Agricultural Commodity Price Forecasting Using Comprehensive Machine-Learning Techniques

Rotem Zelingher

université PARIS-SACLAY

Prévision des prix des produits agricoles à l'aide de techniques d'apprentissage automatique

Agricultural commodity price forecasting Using comprehensive machine-learning techniques

Thèse de doctorat de l'université Paris-Saclay

École doctorale n°581 : agriculture, alimentation, biologie, environnement et santé (ABIES)
Spécialité de doctorat: Sciences économiques
Unité de recherche : Université Paris-Saclay, INRAE, AgroParisTech, Economie Publique, 78850, Thiverval-Grignon, France.
Référent : AgroParisTech

Thèse présentée et soutenue à Paris-Saclay, le 29/11/2021, par

Rotem ZELINGHER

Composition du Jury

Stéphane DE CARA Directeur de Recherche, INRAE Centre IdF-Versailles-Grignon	Président
Eli FEINERMAN Professeur, Volcani Institute (Israel)	Rapporteur & Examinateur
Isabelle PIOT-LEPETIT Chargé de recherche, INRAE, centre Occitanie- Montpellier	Rapporteur & Examinatrice
Iddo KAN Professeur, Hebrew University of Jerusalem (Israel)	Examinateur

Direction de la thèse

Christophe GOUEL Directeur de Recherche, INRAE Centre IdF-Versailles-Grignon	Directeur de thèse & Invité
David MAKOWSKI Directeur de Recherche, INRAE Centre IdF-Versailles-Grignon	Co-Directeur & Examinateur

Thèse de doctorat

NNT : 2021UPASB065

Contents

Acronyms **11**

1 Introduction **31**
1.1 Motivation 31
1.2 Background 34
1.2.1 The Agricultural Problem 34
1.2.2 Agricultural Commodities 35
1.2.3 Food Security - Concept and Strategy 36
1.3 Approach 40
1.3.1 Principle I - Concise Comprehensibility 40
1.3.2 Principle II - Interpretability 41
1.3.3 Principle III - Accessibility 41
1.4 Machine Learning 41
1.5 Data 43
1.6 Scope of Work 43
1.7 Organisation of the thesis 44

2 Literature Survey **45**
2.1 Very short term and short term 47
2.1.1 Non-structural models 50
2.1.2 Structural models 50
2.2 Long term 51
2.2.1 Partial Equilibrium (PE) 54
2.2.2 Computable General Equilibrium (CGE) 55
2.3 Medium term 56

3 Essay I **59**
3.1 Introduction 60
3.2 Materials and method 62
3.2.1 Data 62
3.2.2 Linear and generalised linear models 64

3.2.3 CART . . . 65
3.2.4 Random Forest and Gradient Boosting . . . 66
3.2.5 Models Evaluation . . . 67
3.3 Results . . . 67
3.3.1 Quantitative effects of regional productions on price changes 67
3.3.2 Classification of price increase vs. decrease . . . 70
3.4 Discussion . . . 72
3.5 Conclusions . . . 75

Appendix I **75**
3.A Appendix I . . . 76

4 Essay II **99**
4.1 Introduction . . . 100
4.2 Data . . . 101
4.3 Methods . . . 102
4.3.1 Models 1, 2, and 3 - Machine learning . . . 103
4.3.2 Model 4 - Multivariate linear regression . . . 104
4.3.3 Model 5 - VAR . . . 104
4.3.4 Model 6 - TBATS . . . 105
4.4 Model evaluation . . . 105
4.5 Results . . . 106
4.6 Discussion . . . 111

Appendix I **114**
4.A Appendix II . . . 115
4.A.1 Data information . . . 115
4.A.2 General presentation . . . 118
4.A.3 Breakdown of the price change by inputs and regions . . . 121

5 Essay III **131**
5.1 Introduction . . . 132
5.1.1 Data . . . 134
5.1.2 Models . . . 135
5.1.3 Model evaluation . . . 138
5.1.4 Ranking of the producing regions . . . 139
5.2 Results . . . 140
5.2.1 Maize prices . . . 140
5.2.2 Soybean prices . . . 141
5.2.3 Cocoa prices . . . 142
5.2.4 Most influential producing regions . . . 143

5.3 Discussion 149
5.4 Conclusion 151

Appendix I **151**
5.A Appendix III.A 152
5.A.1 Maize 153
5.A.2 Soybean 155
5.A.3 Cocoa 157
5.B Appendix III.B 160
5.B.1 Maize 160
5.B.2 Soybean 163
5.B.3 Cocoa 166

6 Discussion **169**
6.1 The basic idea behind the study 169
6.2 Empirical Approach 171
6.2.1 Essay I: Assessing the sensitivity of global maize price to regional production using statistical and machine learning methods 171
6.2.2 Essay II: Forecasting global maize prices from regional productions 171
6.2.3 Essay III: Data-driven assessment of the impacts of crop productions on the global prices of maize, soybean and cocoa 173
6.3 Contributions 175
6.3.1 Contribution I - Concise Comprehensible Forecasting Tool . 176
6.3.2 Contribution II - Interpretable Forecasting Tool 176
6.3.3 Contribution III - Accessible Forecasting Tool 177
6.4 Recommendations and future work 178

A Appendix **199**

List of Figures

1.1 Prevalence of moderate or severe food insecurity 32

2.1 Consumer Prices, Food Indices 52

3.1 Time series of global maize price 63
3.2 Relative importance, yield, regression models 69
3.3 PDP, yield in Northern America, regression analysis 70
3.4 AUC, yield, classification analysis 71
3.5 PDP, yield in Northern America, classification analysis 72
3.6 Relative change, output vs. input 76
3.7 CART model results, October price, regression 77
3.8 CART model results, November price, regression 78
3.9 CART model results, December price, regression 78
3.10 Observed price vs. predicted price, yield, regression analysis . . . 79
3.11 Observed price vs. predicted price, production, regression analysis 80
3.12 Relative importance, production, regression analysis 81
3.13 PDP, yield in Southern Africa, regression analysis 82
3.14 PDP, production in Northern America, regression analysis 83
3.15 PDP, production in Southern Africa, regression analysis 84
3.16 CART model results, October price, classification 85
3.17 CART model results, November price, classification 85
3.18 CART model results, December price, classification 86
3.19 Relative importance, yield, classification analysis 87
3.20 Relative importance, production, classification analysis 88
3.21 PDP, yield in Southern Africa, classification analysis 89
3.22 PDP, production in Northern America, classification analysis 90
3.23 PDP, production in Southern Africa, classification analysis 91

4.1 Time series (relative change), output and input data 102
4.2 Price change: observed vs. forecasted 107
4.3 Relative advantage by model 108

4.4 Shapley values, relative extreme December price changes 112
4.5 Correlation: global price and regional production 118
4.6 Price change: observed vs. forecasted, by month, input = production 119
4.7 Price change: observed vs. forecasted, by month, input = yield . . 120
4.8 CART model results, December price 121
4.9 Relative importance, GBM model, January-June, monthly 123
4.10 Relative importance, GBM model, July-December, monthly 124
4.11 Shapley values, extreme price changes, January-June 127
4.12 Shapley values, extreme price changes, July-December 128

5.1 Time series of global crop prices 135
5.2 Time series of crop output 136
5.3 Relative advantage by model, maize 141
5.4 Relative advantage by model, soybean 142
5.5 Relative advantage by model, cocoa 143
5.6 Shapley values and SHAP dependence plots, maize 144
5.7 Shapley values, January maize price, input = regional yield 145
5.8 Shapley values and SHAP dependence plots, soybean 146
5.9 Shapley values, January soybean price, input = regional production 147
5.10 Shapley values and SHAP dependence plots, cocoa 148
5.11 Shapley values, May cocoa price, input = local yield 148
5.12 Relative importance, Maize 161
5.13 Relative importance, Soybean 164
5.14 Relative importance, Cocoa 167
5.15 Time-series of Shapley values, Cote D'Ivoire, Cocoa 168

List of Tables

3.1 RMSE values . . . 67
3.2 Data description, input = production . . . 93
3.3 Data description, input = yield . . . 94
3.4 Summary statistics, linear model, production, regression analysis . 95
3.5 Summary statistics, linear model, yield, regression analysis 96
3.6 Summary statistics, linear model, production, classification analysis 97
3.7 Summary statistics, linear model, yield, classification analysis . . . 98

4.1 Best forecasting method, by month and lags . . . 110
4.2 Variable description and sources . . . 115
4.3 Data composition and summary statistics of inputs . . . 116
4.4 Market year of maize . . . 117
4.5 Decomposition of Shapley values results, input = yield . . . 129
4.6 Decomposition of Shapley values results, input = production . . . 130

5.1 Variable description and data sources . . . 152
5.2 Data description, regional input, maize . . . 153
5.3 Data description, national input, maize . . . 154
5.4 Data description, continental input, maize . . . 154
5.5 Data description, regional input, soybean . . . 155
5.6 Data description, national input, soybean . . . 156
5.7 Data description, continental input, soybean . . . 156
5.8 Data description, regional input, cocoa . . . 157
5.9 Data description, national input, cocoa . . . 157
5.10 Data description, continental input, cocoa . . . 158
5.11 Best forecasting method, maize, by month and lags . . . 160
5.12 Best forecasting method, soybean, by month and lags . . . 163
5.13 Best forecasting method, cocoa, by month and lags . . . 166

A.1 Forecasting cycle for monthly price of Arabica coffee . . . 200
A.2 Forecasting cycle for monthly price of cocoa . . . 201

A.3 Forecasting cycle for monthly price of maize 202
A.4 Forecasting cycle for monthly price of palm-oil 203
A.5 Forecasting cycle for monthly price of rice 204
A.6 Forecasting cycle for monthly price of Robusta coffee 205
A.7 Forecasting cycle for monthly price of soybean 206
A.8 Forecasting cycle for monthly price of wheat 207
A.9 Data sources 1/2 208
A.10 Data sources 2/2 208

Acronyms

AC Agricultural Commodities. 33, 35–45, 47, 48, 50–54, 56–58, 131, 169–173, 175–180

AIC Akaike Information Criterion. 65, 104

ANN Artificial Neural Network. 50

ARCH Autoregressive Conditional Heteroscedasticity. 50

ARIMA Autoregressive Integrated Moving Average. 50, 100

ARMA Autoregressive Moving Average. 50, 101, 105

AUC Area Under Curve. 67, 70, 71, 73

CART Classification And Regression Trees. 65–68, 70, 76, 77, 82–85, 91, 103, 108, 110, 112

CGE Computable General Equilibrium. 54, 55, 57

CME Chicago Mercantile Exchange. 34, 35, 48, 173

CPI Consumer Price Indices. 62

EML Extreme Machine-Learning. 56

FAO Food And Agriculture Organization. 33, 43, 62, 74

FOB Free On Board. 62

GARCH Generelised Autoregressive Conditional Heteroscedasticity. 50

GBM Gradient Boosting Machine. 65–71, 73, 74, 82–84, 91, 99, 103, 104, 106, 108, 110, 112, 174

GDP Gross Domestic Product. 170, 172

GEM General Equilibrium Model. 100

GLM Generalised Linear Model. 65, 67, 73, 91, 97, 98

GLOBIOM Global Biosphere Management Model. 54, 55

GTAP Global Trade Analysis Project. 55

IFPRI International Food Policy Research Institute. 51, 74

LM Linear Model. 66, 67, 70, 82–84, 91, 107–109

LOOCV Leave-One-Out Cross-Validation. 67

LT Long Term. 45

ML Machine-Learning. 40–42, 44, 50, 60, 65, 74, 75, 100, 101, 103, 107, 108, 111, 131, 171, 174–177

MT Medium Term. 32, 42, 44, 45, 56–58, 111, 171, 172

OECD Organisation For Economic Co-Operation And Development. 33, 74

OTC Over The Counter. 35

PDP Partial Dependence Plot. 69, 71, 72, 175–177

PE Partial Equilibrium. 54, 100

RA Relative Advantage. 177

RF Random-Forest. 65–68, 70, 73, 74, 82, 99, 103, 106, 108, 112, 174

RMSE Root Mean Squared Error. 67–69, 73, 81, 105–110, 172

ROC Receiver Operating Curve. 67, 70

RSS Residual Sum Of Squares. 76, 103

SDG Sustainable Development Goal. 31

SHAP Shapley Additive exPlanations. 175–177

ST Short Term. 45, 47–50

SVAR Structural Vector Autoregressive. 50, 51

TBATS Trigonometric Seasonal Box Transformation With Arma Residuals Trend And Seasonal Components. 99, 101, 105–112, 131, 172, 173, 175

TS Time-Series. 50, 56

USDA United States Department Of Agriculture. 48, 57, 100

VAR Vector Autoregressive. 99, 101, 104, 107–109, 112, 172

VST Very Short Term. 45, 47, 48

WASDE World Agricultural Supply And Demand Estimates. 47, 57, 61, 74, 100, 113

WTO World Trade Organisation. 170

Remerciements

Cette thèse a été financée par CLAND et l'Institut national de recherche pour l'agriculture, l'alimentation et l'environnement (INRAE).

Un grand merci à mes directeurs de thèse - David Makowski et Christophe Gouel, qui m'ont accompagné tout au long de mes études et m'ont donné l'opportunité de me développer professionnellement, dans l'espoir de continuer dans le domaine de la recherche. Un merci tout particulier à David, qui a cru en moi jusqu'au bout, m'a permis d'être exposé à de nouveaux domaines et m'a soutenu à tout moment, même pendant le confinement - merci beaucoup!

Au-delà de l'encadrement officiel, je tiens à remercier Nicol Vilmer, ma mentore PSU, qui m'a guidé tout au long de ce parcours et m'a conseillé sur tous les sujets. Nicol m'a soutenu, écouté et guidé depuis notre première "match-up" jusqu'au dernier moment avant la soutenance officielle du doctorat.

Un grand merci à l'équipe de recherche d'EcoPub, INRAE. Merci beaucoup à Stephane De Cara et Estelle Gozlan, qui m'ont toujours donné un coup de main et soutenu.

Elisabeth Maltese, ma professeure de français, qui est bien plus qu'une professeure. Elisabeth m'a appris le français à partir de zéro et m'a soutenu dans les moments difficiles dans un nouveau pays.

Cornelia Mosoiu et Jorge Huerta-Jemio, mes voisins qui m'ont ouvert leur maison et sont devenus une vraie famille - merci pour tous les moments que nous avons passés ensemble, pour les rires et les conversations. Merci merci merci merci!!!

Stéphane Daumas, mon meilleur ami du village. Nous avons passé tellement de bons moments ensemble. Je n'aurais pas survécu sans lui au village!

Elisabeth Dick, qui m'a donné une vraie famille française avec qui passer les fétes.

