Ukraine, deux acteurs centraux dans la production de blé et de Maïs, ce qui entraîne une augmentation du niveau général des prix des denrées alimentaires.[16]. D'après les résultats de l'OCDE[17], "*[...]si les capacités d'exportation de l'Ukraine devaient être réduites à néant et les exportations russes de blé, baisser de 50 %, les prix internationaux du blé pourraient augmenter de 34% durant la campagne 2022/23.*" On comprend rapidement l'impact d'événements géopolitiques majeurs sur le prix des denrées alimentaires, il est donc important de considérer le GPR dans notre analyse.

Cet indice permet de mesurer les risques géopolitiques mondiaux, en prenant en compte plusieurs facteurs comme les conflits militaires, les problèmes sociaux et politiques, les risques de terrorisme ou encore les problème de gouvernance. Dans le

cadre de notre analyse, cet indice appréhende l'impact de tout événement géopolitique majeur sur le prix des denrées alimentaires. A priori, il semblerait que le GPR impact le prix des denrées alimentaires négativement [18], les résultats mettent bien souvent en évidence une relation de cause à effet à sens unique, les facteurs géopolitiques ayant une incidence significative sur les prix des denrées alimentaires.[19]

Figure 4 – **Évolution de l'indice GPR (Geopolitical Risk)**

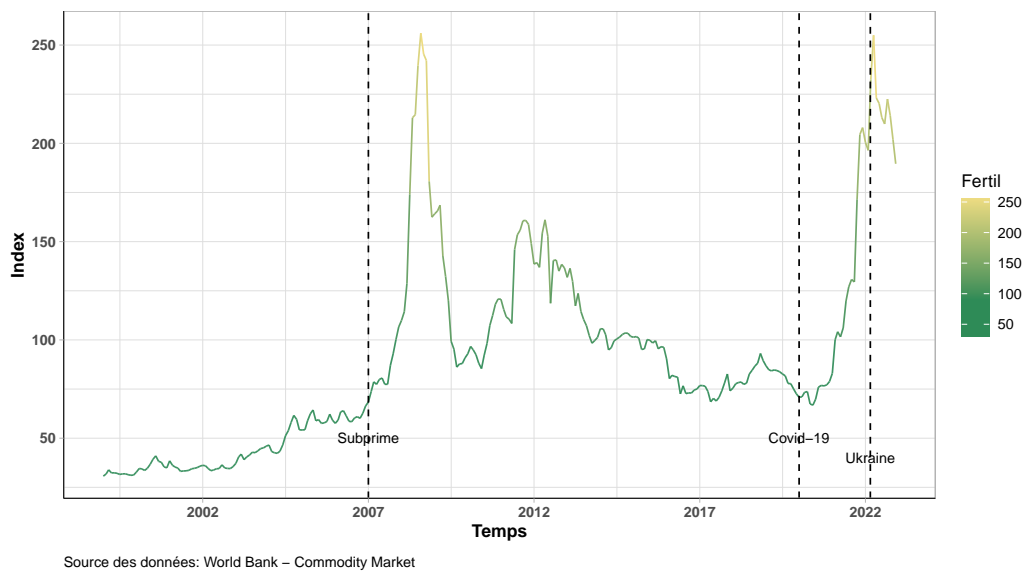


2.2.5 Fertil : Indice des prix des engrais

Le Fertilizers Price Index est un indice qui mesure l'évolution des prix des engrais sur les marchés internationaux. Ce dernier semble avoir un impact sur le FAO Food Price Index [20]. Selon Ott Herve en 2012 [21], le comportement spéculatif sur les marchés des engrais a un impact sur les prix des denrées alimentaires, mais la spéculation sur les marchés dérivés ne peut être considérée comme la cause. D'après cette même étude les prix des engrais et des denrées alimentaires sont étroitement liés et

ont tendance à bouger ensemble. Les prix des denrées alimentaires ont une influence directe sur les prix des engrais, car une hausse de ces prix entraîne une augmentation de la demande d'engrais, ce qui fait augmenter les prix des engrais. En somme, les prix des denrées alimentaires sont une des causes des mouvements des prix des engrais.

Figure 5 – **Évolution du prix des engrais : Fertil**

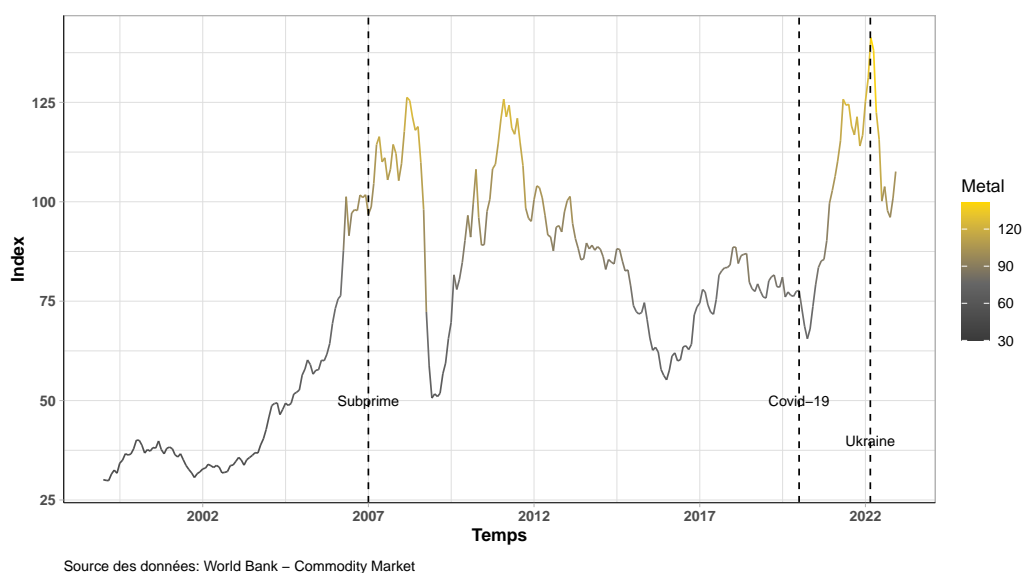


Le secteur de l'énergie est un autre facteur clé qui a déclenché la hausse des prix des engrais. En effet, l'énergie est un élément essentiel de la production d'engrais, et la hausse des prix du pétrole et du gaz naturel a eu un impact significatif sur les nutriments azotés, dont la production dépend fortement de la disponibilité énergétique pour leur fabrication et leur transport. En somme, l'indice du prix des engrais est un indicateur intéressant pour l'étude de l'évolution et des déterminants des prix des denrées alimentaires, car il est étroitement lié aux prix des denrées alimentaires et dépend fortement du secteur de l'énergie pour sa production.

2.2.6 Métaux : Indice des prix des métaux et des minéraux

L'indice des prix des métaux et des minéraux est un indicateur utilisé pour analyser les prix des métaux. Cet indicateur indexé est une moyenne pondérée des prix de l'aluminium, du cuivre, du minerai de fer, du plomb, du nickel, de l'étain et du zinc. L'utilisation de cet indice peut s'avérer pertinent pour l'analyse de l'évolution et des déterminants du prix des denrées alimentaires. En effet, les matières premières métalliques et minérales ont des liens étroits avec les denrées alimentaires, notamment en ce qui concerne l'exploitation des terres agricoles et l'utilisation de machines agricoles.

Figure 6 – Évolution de l'indice "Métaux"



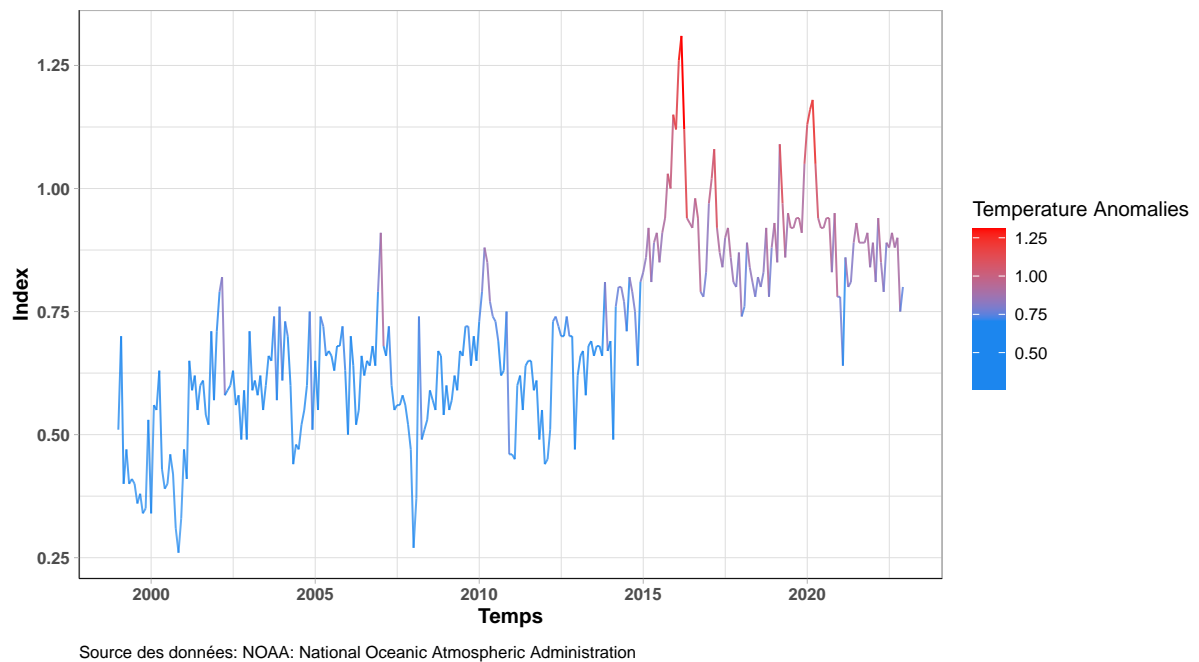
Le prix des matières premières métalliques est souvent influencé par les mêmes facteurs que ceux qui influencent le prix des denrées alimentaires, tels que la volatilité des marchés financiers, les conditions climatiques, les politiques économiques et les

événements géopolitiques[22]. Ainsi, l'utilisation du Metals and Mineral Price Index peut aider à comprendre les liens entre ces variables et leur impact sur les prix des denrées alimentaires, et ainsi potentiellement réaliser des prévisions plus précises.

2.2.7 Temperature Anomalies

Il est crucial de prendre en compte l'impact du changement climatique dans notre analyse, puisqu'il est évident que ce dernier impacte directement les rendements des cultures. En revanche, dans ce domaine les données (*d'échelle mondiale*) sont bien plus difficile à trouver. Dans notre cas, nous utiliserons le niveau "*d'anomalies*" des températures mondiales, des données mensuelles issue de la bien connue NOAA (*National Oceanic and Atmospheric Administration*), créée par le Congrès des États-Unis en 1970. Ces données prennent en compte les températures sur terre mais également des océans, qui ont le sait, se réchauffent rapidement.[23]

Figure 7 – Évolution du niveau d'anomalies des températures mondiales



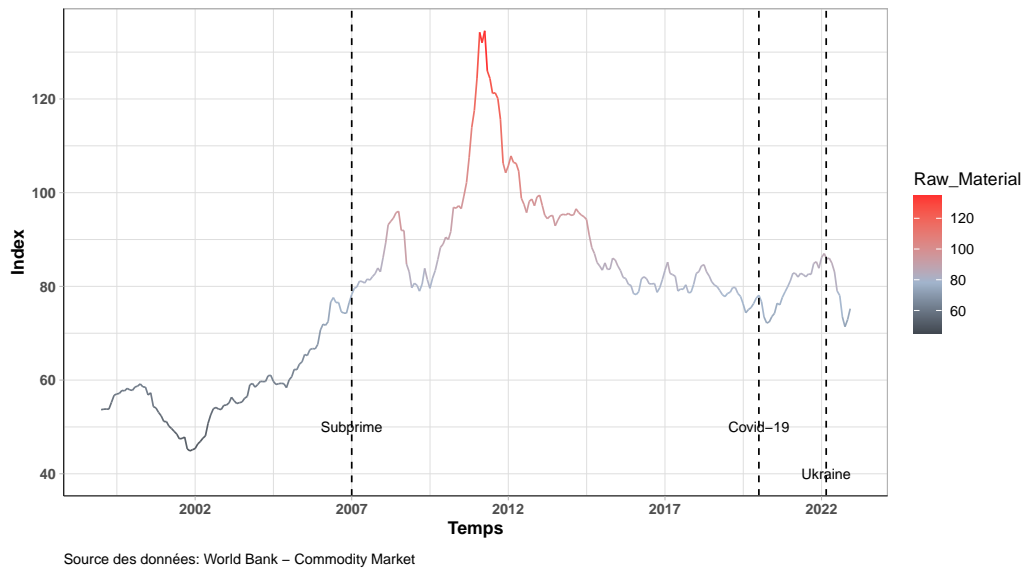
Graphiquement, on voit clairement que les températures n'ont fait qu'augmenter, ou plutôt leur niveau "*d'anomalies*" n'a fait qu'augmenter de 1990 à 2022. Cette augmentation du niveau d'anomalie des températures, entraînant des canicules et sécheresses prolongées, sont des événements très inquiétants et posent de nombreux problèmes pour le rendement des cultures.^[24] En effet, des températures très élevées augmentent le stress thermique des plantes ce qui entraîne souvent des pertes de rendements. On parle également de stress hydrique des plantes quand on observe une réduction de la qualité et quantité de l'eau, nécessaire à la survie et aux bons rendements des cultures. On parle également de prolifération de nombreuses espèces invasives et autres pathogènes, ou encore d'impact direct et indirect sur la santé des

animaux suite au manque d'eau qui entraîne la propagation de nombreuses maladies. Cette liste exhaustive est en réalité bien plus longue et nous permet de facilement comprendre l'importance d'inclure cette variable au sein de notre analyse. On peut s'attendre à ce que les prix des denrées alimentaires augmentent lorsque les conditions météorologiques et climatiques se dégradent (*augmentation des températures et/ou d'évènements climatiques extrêmes*)

2.2.8 Prix des matériaux brut : Raw Materials

Il est important d'inclure l'indice des prix des matières premières brutes dans notre analyse, car il semble avoir une influence sur le coût des produits alimentaires [25]. Cet indicateur est pertinent pour notre étude pour diverses raisons. Premièrement, les matières premières agricoles sont essentielles à la production alimentaire, et leur prix influence directement celui des produits finis. En intégrant cette dimension dans notre analyse multivariée, nous pourrions mieux comprendre la relation entre ces deux variables et saisir comment les changements de prix des matières premières affectent le coût des denrées alimentaires. En outre, les fluctuations des prix des matières premières sont généralement liées à des facteurs tels que les conditions climatiques, les politiques agricoles et les chocs macroéconomiques. En prenant en compte ces facteurs dans notre analyse multivariée, nous pourrions mieux appréhender les mécanismes sous-jacents qui régissent les marchés alimentaires et déterminer si ces facteurs peuvent également expliquer les variations de l'indice des prix des denrées alimentaires de la FAO.

Figure 8 – Évolution de l'indice Raw Materials



2.2.9 Contrat Futures

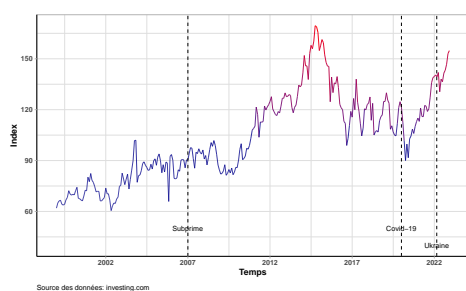
Cette section se distingue des précédentes puisqu'elle intégrera plusieurs variables simultanément, plus précisément tous les contrats Futures sur les commodités qui constituent le Fao Food Price Index. En 2008, l'impact des marchés spéculatifs sur les produits agricoles était un facteur déterminant à considérer pour expliquer les fluctuations historiques de cette période [26]. Loin de "stabiliser" le prix des denrées alimentaires comme promis, la déréglementation des marchés financiers semble avoir l'effet inverse [27]. En effet, il apparaît que la spéculation par les Hedge Funds et autres institutions financières d'investissement sur les denrées alimentaires via des contrats Futures a un effet majeur sur les prix [28]. Mais ce n'est pas tout, la spéculation sur le pétrole (*encore une fois, via des contrats Futures*) semble également être un facteur contributif à l'augmentation marquée des prix des denrées alimentaires

[28]. Nous tiendrons donc compte des contrats Futures suivants :

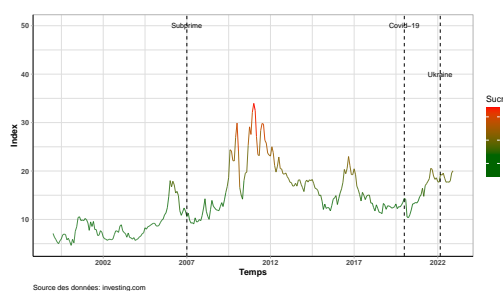
Table 1 – **Contrats Futures utilisés**

Nom	Contrat Futures	Source
Viande	CME Group - Live Cattle	investing.com
Sucre	ICE - No. 11 Sugar Futures	investing.com
Huiles	US Soybean Oil Futures	investing.com
Céréales	US Wheat Futures	tradingeconomics.com

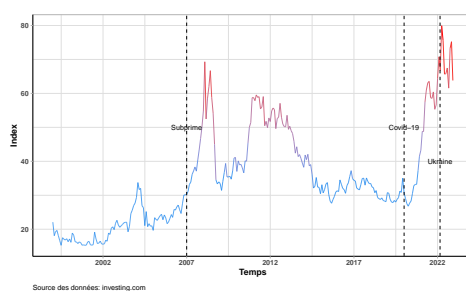
Figure 9 – **Évolution des différents contrats Futures**



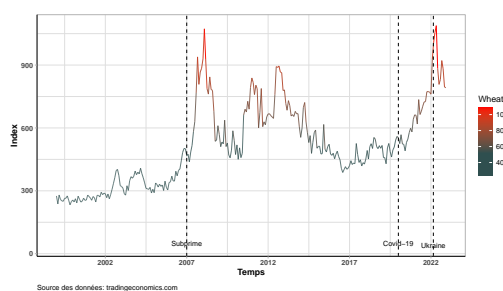
(a) Futures Viande



(b) Futures Sucre



(c) Futures Huiles

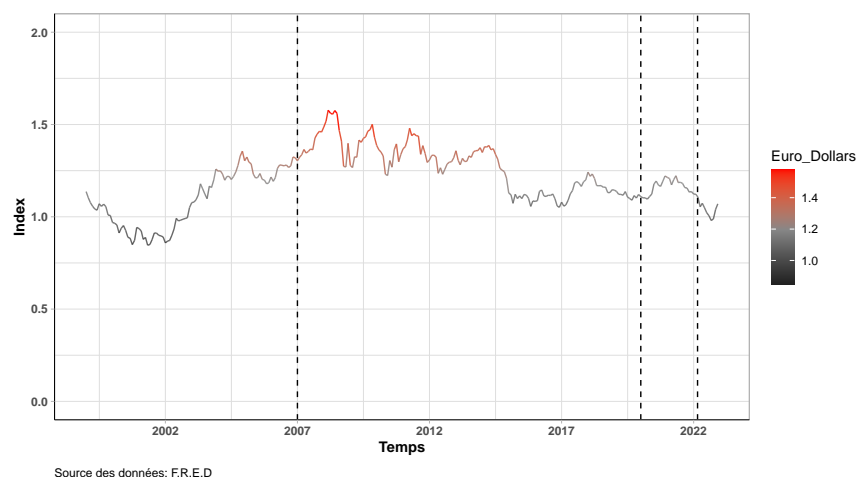


(d) Futures Céréales

2.2.10 Cours Euro-Dollars

La parité euro-dollar, qui se situe parmi les taux de change les plus influents à l'échelle mondiale, joue un rôle crucial dans les transactions internationales, y compris celles concernant les produits alimentaires. Cette variable macroéconomique, souvent exploitée pour mesurer les effets de l'économie globale sur d'autres variables, est fréquemment mobilisée dans diverses études. Elle semble notamment avoir une incidence notable sur les prix des produits alimentaires selon Huchet [29]. La parité euro-dollar est parfois utilisée comme un indicateur de la stabilité économique mondiale, reflétant ainsi les risques économiques et géopolitiques ainsi que les modifications de politique monétaire de la Réserve fédérale des États-Unis et de la Banque centrale européenne. Ces facteurs peuvent avoir une répercussion significative sur le coût des denrées alimentaires.

Figure 10 – Évolution de l'Euro-Dollars



Nous pouvons également penser que les gouvernements peuvent utiliser la politique de change pour influencer les exportations et les importations, ce qui peut avoir

des conséquences sur les prix des denrées alimentaires. Encore une fois, cette variable semble être intéressante dans le cadre de notre étude.

