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

Le Roux Noa - 2 juin 2023

Master 1 Econométrie et Statistique, parcours Econométrie Appliquée

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 prévisions 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.

Time-Series, Forecasting, Machine-Learning, Food Price

1 Cadre de l'étude

1.1 Sujet

Cette étude se concentre sur la prévision de séries temporelles multivariées. L'ambition est de découvrir des variables explicatives significatives et d'exploiter pleinement l'information qu'elles fournissent pour créer un modèle de prévision solide pour la variable dépendante. Dans le cadre de notre recherche, la variable dépendante représente l'indice des prix alimentaires de la FAO, un

indicateur développé par l'Organisation des Nations Unies pour l'agriculture et l'alimentation, qui suit l'évolution mensuelle des prix d'une gamme variée de produits alimentaires.

1.2 Objectifs

Dans cette optique, nous ferons appel non seulement à des modèles économétriques traditionnels, mais également à des modèles de machine learning plus avancés. L'objectif est donc de comparer l'efficacité de ces deux

Note de synthèse

types de modèles dans la prévision de notre variable dépendante, de mars 2018 à décembre 2022 (57 mois), en nous basant sur un total de 286 données.

L'échantillonnage sera effectué de manière non-aléatoire pour maintenir la cohérence des modèles.

1.3 Limites

Notre étude comporte plusieurs limites. En premier lieu, toutes les hypothèses spécifiques à chaque modèle économétrique ne sont pas respectées, à part la stationnarité des séries. Notre objectif n'étant pas d'analyser les relations de causalité entre les diverses variables, mais plutôt de définir le modèle de prévision le plus précis possible. En deuxième lieu, le volume de données que nous avons est assez restreint (286), ce qui pose un défi pour les modèles de machine learning qui fonctionnent généralement mieux avec de grands ensembles de données. Troisièmement, les modèles de machine learning requièrent une expertise considérable pour être utilisés de manière optimale. Le réglage minutieux des hyper-paramètres est crucial pour obtenir des prévisions précises. Malheureusement, notre connaissance limitée nous empêche d'exploiter pleinement leurs capacités, réduisant ainsi leur efficacité. Notre intention est plutôt d'expérimenter avec ces modèles pour mieux les comprendre. Enfin, il est probable que certaines données que nous aurions aimé

utiliser ne soient pas disponibles, ce qui nous empêche de recueillir toutes les informations que nous aurions souhaité pour construire des modèles de prévision robustes. Nous pensons par exemple aux données sur les précipitations mondiales, le coût de l'eau ou différents futures composant notre variable dépendante.

1.4 Apports

Notre analyse se distingue des autres études en incorporant une gamme plus étendue de modèles et de variables, notamment en intégrant des variables financières issues de divers contrats à terme (Futures). L'objectif est d'évaluer l'influence de la spéculation sur les marchés financiers sur les fluctuations des prix des produits alimentaires, en complément des données environnementales. De surcroît, nous employons une diversité plus large de modèles (16 au total), ce qui aide à déterminer quel modèle et quelle méthodologie seraient les plus appropriés pour ce type de données.

2 Méthodologie


2.1 Traitement des variables

Avant de mettre en œuvre nos différents modèles, nous avons préalablement traité toutes nos variables. Initialement, nous avons éliminé les points atypiques, puis nous avons désaisonnalisé et rendu stationnaires nos variables. Finalement, nous avons de

Note de synthèse

nouveau retiré les points atypiques des séries stationnarisées. Cette démarche nous permet d'obtenir des séries plus stables, afin d'éviter de biaiser l'apprentissage des divers modèles en vue d'obtenir des prévisions de qualité satisfaisante.

2.2 Choix des variables

Toutes nos variables ont été validées par le biais de la littérature scientifique, nous avons ainsi commencé notre analyse avec un total de 10 variables : le prix du pétrole, le VIX, l'OVX, le GPR, un indice du prix des engrais, des métaux, des matières premières, le niveau d'anomalies des températures mondiales, le taux de change EUR-USD ainsi que des contrats à terme sur le blé, les céréales, la viande et le sucre. Afin d'éviter de perturber nos modèles de prévision en intégrant des informations superflues (du bruit), nous avons utilisé deux méthodes de sélection de variables. La première est la méthode BestSubSet et la seconde est l'approche GETS, toutes deux mises en œuvre en langage de programmation . Les deux méthodes conduisent à la même sélection de six variables : le prix du pétrole, le prix des matières premières, des contrats à terme sur les huiles végétales, le blé et le sucre.