Un merci spécial à Beth Loubavitch Etudiants, surtout à Mendy et Mushky Lachkar, qui ont été une seconde maison, alors que j'étais si désemparée dans un nouveau pays.

Et, enfin et surtout, Michal et David Klapisch, qui sont toujours avec moi, où que j'aille.

Résumé

Cette thèse se compose de trois chapitres qui traitent de l'analyse/prévision des matières premières agricoles échangées au niveau mondial. Tous, ont utilisé des données et des méthodes accessibles au public qui peuvent être reproduites et qui sont donc accessibles indépendamment de toute limitation budgétaire. En tant que telle, cette thèse propose une nouvelle méthodologie de prévision et d'analyse des prix des produits agricoles de base afin de garantir une grande précision de prévision, tout en étant interprétable et techniquement accessible. Nous montrons que les prix des produits agricoles peuvent être prévus pour des périodes allant d'un mois à un an, tout en maintenant une qualité de prévision élevée et les principes de transparence scientifique.

L'idée centrale cette étude

Transformer la théorie de la prévision des prix des produits agricoles en un outil disponible et accessible serait socialement bénéfique, surtout pour ceux qui n'y ont actuellement pas accès - il s'agit principalement des résidents des pays à faible revenu. Si l'on considère que chaque culture a des valeurs nutritionnelles uniques, la consommation de plusieurs cultures de différents groupes peut créer un régime alimentaire complet et équilibré. Dans le cadre d'une alimentation peu transformée, l'intégration de ces cultures peut permettre de mettre en place un régime alimentaire sain et bon marché. En tant que source bon marché d'énergie et de micro-nutriments, le maïs, associé au soja comme source bon marché de protéines, de graisses et d'autres micro-nutriments, peut promouvoir la sécurité alimentaire des consommateurs à faibles revenus. Parallèlement, et en combinaison avec le cacao, cultivé principalement par les petits agriculteurs des pays en développement, peut contribuer au bien-être en tant qu'outil permettant de sauver les agriculteurs pauvres du cycle de la pauvreté.

Cela a souvent été le rôle historique : le maïs, en tant que culture relativement durable cultivée dans des zones climatiques variées et le soja, qui est un élément important du régime alimentaire chinois depuis des milliers d'années.

Cependant, contrairement à l'énorme potentiel de ces cultures et à leur capacité évidente à nourrir l'ensemble de la population mondiale aujourd'hui (Helms, 2004), leur offre reste limitée dans certaines régions. Bien que la production alimentaire ait plus que triplé au cours des six dernières décennies, l'utilisation croissante des principales cultures du monde comme source d'énergie ou d'alimentation du bétail a augmenté leur consommation dans les régions à haut revenu, au détriment des régions à faible revenu. En outre, l'évolution du régime alimentaire mondial entraîne une hausse de la consommation de viande, même dans les régions les moins riches, ce qui contribue à accroître la demande. La tendance à stocker des denrées alimentaires dans le cadre d'une stratégie (commerciale) de sécurité alimentaire visant à protéger les populations (de gestion des risques) en cas de fluctuations des prix des denrées alimentaires entraîne également une vulnérabilité relativement élevée chez les résidents des pays à faible revenu qui ne possèdent pas de stocks alimentaires suffisants.

Malgré les demandes répétées de l'Organisation mondiale du commerce (OMC) de s'abstenir d'interventions gouvernementales qui pourraient nuire à la compétitivité des marchés (G20, 2020 ; OMC, 2020), la récente crise du coronavirus a prouvé, une fois de plus, que les premiers à souffrir des fluctuations de prix sont les pays à faible PIB par habitant. En outre, la crise sanitaire (et économique) actuelle a révélé les conséquences négatives des tensions internationales, notamment entre les grandes puissances, et du manque de coordination entre les pays commerçants (IFPRI et al., 2020). Les plus vulnérables sont les pauvres, qui n'ont pas supporté l'incertitude, notamment l'impossibilité de vendre leurs produits agricoles ou d'acheter la nourriture. Étant donné que peu de pays ont détenu des réserves alimentaires suffisantes pour trois mois, des millions de personnes ont rejoint celles qui souffraient déjà d'insécurité alimentaire.

Compte tenu du potentiel existant dans le commerce alimentaire international et des affirmations historiques officielles selon lesquelles il permettra d'équilibrer la distribution alimentaire entre tous les pays tout en garantissant un marché compétitif et en équilibrant les fluctuations de prix, cette étude s'est concentrée sur les prix internationaux des produits agricoles. L'ensemble de l'étude a été réalisé dans le but de promouvoir la symétrie des informations concernant les marchés mondiaux des produits agricoles, y compris les facteurs de fluctuation des prix, leur probabilité d'apparition dans un avenir proche (un mois à un an à l'avance) et leur quantification. J'espère que cette étude pourra faire progresser le bien-être social, en particulier celui des personnes défavorisées dans les pays à faible revenu.

Approche empirique

Cette étude est composée de trois essais qui donnent une bonne image de la performance de l'approche empirique proposée. Dans l'ensemble, ils créent un cadre accessible pour l'analyse des marchés internationaux des produits agricoles de base en démêlant les impacts des productions agricoles sur le prix mondial de la même culture.

Essai I: Analyse de la sensibilité du prix mondial du maïs à la production régionale à l'aide de méthodes statistiques et d'apprentissage automatique

Dès l'ouverture, cette thèse tente de retracer deux points essentiels: Le premier est une évaluation de l'impact des zones de production sur les fluctuations de prix du maïs ; le second est l'établissement d'une relation empirique entre le prix mondial du maïs et la variation de la production. Cet article examine le maïs en tant que marché prototype sur près de six décennies. Le caractère unique de cette étude réside dans l'utilisation pionnière de modèles d'apprentissage automatique pour analyser les prix à moyen terme des produits agricoles de base, tout en donnant un aperçu de ce qui se cache derrière les résultats obtenus. Pour chaque modèle, des versions spécifiques à deux mois sont construites : la régression pour quantifier les variations annuelles des prix et la classification, pour évaluer la probabilité d'une baisse ou d'une hausse des prix par rapport à l'année précédente. Tous les modèles ont été évalués par une Leave-One-Out-Cross-Validation. L'étude se concentre sur le quatrième trimestre de l'année, c'est-à-dire au début de la période où le maïs nord-américain (principalement les États-Unis) est physiquement commercialisé sur le Chicago Merchandise Exchange en tant que nouvelle récolte. Les résultats quantifient l'impact de la production de maïs en Amérique du Nord sur le prix mondial du maïs en octobre, novembre et décembre, c'est-à-dire pendant et après la saison de récolte nord-américaine. Les résultats soulignent le potentiel d'utilisation de modèles d'apprentissage automatique pour la prévision des prix, mais ne comparent pas la performance prédictive de cette approche avec les outils de prévision standard.

ESSAI II: PRÉVISION DES PRIX MONDIAUX DU MAÏS À PARTIR DE LA PRODUCTION RÉGIONALE

Le deuxième essai s'inscrit directement dans la continuité du premier en se focalisant sur la prévision. Cet article a été motivé par deux objectifs principaux : Le premier était de prévoir le prix mondial mensuel du maïs dans un horizon temporel à moyen terme, et le second était de trouver la méthode de prévision la plus précise pour différents mois et délais (lags).

Nous comparons les outils d'apprentissage automatique à deux modèles économétriques ; tous deux sont des outils omniprésents dans les études de prévision :

TBATS - un outil auto-régressif qui traite automatiquement les caractéristiques non linéaires et la multisaisonnalité. TBATS a déjà démontré des capacités de prédiction impressionnantes dans des plages relativement courtes pour une variété de sujets, notamment le prix quotidien de l'électricité (Karabiber et Xydis, 2019), la consommation de gaz (Naim et al., 2018), et même les précipitations (Farheen, 2021). Cependant, en ce qui concerne la prévision des prix des matières premières agricoles, c'est la première fois que l'on teste TBATS.

VAR - un modèle auto-régressif multivarié. Le VAR est un outil de prévision largement utilisé et relativement simple, qui revêt une grande importance dans l'élaboration et l'analyse des politiques monétaires. Les modèles VAR s'excellent dans la détection des chocs au sein des données et la combinaison de leurs effets sur la variabilité des principales variables ou, dans notre cas, des prix du maïs. Cependant, si le VAR est un outil efficace pour la prévision de variables telles que l'inflation, la croissance du PIB, le taux de change ou les taux d'intérêt (Bjørnland,2008 ; Kapetaniosetal.,2008), son efficacité n'a pas encore été testée dans le contexte des prix des matières premières agricoles à moyen terme. Le deuxième article compare les modèles de prévision à un benchmark correspondant à une prédiction naïve constante. L'évaluation des modèles comprend un processus de validation croisée glissante, qui produit une erreur de prévision (RMSE) utilisée pour classer tous les modèles et une comparaison avec une prédiction naïve représentée par une valeur moyenne de changement de prix.

Au-delà de la comparaison de l'attrait des modèles pour la prévision des prix du maïs, l'étude comprend une analyse de la nature de la relation entre le niveau de changement de la production annuelle régionale de maïs et le changement de son prix mondial ; l'identification des régions ayant le plus grand impact sur le prix du maïs, ventilé par mois. En outre, l'étude fournit une ventilation précise de la méthode préférée de prévision des prix mensuels du maïs en fonction de l'horizon de prévision (par mois, avec des décalages d'un mois).

L'analyse de l'importance relative, qui cherche à découvrir la pertinence globale de chaque région pour la prévision des prix (König et al., 2021), a confirmé

l'influence relative substantielle de la production de l'Amérique du Nord sur le prix mondial pendant la majeure partie de l'année, à partir du début de l'année de marché en octobre jusqu'en mai. Cependant, l'Asie occidentale a exercé une influence plus substantielle sur les changements de prix du maïs en juillet et en août.

En outre, les valeurs de Shapley ont donné un aperçu des principaux moteurs des fortes et inévitables fluctuations des prix Les résultats montrent en effet que certaines régions ont influencé à l'extrême les fluctuations des prix observées certaines années. Par exemple, Shapley met en lumière l'influence fortement positive des rendements de maïs d'Afrique de l'Est en 2006 sur le prix de novembre de cette même année. Indéniablement, 2006 a été une année de sécheresse extrême dans la région (Solomon et al., 2007), ce qui a nui au secteur agricole (Gebrechorkos et al., 2020), et a entraîné des importations de maïs exceptionnellement élevées, notamment en provenance des États-Unis.

ESSAI III: ÉVALUATION FONDÉE SUR DES DONNÉES, DE l'IMPACT DE LA PRODUCTION AGRICOLE SUR LES PRIX MONDIAUX DU MAÏS, SOJA ET CACAO

Le dernier article applique les connaissances accumulées dans les deux premiers essais en examinant l'efficacité des méthodes de prévision pour deux cultures supplémentaires : le soja et le cacao. Cette analyse explore le caractère générique de l'approche proposée et capture le caractère unique des produits agricoles de base de trois catégories différentes, telles que déterminées par la Banque mondiale : les céréales, pour le maïs ; les huiles et farines, pour le soja ; et les boissons, pour le cacao. En outre, ce chapitre évalue la sensibilité des performances du modèle aux trois échelles géographiques considérées pour les entrées, c'est-à-dire régionale (comme dans les deux premiers essais), continentale et nationale. Enfin, et pour les trois produits de base, nous avons mis en œuvre chaque modèle avec deux ensembles d'entrées : (1) les variations régionales de production ou de rendement ; et (2) les mêmes variables avec l'ajout de la variation annuelle relative du prix de l'année précédente.

En somme, chaque prix mensuel prévu est le résultat du modèle le plus performant, sur 60 (5 algorithmes × 3 échelles géographiques × 4 versions, à l'exclusion de TBATS), et par rapport à la division d'échelle géographique la plus pertinente.

La spécification de trois catégories de marchés de matières premières agricoles et de trois échelles géographiques met en évidence l'importance de la structure économique du marché. Les résultats révèlent l'importance capitale que les structures de marché ont sur le niveau des productions végétales qui

influencent les prix des produits agricoles de base au niveau mondial. En outre, l'étude montre que les changements régionaux dans la production de maïs ont indéniablement des impacts élevés sur son prix, en particulier lorsqu'ils proviennent d'Amérique du Nord - le premier producteur et exportateur mondial de cette culture, et avec une différence considérable par rapport aux autres régions.

L'autre extrême est le marché du cacao. L'Afrique de l'Ouest et l'Amérique du Sud concentrent à elles seules la majeure partie de la production de cacao dans leur région, généralement le fait de petits exploitants dans des fermes familiales situées dans des zones relativement pauvres. Contrairement au maïs et au soja, qui sont principalement négociés sur le marché international situé et géré dans le pays du plus gros producteur, le cacao est négocié principalement du côté des importateurs, à New-York et à Londres, c'est-à-dire loin de son pays d'origine. Ce fait contribue au manque d'information sur le marché parmi les producteurs de cacao et les empêche de contrôler le prix qu'ils recevront pour leur récolte ou la date préférée pour la vendre.

De nombreuses techniques ont été appliquées pour interpréter les résultats des modèles ajustés : analyse de l'importance relative (Greenwell et al., 2020), valeurs de Shapley (Molnar et al., 2018 ; Tianqi et al., 2021 ; Greenwell, 2017 ; Liu et Just, 2020) et analyse de corrélation standard. Pour le cacao, aucune de ces méthodes n'a indiqué une relation forte entre le volume de production du principal producteur (Côte d'Ivoire) et les variations de prix. Les résultats ont cependant montré une absence de pouvoir de marché absolu concentré dans une zone particulière et une distribution assez uniforme de l'impact mensuel par pays sur l'année. En outre, un examen approfondi a mis en évidence une relation assez complexe entre les valeurs de Shapley et les variations de rendement des cultures en Côte d'Ivoire. En se concentrant sur des chocs de prix extrêmes spécifiques, on a constaté une forte contribution du rendement du cacao en Indonésie aux événements de hausses de prix exceptionnellement élevées. Les résultats ont semblé surprenants au départ, car il s'agit du marché le plus concentré parmi les trois marchés examinés, sans compter que la part de marché de l'Indonésie sur le marché mondial du cacao est nettement inférieure à celle de la Côte d'Ivoire. Cependant, une étude approfondie de la littérature sur le cacao a révélé un système complexe dans lequel certains facteurs sapent l'équilibre naturel du marché.