3 Analyse exploratoire

Dans cette partie, nous allons réaliser une analyse exploratoire approfondie sur la variable dépendante (*autrement dit, le FAO Food Price Index*). Nous commencerons par examiner les liens entre les différentes variables (*analyse des corrélations*), avant de repérer et de rectifier les valeurs atypiques dans nos séries de données. Ensuite, nous nous attacherons à supprimer l'influence saisonnière et à rendre ces séries stationnaires. Après ces étapes, nous vérifierons à nouveau l'existence éventuelle de valeurs atypiques. Cette analyse sera exclusivement effectuée sur la série Y (*variable dépendante*). Bien que toutes nos séries de données aient été soumises aux mêmes procédures d'analyse, nous avons choisi de présenter uniquement les résultats pour la série Y afin de ne pas alourdir l'analyse. Les séries corrigées des points atypiques, dé-saisonnalisées et stationnarisées pour toutes les séries étudiées sont disponibles en annexe. (1)

Autrement dit, voici les étapes suivies :

- Analyse des corrélations : 3.1
- Détection des points atypiques : 3.2
- Saisonnalité : 3.3
- Stationnarité : 3.4
- Re-détection des points atypiques : 3.2
- Statistiques descriptives sur la série corrigée 3.5

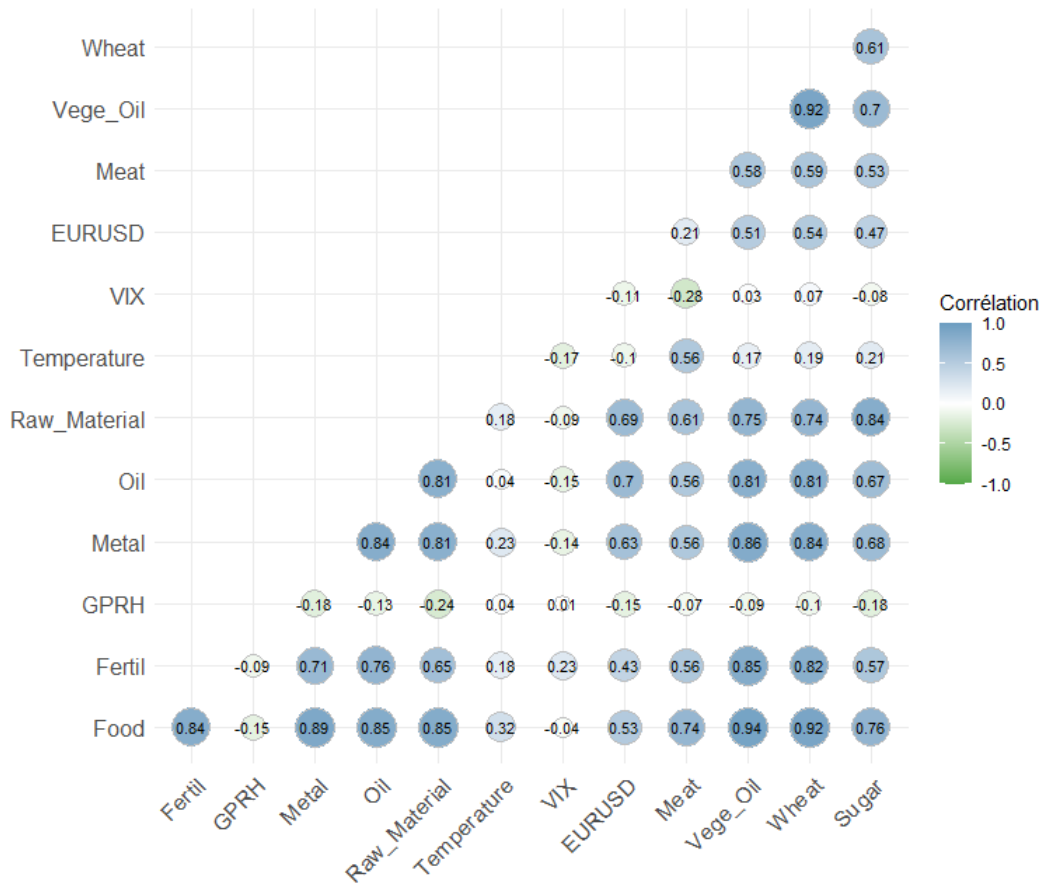
Comme nous l'avons dit précédemment, pour ne pas alourdir l'analyse nous montrons uniquement le processus pour la variable dépendante (Y). Cependant, cette même méthode a été appliquée sur nos variables explicatives (X_i). Les résultats d'ana-

lyse sur les variables explicatives sont visibles en annexe ([4](#),[5](#),[6](#),[7](#),[8](#),[9](#),[10](#),[11](#),[12](#),[13](#),[14](#),[15](#)).

3.1 Corrélation

Analyser les corrélations entre différentes variables est une étape importante dans la modélisation des prévisions. Cela permet d'identifier la force et la direction de la relation entre les variables, ce qui permet d'avoir une meilleure compréhension de leur comportement conjoint et de la manière dont elles peuvent influencer la variable que nous cherchons à prévoir. Ceci est crucial pour construire des modèles de prévision plus robustes et précis. Par ailleurs, la détection de corrélations fortes entre les variables explicatives peut nous aider à éviter les problèmes de multicolinéarité (*variables fortement corrélées*), qui peuvent biaiser les résultats et diminuer la précision de nos prévisions.

Figure 11 – Matrice des corrélations sur les séries brutes



La figure 11 révèle principalement des corrélations positives entre diverses variables. Notre variable Y , "food", affiche une corrélation positive significative avec plusieurs matières premières comme "fertil", "metal", "oil", "raw material", "vege oil" et "wheat", ayant des coefficients de corrélation de 0.84, 0.89, 0.85, 0.85, 0.94 et 0.92 respectivement. Ces corrélations élevées suggèrent que ces matières premières sont très probablement utilisées dans la fabrication d'aliments, expliquant ainsi leur lien avec notre variable dépendante.

De plus, une corrélation notable existe entre "*vege oil*" et "*raw material*" (0.75), impliquant une dépendance mutuelle importante entre ces deux matières. De même, l'interrelation élevée entre "*wheat*" et "*vege oil*" (0.92) pourrait signaler une association étroite dans le contexte de la production alimentaire.

Enfin, une forte corrélation positive entre "*oil*" et "*metal*" (0.84) suggère une utilisation commune dans la fabrication de machines et d'équipements.

Globalement toutes ces relations ne sont pas surprenantes au vu des explications (*basées sur la littérature scientifique*) que nous avons fourni dans la partie précédente(2).

3.2 Détection des points atypiques

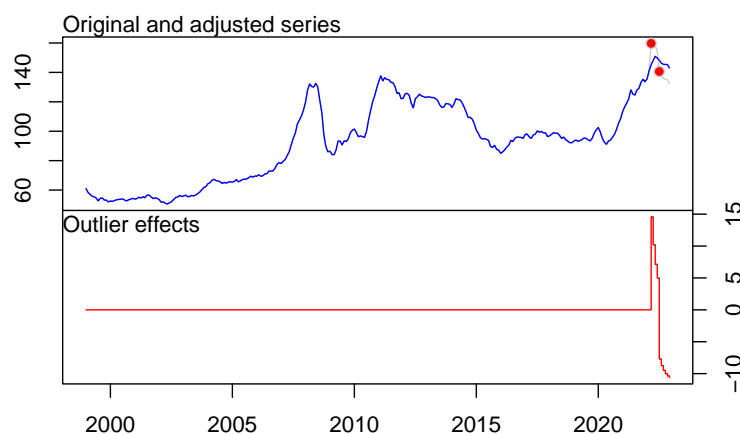
Il est essentiel de traiter les points atypiques d'une série temporelle, car ils peuvent avoir un impact significatif sur les résultats de l'analyse. Les points atypiques peuvent être engendrés par des erreurs de mesure, des événements imprévus ou des changements structurels dans la série temporelle. L'omission ou la mauvaise gestion de ces points peut fausser les résultats de l'analyse et conduire à des conclusions erronées. En effet, les modèles de prévision de séries temporelles sont souvent basés sur l'hypothèse que le passé est un bon indicateur de l'avenir. Si un point atypique est inclus dans l'entraînement du modèle, cela peut fausser les prévisions du modèle. Les points atypiques peuvent donc entraîner une sur-estimation ou une sous-estimation des prévisions futures. De plus, un point atypique peut être considéré comme du bruit, surtout s'il est dû à une erreur de mesure ou à un événement exceptionnel qui est peu susceptible de se reproduire.

La table 2 ainsi que la figure 12 fournissent des indications sur l'atypicité de la série étudiée. En effet la série a deux points atypiques : un point TC (*Transitory Change*) qui est causé par des événements imprévus ou des perturbations temporaires mais également un point LS (*Level Shift*) qui est causé par un changement structurel de la série. Ces points nécessitent une détection et un traitement particuliers pour éviter des biais dans les analyses et prévisions.

Table 2 – **Point atypique sur la série brute**

Type	Période	Coefhat	T-stat
TC	Mars 2022	14.57	9.49
LS	Juillet 2022	-11.19	6.32

Figure 12 – **Détection des points atypiques sur la série brute**



Dans le cas présent, les deux points ont probablement été causé par le déclenchement de l'offensive de la Russie en Ukraine. Cette dernière a eu un impact

significatif sur l'indice des prix de la nourriture en raison de plusieurs facteurs. Tout d'abord, l'Ukraine est un important producteur de céréales, notamment de blé, d'orge et de maïs, qui sont des produits de base utilisés dans la fabrication de nombreux aliments. Les conflits armés, les destructions et les perturbations des transports ont donc eu un impact sur la production et l'approvisionnement de ces produits, entraînant une augmentation des prix. En outre, la guerre en Ukraine a également perturbé les marchés financiers, ce qui a également eu un impact sur l'indice des prix de la nourriture. L'incertitude quant à l'impact de la guerre sur l'économie mondiale a provoqué une volatilité accrue des prix des denrées alimentaires, en raison de la réaction des marchés financiers. La volatilité des prix des denrées alimentaires peut également être attribuée à la spéculation sur les marchés financiers. Selon certaines estimations, les spéculations sur les matières premières agricoles seraient responsables de jusqu'à 40% de l'inflation des prix des denrées alimentaires [30]. Nous pouvons également noter l'impact de nombreux événements climatiques désastreux liés aux changements climatiques [31].

Nous avons également corrigé les points atypiques après traitement des données (*désaisonnalisation et stationnarisation*). Ces étapes seront expliquées dans les parties suivantes. Les résultats de détection des points atypiques sur la série corrigée sont visibles en annexes (3 et 2). Nous remarquons qu'il existe plus de points atypiques après traitement qu'avant traitement. Ceci peut sembler surprenant étant donné qu'il n'y avait que deux points atypiques sur la série brute. Cependant, la différenciation de la série implique une soustraction de chaque observation de la précédente, ce qui peut augmenter la variabilité de la série. Ainsi, les points atypiques peuvent devenir plus évidents après la différenciation.

Nous observons deux TC en 2008, et trois AO en 2011, 2012 et 2021. Concernant les points atypiques de 2008, nous observons une explosion nette des prix ce qui correspond à la période de la crise des subprimes. Cette dernière initie une augmentation brutale et continue des prix des denrées alimentaires, avec un niveau des prix général qui n'est jamais redescendu à son niveau d'avant-crise. Cette hausse sans précédent entraîna plusieurs rationnements et émeutes dans de nombreux pays autour du globe.[32] En effet, cette crise aurait poussé 75 millions de personnes en situation de sous-nutrition et 125 millions en situation d'extrême pauvreté[33], ce qui marque une crise alimentaire majeure. Cette dernière peut s'expliquer par de nombreux facteurs, notamment une demande grandissante chez les pays émergents aux populations très importantes tel que l'Inde ou la Chine, avec simultanément de mauvaises conditions météorologiques (*notamment d'intenses sécheresses en Australie menant à de très mauvaises récoltes*). À cela s'ajoute une chute drastique du stock de céréales, qui en 2008 figurait parmi les plus faibles depuis 1970 [2], ainsi que des coûts énergétiques plus élevés (*notamment et surtout du pétrole*) [5], ce qui augmente le coût de production agricole (prix des engrais, fertilisants, coût des transports). Mais nous pouvons également mentionner le rôle des marchés spéculatifs (*fonds spéculatifs, fonds indiciels et souverains*) sur les matières premières agricoles qui ont été un facteur clé de la volatilité brutale des prix sur les matières premières agricoles [26].

De plus la dérégulation des marchés a mené à une suppression de toutes restrictions quantitatives sur les positions spéculatives des contrats à terme agricoles, exerçant une énorme pression à la hausse sur le niveau des prix. Selon des calculs basés sur des dépôts réglementaires, le montant des fonds investis dans les indices de matières premières est passé de 13 milliards de dollars en 2003 à 260 milliards

de dollars en août 2008. Les prix globaux des matières premières ont augmenté au cours de la même période plus qu'au cours de toute autre période enregistrée dans l'histoire des États-Unis. Un autre événement marquant qui a contribué à l'augmentation globale des prix des denrées alimentaires est le scandale du lait frelaté en 2008 en Chine. Le scandale a atteint son pic médiatique vers la fin de l'été 2008 et a impliqué la contamination du lait en poudre avec de la mélamine, un produit chimique toxique. Cela a conduit à une crise de confiance dans l'industrie laitière chinoise, et les consommateurs se sont tournés vers les produits importés, entraînant ainsi une augmentation de la demande mondiale de lait et, par conséquent, une augmentation des prix.

Durant l'été 2010, une flambée des prix de la nourriture a été constatée. Cette hausse des prix s'explique notamment par l'augmentation des prix du pétrole, qui a engendré une forte demande de production de biocarburants [34]. Par ailleurs, la spéculation sur les marchés financiers ainsi que les conditions météorologiques défavorables ont contribué à cette flambée des prix, révélant un schéma récurrent dans l'évolution des prix des denrées alimentaires. À l'été 2012, ce schéma s'est répété, avec une grave sécheresse aux États-Unis qui a affecté la production de maïs, de soja et d'autres cultures importantes, entraînant une baisse de l'offre mondiale de ces produits et une augmentation des prix des denrées alimentaires. D'autres facteurs tels que la spéculation sur les marchés financiers et l'augmentation de la demande de biocarburants ont également contribué à cette flambée des prix des denrées alimentaires.

Le dernier point atypique datant de mai 2021 peut s'expliquer par l'impact de la pandémie COVID-19 sur l'augmentation significative des prix de la nourriture, comme

illustré sur le graphique 1. De plus, d'après le rapport officiel de la FAO[35], la crise de 2019-2020 liée à la pandémie COVID a entraîné un des plus grands pics de famines connue depuis de nombreuses décennies. À cela s'ajoute des conditions climatiques de plus en plus instables, entraînant sécheresse, inondations ou de violentes tempêtes impactant significativement l'état des récoltes [36]. Les pays les plus touchés sont ceux aux revenus les plus faibles, subissant non seulement la pandémie et les mauvaises récoltes, mais aussi l'augmentation globale du prix des denrées alimentaires. On parle ici de nombreux pays d'Afrique et d'Asie. Si les précédentes crises étaient multifactorielles, cette dernière l'est encore plus. Avec une pandémie mondiale, la chaîne alimentaire se retrouve impactée à tous les niveaux possibles. En effet, la manipulation des aliments peut aggraver la transmission de ce virus, ce qui freine considérablement la production via la mise en place de multiples mesures d'hygiène afin d'éviter tout risque de contaminations [37].

3.3 Saisonnalité

L'étude de la saisonnalité permet d'identifier des modèles cycliques récurrents, ce qui est crucial pour l'analyse des données temporelles. La saisonnalité est un phénomène qui se produit lorsqu'une série temporelle est affectée par des variations périodiques spécifiques. En d'autres termes, c'est une tendance qui se répète à intervalles réguliers sur une période donnée, comme une heure, un jour, une semaine, un mois, un trimestre, une saison ou une année. Ces variations peuvent être dues à divers facteurs tels que les changements climatiques, les jours fériés, les événements sportifs, et les cycles économiques. Les tests de Webel-Ollech [38] et Seasonal Dummies [39] sont souvent utilisés pour détecter la présence de saisonnalité. Le premier

compare la variance de la série originale à celle d'une version où les valeurs saisonnières ont été remplacées par leur moyenne respective, tandis que le second ajoute des variables indicatrices saisonnières à un modèle de régression linéaire. Dans les deux cas, l'hypothèse nulle (H_0) est l'absence de saisonnalité.

Le tableau 3 présente les résultats de ces deux tests. Nous pouvons observer que les p-values associées à chaque test sont supérieures à 0,05, ce qui indique que l'hypothèse nulle de non-saisonnalité ne peut être rejetée pour ces deux tests. En complément de ces tests, l'étude du périodogramme peut également être utilisée pour analyser la saisonnalité. Le périodogramme de notre série est représenté sur le graphique 3 en annexe. L'analyse du périodogramme indique une absence de saisonnalité dans les données étudiées, cela étant corroboré par l'absence de pics significatifs.

Table 3 – **Détection de saisonnalité - Variable Y**

Tests	p-value
Webel-Ollech	0.37
Seasonal dummies	0.35

Nous avons également appliqué ces deux tests de saisonnalité sur les variables explicatives de notre étude. Les résultats de ces tests figurent en annexe 16 et attestent la présence de saisonnalité chez certaines variables. En effet, les variables explicatives dont la p-value des tests est inférieurs à 0.05 indique une potentielle saisonnalité dans les données (H_1 : *présence de saisonnalité*). Pour corriger cela, nous appliquons la méthode STL (*Seasonal and Trend decomposition using Loess*) [40] qui est une méthode de décomposition de séries chronologiques. Comme décrit dans le

cours de M.darne [41], chaque composante est déterminée par une régression locale, aussi appelée :

- lissage **LOESS** (*LOcally Estimated Scatterplot Smoothing*)
- lissage **LOWESS** (*LOcally WEighted Scatterplot Smoothing*)

C'est une méthode non paramétrique, appelée aussi régression polynomiale locale, avec pondération locale, basée sur la méthode des k plus proches voisins (k -NN, *k-nearest neighbors*).

Après application de cette méthode sur les séries présentant de la saisonnalité, toutes nos séries semblent corrigées, les p -value des deux tests étant toutes supérieures à 0.05 17.

3.4 Stationnarité

La stationnarité est un concept important en statistiques, et plus particulièrement en analyse de séries temporelles. Une série temporelle est dite stationnaire si ses propriétés statistiques ne changent pas au fil du temps. En d'autres termes, peu importe le point de départ choisi pour observer la série, les propriétés comme la moyenne, la variance et la structure d'autocorrélation (*c'est-à-dire le degré de corrélation entre les observations à différents intervalles de temps*) restent constantes. On parle également de stationnarité de second ordre (*ou faible*) si elle a une moyenne et une variance constantes, et si la covariance entre deux périodes ne dépend que du décalage entre ces deux périodes et non du moment réel auquel elles sont observées.

Une série temporelle X_t est dite faiblement stationnaire si elle satisfait à ces

trois conditions :

- La moyenne de la série est constante dans le temps :

$$E[X_t] = \mu \quad \forall t$$

où $E[\cdot]$ est l'opérateur d'espérance mathématique (*c'est-à-dire la moyenne*) et μ est un réel constant.

- La variance de la série est constante dans le temps :

$$\text{Var}[X_t] = \sigma^2 \quad \forall t$$


où $\text{Var}[\cdot]$ est l'opérateur de variance et σ^2 est un réel constant.

- La covariance entre deux instants ne dépend que du décalage entre eux :

$$\text{Cov}[X_{t+h}, X_t] = \text{Cov}[X_h, X_0] \quad \forall t, \forall h$$

où $\text{Cov}[\cdot]$ est l'opérateur de covariance.

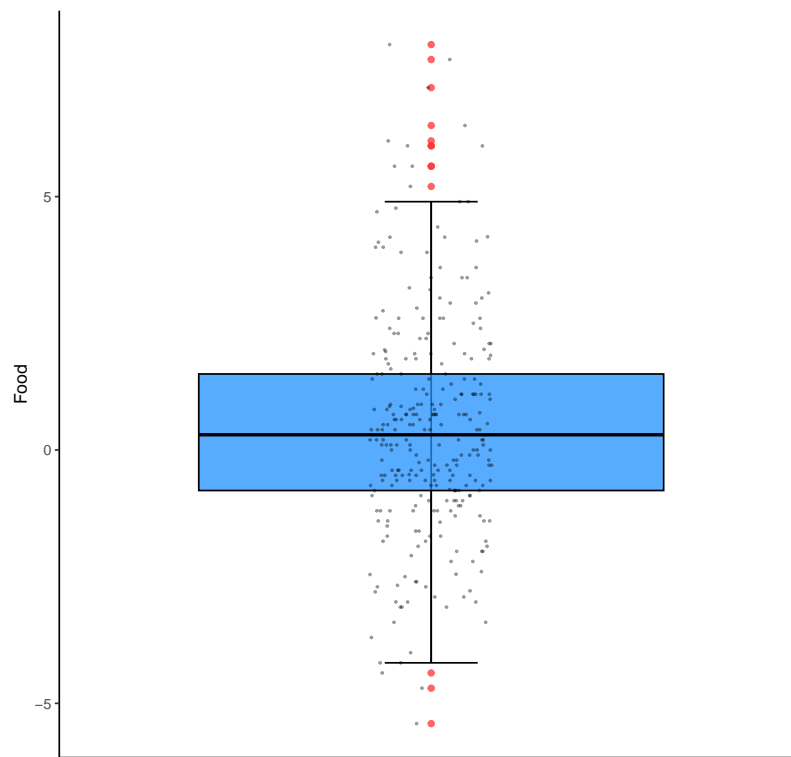
Pour appliquer la plupart des modèles économétriques, il est crucial que les séries temporelles soient stationnaires. En effet, les modèles économétriques reposent souvent sur l'hypothèse que les relations entre les variables restent stables dans le temps (*autrement dit, stationnaires*). Si une série temporelle n'est pas stationnaire, cela signifie que ses propriétés statistiques changent avec le temps, ce qui rend difficile la prévision de ses valeurs futures.