2.3 Choix des modèles

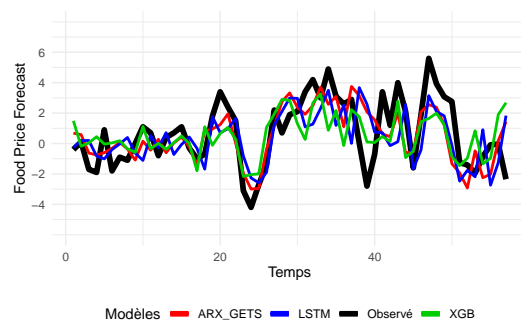
Comme pour nos variables, nos modèles sont validés par la littérature scientifique et par le

cours "*Techniques de prévision & Conjoncture*" de M. Darne [1]. Nous avons utilisé divers modèles économétriques, notamment : ARX, ARX GET, ARMAX, LM, AR(1), GAM. En ce qui concerne les modèles de machine learning, nous avons utilisé : MLP, MARS, SVM, RF, XGB, kNN, LSTM, LSTM-CNN et finalement TDNN.

3 Résultats

3.1 Prévisions

Afin d'éviter d'afficher les 16 modèles sur un seul graphique, nous présentons ici nos trois modèles les plus performants : ARX GET, LSTM et XGB.



Ce graphique est présenté à des fins purement illustratives, il n'est pas réellement possible de tirer des conclusions sur la qualité des modèles à partir de celui-ci.

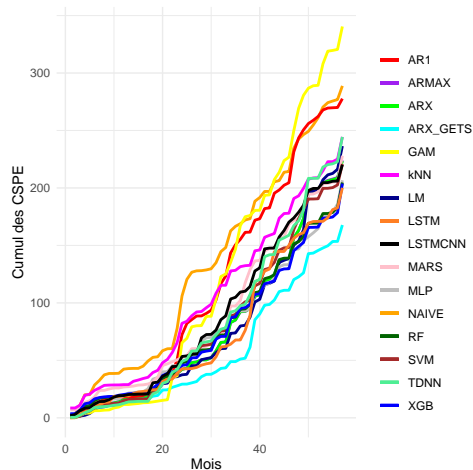
3.2 Evaluation

Modèle	RMSE	CSSD	R^2_{005}
ARX GET	1.72	38.04	0.40
XGB	1.89	62.28	0.27
LSTM	1.87	61.48	0.28

Note de synthèse

Nous avons choisi de présenter ces modèles car ils surpassent les 16 autres modèles en termes de qualité de prévision, selon différents indicateurs de performance de prévision. Comme le montre le tableau précédent, les erreurs quadratiques moyennes (RMSE) de ces modèles sont parmi les plus faibles de tous nos modèles. Il est intéressant de constater que la majorité des modèles de machine learning avaient un RMSE entre 1,90 et 2, tandis que les modèles économétriques avaient souvent un RMSE supérieur à 2. La somme cumulée de la différence d'erreur quadratique (CSSED) ainsi que le R^2 hors échantillon (R^2_{OOS}) affichent les valeurs les plus basses pour le modèle ARX GET. Ce modèle est sans conteste le meilleur que nous avons identifié en termes de qualité de prévision.

3.3 CSPE



Le CSPE (*erreur de prévision cumulée au carré*) représente l'accumulation des erreurs

des différents modèles sur une période prolongée, offrant ainsi une perspective sur la stabilité de leurs prévisions.

Encore une fois, le modèle ARX GET s'avère être nettement le meilleur, car il accumule le moins d'erreurs de prévision sur la période donnée. On peut remarquer que le modèle GAM a produit d'excellentes prévisions pendant les vingt premiers mois, puis a complètement dévié. Ce phénomène est nettement visible sur le CSPE, puisque le modèle GAM est le modèle cumulant le moins d'erreur pendant les vingt premiers mois, puis son CSPE s'envole. Les modèles de machine learning semblent, en moyenne, accumuler moins d'erreurs à long terme que les modèles économétriques.

4 Conclusion

In fine, notre modèle le plus performant pour prédire le prix des denrées alimentaires de mars 2018 à décembre 2022 est le modèle économétrique ARX GET. Bien que, selon tous nos indicateurs de qualité de prévision, les modèles de machine learning aient en moyenne mieux performé que tous les modèles économétriques, notre meilleur modèle reste incontestablement l'ARX GET.

Références

- [1] O. Darne, "[Olivier Darne - Personal webpage - Teachings.](#)"