Pour la Côte d'Ivoire (Cameroun et Nigeria), très peu d'organisations locales collectent la grande majorité de la production des petits exploitants, dont la plupart n'ont pas accès aux informations sur le marché. De fait ils reçoivent, en début de saison, un prix fixé par le gouvernement local en fonction des prix futurs et des bourses (CCI et CNUCED/OMC., 2001). Ce prix est toutefois fixé à un

niveau suffisamment bas pour assurer un retour positif à l'organisme payeur. Au fil des années, le marché d'exportation du cacao en Côte d'Ivoire a été privatisé, de sorte que des sociétés d'exportation privées collectent désormais la production. De ce fait, la situation des petits agriculteurs ne s'est pas encore améliorée (Abbott et al., 2019). Dans ces conditions, les agriculteurs de Côte d'Ivoire prennent la décision critique dès le début de la saison de culture : ils visent à augmenter leur production lorsque le prix qu'ils reçoivent de leur gouvernement augmente et vice versa.

En termes de résultats de prévision, les méthodes d'apprentissage automatique (RF et GBM) sont généralement plus performantes que les autres modèles pour tout horizon supérieur à trois mois dans le futur. Le GBM offre une précision de prévision nettement supérieure pour les mois de la nouvelle récolte de maïs et de soja en Amérique du Nord Dans le secteur du cacao, ces mois, à savoir mars, avril et mai, sont les seuls pour lesquels les modèles d'apprentissage automatique sont pris en compte.

CONTRIBUTIONS PRINCIPALES

Ce projet de recherche offre à la littérature plusieurs contributions sur la prévision des prix. L'essai I a principalement contribué à la présentation de nouvelles méthodes analytiques pour la prévision des prix des produits agricoles de base en identifiant les principaux facteurs de changement des prix du maïs à l'aide de plusieurs techniques économétriques et d'apprentissage automatique. Le premier article, qui était notre première tentative à n'utiliser que des données accessibles et des modèles relativement simples, a révélé que les algorithmes d'apprentissage automatique sont un outil légitime pour la science de la prévision des prix des produits agricoles de base. En outre, il a démontré que ces modèles d'apprentissage automatique ne doivent pas nécessairement être du type "boîte noire" et que leur comportement devient interprétable lorsqu'on utilise de puissantes techniques de visualisation. Le deuxième essai (Essai II) a poursuivi le chemin entamé dans l'essai précédent et a fourni la preuve, par le biais de plusieurs techniques d'interprétation de modèles, suivant laquelle les prix sur le marché du maïs réagissent fortement aux changements de la production agricole en Amérique du Nord, principalement du rendement. Cette dernière s'applique à 10 prix mensuels par an, à l'exception des deux derniers mois de l'année commerciale nord-américaine. Notre technique de classement par importance a révélé une forte avance de l'Asie occidentale au cours de ces deux mois. Cependant, lorsqu'on se concentre sur des événements spécifiques, la valeur de Shapley présente des influences relativement fortes de l'Asie occidentale et de l'Afrique du Nord. Le troisième et dernier essai (essai III) a révélé

des différences notables dans les approches de prévision optimales pour chaque prix de produit agricole unique en développant le deuxième essai. Il a démontré que pour prévoir les prix du maïs avec la plus grande précision, par rapport aux modèles testés dans cette thèse, les rendements régionaux sont l'entrée la plus recommandée à utiliser. Pour le soja, l'impact provient pour l'essentiel de la production régionale, tandis que les prix du cacao sont grandement affectés par le rendement local des six plus grands pays producteurs. En appliquant des techniques d'interprétation multiples (PDP, importance relative, valeur de Shapley. Avec une analyse basée sur le SHAP : Shapley value and PDP) a permis de mettre en évidence l'impact remarquable de chaque unité de production sur le prix mensuel mondial. Elle a ainsi fourni un outil original pour se préparer aux fluctuations extrêmes des prix.

Cette étude a suivi trois principes directeurs, à savoir la concision, la compréhensibilité, l'interopérabilité et l'accessibilité. D'une manière générale, elle a contribué aux efforts de promotion de la sécurité alimentaire mondiale.

Contribution I – OUTIL DE PRÉVISION CONCIS ET COMPRÉHENSIBLE

Nous définissons un modèle idéal capable de prendre en compte les multiples facteurs de fluctuation des prix des produits agricoles tout en restant relativement succinct. Alors que la collecte de données pourrait constituer un obstacle à la réalisation d'un tel modèle, cette étude a réussi à limiter les données d'entrée aux données accessibles au public (production/rendement annuel des cultures), que l'utilisateur peut obtenir en un simple clic. L'utilisateur peut transformer cette information brute en une variable groupée utilisable en téléchargeant les données dans son ordinateur personnel et en exécutant le code. La variable dépendante du modèle (prix mondiaux) est transformée en son style "adapté au modèle" de la même manière.

Contribution II - OUTIL DE PRÉVISION INTERPRÉTABLE

Dans leur article, Coyle et Weller (2020) critiquent le choix par les chercheurs des modèles d'apprentissage automatique comme outil d'analyse et de prévision des questions liées aux politiques. Ils affirment que les modèles d'apprentissage automatique sont souvent non interprétables et empêchent donc leurs utilisateurs de les comprendre et de vérifier la validité de leurs résultats. Pour surmonter ce défi, nous avons choisi de construire les trois articles sur la base de modèles interprétables, puis d'utiliser plusieurs techniques de visualisation non consacrées par les modèles existants (Molnar, 2019). Plus précisément, la

première étape de la réalisation de cette recherche consistait à étudier la relation causale entre les entrées et les sorties du modèle. Pour cette information primaire, nous avons utilisé l'indicateur de causalité de Granger (Granger, 1969). Après l'entraînement des modèles, le niveau de contribution à la précision de la prédiction (RMSE) a déterminé l'importance régionale relative séparément pour chaque algorithme et pour chaque mois. Enfin, les diagrammes de dépendance partielle (PDP) décrivent visuellement les réponses moyennes du prix du maïs aux variations du rendement relatif du maïs dans les régions les mieux classées en fonction de leur contribution à la précision des prédictions des modèles.

Dans le deuxième article, la technique de l'importance relative a montré, une fois encore, l'influence relative des caractéristiques. Plus tard, l'intégration de la valeur de Shapley basée sur la théorie des jeux nous a permis d'évaluer la contribution marginale de chacune des régions productrices aux événements spécifiques des chocs de prix les plus extrêmes, dans les deux sens, positif et négatif. Le troisième article combine plusieurs techniques d'interprétation de modèles. Parmi celles-ci, citons les explications additives de Shapley (SHAP) de Lundberg et Lee (2017). SHAP est une méthode d'interprétation par apprentissage automatique basée sur l'algorithme traditionnel de Shapley. Comme les valeurs de Shapley, SHAP mesure les contributions de chaque caractéristique aux prédictions du modèle. Cependant, le principal avantage de cet algorithme innovant découle de sa capacité à combiner un PDP adapté à Shapley pour une interprétation qui combine à la fois des mesures de quantification et de visualisation.

Contribution III – OUTIL DE PRÉVISION ACCESSIBLE

Pour être accessible le modèle doit être constamment prêt à être adapté par le concepteur stratégique de la sécurité alimentaire, à savoir le responsable politique qui l'utilise. Pour atteindre cet objectif, le décideur doit avoir un accès régulier aux données du modèle. Un tel modèle offre à ses utilisateurs la possibilité de comprendre ce qui se cache derrière et une compréhension globale du marché auquel ils sont confrontés.

Tout d'abord, le programme de prévision comprend un outil d'évaluation des erreurs alternatives appelé "Avantage relatif". Grâce à cet outil, les utilisateurs peuvent déterminer si le modèle de prévision est suffisamment efficace par rapport à la prévision constante et aux autres modèles. En outre, " L'Avantage Relatif" fournit une évaluation dynamique, en fonction du mois requis et du temps restant jusqu'à la date d'échéance. En deuxième lieu les techniques de diagnostic des modèles mentionnées dans la section précédente indiquent quel acteur l'utilisateur du modèle doit examiner avec prudence. Lorsqu'ils l'utilisent,

les décideurs peuvent concevoir leur stratégie jusqu'à un an avant l'achat/la vente des produits agricoles, puis en vérifier l'exactitude à l'approche de la date réelle des échanges.

RECOMMANDATIONS ET TRAVAUX FUTURS

Cette thèse de doctorat fournit un outil de prévision et d'analyse des prix des matières premières agricoles compréhensible, interprétable et accessible pour des perspectives d'analyse de un à douze mois. En outre, il est accessible à quiconque en a besoin sous la forme d'un package R ou python prêt à l'emploi, qui exploite uniquement des données librement disponibles. À ce jour, le modèle examine en détail trois types de cultures différentes faisant l'objet d'échanges internationaux. Cependant, il implique la prévision des prix de huit cultures au total et s'avère donc être un outil nettement plus performant, applicable à d'autres produits agricoles que le maïs, le soja et le cacao.

Les prévisions de prix dans ce projet sont le résultat d'un seul type d'entrée (production ou rendement des cultures). En plus de présenter un résultat final, le modèle fournit une analyse d'erreur qui indique le risque estimé, correspondant à l'erreur de prévision approximative du modèle. La recommandation générale est de considérer à la fois le prix prévu et l'erreur du modèle et de préférer agir les mois où le risque d'erreur est faible. Nous utilisons notre modèle pour mettre en évidence les événements critiques de changement de prix, qui doivent être détectés correctement pour permettre au modèle d'être transparent pour ses utilisateurs. L'un des défis de la prévision des prix des produits agricoles de base est que, si l'on se soucie surtout de prévoir les événements de fluctuations extrêmes, ces événements sont relativement rares. Du point de vue de la sécurité alimentaire, le fait de ne pas identifier des événements de changement de prix extrêmes pourrait être un résultat pire que de manquer des événements de changement de prix modérés. Des techniques telles que l'algorithme de la valeur de Shapley peuvent refléter ces priorités politiques dans le modèle. Notre analyse souligne l'importance de comprendre le compromis entre l'omission de certains chocs de production (rendement) dans des régions influentes et le fait d'accorder, par erreur, plus d'attention à l'offre mondiale totale ou même à la production des régions à faible impact. Les données sur les coûts d'une mauvaise interprétation pourraient aider à évaluer les dommages potentiels de tout choc de production agricole sur le niveau de sécurité alimentaire des régions vulnérables. Au cours de mes recherches, j'ai découvert le domaine du commerce international des produits agricoles de base et l'apprentissage automatique. Au cours des dernières années, j'ai lu de nombreux articles et écouté un nombre incalculable de conférences et d'opinions d'experts de différents do-

maines.

Les chapitres inclus dans ce document de recherche ne montrent pas toutes les tentatives pour maximiser la contribution de notre outil au monde de la prévision des prix des matières premières agricoles. Nous avons examiné différentes variables explicatives, ensemble ou séparément ; nous avons analysé le pouvoir prédictif sur la base de données provenant de diverses sources d'information et nous avons même examiné la dépendance en fonction des saisons de récolte par rapport aux dates des années commerciales locales. En outre, nous avons expérimenté des modèles à partir d'un large éventail de possibilités tout en effectuant différentes versions de l'exécution dans les modèles que nous avons finalement inclus. Pour maximiser la familiarité avec les méthodes actuellement acceptées et l'ensemble des options disponibles, une segmentation des variables explicatives a également été effectuée sur la base d'un travail de recherche approfondi, qui comprenait l'exploration de bases de données économiques et agronomiques parallèlement à une enquête sur la littérature existante.

Cet outil de prévision peut dans l'immédiat ne pas être pertinent pour tous les pays. Les prix internationaux (Banque mondiale) examinés dans cette étude indiquent la valeur mensuelle moyenne payée sur les marchés commerciaux mondiaux directs. Ce prix n'est pas nécessairement un bon indicateur du niveau des prix à la consommation (partie du revenu consacrée à l'alimentation), qui détermine en fin de compte leur niveau de sécurité alimentaire. Comme nous l'avons vu à propos du cacao, ces prix ne reflètent pas toujours le prix payé à l'agriculteur qui l'a produit. C'est ici qu'intervient la grande importance de la nature de l'État importateur ou exportateur de chaque CA.

Comme démontré tout au long de la thèse, alors que de nombreux pays à haut revenu gèrent des programmes bien planifiés pour protéger les consommateurs et les producteurs des fluctuations de prix, les pays à faible revenu ne peuvent pas toujours le faire efficacement. Le résultat de l'existence ou de l'inexistence de tels programmes est le niveau auquel le prix intérieur de chaque pays fluctuera en fonction du prix mondial. Par ailleurs, en attendant que les données soient disponibles de manière suffisamment fiable et riche, il pourrait être bénéfique d'inclure la variation annuelle des stocks de céréales comme l'une des entrées du modèle. En effet, grâce à leur rôle originel, les stocks alimentaires importants peuvent compenser les périodes de mauvaises récoltes ou de prix élevés des produits agricoles de base, et ils fonctionnent donc comme une sauvegarde sociale. Malheureusement, le stockage de nourriture est une question coûteuse qui n'est pas économiquement disponible pour toutes les nations. En outre, des stocks suffisants peuvent atténuer la concurrence pour les produits alimentaires. En revanche, le sur-stockage peut faire sortir les marchés mondiaux de leur phase de stabilisation naturelle, comme cela s'est produit au

début de 2020, lorsque la Chine a reconstitué ses stocks de céréales. Indéniablement, les personnes les moins protégées sont aussi les plus vulnérables. Mais, malheureusement, ce sont aussi ceux qui ne disposent pas des outils nécessaires pour analyser les marchés mondiaux et prévoir le moment optimal pour acheter ou vendre des matières premières agricoles.

Dans un environnement aussi incertain, la volatilité excessive des prix des produits de base a des répercussions négatives tant sur les producteurs que sur les consommateurs. Ce manque d'information se répercute généralement sur les revenus et la production des agriculteurs et conduit à de moins bonnes décisions en matière d'investissement dans les intrants. Les répercussions de l'instabilité des marchés des produits de base peuvent également exacerber les problèmes de pauvreté, notamment dans les zones rurales. C'est ainsi que le manque d'information vient s'ajouter à l'impact négatif sur la sécurité alimentaire dans les pays les plus vulnérables et dépendants des importations.

Une autre question importante non prise en compte ici le rapport entre les prix des différents produits de base. Dans un modèle de prévision des prix, chaque produit agricole est, tout au long de la thèse, considéré comme indépendant des autres. Cependant, dans la pratique, comme la valeur de Shapley l'a montré très visiblement dans l'interprétation des résultats du maïs, il existe une forte relation entre les prix des produits de substitution.

Au niveau nutritionnel, le maïs est un hydrate de carbone et, par conséquent, il est un substitut du blé, du riz et parfois du soja. En effet, en tant qu'aliment pour le bétail, les fluctuations de prix de ces produits agricoles reflètent également les fluctuations de prix d'autres produits agricoles sur le marché international et sont utilisés comme source de protéines : viande et produits laitiers. En ce qui concerne les prix locaux, les prix des œufs évolueront également à terme, en fonction des prix des céréales. En tant que source d'énergie, le maïs est également utilisé comme biocarburant et donc coordonné avec d'autres produits de base, avec les prix des produits de base énergétiques : charbon, pétrole brut et gaz naturel.