Nous avons utilisé la fonction `adf.test` du package `tseries` sur  pour vérifier la stationnarité de notre série. Les résultats des tests présentés dans les tables 1 et 2 situées en annexe montrent d'abord que notre série initiale n'est pas stationnaire mais qu'après avoir effectué une différenciation, nous avons pu obtenir une série stationnaire. Nous appliquons la même méthode pour nos variables explicatives, en différenciant celles qui ne sont pas "*naturellement*" stationnaires. Après différenciation, toutes nos séries semblent stationnaires 18. En effet, leur p-value du ADF test étant inférieur à 0.05 nous pouvons accepter l'hypothèse alternative ($H1$) selon laquelle la série est stationnaire.

3.5 Statistiques descriptives sur la série corrigée

Le boxplot représenté sur la figure 13 montre qu'une dizaine d'observations semblent dépasser l'intervalle de la boîte à moustache. Cela remet en question l'hypothèse précédente selon laquelle notre série Y avait été ajustée pour éliminer les valeurs atypiques potentielles. Ces valeurs atypiques se sont manifestées une fois que la série a été corrigée du premier groupe de points atypiques. De ce fait, les nouvelles valeurs ne sont pas en mesure d'influencer la moyenne de notre série ni d'altérer nos prochains résultats.

Figure 13 – **Boxplot de la série Y corrigée**

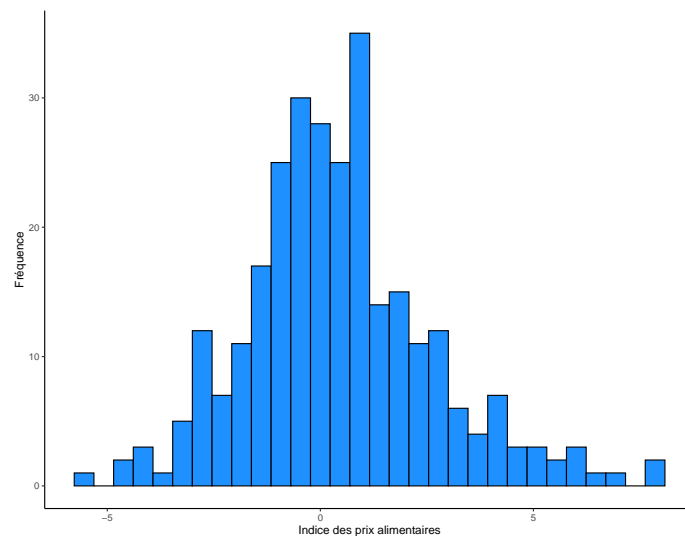


En examinant les données du tableau 19, on constate que la moyenne (0.45) et la médiane (0.30) présentent une différence significative, ce qui suggère une possible asymétrie de la distribution. Toutefois, la variance (4.73) et l'écart-type (2.17) sont significativement élevés, suggérant une dispersion considérable des données et une distribution plutôt étalée.

L'analyse de la distribution des données représentées par l'histogramme présenté dans la figure 14 suggère une distribution qui suit une forme de cloche, caractéristique d'une distribution gaussienne. De plus la kurtosis (0.86) est légèrement élevée, impli-

quant une distribution plus pointue que la distribution normale. La skewness (0.52) est positive, indiquant une légère asymétrie de la distribution vers la droite. Les valeurs minimales et maximales (-5.40 et 8) sont considérablement éloignées l'une de l'autre, tout comme les quartiles (1^{er} quartile -0.80 et 3^{eme} quartile 1,55), témoignant encore de la distribution étalée des données. En somme, cette analyse suggère une distribution étalée des données avec une légère asymétrie vers la droite et une kurtosis élevée.

Figure 14 – **Histogramme de la série ajustée des points atypiques**



4 Sélection des variables

4.1 Approche BestSubSet & Gets

Dans cette partie nous chercherons sélectionner les variables permettant de sortir le meilleur modèle en fonction de plusieurs critères statistiques. Dans un premier temps nous utiliserons l'approche BestSubSet, puis l'approche Gets en enfin nous

comparons les résultats. Comme la grande majorité des méthodes utilisées dans cette étude, ces deux méthodes sont mises en œuvre à l'aide du langage de programmation



4.1.1 Approche BestSubSet

Selon l'approche BestSubSet, la sélection du sous-ensemble optimal de variables implique le test de toutes les combinaisons possibles pour déterminer le modèle le plus performant. Autrement dit, pour chaque nombre possible de variables prédictives (*de 1 à n , où n est le nombre total de variables prédictives*), nous calculons le modèle de régression pour chaque sous-ensemble possible de cette taille. Pour chaque modèle généré, nous calculons un certain critère de qualité de l'ajustement, comme le R^2 , le R^2 ajusté, le critère d'information d'Akaike (AIC), le critère d'information bayésien (BIC).

Le R^2 représente la proportion de la variance de la variable dépendante qui est prévisible à partir des variables explicatives dans le modèle. Autrement un R^2 de 1 indique que le modèle de prédit parfaitement la variable Y . En revanche, un R^2 de 0 indique strictement l'inverse.

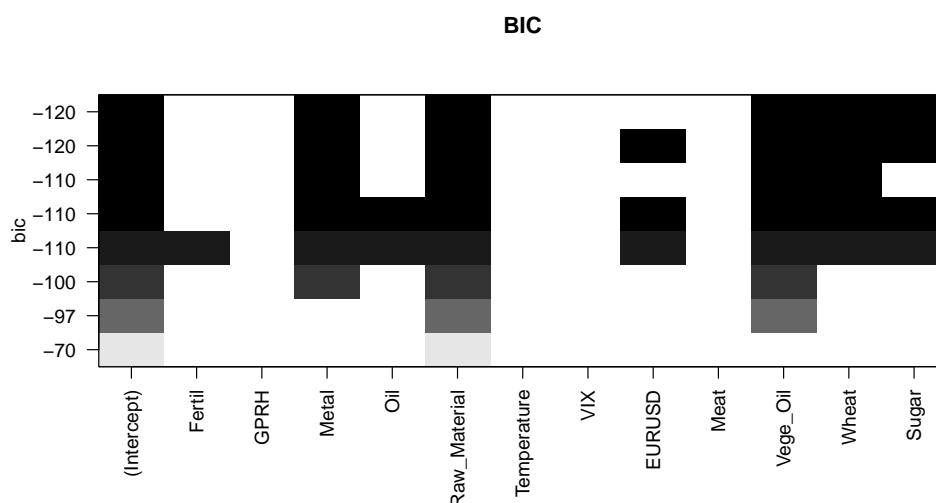
L'AIC est une mesure utilisée pour comparer différents modèles et prend en compte à la fois la qualité de l'ajustement du modèle et le nombre de paramètres utilisés. Un modèle avec un AIC plus faible est généralement préféré car il offre un bon compromis entre la qualité de l'ajustement et la complexité du modèle.

Enfin, le BIC est une mesure utilisée pour comparer différents modèles statistiques. Toutefois, le BIC pénalise davantage l'ajout de paramètres supplémentaires

par rapport à l'AIC. De ce fait, le BIC tend à favoriser des modèles plus simples (ou plus parcimonieux) que l'AIC. De même que pour l'AIC, un BIC plus faible indique un meilleur modèle. Nous privilégions ce dernier dans le choix de notre modèle.

À cet égard, l'analyse a conduit à la sélection de six variables comme étant les plus pertinentes, à savoir : **Metal, Raw Material, Vege Oil ,Wheat,Sugar**. Cette sélection est basée sur la maximisation du critère BIC, comme illustré sur la figure 15. Les autres critères de sélections sont visibles en annexe (4 et 5).

Figure 15 – **Sélection de variables avec le critère BIC**



4.1.2 Approche GETS

L'approche Gets, quant à elle, repose sur l'ajout ou la suppression séquentielle de variables (*autrement dit, une par une*) pour obtenir le meilleur sous-ensemble de variables possible. Dans notre étude, nous avons mis en place une boucle permettant de supprimer la variable la moins significative à chaque itération. En utilisant cette

approche, nous obtenons exactement le même modèle qu'avec la précédente méthode. Ce résultat confirme donc notre choix de variables.

Table 4 – **Sélection de variables avec GETS**

Variable	Coef.	SE	t-value	Pr(> t)
mconst	0.0726889	0.0968183	0.7508	0.4535855
ar1	0.4287796	0.0474120	9.0437	$< 2.2e - 16$ ***
Metal	0.0906157	0.0319542	2.8358	0.0049955 **
Raw Material	0.3975872	0.0742835	5.3523	$2.169e - 07$ ***
Vege Oil	0.1805647	0.0478488	3.7736	0.0002067 ***
Wheat	0.0105620	0.0025176	4.1953	$3.947e - 05$ ***
Sugar	0.2686450	0.0883730	3.0399	0.0026520 **

Pour information, nous avons essayé d'introduire une variable nommée "Food Lag", qui représente notre variable dépendante (*Fao Food Price Index*) décalée d'un mois. Cette variable a été utilisée dans une étude [20] précédente et semblait être la plus contributive à l'analyse, améliorant ainsi divers modèles. L'idée était de capter l'inertie de l'évolution du prix des denrées alimentaires. Cependant, dans notre cas, avec nos données, cette variable a malheureusement faussé les modèles de prévisions, certains affichant des valeurs de RMSE extrêmement élevées. Pour cette raison, nous avons choisi de retirer cette variable de notre étude afin d'avoir des prévisions plus précises.

Il est également important de souligner que nous avons tenté de décaler nos variables explicatives pour éviter tout problème d'endogénéité. En pratique, nous au-

rions prédit la valeur Y de février 1999 à l'aide des valeurs X_i de janvier 1999 pour nos variables explicatives. Malheureusement, cette stratégie n'a pas abouti aux résultats escomptés dans le cadre de notre étude. En effet, aucune de nos variables n'a réussi à passer les tests de sélection de variables. Nous avons donc retiré ce retard.

5 Prévvision & évaluation

Dans la section suivante, intitulée "Prévvision et évaluation", nous allons aborder deux aspects essentiels de notre étude. Dans la partie "Prévvision", nous explorerons chacun de nos modèles, en expliquant leur fonctionnement et en justifiant leur utilisation en nous basant sur les travaux existants issue de la littérature scientifique. La sous-section suivante, "Évaluation", impliquera l'application de divers indicateurs de qualité de prévvision pour évaluer et comparer les performances de nos modèles. De cette manière, nous serons en mesure de déterminer lequel se révèle être le plus efficace pour notre prévvision.

Comme il a été précédemment mentionné, les deux approches que nous avons mises en œuvre ont conduit à une sélection de six variables, que nous intégrerons dans différents modèles. Nous utiliserons dans un premier temps des modèles économétriques pour faire de la prévvision, puis dans un second temps, nous examinerons l'efficacité de modèles basés sur l'apprentissage automatique (*Machine-Learning ou ML*) pour déterminer s'ils surpassent les modèles plus conventionnels en termes de précision dans les prévvisions. Pour tous les modèles de notre analyse nous utilisons 80% des données pour l'entraînement et 20% pour le test.

De plus, l'échantillonnage des données sera non-aléatoire. Normalement, dans une approche de partitionnement de données pour l'entraînement et le test, il est courant d'utiliser un échantillonnage aléatoire pour éviter tout biais dans les modèles. En d'autres termes, chaque observation de l'ensemble de données complet a une chance égale d'être sélectionnée pour faire partie de l'ensemble d'entraînement ou de l'ensemble de test. Avec un échantillonnage non-aléatoire, cela signifie que nous utilisons

80% des "*premières*" données pour l'entraînement et 20% des "*dernières*" données pour le test. Cette approche est nécessaire pour stabiliser nos modèles, puisque avec une base de données de taille $n = 286$, un échantillonnage aléatoire apporte une grande variabilité dans les résultats des modèles de ML. En effet, lorsque le nombre de données est faible, ces derniers ont tendance à être plus sensibles aux fluctuations dans le changement de répartition des données, ce qui peut entraîner une variabilité dans les prévisions. Autrement dit, avec un échantillonnage aléatoire, un subtil changement dans la répartition des données allant dans la base d'entraînement ou test, va changer les prévisions réalisées ainsi que les valeurs des indicateurs de qualité de prévisions ($RMSE$, $CSSD$, R^2 , ...) des modèles de ML. De plus, avec cette méthode, la qualité de prévision se trouve grandement améliorée.

Cependant, cette approche peut introduire un biais dans nos modèles, car ils risquent d'être excessivement ajustés aux données d'entraînement, un phénomène connu sous le nom de sur apprentissage. Cela signifie qu'ils peuvent ne pas être capables de bien s'adapter aux données de test. Bien que dans notre situation spécifique, les prévisions soient nettement améliorées grâce à un échantillonnage non aléatoire, il est important de souligner que nos modèles pourraient rencontrer des difficultés à généraliser efficacement lorsqu'ils seront confrontés à de nouvelles données.

5.1 Prédiction

5.1.1 Modèles économétriques

Dans cette section, nous commencerons par utiliser les modèles économétriques suivants : ARX , $ARX-GETS$, $ARMAX$, LM ainsi qu'un modèle GAM . Nous passerons

ensuite aux modèles de Machine-Learning, à savoir : *MLP*, *MARS*, *SVM*, *Random Forest*, *XGB Boost*, *kNN*, *LSTM* et enfin *LSTM-CNN*. Nous utiliserons également un modèle de référence : un modèle autorégressif de période 1 (*AR1*) ainsi qu'un modèle *NAIVE*. Il convient dans un premier temps d'expliquer le fonctionnement de chaque modèle.

Modèle LM

Nous appliquons le modèle LM (**L**inear **M**odel ou *Modèle Linéaire*) est un modèle statistique qui établit une relation linéaire entre une variable dépendante et un ensemble de variables explicatives. Ce modèle figure parmi les plus simple et est couramment utilisé en statistique.

Ce dernier prend la forme $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \varepsilon$

où :

- Y est la variable dépendante (la variable à prédire),
- $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ sont les coefficients correspondant à chaque variable indépendante X_1, X_2, \dots, X_p ,
- ε représente le terme d'erreur, qui capture les résidus non expliqués par le modèle.

En appliquant ce modèle à nos six variables, nous obtenons un RMSE de 2,03.

Modèle AR(1) (: *forecast*)

Nous appliquons également un modèle autoregressif d'ordre 1 soit AR(1). (*résultats visibles en annexe : 20*) Dans ce modèle, la valeur actuelle d'une variable dépendante

est prédite en fonction de sa propre valeur retardée d'un pas de temps, avec un coefficient associé.

La forme générale d'un modèle AR(1) est la suivante :

$$— Y(t) = \beta_0 + \beta_1 \cdot Y(t-1) + \varepsilon(t)$$

où :

- $Y(t)$ est la variable dépendante à l'instant t .
- $Y(t-1)$ est la valeur retardée de la variable dépendante à l'instant $t-1$.
- β_0 est l'intercept (la valeur attendue de Y lorsque $Y(t-1)$ est nulle).
- β_1 est le coefficient du terme autorégressif, qui mesure l'impact de la valeur retardée $Y(t-1)$ sur la valeur actuelle $Y(t)$.
- $\varepsilon(t)$ est le terme d'erreur, qui capture les résidus non expliqués par le modèle.

Le modèle AR(1) est souvent utilisé comme modèle benchmark ou modèle de référence pour évaluer d'autres modèles plus complexes. Cela est dû à sa simplicité et à sa facilité d'interprétation. De plus, le modèle AR(1) peut capturer des motifs de dépendance temporelle simples et fournir une estimation de base pour la prévision des séries temporelles.

En utilisant un modèle AR(1) comme benchmark, on peut comparer la performance d'autres modèles plus sophistiqués ou complexes. Si un modèle plus complexe ne parvient pas à améliorer significativement les prévisions par rapport au modèle AR(1), cela suggère que le modèle plus complexe n'apporte pas nécessairement une valeur ajoutée significative. En revanche, si un modèle parvient à surpasser le modèle AR(1) en termes de performance prédictive, cela indique que le modèle plus complexe

est potentiellement meilleur pour modéliser et prédire la série temporelle étudiée. Nous utiliserons donc ce modèle comme modèle de référence à titre comparatif.

En appliquant ce modèle à nos six variables, nous obtenons un RMSE de 2,20.

Modèle ARX (: *forecast & Gets*)

Le modèle ARX ("**A**uto**R**egressive with **eX**ogenous inputs" ou *autorégressif avec des variables exogènes*) est une extension du modèle autorégressif (AR) qui ne prend en compte que les valeurs passées de la variable cible (*la variable Y*). Contrairement au modèle AR, le modèle ARX intègre des variables exogènes pour améliorer les prévisions de la variable cible. En incorporant ces variables exogènes, le modèle ARX tente de capturer les relations et les influences des facteurs externes sur la variable cible, ce qui peut conduire à des prévisions plus précises et plus fiables. Concrètement, le modèle prend la forme suivante :

$$Y(t) = \beta_0 + \beta_1 \cdot Y(t-1) + \beta_2 \cdot Y(t-2) + \dots + \beta_p \cdot Y(t-p) + \gamma_1 \cdot X_1(t) + \gamma_2 \cdot X_2(t) + \dots + \gamma_q \cdot X_q(t) + \varepsilon(t)$$

où :

- $Y(t)$ représente la variable cible à l'instant t .
- $Y(t-1), Y(t-2), \dots, Y(t-p)$ représentent les p valeurs passées de la variable cible incluse dans le modèle (termes autorégressifs).
- $X_1(t), X_2(t), \dots, X_q(t)$ représentent les q variables exogènes à l'instant t .
- $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ sont les coefficients des termes autorégressifs de la variable cible.
- $\gamma_1, \gamma_2, \dots, \gamma_q$ sont les coefficients des variables exogènes.
- $\varepsilon(t)$ est le terme d'erreur à l'instant t , qui capture les résidus non expliqués

par le modèle.

Nous utiliserons deux variantes, à savoir un modèle ARX avec la fonction `auto.arima()` ainsi qu'un modèle ARX issu du package `Gets`. La première version détermine automatiquement l'ordre de régression du modèle tandis que la deuxième version est une méthode où nous devons spécifier manuellement l'ordre autorégressif. La fonction `auto.arima()` a sélectionné un modèle ARX avec deux termes autorégressifs. Cela signifie que selon les critères utilisés par la fonction `auto.arima()` (*tels que l'AIC, l'AICc, le BIC, la log-vraisemblance, etc.*), un modèle ARX avec deux ordres de retard a été jugé le plus approprié pour ajuster les données et effectuer les prévisions (21), ce qui n'est pas le cas pour le modèle ARX `Get` dont l'ordre de régression est 1.(22) La première approche nous donne un RMSE de 1,97 tandis que la deuxième 1,71.

Modèle ARMAX (: *forecast*)

Le modèle ARMAX (**A**uto**R**egressive **M**oving **A**verage with **eX**ogenous inputs) est une extension du modèle ARMA (*AutoRegressive Moving Average*) qui inclut des variables exogènes pour améliorer la prévision de la variable cible. Dans un modèle ARMAX, la variable cible est modélisée à l'aide de termes autorégressifs (AR) et de termes de moyenne mobile (MA), tout en prenant en compte des variables exogènes qui peuvent influencer la variable cible. Les termes autorégressifs capturent les dépendances temporelles dans la variable cible, tandis que les termes de moyenne mobile captent les effets résiduels après avoir pris en compte les dépendances autorégressives. Autrement dit, la principale différence entre un modèle ARMAX et un modèle ARX réside dans l'inclusion de la composante de moyenne mobile (MA). Pour ce modèle, nous utilisons également la fonction `auto.arima()`, qui a déterminé un *ARIMA*(1,0,1),

ce qui signifie qu'il comprend un terme autorégressif d'ordre 1, aucun terme de différenciation et un terme de moyenne mobile d'ordre 1. Ce modèle a été choisi car il a été jugé le plus approprié pour capturer les dépendances temporelles et les tendances dans les données, en prenant en compte à la fois les valeurs passées de la variable cible et les variables exogènes. (23)

Nous avons utilisé les modèles précédents puisqu'ils sont directement issus du cours de "*Techniques de prévisions et conjonctures*" de M.Darne.[41] Néanmoins, les prochains modèles n'étant pas issus de ce cours, ils seront tous justifiés via la littérature scientifique.

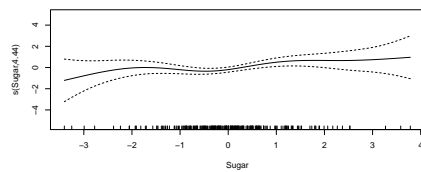
Modèle naïf (: *forecast*)

Le modèle naïf, aussi appelé le modèle de marche aléatoire est l'un des modèles les plus simples pour faire des prévisions de séries temporelles. L'idée de base du modèle naïf est que la meilleure prévision pour le futur est la valeur la plus récente. En d'autres termes, il suppose que la valeur à un instant t est la valeur à l'instant $t-1$. Dans le contexte des séries temporelles, cette approche est souvent appelée "*naïve*" parce qu'elle ne tient compte d'aucune des informations passées autres que la valeur la plus récente. Ce modèle est souvent utilisé comme modèle de référence.

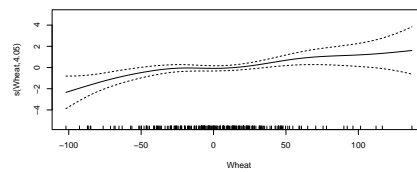
Modèle GAM (: *mgcv*)

Le modèle GAM (**G**eneralized **A**dditive **M**odel ou *modèle additif généralisé*) est une extension du modèle de régression linéaire (*LM*) qui permet de modéliser des relations non linéaires entre une variable dépendante et un ensemble de variables

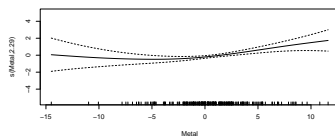
explicatives, tout en prenant en compte des *"effets lisses"* ou *"non paramétriques"*. Ce modèle offre donc une flexibilité plus importante dans l'utilisation, ce qui permet une modélisation plus précise et plus réaliste des données. L'idée principale derrière un modèle GAM est d'ajouter des termes lisses aux variables indépendantes dans le modèle. Ces *"termes lisses"* ou *"fonctions lisses"* permettent de capter les variations non linéaires des variables indépendantes. Autrement dit, elles permettent de modéliser des relations complexes, telles que des courbes, des pics ou des creux, entre les variables explicatives et la variable dépendante. Nous pouvons voir ces relations dans les graphiques ci-dessous.



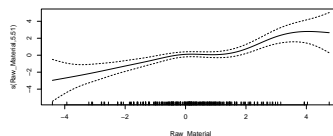
(a) Sucre



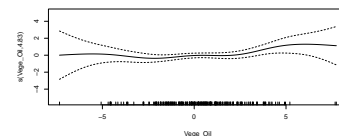
(b) Céréales



(c) Metal



(d) Matériaux bruts



(e) Huiles végétales

Figure 16 – **Fonctions lisses des variables du modèle GAM**

On devine alors que certaines variables ont des formes plus *"linéaires"* que d'autres (*on pense à Céréales et Matériaux bruts notamment*). En général, lorsque

le degré de liberté d'un terme lisse dans un modèle GAM est proche de 1, cela suggère une relation linéaire entre la variable indépendante correspondante et la variable dépendante. On remarque sans surprise que les variables citées précédemment ont effectivement une relation linéaire avec la variable dépendante (*Table 24*).

Les degrés de liberté estimés pour chaque prédicteur sont donnés dans les résultats du modèle (*Table 24*) : ils sont environ 2.29 pour *Metal*, 5.51 pour *Raw_Material*, 4.83 pour *Vege_Oil*, 4.05 pour *Wheat*, et 4.44 pour *Sugar*. Le nombre de degrés de liberté nous donne une indication de la complexité de la fonction lisse utilisée pour chaque prédicteur : un nombre plus élevé de degrés de liberté correspond à une fonction plus complexe (*et potentiellement plus flexible*). Au vu des graphiques précédents, il n'est pas surprenant de constater que les fonctions lisses les plus complexes semblent être celles des variables *Raw_Material* et *Vege_Oil* (*Matériaux brut et huiles végétales*).

Le modèle a un R^2 ajusté de 0.458, ce qui signifie qu'il explique environ 47.3% de la variation de la variable dépendante. Nous pouvons noter que certaines variables ne semblent pas être significatives (*Vege Oil et Sugar*) au seuil de 5%. Cela pourrait expliquer pourquoi nous obtenons un RMSE aussi élevé : 2,44.

5.1.2 Modèles de Machine-Learning Deep-Learning

Les modèles de machine learning (ML) sont des algorithmes et des techniques utilisés pour entraîner des ordinateurs à apprendre à partir de données et à effectuer des prévisions ou des tâches spécifiques sans être explicitement programmés. Ces modèles sont conçus pour extraire des informations et des modèles à partir des données

d'entraînement (*training*), puis les utiliser pour prendre des décisions, effectuer des prévisions ou résoudre des problèmes sur de nouvelles données (*test*).

Les modèles de machine learning sont alimentés par des données d'entrée (*variables indépendantes*) et leurs résultats attendus (*variables dépendantes*). Ces données sont utilisées pour entraîner le modèle en ajustant ses paramètres et en optimisant ses performances. Une fois que le modèle est entraîné, il peut être utilisé pour effectuer des prévisions ou des classifications sur de nouvelles données en se basant sur les modèles et les relations appris pendant la phase d'entraînement. Il existe une différence fondamentale entre ce type de modèles et les précédents qui étaient issu du domaine économétrique. En effet, les modèles économétriques sont souvent utilisés dans le domaine de l'économie et de la finance pour étudier les relations causales et les effets économiques. Leur objectif principal est souvent d'estimer les paramètres des équations économiques afin de comprendre les mécanismes économiques sous-jacents. En revanche, les modèles de machine learning sont plus axés sur la prévision et l'apprentissage à partir de données sans nécessairement se concentrer sur les relations causales. De plus, les modèles économétriques sont souvent basés sur des hypothèses économiques spécifiques et utilisent des spécifications fonctionnelles prédéfinies pour modéliser les relations entre les variables. Ce n'est pas le cas pour les modèles de machine learning, ce qui permet une plus grande flexibilité dans la modélisation des relations en utilisant des approches non paramétriques et en laissant le modèle apprendre les relations directement à partir des données.

Malgré le fait que les modèles de ML ne requièrent pas nécessairement des données nettoyées, corrigées et conformes à certaines hypothèses économétriques, nous opterons néanmoins pour les mêmes données que celles utilisées dans les mo-

dèles économétriques. Cette démarche vise à permettre une comparaison de la capacité prédictive des modèles en utilisant les mêmes ensembles de données. Autrement dit, les données fournis aux différents algorithmes de ML seront les mêmes données nettoyées, stationnarisée et corrigées de toute saisonnalité qui ont été fournies aux précédents modèles économétriques. Cependant, il convient de souligner que nous pourrions également fournir des données brutes afin d'alimenter certains modèles de ML.

Nous commencerons par utiliser des modèles de ML utilisés dans une étude [20] dont le but était également de prédire le prix des denrées alimentaires, mais également dans d'autre études similaires. En effet, les prévisions des denrées alimentaires n'est pas un sujet niche et a eu l'attention de nombreuses études scientifiques.

Modèle MLP (: *neuralnet*)

Le modèle MLP (**M**ulti-**L**ayer **P**erceptron) est un type de réseau neuronal artificiel organisé en plusieurs couches, très utilisé en dans le domaine du Machine-Learning. Ce dernier est composé de plusieurs couches de neurones, y compris une couche d'entrée, une ou plusieurs couches cachées et une couche de sortie. Chaque neurone dans une couche est connecté à tous les neurones de la couche suivante par des connexions pondérées. Chaque connexion est associée à un poids qui détermine l'importance de l'entrée pour la sortie du neurone. Pour commencer, les données sont introduites dans la couche d'entrée du réseau et se propagent jusqu'à la couche de sortie, en passant par toutes les couches intermédiaires. À chaque étape, la somme pondérée des entrées est calculée pour chaque neurone, et une fonction d'activation est ensuite appliquée à ce total pour obtenir la sortie du neurone. Une fois que les

données ont traversé tout le réseau, l'erreur de prévision est calculée.

Cette erreur est simplement la différence entre la sortie prédite par le réseau et la sortie réelle. Cette erreur est ensuite propagée en arrière (*Rétropagation de l'erreur* [42]) à travers le réseau, de la couche de sortie jusqu'à la couche d'entrée. L'objectif de cette étape est d'ajuster les poids des connexions en fonction de leur contribution à l'erreur totale. Pour chaque poids, la dérivée partielle de l'erreur par rapport à ce poids est calculée. Cette dérivée indique dans quelle mesure l'erreur changerait si ce poids était modifié de manière infinitésimale.

Enfin, après avoir calculé les dérivées partielles pour tous les poids, ces derniers sont mis à jour en les ajustant dans la direction qui minimise l'erreur. Typiquement, cela se fait en soustrayant la dérivée partielle (*multipliée par un taux d'apprentissage préfixé ou "learning rate"*) du poids actuel. Ces étapes sont répétées plusieurs fois (*sur plusieurs "époques" ou "epochs"*) jusqu'à ce que l'erreur de prévision soit suffisamment petite, ou jusqu'à ce que le nombre maximal d'époques soit atteint.

La rétropropagation est un processus mathématiquement défini. L'étape clé est le calcul du gradient de la fonction de coût par rapport aux poids du réseau, c'est-à-dire les dérivées partielles de la fonction de coût par rapport aux poids.

Disons que nous avons une fonction de coût $J(w)$ dépendant des poids w du réseau. La règle de mise à jour des poids en utilisant le gradient descend est la suivante :

$$w = w - \alpha \nabla J(w) \quad (1)$$

où α est le taux d'apprentissage et $\nabla J(w)$ est le gradient de $J(w)$.

L'étape clé de la rétropropagation est de calculer ce gradient. Si nous considérons une simple architecture de réseau avec une seule couche cachée, les dérivées partielles pour la mise à jour des poids peuvent être calculées comme suit [42] [43] [44] :

Pour la couche de sortie :

$$\delta_o = (y - \hat{y}) \cdot f'(z_o) \quad (2)$$

$$\Delta w_{oh} = \alpha \cdot \delta_o \cdot h \quad (3)$$

Pour la couche cachée :

$$\delta_h = \delta_o \cdot w_{oh} \cdot f'(z_h) \quad (4)$$

$$\Delta w_{ih} = \alpha \cdot \delta_h \cdot x \quad (5)$$

Dans ces formules :

- w sont les poids du réseau,
- α est le taux d'apprentissage
- $\nabla J(w)$ est le gradient de la fonction de coût J par rapport aux poids w

- y est la valeur cible
- \hat{y} est la sortie du réseau
- $f'(z)$ est la dérivée de la fonction d'activation évaluée à z
- δ est l'erreur
- h et x sont les sorties de la couche cachée et les entrées du réseau, respectivement
- Δw est le changement à apporter aux poids

La dérivée de la fonction d'activation $f'(z)$ est utilisée pour déterminer dans quelle mesure le neurone a contribué à l'erreur de sortie totale. Par exemple, si la dérivée est faible, cela signifie que la sortie du neurone est assez insensible aux changements de ses entrées, donc même si l'erreur est grande, ce neurone en particulier n'est probablement pas le principal coupable. Elle est également utilisée pour "*propager*" l'erreur à travers le réseau lors de la rétropropagation. Quand on calcule l'erreur pour chaque couche, on utilise la dérivée de la fonction d'activation pour déterminer comment diviser l'erreur entre les neurones de la couche précédente.

Autrement dit, plus le modèle est entraîné, plus il devient capable de reconnaître des schémas et de faire des prévisions précises. Cela signifie que notre nombre total de données ($n = 286$) peut être un frein à la capacité prédictive du modèle. Ce dernier est couramment utilisé dans de nombreux domaines de ML pour sa capacité à modéliser des relations complexes et à faire des prévision précises [20] [45].

Cette explication du fonctionnement du modèle nous permet d'expliquer le choix de nos hyper-paramètres. Comme pour tous les prochains modèles, nous avons utilisé la technique de "*recherche sur grille*". [46] Cette approche est très simple, nous

définissons une grille de valeurs d'hyperparamètres possibles, et le modèle est ensuite entraîné et évalué pour chaque combinaison de ces valeurs. La combinaison d'hyperparamètres qui donne la meilleure performance est alors sélectionnée comme la configuration optimale. Grâce à cette méthode, nous choisissons deux couches cachées, un *threshold* de 0.01 et un taux d'apprentissage de 0,01. Nous obtenons avec ce modèle un RMSE de 1,90.

Modèle MARS (: *earth*)

Le modèle MARS [47] (*Multivariate Adaptive Regression Splines*) est une méthode de ML qui combine les avantages des modèles linéaires et non linéaires. Il s'agit ici d'un modèle non paramétrique qui permet de modéliser les relations complexes entre les variables explicatives et notre variable dépendante. MARS est particulièrement bien adapté aux cas où les relations entre les variables sont non linéaires ou impliquent des interactions entre variables.

Ce modèle fonctionne via des "*splines*", qui sont des fonctions mathématiques utilisées pour la modélisation de données. Elles sont définies par morceaux, c'est-à-dire qu'elles sont composées de plusieurs fonctions plus simples, généralement des polynômes, définies sur différents intervalles de la variable indépendante. Elles sont également définies de manière à être aussi lisses que possible. La forme spécifique de la spline (*c'est-à-dire les points où les polynômes se rencontrent*) est déterminée par les données elles-mêmes plutôt que d'être spécifiée à l'avance. En effet, les fonctions sont sélectionnées par un algorithme de type "*forward*" et "*backward*". L'algorithme "*forward*" ajoute des fonctions de base au modèle une à la fois. Ensuite, l'algorithme

"*backward*" élimine les fonctions de base qui n'améliorent pas suffisamment la qualité de l'ajustement, en utilisant un critère comme le critère AIC.

Le modèle MARS va ensuite ajouter un par un les termes minimisant le plus possible les erreurs de prévisions, puis élimine ceux qui contribuent le moins à la prévision. Enfin, le modèle estime chaque coefficient. Ce dernier semble être largement utilisé pour prédire les denrées alimentaires [20] [48] [49] et nous permet d'obtenir un RMSE de 2.

Modèle SVM (*Re1071 & caret*)

Le modèle SVM (*Support Vector Machine*), est un modèle très populaire aujourd'hui en raison de sa capacité à gérer des problèmes de grande dimension et sa flexibilité dans la modélisation de divers types de données [20][50][51]. Il est important de noter qu'un modèle SVM est traditionnellement un algorithme de classification, cependant, avec quelques ajustements il peut également être utilisé pour des tâches de régression.[52] Il cherche à trouver une fonction qui soit capable de capturer au mieux les tendances sous-jacentes de nos données, avec une erreur de prévision aussi faible que possible. Les SVM utilisent ce qu'on appelle des "*fonctions noyau*" pour transformer les données dans un espace de plus grande dimension où elles sont plus faciles à modéliser. On peut également ajouter des hyper-paramètres tel que le paramètre *Cost* qui contrôle le compromis entre la complexité du modèle et la quantité d'erreurs que nous sommes prêt à accepter. Il y a également le paramètre Gamma qui contrôle la complexité de la fonction noyau.

Cette fois encore, nous utilisons la méthode recherche sur grille [46], ce qui nous permet de trouver les hyper-paramètres optimaux adaptés à nos données : un

gamma de 0,1 et un *cost* de 0,35 ce qui nous permet d'obtenir un RMSE de 1,96.

Modèle Random Forest (*randomForest*)

Le modèle Random Forest (RF) est également largement utilisé [20][53][54][55], et est basé sur l'idée d'agréger les prévisions de plusieurs arbres de décision indépendants pour obtenir une prévision finale plus précise et robuste. Premièrement on crée plusieurs arbres de décision. Chaque arbre est formé sur un sous-ensemble différent des données. Ces sous-ensembles sont créés en choisissant aléatoirement des échantillons de nos données avec remplacement (*c'est-à-dire qu'un même échantillon peut apparaître plusieurs fois dans le sous-ensemble : cette technique est appelée "bootstrap" [56]*). Chaque arbre est formé sur un sous-ensemble aléatoire des données d'entraînement et utilise un sous-ensemble aléatoire des caractéristiques à chaque division de l'arbre. Cela permet d'obtenir une variété d'arbres qui sont en quelque sorte décorrélés, ce qui augmente la robustesse du modèle final.

Pour faire une prévision avec un modèle Random Forest, le modèle fait une prévision avec chaque arbre individuellement. Les prévisions de tous les arbres sont ensuite moyennées pour obtenir la prévision finale. L'idée est que, en combinant les prévisions de nombreux arbres, nous pouvons obtenir une prévision finale qui est plus robuste et précise que celle de n'importe quel arbre individuel.

Il faut donc déclarer le nombre d'arbres à utiliser avant de spécifier notre modèle, et encore une fois nous utilisons la méthode recherche sur grille [46], ce qui nous permet de définir un nombre d'arbres de 200 pour obtenir un RMSE de 1,89.

Modèle XGB-Boost (: `xgboost`)

Le modèle XG Boost est un algorithme de Machine-Learning basé sur le principe de boosting d'arbres de décision. L'idée est de combiner les prévisions de nombreux modèles plus simples pour améliorer la robustesse et la précision globale. Le boosting est une méthode d'ensemble qui ajoute de nouveaux modèles de manière itérative, mais au lieu de donner à chaque modèle un vote égal, il pèse chaque modèle en fonction de sa performance. Les modèles qui effectuent des prévisions précises sont pondérés plus lourdement. Cette méthode est plus connue sous le nom de "*gradient boosting*", et l'idée est que chaque nouveau modèle ajouté à l'ensemble tente de corriger les erreurs faites par l'ensemble actuel de modèles. Chaque arbre est construit de manière séquentielle, ce qui signifie que chaque nouvel arbre dépend des arbres précédents.

Pour faire une prévision avec un modèle XGBoost, nous prenons la somme des prévisions de tous les arbres. Chaque prédiction individuelle est une prédiction des résidus, de sorte que la somme des prédictions donne une prédiction du résultat final. Une nouvelle fois nous utilisons la recherche sur grille pour optimiser nos hyper-paramètres. On trouve donc :

- **un taux d'apprentissage de 0,01** : paramètre d'atténuation qui est utilisé pour prévenir l'overfitting. A chaque étape, il réduit la contribution de chaque nouvel arbre ajouté par un facteur de 0.01
- **un gamma de 1** : C'est un paramètre de régularisation qui contrôle à quel point un arbre peut devenir complexe. Plus gamma est élevé, plus les arbres seront simples (c'est-à-dire qu'ils auront moins de divisions).

- **Une profondeur maximale de chaque arbre de 10** : plus cette valeur est élevée, plus les arbres peuvent devenir complexes.
- **Un min child weight de 5** : C'est un autre paramètre de régularisation qui contrôle la complexité des arbres
- **Un subsample = 0.5** : Il s'agit de la fraction d'exemples d'entraînement à utiliser pour chaque arbre. Un échantillon est tiré au hasard sans remise pour chaque arbre. Cela peut aider à prévenir l'overfitting en introduisant plus de variabilité dans les arbres.
- **colsample bytree = 1** : Il s'agit de la fraction de caractéristiques à utiliser pour chaque arbre. Une valeur de 1 signifie que toutes les caractéristiques sont utilisées pour chaque arbre. Comme pour subsample, cela peut aider à prévenir l'overfitting.
- **nrounds = 200** : C'est le nombre d'arbres à construire.

Avec ces paramètres, nous obtenons un RMSE de 1,89.

Modèle kNN (: *class*)

Le modèle k-Nearest Neighbors (kNN) est un algorithme couramment utilisé dans le domaine de la prévision. [20] [57] [58] Son principe repose sur l'approche des k plus proches voisins (ou "*k-nearest neighbors*").[59] L'idée de base derrière k-NN est que des points de données similaires auront probablement des résultats similaires. En d'autres termes, si nous avons une nouvelle observation pour laquelle nous voulons faire une prévision, nous la comparerons aux observations précédentes et nous prédirons qu'elle aura le même ou un résultat similaire à celui de ses voisins

les plus proches. Pour chaque nouvelle observation, l'algorithme calcule la distance euclidienne entre cette observation et toutes les autres observations de l'ensemble de données d'apprentissage, puis la nouvelle observation est généralement assignée à la moyenne des valeurs de ses k voisins. Le choix de k est un aspect important du modèle k -NN. Un k trop petit peut rendre le modèle sensible au bruit dans les données, tandis qu'un k trop grand peut le rendre insensible aux tendances locales.

Grâce à la recherche sur grille, nous trouvons $k = 8$, ce qui nous permet d'obtenir des prévisions avec un RMSE de 2,07.

Modèle LSTM (🧠 : *keras*)

Les LSTM (*Long Short-Term Memory*) sont un type spécial de réseaux de neurones récurrents (*RNN*) qui sont particulièrement efficaces pour apprendre des dépendances à long terme. Ils sont très utilisés pour la prévision des séries temporelles. [20] [60] [61] Un modèle LSTM est composé de cellules LSTM, qui sont les unités de base du réseau. Un modèle LSTM est composé de cellules de mémoire, qui ont la particularité de contenir trois "portes" : une porte d'entrée, une porte d'oubli et une porte de sortie. Ces portes permettent au LSTM de gérer et de maintenir un état interne, qui peut capturer des informations à long terme. Lors de l'entraînement du LSTM, chaque séquence d'entrée est passée à travers le réseau. À chaque pas de temps, la cellule LSTM décide de combien d'information de l'état précédent conserver / supprimer (*porte d'oubli*), de combien d'information du nouvel input ajouter (*porte d'entrée*), et de combien d'information transmettre à la sortie (*porte de sortie*). L'erreur de la prévision est ensuite propagée en arrière à travers le réseau (*rétropropagation*) pour mettre à jour les poids. Autrement dit :

1. **Porte d'entrée (Input Gate)** : décide quelle nouvelle information nous allons stocker dans l'état de la cellule
2. **Porte d'oubli (Forget Gate)** : décide quelle information nous allons jeter de l'état de la cellule. Elle examine l'entrée actuelle et l'état caché précédent (la sortie de la cellule LSTM au pas de temps précédent), et apprend à supprimer les informations inutiles ou non pertinentes de l'état de la cellule.
3. **Porte de sortie (Output Gate)** : décide quelle sera la nouvelle sortie basée sur l'état de la cellule mis à jour. Elle regarde l'entrée actuelle et l'état caché précédent, et crée une sortie pour le pas de temps actuel. Cette sortie sera utilisée comme état caché pour le pas de temps suivant.