Quant au cacao et au café, ces produits agricoles de base ne sont pas essentiels en termes de valeur nutritionnelle pour le consommateur, mais constituent une source de revenus unique ou importante pour de nombreux petits exploitants, notamment dans les pays en développement d'Afrique occidentale. Ces produits agricoles de base sont cultivés principalement sous les tropiques et importés en grande majorité par les pays à revenu élevé.

Les prix des producteurs fluctuent en fonction des cours internationaux et déterminent souvent leur décision en matière d'allocation des terres, passant du cacao au café (Gilbert, 2016). Au-delà de ce qui précède, les résultats obtenus par les techniques d'ouverture de modèle ouvrent la voie à de futures études

qui porteront sur l'amélioration des modèles. Dans ce contexte, il est possible d'explorer d'autres options d'ajout d'une variable explicative ou de conversion à une variable explicative différente, d'examiner la qualité des prévisions de modèles supplémentaires ou de construire une prévision basée sur l'exécution de plusieurs modèles simultanément. Un autre conseil est d'analyser les possibilités de combiner les différents algorithmes pour créer un modèle couvrant plusieurs cultures.

En conclusion, ce travail offre un outil complet et disponible pour l'analyse et la prévision des prix des produits agricoles de base dans des plages de temps allant d'un mois à un an. S'il est utilisé correctement, le mécanisme proposé peut contribuer à la sécurité alimentaire et économique des ménages, des agriculteurs ou d'autres entités dans le besoin. Cependant, comme déjà mentionné dans le premier paragraphe de ce travail, cet outil apportera un bénéfice maximal s'il est incorporé dans le cadre d'un plan multidisciplinaire de sécurité alimentaire.

Chapter 1

Introduction

```
A person's ability to avoid starvation
   will depend both on his ownership
   and on the exchange entitlement mapping
   that he faces.
```

– Amartya Sen, Poverty and Famines (p.4)

1.1 Motivation

My voyage into higher education had begun out of a genuine desire to bring about "Tikun Olam" - repair the world and lead it to a better place. Over time, I realised that my capacity was limited and that my efforts were unlikely to lead to significant change. However, I also realised that if I learn and persevere, I can take a small part in a big project, striving for the same goal. The purpose of this thesis is to contribute to the second Sustainable Development Goal (SDG) of "End hunger, achieve food security and improved nutrition and promote sustainable agriculture" through its first criteria of correcting and preventing trade restrictions and distortions in the world agricultural markets.

The prevailing view amongst many economists was that the solution to the food insecurity problem in the world lies in the existence of world trade in agricultural commodities. According to them, international trade can avoid costly food surpluses (or deficiencies) in certain zones by transporting them to areas with a shortage. The argument was that international trade would help maintain a rich diet throughout the year and at a relatively stable price level (Costinot and Rodríguez-Clare, 2018; van Meijl et al., 2017).

Then, toward the end of 2019, the coronavirus pandemic erupted, and the normal local-global food prices equilibrium distorted drastically (Schmidhuber

et al., 2020). If until that point, food shortage problems were the domain of low-income countries, the COVID-19 crisis had brought, amongst other things, a sharp rise in food prices of some agricultural commodities (AC), while the price of others have declined due to adverse demand conditions[1]. These have led to an increase in food insecurity in most of the world. Whereas the highest increases were in medium-low income countries, an increase is also apparent in upper-middle-income economies (see Fig 1.1).

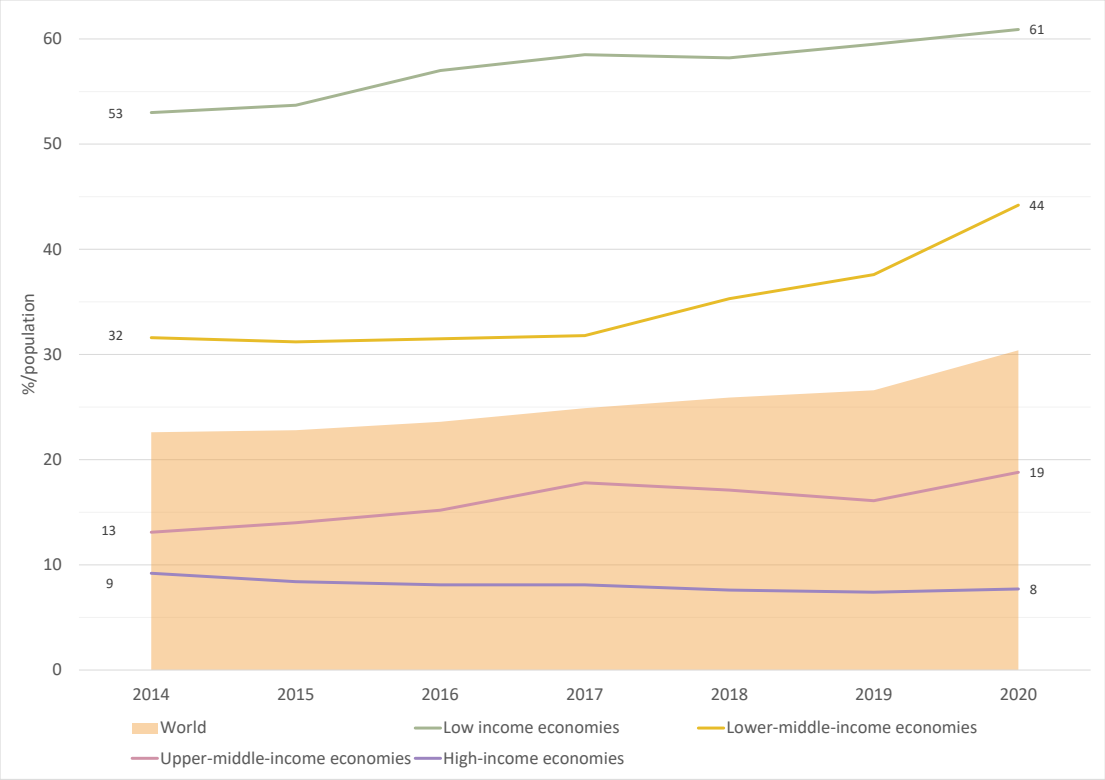


Figure 1.1 – Prevalence of moderate or severe food insecurity (%/population)
Source: FAO (2020)

In this thesis, I have chosen to explore the possibility of improving the credibility of global food price forecasts and making them accessible to all those who need them. The main drive for this work is the conclusion that agricultural commodity price plays an essential role in food security (FAO, 2018). However, today there is a lack of accessible tools for forecasting their variations in the Medium

[1]Part of my doctoral studies has involved a monthly media monitoring on COVID-19 impacts on food and agricultural products and price. All the information has been published in a newsletter format and is available at cland.lsce.ipsl.fr/covid19

Term. We assume that such tools if existed at the beginning of the Covid-19 crisis, could mitigate the deteriorating quality of life of those 118 million people added to the global hunger map in 2020 (World-Bank, 2021b). This assumption has encouraged me to find a solution to this problem. Furthermore, given the development of the food insecurity phenomena, both levels of quantity and severity, this problem is now the concern of populations who, until recently, have not experienced food shortages. In the USA, for example, the number of households that needed assistance in obtaining food increased in 2020-2021 (Coleman-Jensen et al., 2021; USDA, 2021). Similarly, in Israel, the number of families who have become dependent on food associations has increased significantly (Mayzel, 2021; Latet, 2021).

The instability of the global AC prices is a long last common topic in both food policy literature (Taylor, 1919; Brorsen and Irwin, 1996) and forecasting theory (Allen, 1994; Brandt and Bessler, 1983). This issue is a topic of concern for several major international organisations such as the FAO, the OECD and the World Bank. However, none of the existing studies has provided a medium-term forecast that enables users to reproduce the model and glimpse into the prediction algorithm's decision-making process. Beyond that, using machine learning models to predict agricultural commodity prices in the non-short term (as will be expanded below) is also pioneering in its field. This research project aims to bridge this gap by developing a novel tool to forecast AC prices. This tool would hopefully incorporate high reliability, accessibility and replicability.

In conclusion, this complex situation led us to ask three questions that have constructed this dissertation together. The research questions are:

1. Which are the main drivers to AC price fluctuations?

2. Are there differences between the market influence of each player?

3. We define improvement as a forecasting tool that provides reliability, accessibility, comprehensibility and interpretability. According to this definition, can we improve the existing forecasting performance of international AC prices up to a year ahead?

Answering these questions would enable turning price forecasting into an accessible tool that could help improve food security worldwide.

1.2 Background

1.2.1 The Agricultural Problem - An adequate agricultural production does not guarantee access to food

For thousands of years, man has lived in the shadow of scarcity and daily worries about an adequate food supply. This concern was first scientifically expressed by Malthus (1789), who demonstrated by a simple economic model how, given fixed factors of production and declining marginal output, the amount of food per capita would decrease with each increase in the number of consumers. Malthus assumed that the food production increases logistically, while the world population, i.e. the demand side, increased exponentially. Under the minimal food required for human survival, the result of this model is an equilibrium, where the marginal amount of food produced meets the minimum amount of food consumed. From this point on, neither the total food supply increase nor the world population. The clear conclusion of this theory is that man is doomed to continue living under food shortage while consuming the necessary minimum and that the demographic growth will come to an alt.

Since the development of the Malthusian theory in the late 18th century, agricultural production has utterly changed, mainly due to the shift to mechanism-based production and technological improvement, which have led to a significant increase in the amount of food produced all over the world, notably in industrialised countries. At the same time, the world population continued growing to a level where the production curve seemed to be reversed, as the growth rate of food production outpaced the demographic growth (Daily et al., 1998). As a result, food prices fell, and the use of agricultural output changed. Apparently, throughout history, maize has been cultivated and consumed as a crucial component in the daily diet of Native Americans (Ranum et al., 2014). However, today only 8-10% of the maize produced in the US (the world's largest producer) is used for human consumption, while the rest is for livestock feeding (55-60%) and ethanol (35-40%)[2].

Thus, although the Malthusian theory has not stood the test of time, there exist populations who suffer from insufficient food or essential nutrients. According to the World-Bank (2021b), during the year 2020, about 30% of the world's population did not have access to food at an adequate nutritional level. Despite that, 21st-century's hunger is not due to scarcity of food supplies but due to a shortage of tools that allow access to it, that is, low-income (as demonstrated in Fig 1.1).

[2]According to the CME Education page.
source: www.cmegroup.com/education/courses.

1.2.2 Agricultural Commodities

Agricultural commodities (AC) have strong links with the financial world. These products are derived from agriculture and considered necessities, and they include staples such as wheat, maize, rice and soybeans, and wood, cotton, cocoa and coffee. These AC are quoted on the Chicago Mercantile Exchange (CME) or other big global markets exchanges. They are also priced in future contracts and Over The Counter (OTC).

As it comes, only a minor part of futures contracts are delivered physically (CME-Group, 2021), while their role is mainly to be a risk managing tool for producers and middle operators. At the same time, contracts reflect the situation in the cash market rather accurately, especially when the due date comes closer.

In the socio-economic aspect, changes in the agricultural commodity markets impact food supply chains through production volumes. On the demand aspect, the world population is constantly growing, especially in developing areas, where the agricultural sector is particularly vulnerable (traditional agriculture techniques combined with worsening climatic conditions). Parallel, the growing consumption of animal-based products has led to increased grain use, which serves as livestock feed.

Most generally, AC prices tend to be particularly volatile due to their natural dependence on three unstable market elements (FAO, 2012): 1. Agricultural supply is subject to exogenous natural shocks (weather, water and soil quality, diseases and pests), which affect both quality and quantity. Therefore, **agricultural production varies greatly**, not seasonally but also annually; 2. The relatively long and limited production time causes a **low price elasticity of supply**, at least in the short term; and 3. most of the world trade in agriculture is concentrated around products whose regular supply is essential anywhere and at any point in time. In the absence of an adequate alternative supply (usually grain stocks), the consumer will be willing to pay a high price, provided he can consume food. That is, in the short term, the **price elasticity of demand is low**.

In the context of international trade, the topic is perceived often as a tool for alleviating production shocks challenges to food security. More precisely, it creates constant trade flows that may contribute to an all year balanced food supply and diets worldwide (Costinot and Donaldson, 2016; van Meijl et al., 2017).

Due to their high dependency on exogenous factors, the volumes of production of different regions often show abrupt variations which may impact the global market (Abbott et al., 2009). Circumstantially, AC prices fluctuate differently from other internationally traded commodities (World-Bank, 2020a).

Local factors, such as weather or national policies, do not systematically affect trades at a global level. However, in some cases, these local factors have significant impacts on commodity prices (Abbott et al., 2009). Undoubtedly, his-

torical evidence suggests that global AC price shocks, some of which led to large-scale food crises, resulted from changes in local environmental or socio-economic conditions. In their studies, Headey and Fan (2008) first show how the 2005-2008 food crisis came in the wake of severe climatic events and droughts that resulted in poor harvests of several major AC (maize, wheat and rice) in specific key-exporting countries. In a later review, Headey and Fan (2010) argue that political factors in those countries were also central factors in those price increases, as China and India grain stocks declined significantly for almost a decade before the crisis began. Specifically in India, the government's decision to ban exports and to import massive amounts of rice stemmed mainly from public pressure before the upcoming elections, as the country's cereals reserves were still sufficiently high. Specifically to rice, prices increased by almost 300% in only three months (FAO, 2008). The recent trade war between the US and China, for instance, led to mutual tariff increases, which caused high volatility and declining demand from the US, mainly in grain commodities. Another example, at the height of the first COVID-19 wave, Russia prohibited its agricultural export. This decision, which was taken shortly after by 23 other countries, led to a rise in prices despite the decline in world demand at the same time (ITC, 2021). The COVID-19 pandemic has caused significant shocks in grain production, which were very apparent in vulnerable regions where agriculture is based mainly on labour work (Schmidhuber et al., 2020). As sometimes these regions are also big world-producers, food prices have soared high worldwide.

All the price shocks mentioned above were unpredicted; their impacts were substantial and have led to an increase in food insecurity and worsened diet globally, whereas the over-whole influences are yet to come (Laborde et al., 2021). Similarly, Mundlak and Larson (1992) show that most of the changes in world prices pass on to household (consumer) prices.

To understand the AC markets, it is not necessary to know all the influencing factors but, preferably, recognise the most influential ones relative to the price forecasting horizon.