Nous pouvons donc faire une prévision en prenant en réseau une séquence d'entrée (*qui peut être une séquence précédente de la série temporelle*), et le réseau produit une prévision pour le prochain pas de temps. Nous pouvons ensuite utiliser cette prévision comme une partie de l'entrée pour prédire le pas de temps suivant, et ainsi de suite. Pour alimenter un modèle LSTM, il faut dans un premier temps régler les hyper-paramètres, à savoir :

1. **Nombre d'unités dans les couches LSTM** : L'hyperparamètre qui contrôle le nombre d'unités ou de neurones dans les couches LSTM. Un nombre plus élevé d'unités permet généralement au modèle de capturer des motifs plus complexes dans les données, mais cela augmente également la complexité du modèle et peut entraîner un temps d'entraînement plus long.
2. **Nombre d'epochs** : L'hyperparamètre qui détermine le nombre d'itérations ou "*d'époques*" pendant lesquelles le modèle est entraîné sur les données. Une

valeur trop faible peut conduire à un sous-apprentissage, tandis qu'une valeur trop élevée peut conduire à un sur-apprentissage.

3. **Taille du lot (batch size)** : L'hyperparamètre qui spécifie le nombre d'échantillons d'entraînement utilisés pour mettre à jour les poids du modèle à chaque étape de l'entraînement. Une taille de lot plus grande peut accélérer l'entraînement, mais nécessite également plus de mémoire.

Comme pour les précédents modèles, nous avons utilisé la recherche sur grille pour optimiser les hyper-paramètres de la prévision. Dans notre cas nous avons donc un modèle à 50 unités de neurones, 100 epochs et une taille de lot s'élevant à 8. Les prévisions de ce modèle atteignent un RMSE de 1,87.

Modèle LSTM-CNN (🧩 : *tensorflow, sklearn, keras*)

Un modèle LSTM-CNN est une combinaison d'une architecture LSTM (*Long Short-Term Memory*) et d'une architecture CNN (*Convolutional Neural Network*). Il s'agit d'une approche hybride qui combine les avantages des deux architectures de réseaux de neurones profonds, et est couramment utilisée pour la modélisation de séries temporelles. [20] [62] Les CNN sont composés de couches de convolution qui peuvent apprendre à reconnaître des motifs locaux dans les données (*des tendances, des saisons*). Les couches de convolution sont le bloc de construction fondamental d'un CNN. Elles effectuent une opération appelée convolution [63], qui est une sorte de multiplication matricielle spéciale conçue pour traiter des données structurées en grille. L'avantage des couches de convolution par rapport aux couches entièrement connectées (*qui sont utilisées dans les réseaux de neurones traditionnels*) est qu'elles sont capables de reconnaître ces caractéristiques locales indépendamment de leur

position dans l'entrée. De plus, elles partagent les poids entre toutes les positions, ce qui permet de réduire le nombre de paramètres à apprendre.

Modèle TDNN (🧠 : *tensorflow, sklearn, keras*)

Un TDNN (*Time-Delay Neural Network*) est un type de réseau de neurones artificiel issu d'une version des réseaux de neurones à convolution (*CNN*). L'avantage du TDNN est qu'il cherche des caractéristiques dans différentes parties d'une séquence temporelle. Par exemple, cet algorithme peut être conçu pour détecter un pic dans une série temporelle, et il le ferait indépendamment du moment où ce pic se produit.

Pour ce faire, un TDNN utilise ce que l'on appelle des "*fenêtres de retard*" [64]. Une fenêtre de retard est un sous-ensemble de points de données consécutifs dans la série temporelle. Le modèle applique son filtre à différentes fenêtres de retard dans la série temporelle pour produire une nouvelle séquence de valeurs, qui peuvent ensuite être utilisées pour faire une prévision.

Un des principaux avantages des TDNN est qu'ils peuvent gérer des séries temporelles de longueurs différentes et qu'ils sont capables de reconnaître des caractéristiques temporelles indépendamment de leur position dans la série. Cela peut être très utile pour la prévision de séries temporelles, et est déjà très utilisé dans ce cadre-là. [20] [65] [66] [67] [68]

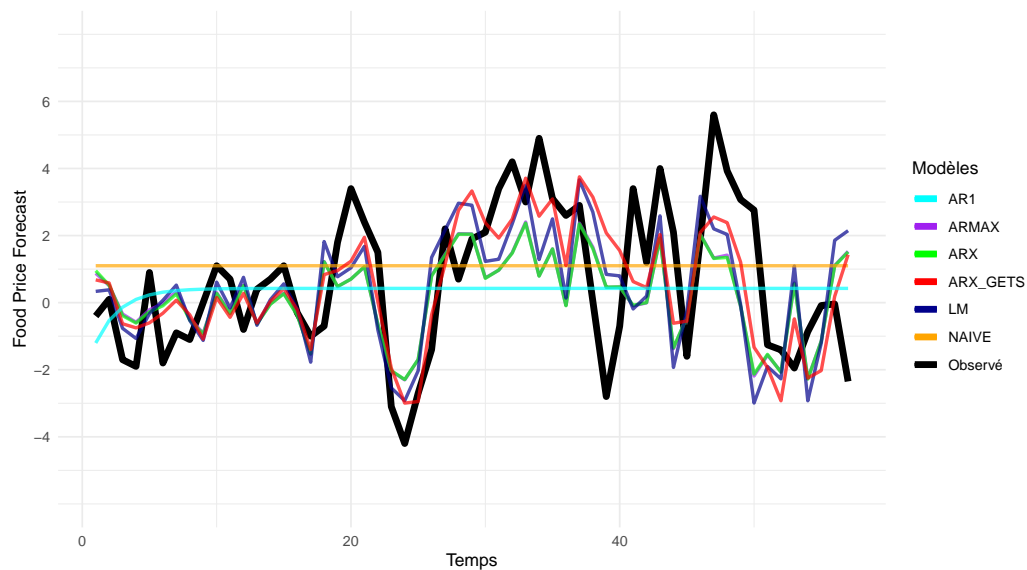
5.2 Prévisions Évaluation

5.2.1 Prévisions

Graphiques des modèles économétriques

Nous pouvons déjà présenter le graphique de prévisions des différents modèles économétriques. Nous analyserons le résultat de l'ensemble des modèles dans la section 5.2.2

Figure 17 – **Prévisions des modèles économétriques**



Il est intéressant de constater que parmi les modèles examinés, le modèle ARX Get semble fournir les meilleures prévisions en suivant de près la tendance des valeurs réelles. Ce n'est pas surprenant quand on sait qu'il est le modèle avec le RMSE le plus faible (1,71). On constate également que les modèle naïf et AR(1) font une ligne droite. Ce n'est pas étonnant puisque le modèle AR(1) est un modèle simple qui prédit la valeur à l'instant t comme une fonction linéaire de la valeur à l'instant t_{-1} . En effet, lorsque l'on fait une prévision multipériodes à l'aide d'un modèle AR(1), la prévision pour t_{+1} est basée sur l'observation à t , la prévision pour t_{+2} est basée sur la prévision pour t_{+1} , et ainsi de suite. Par conséquent, après quelques périodes, les prévisions tendent à converger vers une constante (*la moyenne de la série*), ce qui

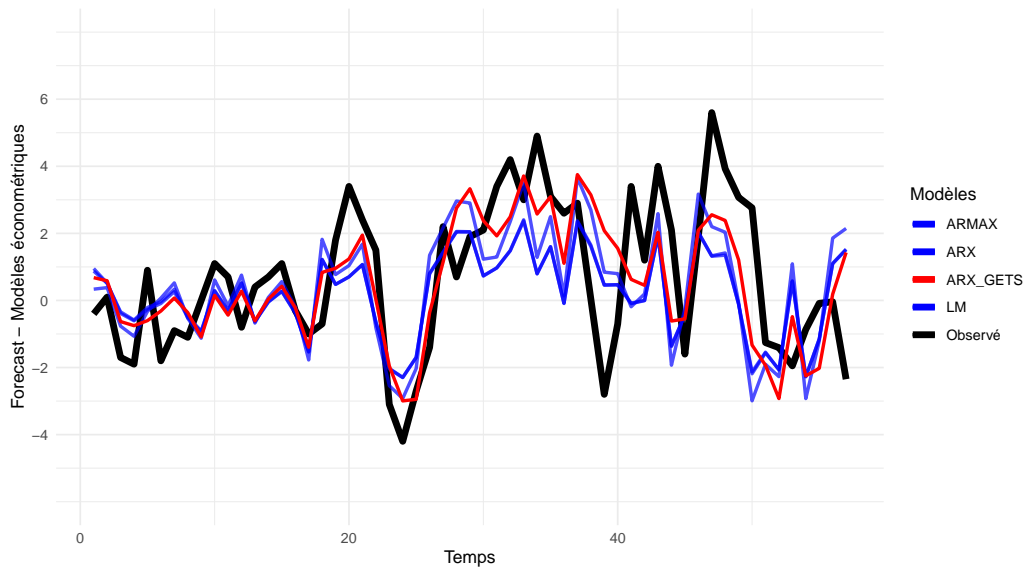
peut ressembler à une ligne droite, surtout si la valeur de l'autocorrélation est proche de 1. Cette explication se confirme quand on sait que la moyenne de "*Food*" (soit "*Observé*" sur le graphique) est de 0.667529.

Concernant le modèle naïf, cela n'est pas étonnant non plus puisque la prévision pour chaque période future est simplement la dernière observation dans les données d'entraînement. Par conséquent, si en faisant une prévision sur plusieurs périodes, nous obtenons une ligne horizontale droite à la valeur de la dernière observation.

Nous analyserons les indicateurs de qualité de chaque modèle dans la section [5.2.2](#). Concernant les modèles économétriques, il est important de souligner que nous n'avons pas vérifié toutes les hypothèses sous-jacentes à chacun d'entre eux. Certes, les données utilisées étaient corrigées des points atypiques, désaisonnalisées et stationnarisées mais il reste d'autres hypothèses importantes. Par exemple, certaines de ces hypothèses peuvent concerner la distribution normale des erreurs du modèle ou l'homoscédasticité (*c'est-à-dire, la constance de la variance des erreurs*). Si notre objectif était de comprendre les relations et les causalités entre différentes variables, il serait crucial de vérifier ces hypothèses. Cependant, l'objectif de cette étude est différent : nous cherchons principalement à construire le meilleur modèle de prévision possible.

À titre de comparaison, nous pouvons représenter les modèles économétriques (*en retirant le modèle AR(1) et Naïf qui ne sont que des lignes droites*) pour mettre en évidence le meilleur modèle parmi ces derniers : le modèle ARX Get, dont le RMSE est le plus faible.

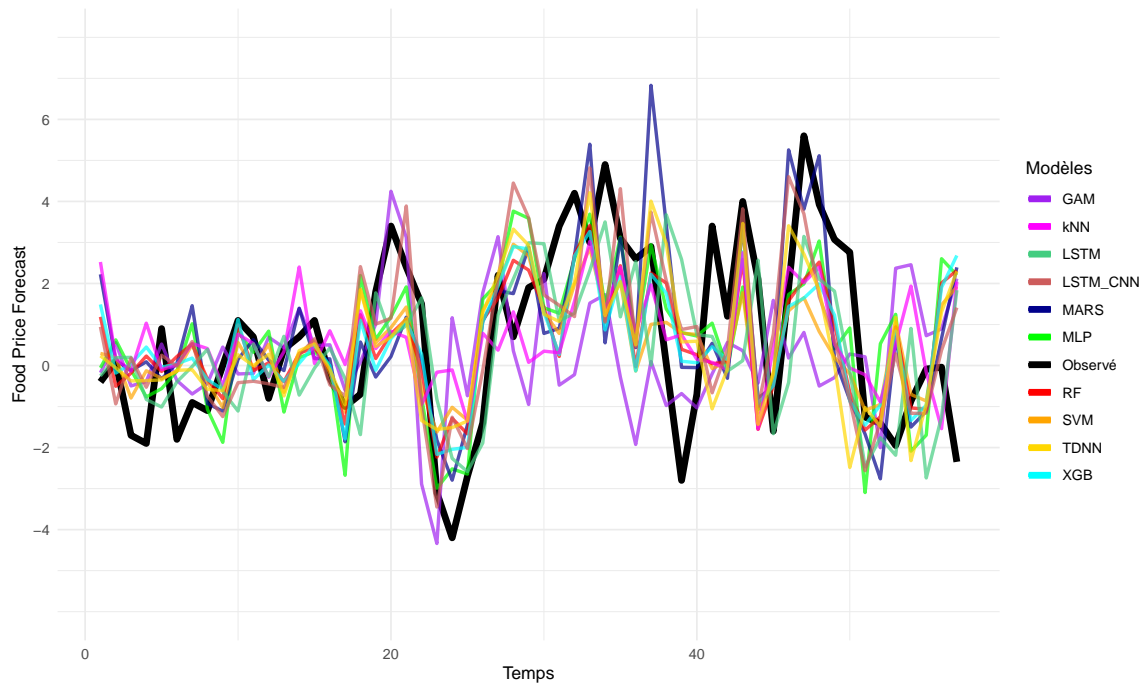
Figure 18 – **Prévisions des modèles économétriques - simplifié**



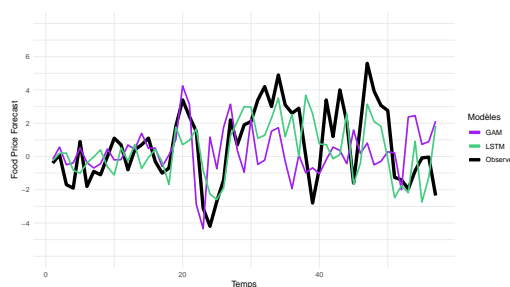
Graphiques des modèles ML

Nous pouvons déjà présenter le graphique de prévisions des différents modèles de machine-learning. Nous analyserons le résultat de l'ensemble des modèles dans la section [5.2.2](#)

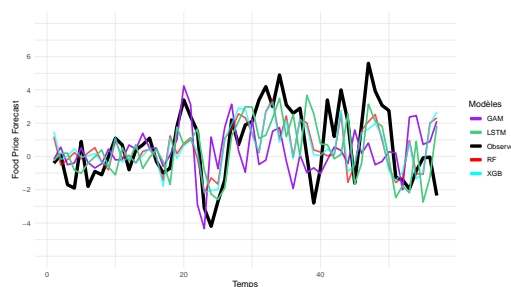
Figure 19 – Prévisions des modèles ML



Étant donné la multitude de modèles représentés sur ce graphique, une analyse précise se révèle complexe. Pour simplifier, nous pourrions représenter - *dans le graphique ci-dessous* - notre meilleur modèle d'apprentissage automatique, à savoir le modèle LSTM, qui obtient une valeur RMSE de 1,87. En contraste, notre modèle le moins performant, le modèle GAM, présente un RMSE de 2,44. Dans un second graphique, nous mettons en avant deux autres modèles performants de ML. Il s'agit du modèle XGB Boost, qui obtient une valeur RMSE de 1,8909, ainsi que du modèle Random Forest, dont le RMSE s'élève à 1,8951.



(a) Modèle GAM et Modèle LSTM



(b) Modèle GAM, LSTM, XGB et RF

On observe que le modèle LSTM, comme anticipé, épouse étroitement la courbe des valeurs réelles. Toutefois, il manque quelques pics anticipés. Le plus notable est celui du 40ème mois que le modèle LSTM avait prévu, mais qui, en réalité, ne s'est pas produit. Quant au modèle GAM, il semble performant au début de la période de prévision, mais après le 20ème mois, la précision de ses prévisions se dégrade. On peut imaginer qu'une hausse significative du CSPE (*Somme des Erreurs de prévision au Carré*) aura lieu à partir de ce point, car non seulement les prévisions s'éloignent des valeurs réelles, mais les pics prévus ne coïncident pas du tout.

Il est intéressant de constater que les prévisions des modèles XGB et RF sont étonnamment similaires, avec des tendances identiques et des pics prévus qui se ressemblent presque parfaitement. Cette similitude est surprenante compte tenu de la structure profondément distincte de ces deux modèles, comme décrit dans les sections 5.1.2 et 5.1.2. Dans l'ensemble, ces deux modèles suivent assez fidèlement les tendances et les pics prédits par le modèle LSTM, à l'exception d'un léger décalage temporel. En effet, les pics prévus par les modèles XGB et RF se produisent légèrement plus tôt que ceux prévus par le modèle LSTM.

5.2.2 MSE, CSSED, R² OOS

Le tableau 5 présente une comparaison des performances de divers modèles de séries temporelles en se basant sur plusieurs critères : la Mean Squared Error (MSE), le Root Mean Squared Error (RMSE), la somme cumulative des différences des erreurs quadratiques (CSSED) et le coefficient de détermination hors échantillon (R² OOS).

- La **MSE** est une mesure qui évalue la qualité d'un estimateur ou d'un prédicteur. Elle mesure l'espérance des carrés des erreurs ou des écarts. C'est une mesure de risque quadratique. Plus la MSE est faible, plus l'estimation ou la prévision est précise.
- Le **RMSE** est une autre métrique d'évaluation des modèles. Elle est simplement la racine carrée de la MSE. Le RMSE a la même unité que la quantité d'intérêt, ce qui permet une interprétation plus facile. Tout comme la MSE, un RMSE plus faible indique une meilleure performance du modèle.
- La **CSSED** est une mesure du cumul des erreurs. Elle donne une idée de l'accumulation des erreurs sur l'ensemble de la période d'estimation ou de prévision. Une CSSED plus faible signifie que le modèle est plus précis sur l'ensemble de la période d'intérêt.
- Le **R² OOS** est une version de la statistique R² qui s'applique lorsque nous faisons des prévisions sur des données non utilisées pour entraîner le modèle. C'est une mesure de la qualité de la prévision. Un R² OOS plus élevé indique que le modèle explique une plus grande proportion de la variabilité dans les données hors échantillon.

En résumé, ces critères permettent d'évaluer la performance des modèles sur

plusieurs dimensions : la précision des prévisions, l'accumulation des erreurs et la qualité de la prévision sur des données hors échantillon. Ils sont essentiels pour comparer les modèles et choisir le plus adapté à nos données et à notre objectif de prévision.

Table 5 – **Tableau des MSE, CSSED, RMSE et R² OOS**

Type	Modèle	MSE	RMSE	CSSED	R ² OOS
Modèles économétriques	ARMAX	3.93	1.98	103.28	0.19
	ARX	3.92	1.98	105.45	0.20
	LM	4.15	2.04	57.10	0.15
	NAIVE	5.07	2.25	163.89	-0.04
	ARX GETS	2.94	1.72	38.04	0.40
	AR1	4.87	2.21	96.83	0.00
	GAM	5.97	2.44	121.04	-0.23
Modèles Machine-Learning	MLP	3.63	1.90	40.90	0.26
	MARS	4.00	2.00	42.62	0.18
	SVM	3.87	1.97	65.82	0.20
	RF	3.59	1.90	58.45	0.26
	XGB	3.58	1.89	62.28	0.27
	kNN	4.29	2.07	72.05	0.12
	LSTM	3.51	1.87	61.48	0.28
	LSTM CNN	3.51	1.96	45.31	0.20
	TDNN	4.28	2.07	43.69	0.11

Dans ce récapitulatif, nous avons mis en évidence les trois modèles les plus performants de notre étude (ARX Gets, XGB, LSTM) en les marquant en vert. Cette

synthèse offre un aperçu global de la performance de tous nos modèles et révèle plusieurs points d'intérêt.

En premier lieu, si nous nous penchons sur le RMSE, les modèles de Machine Learning (ML) semblent surpasser leurs homologues économétriques. Pour la plupart, ces derniers affichent un RMSE inférieur à 2, contrairement aux modèles économétriques où ce chiffre est franchi pour quatre d'entre eux. Cependant, malgré une performance moyenne en RMSE plus favorable aux modèles de ML, le modèle le plus performant se trouve parmi les modèles économétriques. En effet, le modèle ARX Gets demeure notre meilleur outil de prévision avec un RMSE de seulement 1,72. Le seul modèle qui se rapproche de cette performance est le LSTM, avec un RMSE de 1,87, une valeur qui reste, nous en conviendrons, significativement supérieure à 1,72.

De façon surprenante, le CSSED ne présente pas la même hiérarchie de performance que le RMSE. Hormis pour le modèle ARX Gets, qui conserve la première place, les modèles XGB et LSTM ne figurent pas parmi ceux avec le CSSED le plus bas. En réalité, ce sont les modèles MLP et MARS qui semblent accumuler le moins d'erreurs sur la période étudiée. Ce constat est particulièrement surprenant lorsque nous superposons ces modèles sur un même graphique (*voir 6*). Effectivement, les modèles MLP et MARS (*et particulièrement MARS*) ne donnent pas l'impression de mieux suivre la tendance ou de mieux prédire les pics que XGB et LSTM. Au contraire, l'inspection visuelle du graphique pourrait laisser penser que les modèles MLP et MARS engendrent davantage d'erreurs.

Enfin, le R^2_{OOS} nous renvoie aux mêmes résultats que le MSE et RMSE : les meilleurs modèles sont ARX Gets, XGB et LSTM. En effet, les R^2_{OOS} les plus élevés

sont ceux de ces modèles, ce qui montre donc que ces derniers ont la meilleure qualité de prévisions en permettant d'expliquer une plus grande proportion de la variance dans les données hors échantillon.

Enfin, il est important de noter que les graphiques peuvent parfois être trompeurs. Même si un modèle semble graphiquement faire moins d'erreurs, il est possible qu'il accumule plus d'erreurs sur l'ensemble des données, d'où l'importance de mesurer quantitativement la performance des modèles.

En ce qui concerne les différences entre les indicateurs, le CSSED et le R^2_{OOS} évaluent des aspects différents de la performance du modèle. Le CSSED prend en compte l'accumulation des erreurs, tandis que le R^2_{OOS} mesure la capacité du modèle à expliquer la variabilité des données qui n'ont pas été utilisées pendant l'entraînement.

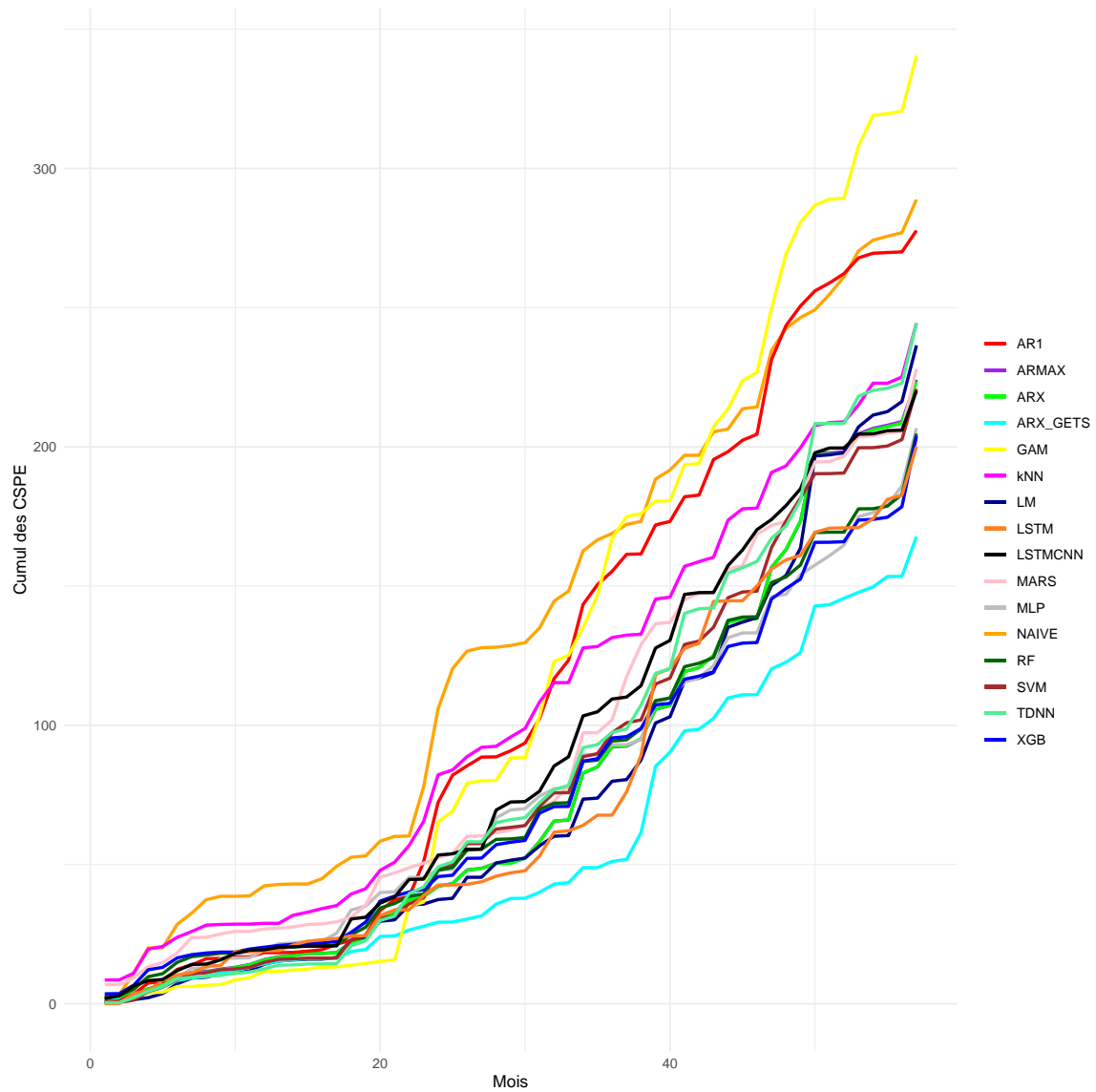
Cela pourrait expliquer pourquoi les modèles MLP et RF se démarquent en termes de CSSED : même si ces modèles peuvent accumuler plus d'erreurs au fil du temps, ils pourraient être plus stables et donc accumuler moins de grandes erreurs, ce qui se traduit par un CSSED plus faible. D'autre part, les modèles ARX Gets, LSTM et XGB pourraient être capables d'expliquer une plus grande variabilité des données, ce qui se traduit par un meilleur R^2_{OOS} .

5.2.3 Erreurs de prévision cumulées au carré (CSPE)

La Figure 21 illustre l'évolution des erreurs de prévision cumulées au carré (CSPE) sur une période mensuelle, permettant ainsi d'évaluer la stabilité de la précision des prévisions pour chaque modèle. Nous avons précédemment vu les capacités de chaque modèle, via les graphiques de prévisions mais également le tableau 5 avec

ces différents indicateurs. Cependant il manque encore un indicateur important, le CSPE nous permettant de voir quel modèle accumule le moins d'erreur au fil du temps. Autrement dit, cet indicateur illustre la stabilité / qualité de prévisions des modèles sur le long terme.

Figure 21 – Erreurs de prévision cumulées au carré (CSPE)



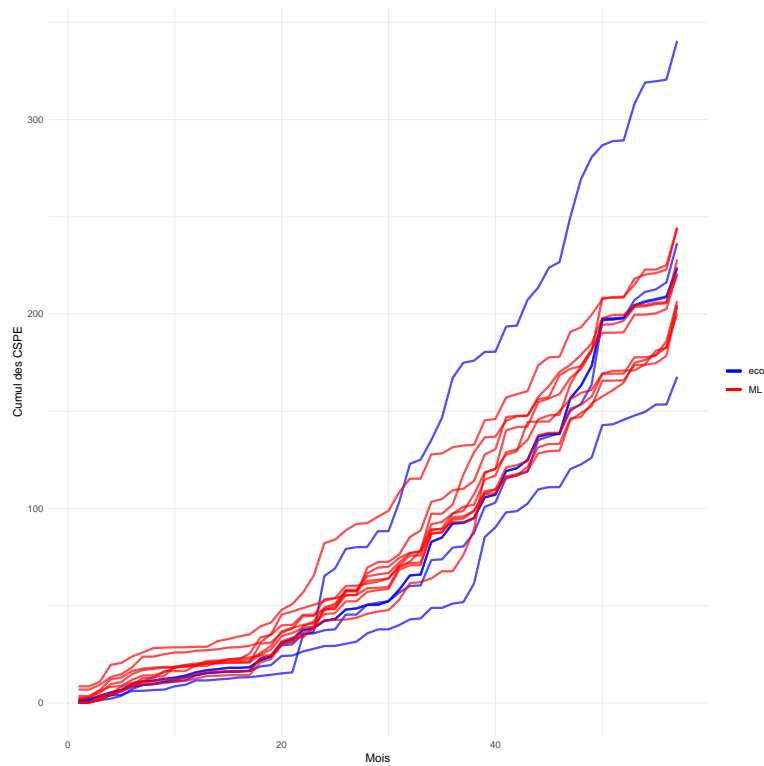
Cette représentation graphique souligne plusieurs points d'intérêt. Premièrement, elle confirme nos suppositions initiales au sujet du modèle GAM. Comme nous

l'avions observé précédemment (voir [20a](#)), ce modèle semblait produire des prévisions d'excellente qualité durant les 20 premiers mois, pour ensuite perdre en précision. L'indicateur CSPE semble corroborer cette tendance. En effet, pendant les 20 premiers mois, l'erreur quadratique cumulée du modèle GAM est la plus basse de tous les modèles. Cependant, à partir du 21ème mois, cette erreur cumulée commence à s'accroître de manière significative, atteignant finalement le niveau d'erreur cumulée le plus élevé parmi tous les modèles examinés.

Après le 20ème mois c'est donc le modèle ARX GETS qui semble prendre le relais, devenant le modèle accumulant le moins d'erreurs, et il le restera jusqu'à la fin de notre étude. On peut soupçonner un sur-ajustement aux données d'entraînement, mais il est également possible que les données aient subi un changement structurel ou une rupture après les 20 premiers mois, ce qui pourrait expliquer pourquoi le modèle GAM performe bien initialement puis perd de sa précision.

Il est également intéressant de constater que les modèles de ML semblent en moyenne engendrer moins d'erreurs sur la durée (*voir annexe 7 pour plus de clarté*). On pourrait supposer qu'avec un volume de données plus conséquent, les modèles de ML pourraient potentiellement devenir ceux qui commettent le moins d'erreurs, étant donné qu'ils génèrent des prévisions plus stables sur le long terme.

Figure 22 – **CSPE - Modèles économétriques & Modèles ML**



Ce graphique nous permet d'observer que, sur la durée, les modèles de ML tendent à surpasser légèrement les modèles économétriques. Il est important de souligner que nous avons écarté les modèles NAIVE et AR(1), considérés comme des modèles de référence et non destinés à fournir des prévisions très précises. Il semble donc que, à long terme, les modèles de ML génèrent moins d'erreurs, cependant, le modèle ARX GETS demeure le meilleur, comme nous pouvons le constater clairement.

5.2.4 Test de Diebold-Mariano

Le test de Diebold-Mariano est une méthode statistique permettant de comparer les performances prédictives de deux modèles de prévision. Dans notre cas, nous utilisons l'option *alternative = "less"* lors de l'utilisation de la fonction *dm.test* permettant de réaliser le test de Diebold Mariano. Cela signifie que nous testons si les erreurs de prévision du modèle AR1 sont inférieures à celles du modèle comparé.

Le test consiste à calculer la p-value, qui permet d'évaluer la force de la preuve contre l'hypothèse nulle. Il compare les erreurs de prévision des deux modèles en calculant la différence entre elles, puis en vérifiant si l'espérance de cette différence est statistiquement différente de zéro. Si c'est le cas, on peut conclure que les deux modèles ont une précision de prévision statistiquement différente.

Pour le dire plus simplement :

- **Si $p < 0.05$** , nous rejetons H_0 et acceptons H_1 , c'est-à-dire que nous concluons que les erreurs de prévision du modèle AR1 sont significativement inférieures à celles du modèle comparé, ce qui signifie que le modèle AR1 est meilleur.
- **Si $p > 0.05$** , nous ne rejetons pas H_0 , c'est-à-dire que nous concluons que les erreurs de prévision du modèle AR1 et du modèle comparé sont équivalentes, ce qui signifie que les modèles ont la même qualité de prévision.

Table 6 – Test de précision de Diebold-Mariano pour l'ensemble des modèles

ARX	0.9215	NAIVE	0.3186	SVM	0.9566	LSTM	0.947
ARX GET	0.1569	GAM	0.06984	RF	0.9658	LSTM CNN	0.9029
ARMAX	0.9161	MLP	0.9461	XGB	0.965	TDNN	0.7513
LM	0.7824	MARS	0.8462	kNN	0.8286		

Les résultats du tableau 6 suggèrent qu'à l'exception du modèle GAM, l'hypothèse nulle n'est pas rejetée pour tous les autres modèles. En termes de qualité de prévision, cela signifie que ces modèles ne présentent pas de différence significative par rapport au modèle AR1. Il est important de noter que lorsque nous disons que les modèles ont des qualités de prévision "similaires", cela ne signifie pas nécessairement qu'ils sont équivalents en termes de performance. Cela signifie simplement que, d'après les données dont nous disposons et les hypothèses du test de Diebold-Mariano, nous ne pouvons pas conclure de manière statistiquement significative qu'un modèle est supérieur à l'autre.

D'autre part, pour le modèle GAM, l'hypothèse alternative est acceptée, indiquant que la qualité de prévision du modèle GAM est significativement inférieure à celle du modèle AR1. C'est le seul modèle pour lequel nous pouvons conclure avec une certaine confiance statistique qu'il est moins performant que le modèle AR1 en termes de qualité de prévision.

Ces résultats sont en accord avec nos observations précédentes, comme illustré dans les graphiques 8 et 21. Nous avons déjà constaté que le modèle GAM affiche d'excellentes performances durant les 20 premiers mois, avant de connaître une baisse

significative de sa qualité de prévision. On observe une augmentation correspondante du CSPE à partir de ce moment-là. Il n'est donc pas surprenant que les erreurs de prévision du modèle AR1 soient statistiquement plus faibles que celles du modèle GAM. Cette situation met en évidence l'importance de la constance dans les performances de prévision, car une forte performance initiale peut être compromise par des résultats ultérieurs moins satisfaisants.

6 Conclusion & Discussion

Notre étude a été menée avec l'objectif d'établir des modèles prédictifs pour l'indice des prix des denrées alimentaires de la FAO. À cet effet, nous avons sélectionné dix variables mensuelles de janvier 1999 à décembre 2022, justifiées par une revue de la littérature scientifique. Ces variables ont été minutieusement traitées (*correction des valeurs aberrantes, désaisonnalisation, stationnarisation*) pour assurer leur compatibilité avec nos modèles. L'analyse exploratoire qui a précédé était une étape nécessaire pour comprendre et préparer nos variables avant d'appliquer les modèles prédictifs. Nos méthodes de sélection de variables ont réduit leur nombre à six, nous permettant de prédire l'indice des prix des denrées alimentaires pour les 57 derniers mois de notre période d'étude. Pour assurer une comparaison adéquate, tous les modèles ont été formés avec la même répartition des données : 80% pour l'apprentissage et 20% pour les tests, avec un échantillonnage non-aléatoire. Après avoir expliqué le fonctionnement de chaque modèle, nous avons pu les appliquer à notre sélection de variables.

Cette étude nous aura permis de mettre en lumière un élément important : les modèles les plus sophistiqués ne sont pas nécessairement les plus performants en matière de prévision. En fait, malgré l'application de plusieurs modèles de ML avancés, notre modèle ARX GET s'est avéré être le meilleur, avec la plus faible RMSE et le plus faible CSPE. De plus, ce modèle présente la plus faible CSSD et le R^2_{OOS} le plus élevé. Selon tous nos critères de qualité, ce modèle est notre meilleur outil de prévision. La conclusion de notre étude n'est pas surprenante car il est bien établi que les modèles les plus complexes ne sont pas systématiquement les plus efficaces, la qualité des prévisions reposant largement sur les données utilisées. Certains modèles peuvent

être - *par exemple* - mieux adapté à des données journalières, volatiles ou encore à des données sans saisonnalité. Cependant, nous pouvons noter qu'en moyenne les modèles de ML semblaient plus performants que les modèles économétriques plus "*classiques*". De plus, ces derniers ne nécessitent pas tout le temps de vérification d'hypothèses avant application.

À ce sujet, il y a certaines contraintes inhérentes à notre approche. En effet, nous n'avons pas vérifié les hypothèses propres à chaque modèle économétrique, sauf celle de la stationnarité. Cela est dû au fait que notre objectif principal était de construire le modèle de prévision le plus efficace, plutôt que d'explorer les liens de causalité entre différentes variables par leurs coefficients. Cependant, cette négligence peut entraîner une distorsion des résultats de nos modèles, pouvant altérer la précision de nos prévisions. D'abord, l'intégration de décalages temporels dans nos variables explicatives aurait peut-être été judicieuse (*par exemple, utiliser les valeurs de janvier pour prédire l'indice "Food" de février*). Toutefois, après l'application d'un tel décalage, nos variables ne passaient plus les tests de sélection.

En ce qui concerne les modèles de machine learning, nous n'avons qu'effleuré leur potentiel prédictif. En effet, ces modèles nécessitent une compréhension bien plus approfondie de leurs mécanismes et une optimisation soignée de leurs hyperparamètres. Chaque modèle doit être finement ajusté pour maximiser ses performances prédictives, ce qui peut exiger des ressources de calcul conséquentes, une expertise technique pointue ainsi qu'une grande quantité de données. Plus la taille du jeu de données est grande, plus les modèles de machine learning peuvent extraire et comprendre les tendances et les schémas complexes dans les données historiques, améliorant ainsi leur capacité à réaliser des prédictions précises et robustes sur de nou-

velles données. Habituellement, ces modèles sont déployés sur des jeux de données beaucoup plus volumineux et, dans notre cas, le nombre limité de données (*286 en tout*) a potentiellement limité leur performance. Cependant, nous tenons à souligner à nouveau notre manque d'expertise flagrant dans ce domaine spécifique. Nous ne prétendons pas avoir une maîtrise parfaite de ces modèles, l'objectif de cette étude était simplement de les expérimenter afin d'essayer de construire des modèles de prévision aussi performants que possible, selon nos capacités.

Enfin, bien que nous aurions préféré disposer de plus de données et de plus de variables pour notre analyse, l'accessibilité des données recherchées a été un obstacle. Il est difficile de trouver gratuitement des données mensuelles qui couvrent la période de janvier 1999 à décembre 2022. Par exemple, nous aurions aimé avoir des données sur les précipitations mondiales (*qui pourraient influencer les rendements des récoltes*), des données sur d'autres Futures à terme, ou des données sur le prix mondial de l'eau, mais ces informations ne sont actuellement pas disponibles.

La conclusion majeure de notre analyse est que le modèle ARX Get se distingue comme le modèle de prévision le plus performant. Toutefois, les modèles de machine learning LSTM et XGB Boost le suivent de très près en termes de performance.

Dans l'optique d'améliorer encore nos modèles, nous suggérons d'exploiter un ensemble de données plus large pour tirer pleinement parti des modèles de Machine Learning, et de développer une compréhension plus approfondie de ces modèles. Il serait aussi bénéfique d'employer une plus grande variété de modèles économétriques et de peaufiner certains d'entre eux - en particulier le modèle GAM, qui a donné d'excellentes prévisions sur les vingt premiers mois. L'intégration de données environ-

nementales supplémentaires serait aussi précieuse, compte tenu du contexte environnemental actuel qui se détériore et qui est susceptible d'affecter de plus en plus le prix des denrées alimentaires à l'échelle mondiale.

Références

- [1] C. Emsden, "[L'Indice FAO des prix des produits alimentaires atteint un niveau record en février](#)," 2022.
- [2] J. Clapp and M. J. Cohen, "[The global food crisis: Governance challenges and opportunities](#)," 2009.
- [3] L. B. Mondiale, "[LE POINT SUR LA SÉCURITE ALIMENTAIRE](#)."
- [4] EuroStat, "[Indice des prix de Laspeyres](#)."
- [5] M. Alghalith, "[The interaction between food prices and oil prices](#)," 2010.
- [6] M. Roman, A. Górecka, and J. Domagała, "[The linkages between crude oil and food prices. Energies, 13 \(24\), 6545](#)," 2020.
- [7] M. S. Khan, "[The 2008 Oil Price "Bubble"](#)," Policy Briefs PB09-19, Peterson Institute for International Economics, Aug. 2009.
- [8] P. Davidson, "[Crude Oil Prices: \"Market Fundamentals\" or Speculation?](#)," *Challenge*, vol. 51, no. 4, pp. 110–118, 2008.
- [9] J. Peng, Z. Li, and B. M. Drakeford, "[Dynamic Characteristics of Crude Oil Price Fluctuation—From the Perspective of Crude Oil Price Influence Mechanism](#)," *Energies*, vol. 13, no. 17, 2020.
- [10] C. Baumeister and L. Kilian, "[Understanding the Decline in the Price of Oil since June 2014](#)," *Journal of the Association of Environmental and resource economists*, vol. 3, no. 1, pp. 131–158, 2016.

- [11] M. I. Khan, T. Yasmeen, A. Shakoor, N. B. Khan, and R. Muhammad, “[2014 oil plunge: Causes and impacts on renewable energy](#),” *Renewable and Sustainable Energy Reviews*, vol. 68, pp. 609–622, 2017.
- [12] N. Devpura and P. K. Narayan, “[Hourly oil price volatility: The role of COVID-19](#),” *Energy Research Letters*, vol. 1, no. 2, 2020.
- [13] I. Appiah-Otoo, “Russia–ukraine war and us oil prices,” *Energy Research Letters*, vol. 3, no. Early View, 2022.
- [14] G. Çınar and A. Uzmay, “[Does fear \(VIX index\) incite volatility in food prices?](#),” *International Journal of Food and Agricultural Economics (IJFAEC)*, vol. 5, no. 1128-2018-069, pp. 69–78, 2017.
- [15] E. Sentana, “Volatility, diversification and contagion,” 2018.
- [16] C. de l’Union Européenne, “[Impact de l’invasion de l’Ukraine par la Russie sur les marchés: réaction de l’UE](#).”
- [17] OCDE, “[Effets de l’agression russe contre l’Ukraine sur les marchés agricoles et conséquences pour l’action publique](#).”
- [18] “[How world uncertainties and global pandemics destabilized food, energy and stock markets? Fresh evidence from quantile on quantile regressions](#), author=Chowdhury, Mohammad Ashraful Ferdous and Meo, Muhammad Saeed and Aloui, Chaker, journal=International Review of Financial Analysis,” vol. 76, p. 101759, 2021.
- [19] F. Saâdaoui, S. B. Jabeur, and J. W. Goodell, “[Causality of geopolitical risk](#)

- on food prices: Considering the Russo–Ukrainian conflict,” *Finance Research Letters*, vol. 49, p. 103103, 2022.
- [20] T. Ulussever, H. M. Ertuğrul, S. Kılıç Depren, M. T. Kartal, and Ö. Depren, “Estimation of impacts of global factors on world food prices: a comparison of machine learning algorithms and time series econometric models,” *Foods*, vol. 12, no. 4, p. 873, 2023.
- [21] O. Herve, “Fertilizer markets and their interplay with commodity and food prices,” *Research Papers in Economics*, p. 66, 2012.
- [22] E. Woertz, E. Soler, O. Farrés, and A. Busquets, “The Impact of Food Price Volatility and Food Inflation on Southern and Eastern Mediterranean Countries,” 2014.
- [23] J. P. Abraham, M. Baringer, N. L. Bindoff, T. Boyer, L. Cheng, J. A. Church, J. L. Conroy, C. M. Domingues, J. T. Fasullo, J. Gilson, *et al.*, “A review of global ocean temperature observations: Implications for ocean heat content estimates and climate change,” *Reviews of Geophysics*, vol. 51, no. 3, pp. 450–483, 2013.
- [24] s. f. b. s. l. c. c. Auteurs de Climate.be, “Changement Climatique - Conséquences sur l'agriculture.”
- [25] D. of Agricultural and C. E. U. of Illinois, “Basic Facts about Food Price Inflation in the U.S.).”
- [26] A. Mittal *et al.*, *The 2008 food price crisis: rethinking food security policies*. UN, 2009.