1.2.3 Food Security - Concept and Strategy

Food security is a situation where all people have regular and constant availability, stability, utilisation, and access to a healthy diet that come in hand with their food preferences (Ghorbani and Zou, 1996). Food insecurity, on the other hand, exists when a person lacks at least one of the four components that define food security (FAO and WFP, 2010).

In general, the higher the income level, the lower the share of food expenditure. Indeed, in countries with a high per capita income, the average volume of

food expenditure is lower relative to total household expenditure (6% in the US, 12% in France and 16% in Israel)[3]. In contrast, food expenditure in countries with low per capita income counts as a big part of the monthly expenditure (52% in Kenya, 59% in Nigeria). Therefore, any change in food prices impacts the overall living system in low-income countries, while high-income countries are less vulnerable.

Food price changes relationships with global food security levels have been the subject of many studies over the years. The issue of price volatility is still central today for countries that still rely heavily on commodity exports (FAO, 2002; UN, 2019). Although instability of international AC prices can be determined by demand fluctuations, their physical price reflects an equilibrium between the two. As such, if the price elasticity of demand is low, even small supply shocks can result in significant price fluctuations and cause profound impact in terms of food security (Smith and Subandoro, 2007; Smith et al., 2014). It is important to emphasise that the measure of the food supply is a sum of the amount produced and consumed in an agricultural year, the annual change in the crop stocks, minus the depreciation originating from the supply chain (FAO, 2021). Moreover, it is often acceptable to use a country's food stocks level for indicating its food security situation (Christian and Marco, 2006) or local food prices (Gouel et al., 2016). In recent decades, several extreme changes in commodity markets have put both producers and consumers in dire straits. Moreover, the international context seems to be increasingly unfavourable to producers. Let us point out two problems:

High instability of global AC prices

Most generally, AC prices tend to be particularly volatile for several reasons. First, agricultural supply is subject to exogenous natural shocks (weather, water and soil quality, diseases and pests), which affect both quality and quantity. Therefore, it varies not only seasonally but also annually. The relatively long and limited production time causes low production elasticity. As for the food crisis of 2008-2009, there was an exceptionally high price increase, which, in some countries, resulted in serious hunger riots. Looking into details, both China and India started reducing their excessive grain stocks since the beginning of the 2000s following strategic decisions. As the global grain stocks slowly went down, prices of the world's major crops have mounted gradually (Headey, 2011). These decisions and a dietary transition toward higher meat consumption (mainly in China) triggered higher demand for oilseeds, maize, and soybean. Parallel, fuel

[3]According to 2017 data. For more information, visit `https://ourworldindata.org/grapher/share-of-consumer-expenditure-spent-on-food?tab=table`

prices started augmenting, which stimulated higher local bio-energy consumption in the USA (maize origin bio-fuel) and thus lowered maize exports, which pushed prices even higher. Then, in 2006-2008 severe climatic conditions have resulted in meagre harvests globally. To cope with the challenging market conditions and protect their population from food shortage, several of the world big exporting countries have used export bans or restrictions (Childs et al., 2009). The combination of all these problems has caused massive price soars, which, in turn, resulted in the food crisis.

Similar to the crisis described earlier, in the Covid-19 crisis, too, we see a global process that began before the final explosion. The cease of the African swine fever, which started immediately after the swine flu, has forced China, the world's largest grain consumer, to make massive grain imports to refill its stocks. Towards the end of 2019, China and the United States signed an agreement on ending the "trade war" under which China purchased, among other commodities, American AC for about 32 billion USD, mainly grains and soybeans (Horne et al., 2020). [4] The adverse weather events, which hit both the US and Brazil, damaged grain production and thus negatively impacted the already low grain supply. All of these things lead to a sharp price shock.

Shortly after, severe floods hit several Asian countries. The extreme weather damaged crop yield and prevented labour access to the fields. Moreover, the parallel breakout of the COVID-19 and the low AC supply in the global markets have triggered export restrictions by the world biggest exporters, notably Russia, leaving big importers with limited access to their staple food. By the third month of 2020, many countries were already under extreme mobility restrictions, which had affected the entire global trade system. Among those most affected are the basic and essential commodities, such as foods. At that time, however, AC prices went down. The sudden closure of most of the world's borders and the severe mobility restrictions within the countries has caused a sharp drop in demand for fuels, which led energy prices to collapse. As explained above, a large part of the AC is also an energy source (biodiesel). Indeed, commodity prices used as an energy substitute have fallen following oil prices (World-Bank, 2020b). In this context, it is maize, soybean, palm oil, and even sugar. However, this situation did not last long, and shortly after, the agricultural commodity prices skyrocketed. The crisis affected the whole food system abruptly, moving through the entire value chain straight up to the consumers. Moreover, from the supply side, the strict regulations have resulted in an immediate shortage of workers

[4] China signed an agreement with the US in January 2020. At the beginning of 2020, there was political tension between the Chinese and the Australian governments. This tension drove China to import grains from the US rather than from Australia (Liangyue and Greenville, 2020). This enormous change contributed enormously to the decreased grain supply of the largest exporter (the US).

in the agriculture industry, notably among labour-intensive sectors and areas. In addition to all these, the La Niña events that plagued large parts of the world during the summer and fall months caused severe damage to crops, especially in Northern America. The triple combination of uncertainty about the virus, production slowdown and speculation about a possible food shortage led to panic among governments, who wanted to ensure the availability of food staples to consumers in their countries. As a result, many governments have imposed export restrictions on food products, thus preventing import-dependent countries from having a regular food supply. However, those changes were local and had varied depending on product and country. Therefore, the normal local-global food prices equilibrium distorted drastically (Schmidhuber et al., 2020). Parallel, the demand side had shifted abruptly. In high-income countries, demand transposed from big food suppliers, such as restaurants and public distributors, to private households, causing a sudden increase in demand for high-quality food, such as meat, dairy, fruits and vegetables (Laborde et al., 2020). On the other hand, in low- and middle-income societies, the vast loss of income had forced the poor to increase the already high staple food consumption while giving up other nutritional sources. Putting all these changes together, local scarceness in food products, and the insurance regarding the about to come demands had driven a global food security crisis.

Producers are not always sufficiently protected from price fluctuations

Secondly, price stabilisation mechanisms, mainly in low-income countries, are not always sufficient for protecting producers from extreme price shocks (Gouel, 2011, 2012). Consequently, in times of sharp price fluctuations, many farmers find themselves financially exposed to income losses (Gilbert and Varangis, 2003). Additionally, many countries have a weak capacity to manage the consequences of price instability: market risk management instruments, such as sufficiently large food stocks, are only used in a few developing countries and do not yet constitute a comprehensive solution to the price instability problem, which harms both producers and consumers. Given that those citizens of low-income countries spend a large portion of their total income on food, the negative impact of any price shocks on their daily life becomes a real threat. This finding leads us to examine the possibility of forecasting AC price changes in a way that would be both accurate and accessible, even for users with low economic capacity.

1.3 Approach - Science Without Borders

The doctoral thesis was made possible by generous funding from CLAND, aiming to assist the global efforts to promote food security. Being part of the CLAND project, this study advocates the idea of open science - to be used by others. I strongly hope it will be used by those who can improve it, but most of all by those who need it. Intrinsically, this project works under three main principles.

1.3.1 Principle I - Concise Comprehensibility

The first principle that led this work derives from three specifications stated in section 1.2.2 and 1.2.3: 1. AC prices are highly volatile due to their dependency on unstable market conditions. In the absence of sufficient food stocks and the more basic food consumption (i.e., food that is not highly processed), sharp fluctuations in AC prices often cause unstable food prices; 2. In low-income countries, the average volume of food expenditure is high relative to total household expenditure. Also, in these countries, the food reserves and the governmental tools for food price stabilisation are relatively limited, so that imported food prices are significantly affected by changes in AC prices in the global markets; and 3. Finally, at relatively low-income levels, consumption of staples out of the total diet is relatively high compared to consumption of processed or animal products (FAO, 2012).

This led to our perspective that food security is feasible under the conditions of stable food prices, especially for those of low-income levels. Furthermore, machine learning models can contain numerous explanatory variables and analyse complex situations while maintaining simplicity in their operation. As such, choosing them as the leading research method in this study was a natural choice. Advances have been made in transparent ML models intended for predicting various quantities of interest for food security (Beillouin et al., 2020; Blumenstock et al., 2015; Lentz et al., 2019). While their link to food security is unquestionable, none of them addresses AC prices directly.

This thesis restricts the independent variables to local (domestic, regional and continental) productions and yields. These variables inform directly on the level of commodity supply, which is usually an unstable component of the market, and have a significant impact over AC prices (Headey and Fan, 2010). Although these variables can be potentially relevant for crop price predictions, they are rarely used as a sole predictor in econometric-based studies due to the risk of endogeneity.

1.3.2 Principle II - Interpretability

The second principle that leads this study states that ML models should be interpretable (Molnar, 2019). To date, as will be further detailed in Chapter 2, models that have focused on AC price forecasts through machine learning have either lacked transparency or have used complex analysing methods either produced only final results and thus kept the logic behind the algorithm decision unrevealed.

Here, we train the ML models to analyse the relationships between inputs describing local crop production and yield variations and outputs representing crop price variations. The proposed models use several techniques to rank the producing units according to their level of influence over the global market and quantify the effect that any change in the annual regional productivity has on global price changes.

1.3.3 Principle III - Accessibility

The third and final principle is accessibility. To benefit all policymakers, the model must rely solely on reliable and entirely accessible data. Regular access to data is a key to developing a flexible food security strategy that can keep up to date and adapt to any change in the field.

Under this concept, this research uses only publicly available yearly production data (FAO, 2020) and monthly price data (World-Bank, 2021a) and open-access software.

1.4 Machine Learning as a price forecasting method

Predicting AC prices serves essential needs for any time frame, from the shortest few seconds to years ahead. Chapter 2, which reviews the existing price forecasting literature, illustrates the richness of the extant economic models for forecasting prices over different time frames. In short terms, a wide range of statistical methods can analyse or predict AC prices at a high-reliability level. However, there seems to be a gap between the current forecasting methods and the need to predict AC prices over time ranging from a few months and up to a year ahead, in a way that will make it possible to analyse the forecasted results and understand their causes. Specifically, a method has not yet been found that will make it possible to accurately predict AC prices while relying on exogenous explanatory variables without involving economic theories.

Recent developments in technology and research skills have accelerated the application of statistical and machine learning algorithms. These models solve

complex problems using relatively simple methods while providing prediction results of high accuracy, even when compared to particularly advanced models (Storm et al., 2019; Lobell and Burke, 2010). As a result, these models have become more frequent when forecasting complicated processes.

Most generally, ML can treat two types of problems:

Supervised, in which data with existing labelling or classification accompany the model to create new observations, i.e., forecasts, based on the information known; And

Unsupervised, in which the objective is to divide a high-dimensional raw set of data into clusters based on similar internal structures and patterns, as identified by the algorithm.

Here, we try to predict a response variable (AC prices) as a function of several inputs and use a set of observed input-and-output values. Therefore, we use supervised learning techniques.

In supervised learning, X is a data set of observed inputs variables, and Y is a set of the observed output variable to be predicted. The primary goal of this method is to find a model that could use the explanatory variables to predict the value of the dependent variable. Supervised learning can generate accurate and reliable forecasting results if integrated with the appropriate model. Following the assumptions of independent variables and a stable data generating process across training, and the application procedure in this thesis is as follows:

1. Define the Training-set (*in-sample*): a collection of m X_i's and Y_i's observations, $i = 1, 2, ..., m$.
   Where X_i is a row vector such that $X_i = (x_{i,1}, x_{i,2}..., x_{i,K})$;

2. Train several algorithms to forecast the output (*out-of-sample*) from the inputs, using the m observations from the training dataset. Repeating the process for several algorithms;

3. Compare the forecasting results based on a test dataset to select a model, aiming to minimise the forecasting error.

One of the salient advantages of ML is the ease of applying it to a wide range of data and research methods. Because the models examined in this study have not yet served for Medium Term AC global prices forecasts, this advantage is particularly significant.

The three ML methods considered in this study are decision-tree based algorithms. Tree-based approaches are embedded as data-partition predictors of if-then forecasts. Moreover, they stratify the predictor space into simple and homogeneous domains and use splitting rules that are easy to implement.

1.5 Data

This project uses monthly prices data of three agricultural commodities (maize, soybean and cocoa). The prices are publicly available on the World-Bank website (World-Bank, 2021a), starting in January 1960 to the present. This time range is large enough to take inflation into account. Thus, the first step was to deflate the nominal prices to bring them down to the same scale. For that, we used the monthly nominal published by the World Bank. Then, we replicated the process on 12 different price indices, searching for the index that will bring prices closest to the real AC prices, also published on the World Bank website, but on an annual basis.

Let us define $q^n_{m,y}$ as a nominal price relative to a month m in a year y, $q_{m,y}$ as the deflated prices and $In_{m,y}$ as the price index, both relative to the same period. Setting 2010 as the year of basis ($In_{m,2010} \approx 100$) the deflation was as follows:

$$q_{m,y} = \frac{q^n_{m,y} \times In_{m,2010}}{In_{m,y}} \tag{1.1}$$

Finally, we chose the agricultural price index to serve as the deflator. The decision derived from three factors: The first derives from Tadasse et al. (2016) indicating that, although widely used, the US consumer price index (CPI) could be a biased deflator when dealing in a global market that includes both developed and developing countries. The second reason is a relatively lower gap (measured in terms of Root Mean Square Error) between the AC annual real prices, as published by the World Bank. Third, to reassure this decision, I directly asked the World Bank's commodities team, who approved this decision.

Relative to the models' input, we extracted annual crop yield and production time series from the FAO data website (FAOSTAT) for all years available (1961 to 2019), relative to three geographic scales: continental, regional and national (local).

1.6 Scope of Work

This doctoral thesis examined and forecasted the monthly prices of several AC,[5] which are of high importance in terms of trade and food security. All along, the research relies on publicly available data of annual production, according to three types of geographical divisions. This thesis offers a double contribution:

[5]The complete price forecasting tool includes three beverage commodities (cocoa, Arabica coffee and Robusta coffee); two oils and meals commodities (soybean and palm oil); and three grain commodities (maize, rice and wheat)

on the academic side, it is the pioneer in performing a Medium Term price forecasting of agricultural commodities using ML. It also detects the main drivers for AC price changes through investigation of the ML algorithms. Second, it offers a practical, non-academic contribution - it provides AC price forecasting tools that can benefit policymakers who lack the tools that are required for trading in the global AC markets in an optimal manner. As such, although the thesis incorporates diverse and advanced research approaches, the final model will supply policymakers with a ready for use product: it is accompanied by detailed and comprehensive explanations, under the promise to be accessible, even for non-specialists. Moreover, as part of the transparency agenda that accompanies this work, all the data and techniques assisting it are accessible to the public, free of charge, and have a high level of reliability.

1.7 Organisation of the thesis

The current chapter presents and describes the challenges in predicting the price of agricultural commodities. The chapter sheds light on the importance of maintaining stable prices and pre-prepare for extreme changes and describes the scope and purpose of the work. Chapter 2 presents key literature references and past studies that all together created the basis for this thesis. Chapters 3, 4 and 5 are articles written and submitted (or published) in scientific journals. Chapter 3 analyses the maize market and identifies which regions drive price shifts in the global maize market through changes in their annual maize production. In addition, it offers a ranking of the relative impact of 17 market players. Chapter 4 also deals with the maize market. This article directly continues the project presented in the third chapter. Again, it relies on its conclusions while attempting to trace the optimal model for forecasting global maize prices in periods of up to one year ahead. Here, too, all models are opened to analyse specific extreme price fluctuations events and understand the main drivers for their occurrence. Next, Chapter 5 is a natural continuation of two previous studies. It researches and examines additional ways of applying forecasting models for two other major crops, each of different market characteristics - soybean and cocoa beans. All three commodities are of high global importance, and each is associated with a different group of AC, according to the official division as done by the World Bank. Finally, the 6th and final chapter summarises and discusses the main finding and conclusions derived from this doctoral thesis.

Chapter 2

Literature Survey of AC Price Fluctuations Research

```
The end of starvation
   reflects a shift
   in the entitlement system
```

– Amartya Sen, Poverty and Famines (p.82)

The term *price fluctuations* defines a change in a commodity price relative to a given period under the research needs as defined by the researcher. For example, Adjemian and Irwin (2018) analyse by-minute price changes of three AC and therefore refers to price fluctuations as a single minute change (very-short term). Other studies define price fluctuations according to a day (Karali et al., 2019), a year (Haile et al., 2017) and even of decades, as in Fuss et al. (2015).

The literature concerning AC price fluctuations consists of numerous methods. Following Popkin (1977) and Piot-Lepetit and M'Barek (2011), the AC prices literature had been categorised into four time-horizons: very short term (VST), for time frames of a few hours and even minutes ahead; short-term (ST), for periods of one day to three months ahead; medium-term (MT) for periods of up to 18 months into the future; and long-term (LT) for any further periods. As will be seen during the review, the longer the analysis range, the greater is the complexity of the model.

Beyond the time-frame differences, the richness of approaches concerning AC price fluctuations have led to the development of a wide range of methods for analysing them. On that account, we also distinguish between two analysis frameworks of AC price fluctuations: Statistical (non-structural) methods; and structural methods, which are based on economic theories.

Structural methods assume a theory that correctly describes the actual economic behaviour of prices and serve for non-short periods analysis. We classify this group into two:

- Equilibrium models: These economic models represent market equilibrium in various determinants. Equilibrium models vary by complexity, as detailed in this literature survey. These models rarely serve for direct forecasting but are mainly used for price analysis (Deaton and Laroque, 1992) or in counterfactual simulations, which indicate the role of each variable (e.g., changes in policy, climate, etc.) on the price behaviour. Although considered too complex for direct forecasting, equilibrium models can serve as an indirect forecasting tool after calibration on different sources (Valin et al., 2014; Kan et al., 2018).

- Statistical methods: Structural Vector Autoregressive (SVAR) models are based on a statistical estimate of several series, assuming long-term and short-term relationships derived from economic theory. When estimating price fluctuations using SVAR, the researcher assumes that shocks in a particular explanatory variable are neutral in the long run. In this sense, shocks in other explanatory variables and the explanatory variable itself (price) are the ones that ultimately determine the potential price.

The **non-structural methods** group includes all the analysis models based on a particular statistical process rather than an explicit economic theory. These models are used concerning price fluctuations ranging from a few minutes and may reach time horizons of up to about a year and a half ahead. As before, we distinguish between two non-structural approaches:

- Causal inference methods: This type of literature usually aims to estimate the causal effects of certain variables on price fluctuations. The estimation of the causal effects is made by observed prices, using direct measurement using relatively simple linear or smoothed trends. In the short term, the crop production, the total supply and demand for agricultural products are given and fixed. Then, assuming they are a function of a wide range of constraints, i.e., production, policies, transportation costs and suchlike, the researcher calculates the observed price directly, using a time series of affecting variables. As such, these models mainly serve for relatively short-term analysis, such as future prices (Colling et al., 1996), spot prices (Weymar, 1965) or price in practice (Jichlinski, 1983).

- Forecasting methods: These methods serve for forecasts in periods ranging from a few minutes up to relatively one year. They include time series models, regression and classification based methods. During the last years,

forecasting models have become common in forecasting different topics such as yield (Laudien et al., 2020), production (Beillouin et al., 2020) and even nutritional needs (Zeevi et al., 2015). However, concerning prices, these are only short-term forecasts studies that have applied these models up to this day.

Despite this precise distinction between structural and non-structural approaches, there are models, such as WASDE SAP, which combine in their forecasts several methods from the two groups (Frederic and Gerald, 1999; Hoffman et al., 2018).

Whilst the thesis aims to forecast AC prices; this survey covers both models of prediction and models of causal inference. Indeed, although forecasting is very useful, economists mainly study causal inferences: how oil price affects world food prices (Abdoulkarim and Zainab, 2011); what is the effect of global AC prices on local agricultural production (Haile et al., 2016); or how wine prices react to changes in wine stocks (Bukenya and Labys, 2007)?

2.1 Very short term and short term agricultural commodity price research

Despite the dissimilarity between some of the analyses regarding price valuation in the VST (of about up to one-day intervals) and ST (time intervals of up to three months), it is sometimes difficult to differentiate between them. Therefore, this section surveys methods of both time frames while indicating to which of them each method belongs. In very-short terms, any data regarding supply and demand is unchanged, so that AC prices are analysed using relatively simple **non-structural** methods and can be forecasted directly from historical data. Most generally, any period added to the analysis model enables the examination of higher complexity and added relationships or inter-relationships at different levels. However, it comes with the price of a decrease in the analysis ability of the model.

AC are soft commodities but, like other goods traded in international markets, are often traded according to contracts and priced "Real-time", with no breaks. AC are known for their high volatility and thus subjected as risky assets (Clapp and Isakson, 2018). One of the major factors for AC price change is unexpected drastic fluctuations that, although not frequent, has a critical role in AC trade (ITC and UNCTAD/WTO., 2001).

The literature aims to analyse AC prices in the VST studies the causal effect of certain variables on AC prices, i.e., to analyse them. One of the questions of

interest in this literature is if and how USDA announcements affect AC futures prices.

Adjemian and Irwin (2018), for example, analyse the impact of the USDA reports on the prices of maize, soybean and wheat in the Chicago Mercantile Exchange (CME). They compare AC price fluctuations before and after May 2012, i.e., during the period in which trading ceased on the morning of the publication of the government report, as opposed to a period in which trading is continuous. To do this, they examine the minute fluctuations according to three versions of price fluctuations, all three of which use the observed price by the minute. In addition, they test the price volatility according to three equations: the maximum daily difference according to log (first equation), an average daily difference (second equation) or the difference between the last price observed before the publication of the report (third equation). Nevertheless, the price patterns analysis always directly estimates one variable - price fluctuations explained by the number of contracts issued per minute.

Another study conducted in recent years (Karali et al., 2019) also questions the impact of US government reports on the daily price of the same commodities, i.e., maize, soybeans and wheat. In this study, the researchers used USDA reports on yields and crop projections and compared them to forecasts published by private analysts. The definition for the difference between them (in per cent) is the degree of surprise of the market. Hence, the higher is the difference between the forecasts, the higher is the relevance of the USDA forecasts (as aforementioned, the process of all USDA reports is done under a cloak of secrecy). Moreover, Karali et al. (2019) performed their analysis for a slightly longer time frame compared to Adjemian and Irwin (2018) and thus investigated the AC prices with somewhat higher complexity.

Finally, for VST, a relatively modern method of analysing commodity prices is a machine learning-based approach of data clustering, i.e. unsupervised learning. When used by financial institutions and entities, or governments, the researcher obtains the information from selected sources (usually for a fee). It reaches millions of records a year to create a rich and detailed database - Big Data. As unsupervised learning is more flexible than supervised learning, in the sense that it does not require pre-definition of explanatory variables, it is a convenient tool for VST and ST price analysis. The researcher defines several main variables or points within the dataset to analyse the data. Next, the algorithm creates new variables according to shared characteristics it identifies through the learning process. This set of variables can express almost 90% of the total existing variance (Hativ and Mazouz, 2021). According to the data's centres of gravity, a further division of the information using the k-Means method (an unsupervised learning model) can create even more efficient learning. This type of

learning is considered particularly effective for detecting market opportunities and players associated with the market. k-Means is also efficient in identifying outliers activities in the studied sector. However, on the downside, it demands high acquaintance with the learned market and careful data organisation and filtration. Also, handling Big-Data requires using cloud tools to store and process this amount of data. That is, a machine strong enough to support these requirements, as well as a budget for maintaining a cloud with large memory.

Analysis of agricultural markets through unsupervised learning exists in short term analysis in the scientific literature, although it is somewhat rare. Kim and Dharmasena (2018), for example, analysed the US pecan price by identifying interactions between four leading countries in this market. First, to define the most relevant variables, they apply the graphical-based detection algorithm directed acyclic graph (DAG) (Pearl, 2009). Next, an autoregressive model captured the causality structures that best determine bi-weekly price fluctuations' drivers.

Another academic study, Deng and Yǔ (2019) explored patterns and internal relationships over slightly longer time frames in another academic study. Using a time series of production costs per acre from all fields in the Chinese Heilongjiang province, the researchers explained price fluctuations in soybean prices. In the first stage of identifying the factors for the price fluctuations, a Dynamic Programming (DP) algorithm runs in iterations on the time series data. Next, the Toeplitz Inverse Covariance-based Clustering (TICC) model analyses the data, using these variables to find the proper inter-connections within the price. All in all, the researchers discovered four patterns over monthly time frames. Although ST is relatively short (up to three months intervals), it is a long enough time to account for the impacts of external factors variation on the price of AC. Such fluctuations are frequently related to one or more of the following:

1. Adverse weather events;

2. Export bans or restrictions, notably when posed by big exporters;

3. Enormous purchases (e.g, the 32 billion USD purchase made by China in early 2020 (Horne et al., 2020)) or panic buying behaviour of numerous importers/consumers, especially of staples (Hobbs, 2020; Wright, 2008) ; and

4. Rapid and sharp changes in currency, usually the USD (Headey and Fan, 2010).

To enable the performance of causal inference but also of forecasting. Moreover, this time frame allows the application of structural models and non-structural analysis.

2.1.1 Non-structural models

When forecasting AC prices for short terms (ST), which are slightly longer than the VST, non-structural big-data analysis is again prevalent for decision making. The data are usually presented as time-series (TS), reflecting the development of activity over time (Mantzura, 2016). As here time frames are a bit longer, analysis for food prices, which do not change by the minute, become relevant, along with AC.

Li et al. (2010) compared the forecasting accuracy of an Artificial Neural Network (ANN) model and ARIMA model on the tomato price in China for three forecasting horizons: daily, weekly, and monthly. Both ANN and ARIMA have been proved to provide good ST time series forecasting (Zhang, 2003; Ratnayaka et al., 2015). They successfully obtained high forecasting accuracy when using the ARIMA model for a forecast period of one day to one week. The authors show significant rapid growth in the relative error, and indeed, for a price forecast for the term of one month ahead and further, ANN had become a preferable model to use.

The use of several time-series based models also serves the European dairy market. In their study, O'Connor and Keane (2011) have applied Conditional Heteroscedasticity Models (ARCH and GARCH) and Time Series Models (ARMA and ARIMA) on the price of dairy products in Europe. Rather than declaring the best forecasting model, the authors point out the strong influence of government decisions over many agricultural markets. According to this study, government decisions undermine the normalisation that guides many of today's models. That is, market-related predictions might become more realistic if a model is used that does not require forced research assumptions.

Similar to them, many more studies have been examining AC price fluctuations in the short term. The common point of these studies is the integration of TS or ML models with big-data information. However, while these works present low forecasting errors, they do not explain the source of those price fluctuations, i.e., they do not diagnose the results obtained by the forecasting algorithms.

2.1.2 Structural models

All the models presented have operated in a purely statistical method. However, for not VST, structural methods that assume that a particular theory describing the actual economic behaviour are also possible. In relatively short terms, SVAR (Blanchard and Quah, 1988) is considered a reliable method. Although originally presented concerning the labour market and Gross National Product (GNP), SVAR is now a widespread tool in other contexts, agricultural commodity prices among them. These models assume that long-term and short-term re-

lationships stem from economic theory. The central concept of SVAR is that all the shocks in the economy affect all the variables. If, for example, the dependent variable is price and the explanatory variables are the regional production, SVAR assumes that the price is affected by changes in all variables, including itself. Similarly, the regional productions affect each other and are affected by the price. The model assumes that the effects of the explanatory variables on the dependent variable reset in the long run.

An excellent example would be the study conducted by Hao et al. (2017) on the relationship between US ethanol prices and corn prices in developing countries. Beyond the obvious fact that prices in the US market are stable relative to prices in emerging markets (see explanation in the Discussion 6.4), the researchers found that most of the effect of ethanol price fluctuations on maize prices in the countries surveyed occurred within a period of up to three months.

2.2 Long term agricultural commodity price research

In the long run, AC prices fluctuate primarily as a function of trends or long-time changes in trade markets, the natural environment (climate), or politics and society. Long-term AC price forecasts play an important role when building long-term strategies. These could be global cross-sectorial, such as the Sustainable Development Goals of the UN (2021); IFPRI's food strategy (IFPRI, 2018) or even local land allocations between agricultural activities (Zelingher et al., 2019). Due to the high variability and number of possible scenarios, long-term strategies tend to be highly flexible and usually change over time to adapt to actual changes. Following Headey and Fan (2010) and Swinnen and McDermott (2020), the main factors for long-term price shocks are:

1. Economic growth, mainly in developing countries;
2. Price increase of substitutes (food or other commodities);
3. Change in utilisation of AC, such as for energy or livestock feed;
4. Energy/Fertilisers price spikes;
5. Gradual and continuous decrease in AC stocks; and
6. Technological growth recession, especially concerning the sharp rise in crop yields, as has occurred in recent decades.

An example of a demand-side market that significantly affects the prices of agricultural commodities, in the long run, could be the rise in the income level in developing economies. The recent rapid price rise observed in many developing countries is a serious concern, especially when food prices are concerned. Food prices in emerging economies have risen sharply over the past decade, as seen in the local consumer food price index compiled by FAO (2020) in Fig 2.1.