- [27] J. Ghosh, J. Heintz, and R. Pollin, "[Speculation on commodities futures markets and destabilization of global food prices: exploring the connections](#)," *International Journal of Health Services*, vol. 42, no. 3, pp. 465–483, 2012.
- [28] P. Wahl, "[Food speculation: The main factor of the price bubble in 2008](#)," *Briefing Paper*, 2009.
- [29] M. Huchet-Bourdon, "[Agricultural commodity price volatility: An overview](#)," 2011.
- [30] M. F. France Bleu, "[ENQUÊTE - Inflation : les spéculateurs financiers, l'autre cause de l'augmentation des prix.](#)"
- [31] E. Peterson, "[The Coming Global Food Crisis](#)," 2022.
- [32] R. Chand, "[The global food crisis: causes, severity and outlook](#)," *Economic and Political Weekly*, pp. 115–122, 2008.
- [33] E. C. Gürcan, "[Food crisis and beyond: locating food-sovereign alternatives in a post-neoliberal context](#)," *Kasarinlan : Philippine Journal of Third World Studies*, vol. 26, no. 1-2, pp. 482–496, 2011.
- [34] T. Johnson, "[Food price volatility and insecurity](#),"
- [35] FAO, "[The state of food security and nutrition in the world, 2021.](#)"
- [36] GIEC, "[Sixième rapport d'évaluation du GIEC : changement climatique 2022.](#)"
- [37] D. de Paulo Farias and M. G. dos Santos Gomes, "[COVID-19 outbreak: what should be done to avoid food shortages?](#)," *Trends in Food Science & Technology*, vol. 102, p. 291, 2020.

- [38] D. Ollech, "[Ollech and Webel's combined seasonality test.](#)"
- [39] D. Ollech, "[Seasonality Tests - Package 'seastests'.](#)"
- [40] A. Insights, "[Méthode STL \(Seasonal-Trend decomposition using LOESS\).](#)"
- [41] O. Darne, "[Olivier Darne - Personal webpage - Teachings.](#)"
- [42] B. Krose and P. v. d. Smagt, [An introduction to neural networks - p33](#). The University of Amsterdam, 1996.
- [43] J. Schmidhuber, "[Deep learning in neural networks: An overview - p85 à 117,](#)" *Neural networks*, vol. 61, pp. 85–117, 2015.
- [44] G. Stuart, N. Spruston, B. Sakmann, and M. Häusser, "[Action potential initiation and backpropagation in neurons of the mammalian CNS,](#)" *Trends in neurosciences*, vol. 20, no. 3, pp. 125–131, 1997.
- [45] D. Das and S. Chakrabarti, "[Forecast Model Development of Some Selected Wholesale Price Index of India Using MLP,](#)" in *Proceedings of International Conference on Computational Intelligence, Data Science and Cloud Computing : IEM-ICDC 2020*, pp. 217–230, Springer, 2021.
- [46] M. Claesen and B. De Moor, "[Hyperparameter search in machine learning,](#)" *arXiv preprint arXiv :1502.02127*, 2015.
- [47] J. H. Friedman, "[Multivariate adaptive regression splines,](#)" *The annals of statistics*, vol. 19, no. 1, pp. 1–67, 1991.


- [48] Y. E. Shao and J.-T. Dai, "Integrated feature selection of ARIMA with computational intelligence approaches for food crop price prediction," *Complexity*, vol. 2018, pp. 1–17, 2018.
- [49] B. Nayana, K. R. Kumar, and C. Chesneau, "Wheat Yield Prediction in India Using Principal Component Analysis-Multivariate Adaptive Regression Splines (PCA-MARS)," *AgriEngineering*, vol. 4, no. 2, pp. 461–474, 2022.
- [50] O. Hegazy, O. S. Soliman, and M. A. Salam, "Comparative study between FPA, BA, MCS, ABC, and PSO algorithms in training and optimizing of LS-SVM for stock market prediction," *International Journal of Advanced Computer Research*, vol. 5, no. 18, pp. 35–45, 2015.
- [51] S. Qiu and J. Wang, "The prediction of food additives in the fruit juice based on electronic nose with chemometrics," *Food chemistry*, vol. 230, pp. 208–214, 2017.
- [52] H. Drucker, C. J. Burges, L. Kaufman, A. Smola, and V. Vapnik, "Support vector regression machines," *Advances in neural information processing systems*, vol. 9, 1996.
- [53] M. Rakhra, P. Soniya, D. Tanwar, P. Singh, D. Bordoloi, P. Agarwal, S. Takkar, K. Jairath, and N. Verma, "Crop price prediction using random forest and decision tree regression:-a review," *Materials Today : Proceedings*, 2021.
- [54] K. O. Nti, A. Adekoya, and B. Weyori, "Random forest based feature selection of macroeconomic variables for stock market prediction," *American Journal of Applied Sciences*, vol. 16, no. 7, pp. 200–212, 2019.


- [55] C. Browne, D. S. Matteson, L. McBride, L. Hu, Y. Liu, Y. Sun, J. Wen, and C. B. Barrett, "[Multivariate random forest prediction of poverty and malnutrition prevalence](#)," *PloS one*, vol. 16, no. 9, p. e0255519, 2021.
- [56] T.-H. Lee, A. Ullah, and R. Wang, "[Bootstrap aggregating and random forest](#)," *Macroeconomic forecasting in the era of big data : Theory and practice*, pp. 389–429, 2020.
- [57] S. Hasmita, F. Nhita, D. Saepudin, and A. Aditsania, "[Chili commodity price forecasting in bandung regency using the adaptive synthetic sampling \(adasyn\) and k-nearest neighbor \(knn\) algorithms](#)," in *2019 international conference on information and communications technology (icoiact)*, pp. 434–438, IEEE, 2019.
- [58] S. K. Mitra and M. Chattopadhyay, "[The nexus between food price inflation and monsoon rainfall in India: exploring through comparative data mining models](#)," *Climate and Development*, vol. 9, no. 7, pp. 584–592, 2017.
- [59] F. Martínez, M. P. Frías, M. D. Pérez, and A. J. Rivera, "[A methodology for applying k-nearest neighbor to time series forecasting](#)," *Artificial Intelligence Review*, vol. 52, no. 3, pp. 2019–2037, 2019.
- [60] M. S. Ahnaf, A. Kurniawati, and H. D. Anggana, "[Forecasting pet food item stock using arima and lstm](#)," in *2021 4th International Conference of Computer and Informatics Engineering (IC2IE)*, pp. 141–146, IEEE, 2021.
- [61] "[Comparing ML Models for Food Production Forecasting](#), author=Alkaabi, Nouf and Shakya, Siddhartha, booktitle=Artificial Intelligence XXXIX : 42nd SGAI International Conference on Artificial Intelligence, AI 2022, Cambridge, UK,

December 13–15, 2022, Proceedings, pages=303–308, year=2022, organization=Springer,”

- [62] R. Murugesan, E. Mishra, and A. H. Krishnan, “[Forecasting agricultural commodities prices using deep learning-based models: basic LSTM, bi-LSTM, stacked LSTM, CNN LSTM, and convolutional LSTM,](#)” *International Journal of Sustainable Agricultural Management and Informatics*, vol. 8, no. 3, pp. 242–277, 2022.
- [63] R. Collobert and J. Weston, “A unified architecture for natural language processing : Deep neural networks with multitask learning,” in *Proceedings of the 25th international conference on Machine learning*, pp. 160–167, 2008.
- [64] W. Ji and K. C. Chee, “Prediction of hourly solar radiation using a novel hybrid model of arma and tdnn,” *Solar energy*, vol. 85, no. 5, pp. 808–817, 2011.
- [65] M. RL and A. K. Mishra, “[Forecasting spot prices of agricultural commodities in India: Application of deep-learning models,](#)” *Intelligent Systems in Accounting, Finance and Management*, vol. 28, no. 1, pp. 72–83, 2021.
- [66] T. Xiong, C. Li, and Y. Bao, “[Seasonal forecasting of agricultural commodity price using a hybrid STL and ELM method: Evidence from the vegetable market in China,](#)” *Neurocomputing*, vol. 275, pp. 2831–2844, 2018.
- [67] G. K. Jha and K. Sinha, “[Agricultural price forecasting using neural network model: An innovative information delivery system,](#)” *Agricultural Economics Research Review*, vol. 26, no. 347-2016-17087, pp. 229–239, 2013.

- [68] P. Prakash, D. Jaganathan, S. Immanuel, A. Lama, J. Sreekumar, and P. Sivakumar, "Forecasting of sweet potato (*ipomoea batatas* L.) prices in india," *Indian Journal of Extension Education*, vol. 58, no. 2, pp. 15–20, 2022.
-

Package  : tseries, tsoutliers, ggplot2, moments, seastests, TSA, RJDemetra, ggcorrplot, ggplot2, gets, leaps, olsrr, forecast, tidyr, dplyr, ggdist, ggirdges, forecast, Gets, mgcv, neuralnet, earth, e1071, caret, randomForest, xgboost, class

Package  : keras, pandas, numpy, tensorflow, sklearn

Cours Master ECAP : "Techniques de prévision et conjoncutre" de Olivier Darne, "R-avancés et Github" de NEDELLEC Raphael, "Introduction au logiciel Python" de CHEVALEYRE Guillaume

7 Annexe

Figure 1 – Graphiques des séries corrigées

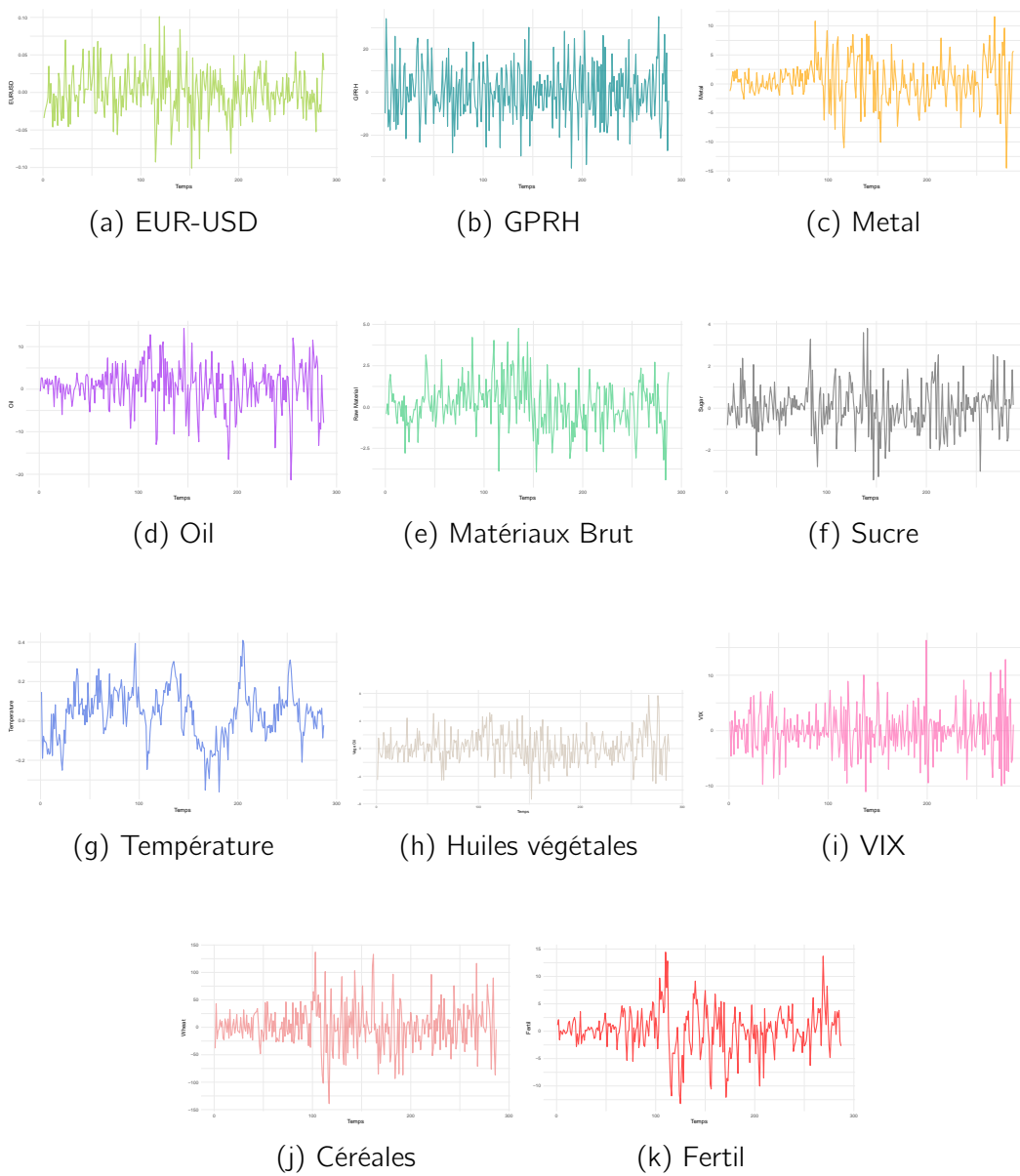


Table 1 – **Test de Dickey-Fuller sur la série brute**

Test	Statistique	p.value	alternative
Dickey-Fuller	-2.4897	0.3698	stationary

Table 2 – **Test de Dickey-Fuller sur la différenciée**

Test	Statistique	p.value	alternative
Dickey-Fuller	-5.9122	0.01	stationary

Table 3 – **Points atypiques sur la série corrigée**

Type	Période	Coefhat	T-stat
TC	Aout 2008	-7.516	-4.263
TC	Octobre 2008	-9.115	-5.161
AO	Mars 2011	-5.956	-3.923
AO	Juillet 2012	6.997	4.617
AO	Mai 2021	5.956	3.931

Figure 2 – **Points atypiques sur la série traitée/corrigée**

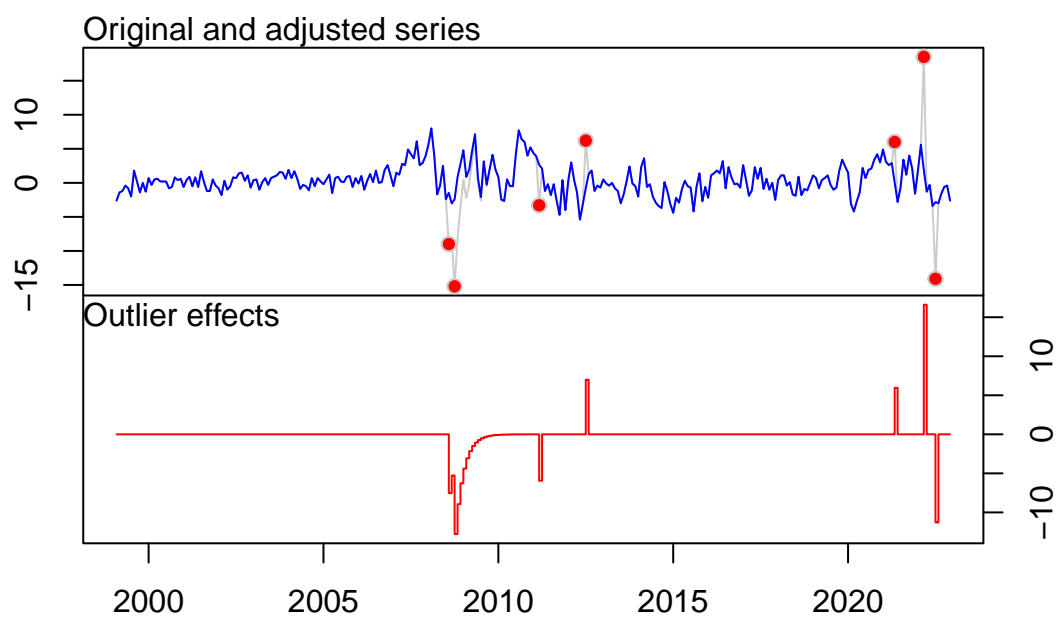


Table 4 – **Points atypiques sur la série Fertil**

Type	Ind	Période	Coefhat	T-stat
AO	114	2008 :07	-16.70	-6.398
TC	116	2008 :09	17.18	4.537
LS	119	2008 :12	-58.08	-12.653
TC	120	2009 :01	-14.46	-3.628
LS	150	2011 :07	33.03	7.553
AO	163	2012 :08	-26.16	-10.362
AO	273	2021 :10	-20.91	-7.897
LS	275	2021 :12	21.85	4.932
LS	279	2022 :04	39.45	8.823
AO	280	2022 :05	26.95	10.185
TC	285	2022 :10	17.00	4.448

Table 5 – **Points atypiques sur la série GPRH**

Type	Ind	Période	Coefhat	T-stat
TC	33	2001 :10	236.87	20.048
AO	34	2001 :11	63.79	6.306
TC	51	2003 :04	118.43	9.972
TC	53	2003 :06	-46.59	-3.986
AO	79	2005 :08	40.81	4.070
AO	253	2020 :02	45.54	4.546
TC	279	2022 :04	60.57	5.152

Table 6 – **Points atypiques sur la série Metal**

Type	Ind	Période	Coefhat	T-stat
AO	89	2006 :06	10.905	4.902
AO	136	2010 :05	8.478	3.804

Table 7 – **Points atypiques sur la série Oil**

Type	Ind	Période	Coefhat	T-stat
LS	118	2008 :11	-20.59	-4.247
AO	279	2022 :04	11.91	4.031

Table 8 – **Points atypiques sur la série Raw Material**

Type	Ind	Période	Coefhat	T-stat
TC	146	2011 :03	7.413	5.404
LS	155	2011 :12	-8.131	-5.192

Table 9 – **Points atypiques sur la série Temperature Anomalies**

Type	Ind	Période	Coefhat	T-stat
AO	111	2008 :04	0.3151	4.226

Table 10 – **Points atypiques sur la série VIX**

Type	Ind	Période	Coefhat	T-stat
TC	118	2008 :11	27.72	6.922
AO	153	2011 :10	14.05	4.215
TC	254	2020 :03	21.94	5.799
AO	255	2020 :04	17.70	5.399
AO	262	2020 :11	14.69	4.495

Table 11 – **Points atypiques sur la série EUR-USD**

Type	Ind	Période	Coefhat	T-stat
TC	118	2008 :11	-0.1290	-4.481
TC	137	2010 :06	-0.1038	-3.604
TC	143	2010 :12	-0.1065	-3.698

Table 12 – **Points atypiques sur la série Meat**

Type	Ind	Période	Coefhat	T-stat
AO	83	2005 :12	-24.60	-6.594
AO	201	2015 :10	-17.70	-4.747
TC	220	2017 :05	20.53	4.222

Table 13 – **Points atypiques sur la série Vege Oil**

Type	Ind	Période	Coefhat	T-stat
AO	110	2008 :03	15.67	8.316
LS	118	2008 :11	-10.44	-3.918
LS	282	2022 :07	-9.90	-3.715
AO	288	2023 :01	-11.42	-4.285

Table 14 – **Points atypiques sur la série Wheat**

Type	Ind	Période	Coefhat	T-stat
AO	105	2007 :10	151.5	5.182
TC	110	2008 :03	189.7	4.976
AO	125	2009 :06	119.6	4.090
LS	139	2010 :08	182.1	4.405
TC	150	2011 :07	-179.1	-4.699
AO	198	2015 :07	128.6	4.400
LS	278	2022 :03	166.7	4.033
LS	282	2022 :07	-204.0	-4.934

Table 15 – **Points atypiques sur la série Sugar**

Type	Ind	Période	Coefhat	T-stat
LS	128	2009 :09	5.780	5.023
TC	132	2010 :01	4.789	4.464
AO	133	2010 :02	4.576	5.613
LS	135	2010 :04	-6.836	-5.886
LS	141	2010 :10	5.633	4.894
LS	147	2011 :04	-5.390	-4.684
LS	150	2011 :07	5.183	4.504

Figure 3 – **Périodogramme de la série corrigée**

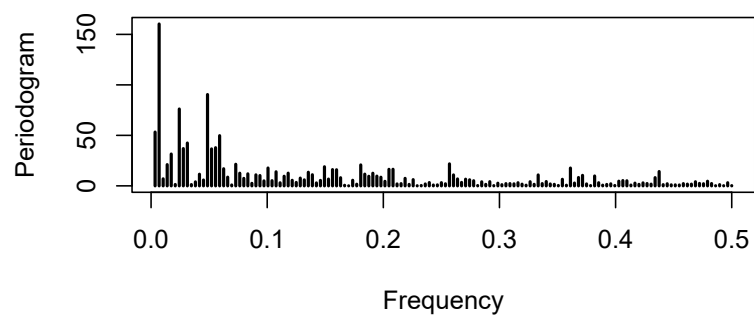


Table 16 – **Détection de saisonnalité pour les variables explicatives**

Variable	Webel-Ollech	Seasonal Dummies
Fertil	0.285	0.869
GPRH	0.300	0.660
Metal	0.046	0.370
Oil	0.331	0.311
Raw Material	0.019	0.038
Temperature Anomalies	0.008	0.005
VIX	0.562	0.181
EUR-USD	0.712	0.549
Meat	8.65e-05	3.46e-05
Huile	0.338	0.038
Wheat	0.942	0.968
Sucre	0.018	0.012

Table 17 – **Détection de saisonnalité sur les variables explicatives après application de la méthode STL**

Variable	Webel-Ollech	Seasonal Dummies
x1	0.285	0.869
x2	0.300	0.660
x3	0.994	0.999
x4	0.331	0.311
x5	0.983	1
x6	0.999	1
x7	0.562	0.181
x8	0.712	0.549
x9	0.997	1
x10	0.939	1
x11	0.942	0.968
x12	0.999	1
x13	0.604	0.927

Table 18 – **ADF test sur toutes les variables**

Variable	Statistique Dickey-Fuller	p-valeur
Food	-4.7971	0.01
Fertil	-4.506	0.01
GPRH	-8.054	0.01
Metal	-6.575	0.01
Oil	-6.614	0.01
Raw Material	-5.455	0.01
Temperature Anomalies	-3.907	0.014
VIX	-8.977	0.01
EUR-USD	-6.184	0.01
Meat	-7.368	0.01
Vege Oil	-5.658	0.01
Wheat	-5.760	0.01
Sugar	-5.438	0.01

Table 19 – **Statistiques descriptives de la série Y corrigée**

Moyenne	0.45
Médiane	0.30
Variance	4.73
Ecart-type	2.17
Kurtosis	0.86
Skewness	0.52
1 ^{er} quartile	-0.80
3 ^{eme} quartile	1.55

Figure 4 – **Critère de Mallow**

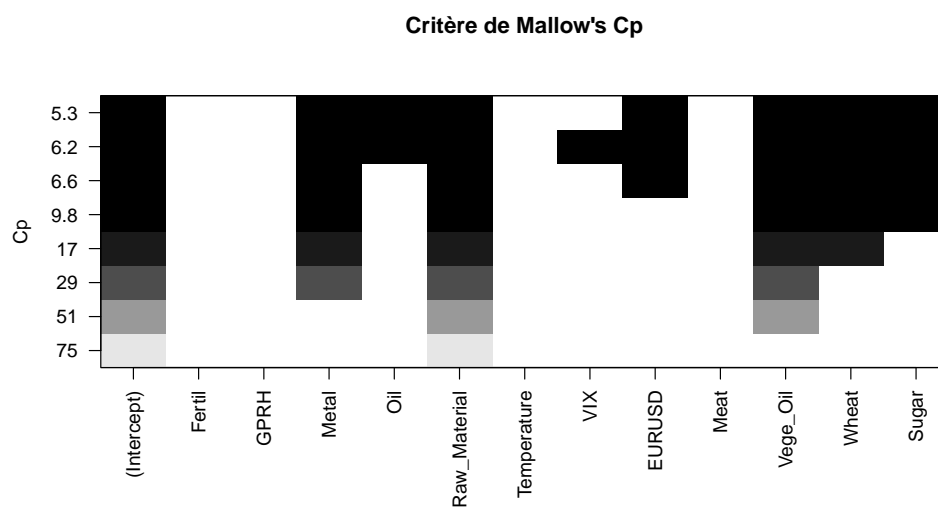


Figure 5 – R^2 ajusté

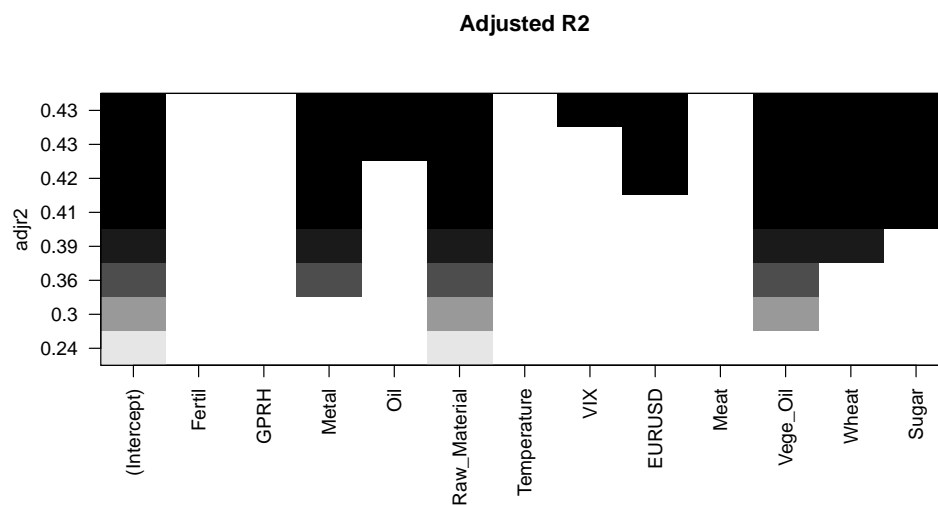


Table 20 – **Résultats du modèle AR(1)**

Coefficients	Valeurs	Erreurs standard
ar1	0.5690	0.0496
mean	-0.5499	0.2550
σ^2		3.378
Log likelihood		-556.8
AIC		1119.59
AICc		1119.68
BIC		1130.44

Table 21 – Résultats du modèle ARX avec auto.arima()

	Coefficients	Erreur standard
AR1	0.4164	0.0685
AR2	0.1385	0.0666
Metal	0.111	0.032
Raw_Material	0.4388	0.0725
Vege_Oil	0.1139	0.0436
Wheat	0.0083	0.0024
Sugar	0.2105	0.0859
σ^2	2.222	
Log vraisemblance	-414.78	
AIC	845.56	
AICc	846.22	
BIC	873.07	

Table 22 – Résultats du modèle ARX GET

	Coefficients	Erreurs standard	T-stat	p-value
mconst	0.0764390	0.0965001	0.7921	0.4291403
ar1	0.4300363	0.0473045	9.0908	< 2.2e-16 ***
Metal	0.0878499	0.0316037	2.7797	0.0059070 **
Raw_Material	0.4019346	0.0738565	5.4421	1.388e-07 ***
Vege_Oil	0.1811268	0.0477748	3.7913	0.0001932 ***
Wheat	0.0104420	0.0025068	4.1654	4.451e-05 ***
Sugar	0.2682796	0.0882499	3.0400	0.0026499 **

Table 23 – Résultats du modèle ARMAX

Coefficients	Valeurs	Erreurs standard
ar1	0.7355	0.0885
ma1	-0.3424	0.1255
Metal	0.1159	0.0317
Raw_Material	0.4335	0.0729
Vege_Oil	0.1123	0.0437
Wheat	0.0084	0.0024
Sugar	0.2127	0.0858

Table 24 – **Modèle GAM**

Termes	Estimate	Std. Error	t-value	Pr(> t)
(Intercept)	0.47299	0.09702	4.875	1.89e-06 ***
s(Metal)	2.314	2.953	9.588	8.26e-06 ***
s(Raw_Material)	1.000	1.000	51.162	< 2e-16 ***
s(EURUSD)	4.358	5.385	3.475	0.00350 **
s(Vege_Oil)	1.789	2.248	4.889	0.00667 **
s(Wheat)	1.000	1.000	11.610	0.00076 ***
s(Sugar)	2.128	2.732	5.184	0.00326 **
Indicateurs	Valeurs			
R-sq.(adj)	0.458			
Deviance expliquée	48.3%			
GCV	2.7136			
Estimation de l'échelle	2.579			
<i>n</i>	274			

Figure 6 – **Modèle MLP,MARS,XGB et LSTM**

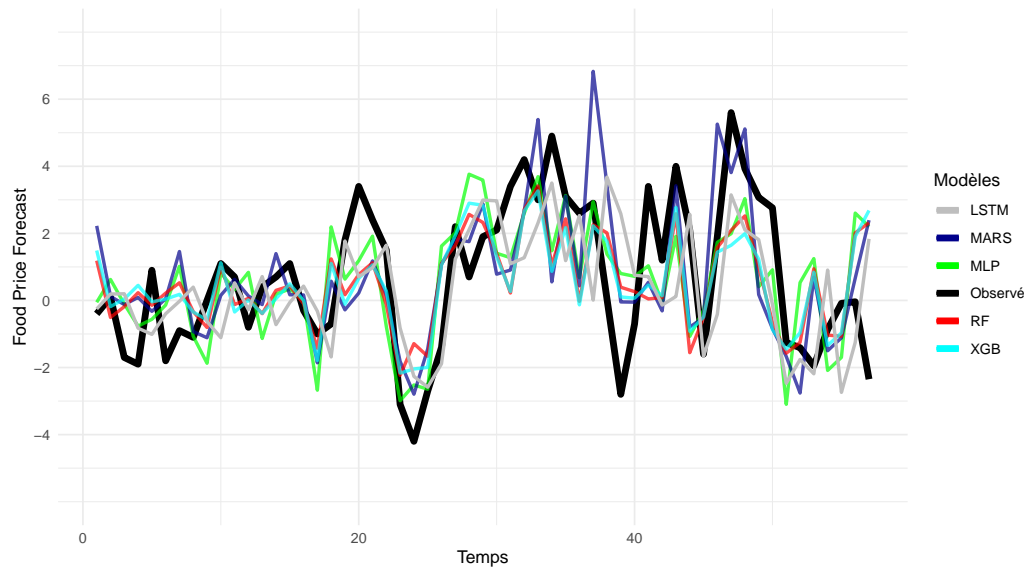


Figure 7 – CSPE - Modèles ECO vs ML

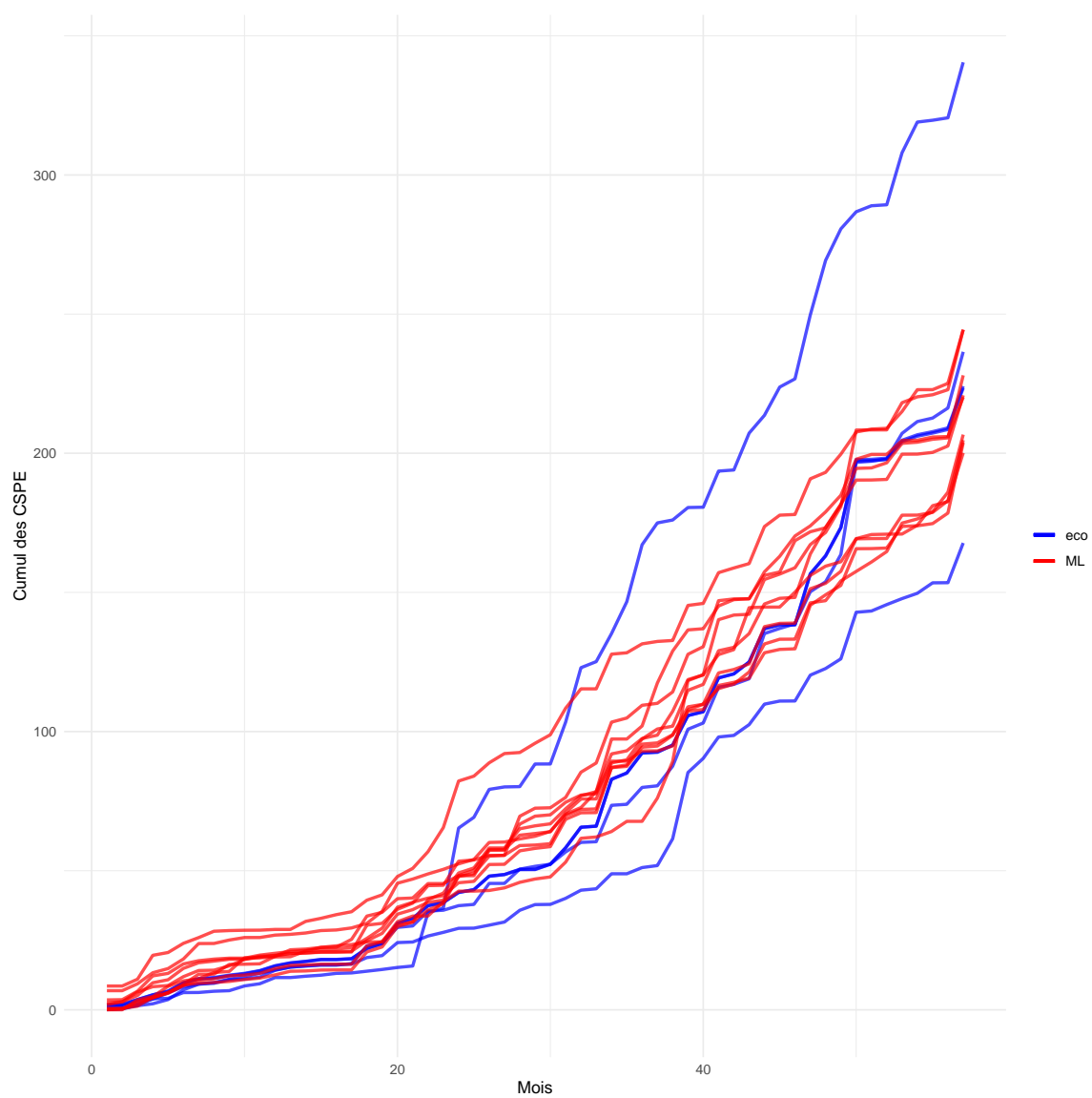
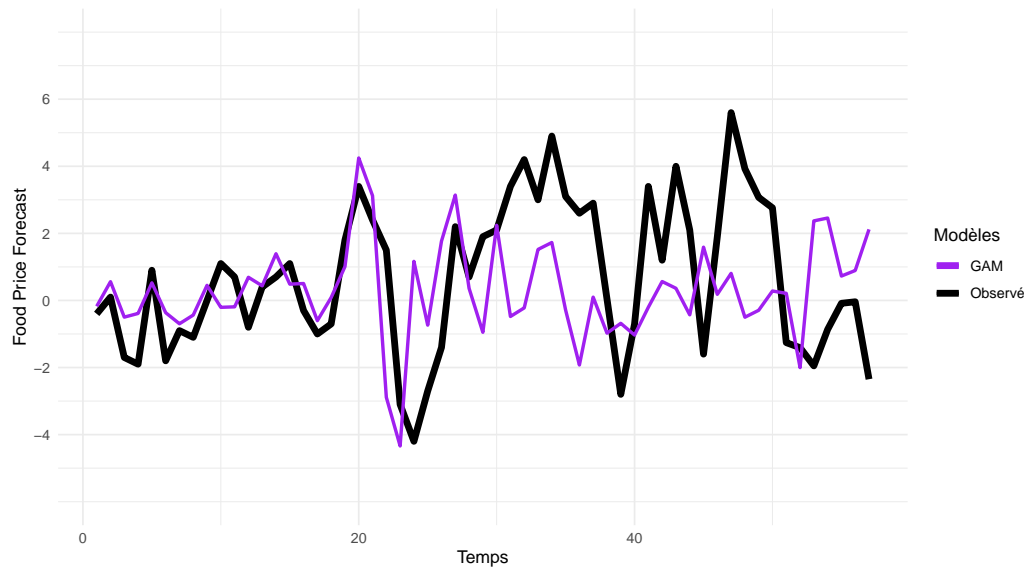


Figure 8 – **Prévisions - Modèle GAM**



8 Source des données

Table 25 – **Source des données utilisées**

Variable	Nom	Source
"Food"	Fao Food Price Index	fao.org
"Oil"	Crude Oil Price WTI	fred.stlouisfed.org
"VIX"	CBOE Volatility Index	finance.yahoo.com
"OVX"	CBOE Crude Oil Volatility Index	finance.yahoo.com
"GPR"	Geopolitical Risk (GPR) Index	matteoiacoviello.com
"Fertil"	Monthly Indices - Fertilizers	worldbank.org
"Metals"	Monthly Indices - Metals	worldbank.org
"Temp"	Temperature Anomalies	noaa.gov
"Raw"	Raw Materials	worldbank.org
"Sugar"	ICE - No. 11 Sugar Futures	investing.com
"Wheat"	US Wheat Futures	tradingeconomics.com
"Vege Oil"	US Soybean Oil Futures	investing.com

Table des matières

1	Introduction	6
1.1	Contexte	6
2	Présentation des variables	10
2.1	Variable dépendante : FAO Food Price Index	10
2.2	Choix des variables explicatives	13
2.2.1	Cours du pétrole	13
2.2.2	VIX : Indice de volatilité	15
2.2.3	OVX : La volatilité du pétrole	16
2.2.4	GPR : Indice de risque géopolitique	17
2.2.5	Fertil : Indice des prix des engrais	18
2.2.6	Metals : Indice des prix des métaux et des minéraux	20
2.2.7	Temperature Anomalies	21
2.2.8	Prix des matériaux brut : Raw Materials	23
2.2.9	Contrat Futures	24
2.2.10	Cours Euro-Dollars	26
3	Analyse exploratoire	28
3.1	Corrélation	29
3.2	Détection des points atypiques	31
3.3	Saisonnalité	36
3.4	Stationnarité	38
3.5	Statistiques descriptives sur la série corrigée	40

4	Sélection des variables	42
4.1	Approche BestSubSet & Gets	42
4.1.1	Approche BestSubSet	43
4.1.2	Approche GETS	44
5	Prévision & évaluation	47
5.1	Prévision	48
5.1.1	Modèles économétriques	48
5.1.2	Modèles de Machine-Learning Deep-Learning	55
5.2	Prévisions Évaluation	69
5.2.1	Prévisions	69
5.2.2	MSE, CSSED, R^2 OOS	75
5.2.3	Erreurs de prévision cumulées au carré (CSPE)	78
5.2.4	Test de Diebold-Mariano	83
6	Conclusion & Discussion	86
7	Annexe	99
8	Source des données	118

Liste des tableaux

1	Contrats Futures utilisés	25
2	Point atypique sur la série brute	32
3	Détection de saisonnalité - Variable Y	37
4	Sélection de variables avec GETS	45
5	Tableau des MSE, CSSE, RMSE et R^2 OOS	76
6	Test de précision de Diebold-Mariano pour l'ensemble des modèles	84
1	Test de Dickey-Fuller sur la série brute	100
2	Test de Dickey-Fuller sur la différenciée	100
3	Points atypiques sur la série corrigée	100
4	Points atypiques sur la série Fertil	102
5	Points atypiques sur la série GPRH	102
6	Points atypiques sur la série Metal	103
7	Points atypiques sur la série Oil	103
8	Points atypiques sur la série Raw Material	103
9	Points atypiques sur la série Temperature Anomalies	103
10	Points atypiques sur la série VIX	104
11	Points atypiques sur la série EUR-USD	104
12	Points atypiques sur la série Meat	104
13	Points atypiques sur la série Vege Oil	105
14	Points atypiques sur la série Wheat	105
15	Points atypiques sur la série Sugar	106
16	Détection de saisonnalité pour les variables explicatives	107

17	Détection de saisonnalité sur les variables explicatives après application de la méthode STL	108
18	ADF test sur toutes les variables	109
19	Statistiques descriptives de la série Y corrigée	110
20	Résultats du modèle AR(1)	111
21	Résultats du modèle ARX avec auto.arima()	112
22	Résultats du modèle ARX GET	113
23	Résultats du modèle ARMAX	113
24	Modèle GAM	114
25	Source des données utilisées	118

Table des figures

1	Évolution de l'indice des prix de la nourriture de janvier 1990 à décembre 2022	12
2	Évolution du cours du pétrole WTI	13
3	Évolution du VIX	16
4	Évolution de l'indice GPR (Geopolitical Risk)	18
5	Évolution du prix des engrais : Fertil	19
6	Évolution de l'indice "Metals"	20
7	Évolution du niveau d'anomalies des températures mondiales	22
8	Évolution de l'indice Raw Materials	24
9	Évolution des différents contrats Futures	25
10	Évolution de l'Euro-Dollars	26
11	Matrice des corrélations sur les séries brutes	30

12	Détection des points atypiques sur la série brute	32
13	Boxplot de la série Y corrigée	41
14	Histogramme de la série ajustée des points atypiques	42
15	Sélection de variables avec le critère BIC	44
16	Fonctions lisses des variables du modèle GAM	54
17	Prévisions des modèles économétriques	70
18	Prévisions des modèles économétriques - simplifié	72
19	Prévisions des modèles ML	73
21	Erreurs de prévision cumulées au carré (CSPE)	80
22	CSPE - Modèles économétriques & Modèles ML	82
1	Graphiques des séries corrigées	99
2	Points atypiques sur la série traitée/corrigée	101
3	Périodogramme de la série corrigée	106
4	Critère de Mallow	110
5	R^2 ajusté	111
6	Modèle MLP,MARS,XGB et LSTM	115
7	CSPE - Modèles ECO vs ML	116
8	Prévisions - Modèle GAM	117