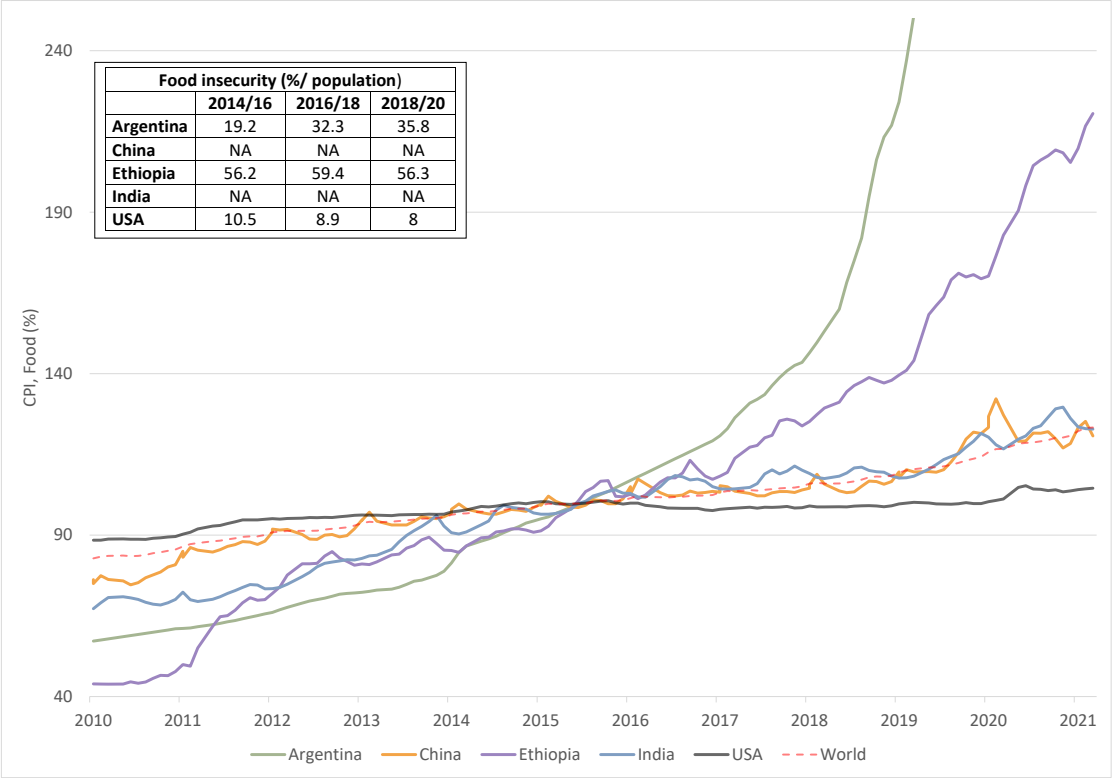


Figure 2.1 – Consumer Prices, Food Indices (2015 = 100 %)
Source: FAO (2020)

In the long term, a key driver to the rapid surge in food prices is the rise in revenues in emerging economies, notably China, which accounts for about one-fifth of the world population. The increase in revenue has allowed consumers to reduce their traditional grain-based diets in favour of a more expensive menu, containing higher rates of animal source products, fruits and vegetables. This trend has, of course, affected the rise in world food prices (World-Bank, 2014). Nonetheless, in developing countries, a high proportion of the population has a low-income capacity and thus spend a large portion of their income on food purchases. As a result, a local rise in food prices is much more significant than a similar rise in high-income countries, i.e. the US or the EU. The growing consumption of animal-based food leads to price increases of livestock feed, including maize and soybean. In the long term, this inflation is an incentive to expand the growing areas of these crops, often at the expense of less rewarding grains, such as rice, which are consumed only by humans (Childs et al., 2009; Headey and Fan, 2010).

An example of a variable that affects AC prices in the long run, this time on the supply side, is energy price increment (World-Bank, 2016). Although not agri-

cultural products, fuels are involved in the entire agricultural supply chain, from the early stage of the purchase/transportation of inputs until the final delivery to the point of sale. Farmers who irrigate their crops are particularly affected by high pumping costs, while those who use modern seeds face higher prices of fertilisers [1]. Moreover, the fact that many countries subsidise fuels for their residents, or local farmers (Nwachukwu and Chike, 2011), causes a non-reduction in the use of the same inputs by farmers, which places an economic burden on the entire country. As will be seen in the case of an increase in the general income level in China, here too, an increase in the demand for petroleum-based materials or oil has a worldwide effect on the prices of agricultural produce in the long run. Another side-effect of the rise in fuel prices impacts international trade. As aforementioned, rising fuel prices increase the transportation of agricultural commodities, making their final price for importing countries to elevate: the higher is the price of energy, the higher is the additional price importing countries must pay for their imports. Facing such situations, governments of importing countries may try to lessen imports by encouraging local farmers to produce more of the imported crop through subsidies, reduced producer taxes or other supports (Nidhiprabha, 2019). In this respect, the cost of all production will increase, as importing efficient producers (e.g., the USA for maize, Thailand for rice or non-agricultural products) will become less and less cost-effective when imported. The primary way to deal with this dependence is to use crops as a source of bio-fuels, the most common of which are ethanol (made from maize, sugar beet or sugar cane) and bio-diesel (made from canola, soy or palm oil). Indeed, over the last few decades, the use of biofuels has been steadily rising, pushing the prices of those goods up and shifting their use even further from their most basic purpose, as food for humans (HLPE, 2013). Moreover, given that these crops also make up the bulk of the livestock nutrition, feed costs also rise, leading to increased dairy, meat and eggs prices. As expected, the high commodity prices are an incentive for farmers to allocate more significant parts of their land in the long run. This land reallocation comes in favour of growing fuel-substitution plants, at the expense of products used mainly for human consumption, such as wheat and rice (Abdoulkarim and Zainab, 2011; Tadasse et al., 2016).

Another important factor is long-term exchange rate fluctuations. Most of the AC global trade prices are of USD units. As each country has to convert its currency into USD, each fluctuation in the USD-local currency ratio will have a strong influence over the real-price value of the commodity (FAO et al., 2020).

[1] Modern seeds produce high yields but also require the use of fertilisers (Hamzei and Seyyedi, 2016). The production process of fertilisers involves the use of fuels. Therefore, fuel price increase leads to elevation production costs of fertilisers and hence to higher prices.

This factor, of course, is of decisive influence on the prices of other commodities. Compared with Euro, changes in the USD value have influenced AC prices at higher levels Gilbert (1989); Mitchell (2008).

Owing to the nature of the possible changes over long periods, models that analyse AC price changes tend to be more complex compared to short-term forecasting models (Piot-Lepetit and M'Barek, 2011). Generally speaking, two types of models are most known to serve for long-run AC price predictions, both are structural: partial equilibrium (PE) models and computable general equilibrium (CGE) models (Thomsen, 2021).

2.2.1 Partial Equilibrium (PE)

PE models traditionally describe a single or few market(s) while excluding income effects and feedback from other markets. However, the chosen market(s) can include many commodities. PE models use observed data to produce an outcome from a series of individual equations. As such, PE models generate a mathematical equilibrium in the desired market.

One could distinct between PE models by differentiation of the production/demand spatial resolution units of the production/demand units (e.g., countries, regions, grids); term of demand (e.g., caloric consumption, price and income elasticity, bio-fuels demand), and by the type of explanatory variables for the dependent price factor. The latter varies greatly and could include many variables - land and non-land inputs, yields, investment in agricultural R&D amongst them.

The Global Biosphere Management (GLOBIOM) model (Havlík et al., 2011) is an example of a global PE economic model. The principal data sources used are mainly from FAO. GLOBIOM provides policy analysis on global issues related to land-use competition between vegetative farming (18 major crops globally + 9 crops for the EU), grassland (as livestock feed of 7 animals), bioenergy crops (short-rotation crops) and managed forest. The model considers relationships between all products and land use and enables changing land allocation concerning exogenous factors such as price and productivity changes. The regional coverage consists of 37 regions (globally), which represent the global trade and demand. The demand side representation in GLOBIOM acts as an endogenous system specified by price elasticity increasing functions, relative to gross domestic product (GDP) per capita, population (exogenous variables) and a price for each product (endogenous). On the other hand, the supply side of the model is the productivity of three land-use sectors: agriculture, forestry and bioenergy, presented by production functions, along with related environmental parameters such as greenhouse gases (GHG) emissions and production/supply costs. The overall equilibrium of the agricultural and forestry markets is the maxim-

isation of the total producer and consumer surpluses under constraints of resources, technology and policy. A detailed description of the GLOBIOM is available at Havlík et al. (2018).

2.2.2 Computable General Equilibrium (CGE)

While based on the same mathematical equilibrium as PE, CGE are cross-sectorial models, which are commonly used in macro-economic research to take into account a large number of factors and parameters. As such, the price elasticity of demand in CGE models is relatively high, whereas price responses to supply shocks are comparably low (Burfisher, 2021; Valenzuela et al., 2007). The demand side in CGE is an aggregation of several factors like GDP, oil price, and even governmental policies (Hertel et al., 2016).

When talking about global trade in agriculture, one of the leading CGE models is the Global Trade Analysis Project (GTAP), initially developed by Hertel (1997). It functions as a framework for many other works and models, including the Agricultural Model Intercomparison and Improvement Project (AgMIP) (Rosenzweig et al., 2013), the global Modular Applied General Equilibrium Tool (MAGNET) (van Meijl and Woltjer, 2012) and local models such as the Israeli General Equilibrium Model (IGEM) of Palatnik and Shechter (2008). GTAP is a general equilibrium framework applied at the global level. By definition, it captures all economies and under the assumption of free trade flows (perfectly competitive) between them, including interactions between sectors and markets (Corong et al., 2017). GTAP consists of a large set of data, counting 114 countries, where each one has its local industries and products, and based on publicly available information. In addition, it recognises five national income levels and represents economic behaviour within economies (forms of production and preference functions, or calibration of elasticities) This greatness is also the drawback of the framework, making it highly complex. Specifically for agriculture, GTAP, on its GTAP-Agr version, is a static model.

The model seeks an equilibrium point: The producers (firms) are profit maximisers under the constraints of production technology (given prices) and input of labour, capital and intermediate deliveries. On the other hand, the demand side aggregates all the private households (as a single consumer), which maximise their present value of current and future utility under a limited budget and given prices. They are considered as having the initial endowments, and thus, they own all the production factors and consume the final products, according to a consumption function. As GTAP-Agr is static, technology is assumed to be fixed in the short term.

2.3 Medium Term agricultural commodity price research

Medium Term AC price changes (time intervals of three to 18 months) are most often associated with shocks in commodity markets; local changes, especially those that have a high impact on the market (Hertel et al., 2016); or catastrophes on a global scale. On the scientific side, it is customary to examine a change in AC prices mainly in the context of price cycles in the AC itself or as a result of changes in the prices of other commodities. Consistent with Labys (2003) argument, the in-depth literature review conducted throughout the making process of this research project shows that most of today's studies use one or a combination of non-structural techniques. e.g., TS based models, including vector autoregression (VAR), exponential smoothing or models of heteroscedasticity. However, structural models, too, are widely used in the academic literature, albeit for price analysis, as will be described below.

When dealing with the analysis or forecasting of AC prices in Medium Terms, it is possible to use models based on time series, as was done for the short term. In this case, the option to apply interpretable models, i.e. techniques that are not necessarily "black box", is also open. Despite the rich academic literature, when it comes to AC price studies using machine-based methods, the vast majority of them adhere to a relatively limited number of models, those that are not interpretable. A unique work conducted by Xiong et al. (2018) predicted vegetable prices in the Chinese market in time frames ranging from a month to six months ahead, based on monthly price series of up to 12 years. The uniqueness of this study stems from the attempt to understand the nature of the markets in-depth before operating the forecasting model. As part of this approach, the authors combine several seasonal-trend decomposition procedures and extreme learning machines (ELM), where each comes at a different stage with a different role. At the first step of the learning process, the input data (prices) are loaded into a seasonal-trend decomposition time series algorithm, based on a smoothing technique that deals with missing data: STL (Cleveland et al., 1990). STL decomposes all five price series into detailed seasonal, trend and remainder (residual) components. This step is of great significance since, in the next step, an EML performs the forecasting task. EML is a neural network fast machine learning technique, which is known for its high forecasting accuracy (Huang et al., 2006). However, these models cannot be opened and thus are not interpretable. The initial data decomposition allows a separate prediction of each of the three components to predict that it is both more accurate and more understandable in terms of the composition of the result. While there is no significant opening up of the "black box", this study represents a breakthrough in attempting to inter-

pret AC price prediction results through machine learning.

Although somehow limited in the academic aspect, AC price forecasting in the non-academic world is very developed. Currently, several international organisations periodically publish Medium Term AC price forecasts. The World Bank, for example, publishes price forecasts (medium and long period) of 20 AC, including three animal-based products. The forecast is published twice a year, and the results are publicly available. However, an in-depth search has not yielded up-to-date publications regarding the technique used to predict those prices. Existing publication regarding the sugar market (Vries, 1980b) presents a CGE model that is probably no use for today's Medium Term forecasts. Another publication presents the banana market (Vries, 1980a) in a manner that could indicate the current forecasting methodology. The Commodity Price Index is published once a month by the World Bank in two days lag of the month measured. Twice a year (every April and October), the World Bank produces general price indices forecast along with the price of each commodity. The purpose of which is to enable the analysis of the information. The agricultural commodities item (accounts for almost 65% of the non-energy commodity index), consists of three groups (beverages, food and raw materials), is highly volatile. As such, its potential contribution to the forecast error is significant. The monthly prices and indices are built based on commodity prices, collected regularly from the world largest international markets. In the case of bananas, these are the markets in Hamburg and US Gulf, under the guiding assumption of pure competition. The global AC price is an aggregation of the prices obtained from each producing country, according to its weight from all weights in the relevant month. For each such market, the price (in USD terms) is a function of the local yield (extrapolated based on their historical growth rate), an international price index and relative market pressure in the US market (defined by export criteria). The model also accounts for trends and seasonality.

The USDA uses the World Agricultural Supply and Demand Estimates (WASDE) model to forecast the future prices of numerous grains, meals and oils, sugars, animal-based products (milk and meat). The forecasts results are made public every month for up to 18 months ahead. However highly used by governments and policymakers (US-HR, 2009), these models are not only complex (Hoffman et al., 2018) but are also un-interpretable, as their developers keep the methodology and the market data under "Lock-up Conditions" (Mallory, 2021).

To summarise this survey, as of today, various models provide price forecasts of AC in the Medium Term, which is the most critical for the determination of an immediate food security strategy. However, these models do not provide sufficient information about the factors that led to the final forecast results. Thus, these models do not allow policymakers to assess the risks they are facing nor

understand the market in which they operate.

In order to fill this research gap, this doctoral dissertation provides AC price forecasts in the Medium Term. Unlike current models, price forecasts are made solely in a way that will also allow for minimal forecasting error. At the same time, they also provide the users with a glimpse of the model decision process. If will follow the detailed and comprehensible instructions, even non-specialised users will be able to replicate the forecasting process and understand the forces operating in each relevant market.

Chapter 3

Assessing the sensitivity of global maize price to regional productions using statistical and machine learning methods

Co-authors: David Makowski, Thierry Brunelle [1]

Abstract

Agricultural price shocks strongly affect farmers' income and food security. Therefore, it is essential to understand and anticipate their origins and occurrence, particularly for the world's primary agricultural commodities. This study assesses the impacts of yearly variations in regional maize productions and yields on global maize prices using several statistical and machine learning (ML) methods. Our results show that Northern America is by far the most influential of all regions considered. More specifically, our models reveal that a yearly yield gain of +8% in Northern America negatively impacts the global maize price by about -7%, while a decrease of -0.1% is expected to increase global maize price by more than +7%. Our classification models show that a slight decrease in the maize yield in Northern America can inflate the probability of a global maize price increase. The maize productions in the other regions have a much lower influence on the global price. Among the tested methods, random forest and gradient boosting perform better than linear models. Our results highlight the interest

[1]Chapter published in the Frontiers in Sustainable Food Systems journal (DOI: 10.3389/fsufs.2021.655206)

of ML in analysing global prices of major commodities and reveal the strong sensitivity of maize prices to minor variations of maize production in Northern America.

3.1 Introduction

Over the past decade, the four components of food security - availability, stability, utilisation, and access - have become significant sources of concern. At the turn of 2010, prices of main food crops in the international markets have shown high variability, sometimes doubling in a short time frame (Headey and Fan, 2010). For example, the price of maize increased by 75% from September 2007 to May 2008 (Headey, 2011). Poor harvests and rising prices of agricultural commodities had contributed to triggering the hunger riots of 2007-2008 and the Arab Spring of 2011 (Headey and Martin, 2016). High levels of volatility in the food prices are now recognised to affect food security for a growing number of households (Rosenzweig et al., 2001; Schmidhuber and Tubiello, 2007).

Several reasons can further explain the food crises at the turn of the decade: low levels of food stocks, rising prices of inputs - particularly fertilisers - and growing demand for bio-fuel (Headey and Fan, 2008). One of the most frequently cited is idiosyncratic shocks on agricultural production at the regional level. In 2007 and 2010, for example, extreme local environmental conditions (e.g., droughts in Russia and extensive wildfires in Australia) and resultant declines in regional production significantly contributed to the spike in global food prices (Tadasse et al., 2016). For example, the heatwave in Russia in the summer of 2007 and 2010 led to a significant drop in local wheat production, which resulted in export restrictions and subsequent tensions on international markets (Wegren, 2011). Restrictions on rice exports in India and Vietnam in 2007/2008 also led to substantial price increases on international markets (Headey, 2011).

Increased inter-connectivity in global food markets can be a source of resilience, as seen in the recent Covid-19 outbreak, but also of vulnerability, particularly when the agricultural production of a major exporter is affected. Least developed countries are particularly vulnerable as they may suffer more significant import losses through their strong dependence on imports for staple foods (Puma et al., 2015). In this case, we speak of teleconnected supply shocks (Bren d'Amour et al., 2016). Bren d'Amour et al. (2016) find that the Middle East is most sensitive to teleconnected supply shocks in wheat, Central America to supply shocks in maize, and Western Africa to supply shocks in rice. In the future, climate change and the increasing frequency of extreme weather events could make the food system even more vulnerable to such teleconnected shocks. Several works study the transmission of prices and price volatility from international

to domestic markets (Baquedano and Liefert, 2014; Kalkuhl, 2016). However, to our knowledge, no article has so far attempted to quantify the inverse link, namely the sensitivity of the world price to supply shocks at the regional level.

The international maize market is a highly relevant case study because maize is one of the most traded crops and plays an essential role in food security in many countries. Accurate identification of the most influential maize producing regions would potentially be helpful for decision-makers who need to optimise both their dates of commodity purchases and their stock usages (World-Bank, 2005). Although maize is the most widely traded crop globally, only a few countries export their maize productions, suggesting that the production of a small number of regions might impact maize prices. As some countries rely heavily on maize imports to ensure food security (Wu and Guclu, 2013; Rouf Shah et al., 2016), it is essential to be able to anticipate price shocks for this commodity. Models that provide relatively short-term maize price projections are relevant to many stakeholders. For example, the WASDE forecasts are helpful for risk calculation and for designing the federal US crop insurance program (US-HR, 2009). However, these models were criticised because of their complexity (Hoffman et al., 2018) and, sometimes, because of their lack of accuracy (Hoffman, 2011; Warr, 1990; Hoffman et al., 2015; Lusk, 2016). Other forecasting models serve private institutions, particularly companies specialising in commodity trading. Auto-regressive methods are widely known to forecast food prices in the academic literature (Li et al., 2010; Shively, 1996). Although all these tools are undoubtedly helpful for forecasting maize prices, they provide little insight into the effects of regional maize production variations on global maize prices.

Although it is difficult to predict precisely the extent to which global scale price variations could affect local prices, it has been previously shown that shifts in international prices can transmit into regional domestic prices (Headey and Fan, 2010). In more recent research, Kalkuhl (2016) suggests that there is a strong relationship between international prices and domestic ones, even when the global market trades with futures.

The objective of our study is (i) to identify the maize-producing regions having the most significant influence on the global price of maize through their production and (ii) to quantify the effects of regional production changes on global price changes. Under the assumption that regional production shifts primarily drive shifts in maize prices (Hertel et al., 2016), we train several statistical and machine learning models using publicly available regional yearly production data and monthly price data. Monthly price data are pertinent because maize prices do not tend to change on a daily or weekly basis but rather monthly (Ochieng and Baulch, 2019; Dorosh et al., 2004). Furthermore, our input variables, i.e. regional maize productions or yields, directly inform on the level of commodity

supply, which is usually an unstable component of the market. Therefore, the trained models are used to analyse the relationships between regional maize production (or yield) and global prices, to identify the most and least influential producing regions in the global maize market, and finally to quantify the effect of regional production (or yield) changes on global price changes.

In our study, we chose to use various statistical and machine learning methods. The use of different methods has several advantages. First, it allows us to study the robustness of the main conclusions to the data analysis method implemented. Second, it makes it possible to compare different methods' precision and determine the most efficient ones. Third, our comparison of models thus contributes to improving our understanding of maize price determinants and developing operational and accessible predictive tools. In this way, our study is relevant for designing food security policies.

3.2 Materials and method

3.2.1 Data

Historical annual yield (hectograms per hectare) and production (tonnes) data were obtained from the FAO data website (FAOSTAT) for all years available (1961 to 2018) for 19 regional entities (defined by FAO) covering 242 countries. For further data definitions and the sources of the variables included in our models, see tables 3.2 and 3.3 in Appendix 3.A.

We extracted data on maize global monthly prices from the World Bank's commodity markets database as the U.S. No. 2 yellow free on board (FOB) Gulf of Mexico, U.S. nominal price, per metric tonne units. Although this price is the traditional representative price for the maize produced in the U.S., this quotation is today's leading benchmark price for the international maize trade (FAO, 2021).[2]

The time series summarises the monthly price of maize, as globally traded in FOB U.S. Gulf ports, from January 1960 to December 2019. We converted these prices into real 2010 U.S. Dollars, using the monthly agricultural index of the World-Bank[3] (Fig 3.1).

[2]We found a strong correlation between the series of relative yearly maize price changes used in this paper and the relative maize price changes of other countries. For example, Argentina Ukraine correlation is about 0.75, according to the data made available in the GIEWS database of the FAO.

[3]Although the most frequently use price index is the American CPI, we chose to use the World-Bank monthly agricultural price index. We base our decision on two factors: The first derives from Tadasse et al. (2016) indicating that the U.S. CPI could be a biased deflator when dealing in a global market that includes both developed and developing countries. The second reason is a relatively lower gap (RMSE) between the maize annual real prices published by the World Bank

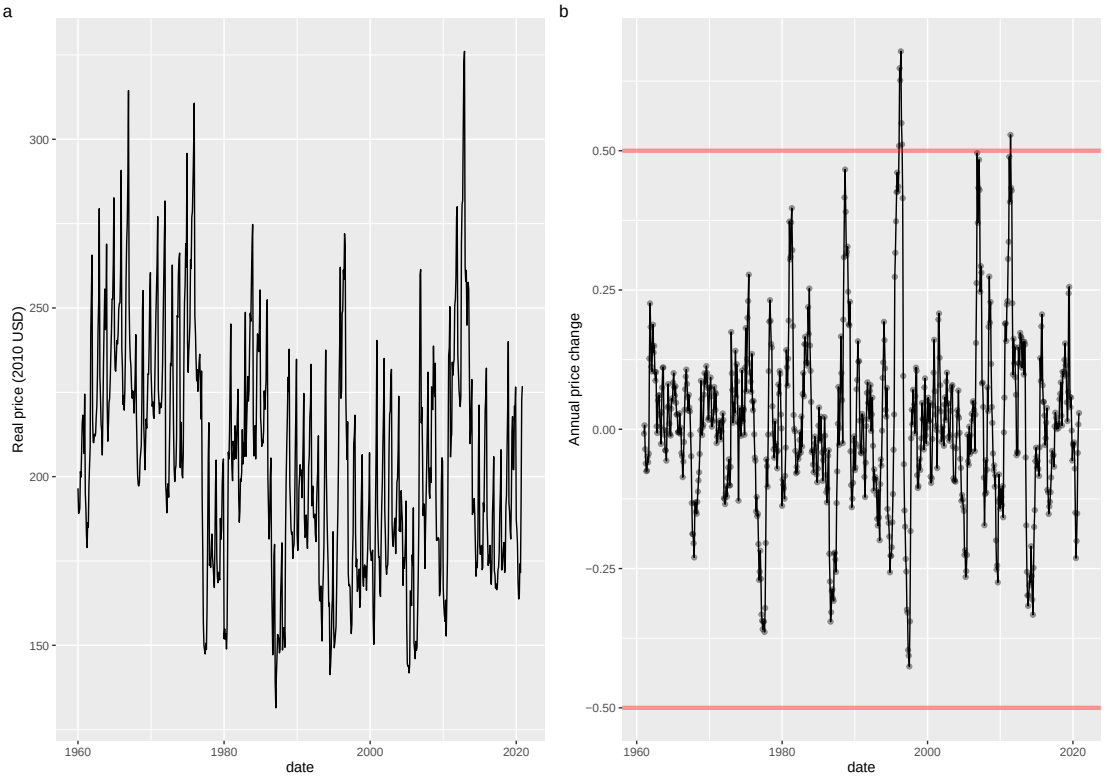


Figure 3.1 – Time series of global maize price. (a) Real terms in 2010 US Dollars. (b) Real terms in relative change from the same month of the previous year (ratio)

The deflated prices are further denoted as $q_{m,y}$, where m and y are the months and year indices, respectively. Exportable maize is usually harvested once a year, during the main harvest season, and levels of maize production can thus potentially have substantial effects on yearly price changes. For this reason, the dependent variable in our analysis is defined as the relative price difference of maize expressed relatively to the same month of the previous year. It is defined as

$$p_{m,y} = \frac{q_{m,y} - q_{m,y-1}}{q_{m,y-1}} \tag{3.1}$$

and their values are shown in fig 3.1.b. From the series of $p_{m,y}$, we define a binary variable $p^b_{m,y}$ equal to one in case of price increase ($p_{m,y} > 0$) and to zero otherwise.

Maize prices for month m in year y are estimated as a function of relative production (or yield) changes between the month m in year y and the same

to the deflated maize global monthly price calculated for this study.

month in year $y-1$. To accomplish this, we transformed regional yield (grain weight per unit of the cropping area, in hectograms per hectare) and production (total regional grain weight, in tons) data into relative changes compared to the previous year, as follows:

$$x_{k,y} = \frac{z_{k,y} - z_{k,y-1}}{z_{k,y-1}} \tag{3.2}$$

Where $z_{k,y}$ is the production (or yield) in a region k (k=1, ..., 19) and year y, and $x_{k,y}$ is the relative production (or yield) change in the same region and the same year.

We predict prices during the last quarter of each year, that is, in October, November, and December ($m \in$10,11,12), i.e., when all regions have finished (or almost finished) their maize harvest and reported the yearly production and yield obtained. For a given year, it is indeed possible to obtain accurate estimates of maize yield and production from October onward and to use them to predict price shocks of the same year.[4]

In the next sections, we present and compare several methods to estimate $p_{m,y}$ and $p^b_{m,y}$ at $m \in$10,11,12 as a function of $x_{k,y}$, $k \in$1,...,19. Each method is implemented twice; first using relative changes in regional productions as input variables and then using relative yield changes.

3.2.2 Linear and generalised linear models

Although the relationships between price and production or yield changes may be non-linear, we use a linear regression model as a benchmark to estimate price fluctuation as a function of changes in regional productions or yields. Our linear model (LM) is defined as follows:

$$p_{m,y} = \alpha + \sum_{k=1}^{19} \beta_k x_{k,y} + \epsilon_{m,y} \tag{3.3}$$

where α and β_k are regression parameters and $\epsilon_{m,y}$ is the residuals. Additionally, we define a variant of this model including the price change of year $y-1$ (i.e., $p_{m,y-1}$) as a supplementary input. This serves for investigating Granger causal relation between $p_{m,y}$ and $x_{k,y}$ (Granger, 1969). The significance of the effects of $x_{k,y}$ are tested with and without using $p_{m,y-1}$ as an additional input in the regression model. If some of the $x_{k,y}$ are still significant while taking $p_{m,y-1}$ into account, one can be considered that there is a Granger causal relation between $p_{m,y}$ and these $x_{k,y}$.

[4] http://www.amis-outlook.org/amis-about/calendars/maizecal/en/, retrieved 23 March 2020

For classification, we use a generalised linear model (GLM) with a binomial family and a logit link. This model computes the probability that $p^b_{m,y}$=1 (i.e., price increase), given the values of the regional production (or yield) changes $x_{k,y}$, $k \in$1,...,19.

Both models are implemented with the GLM function of R (R-Core-Team, 2020). As done with the other methods, we fit linear models for each month (October, November, December) using production and yield changes as inputs. The most influential inputs were selected using a stepwise procedure implemented with the AIC (step function of R).

3.2.3 CART

The three ML methods considered in this study are decision-tree based algorithms: Classification and regression trees (CART), random forests (RF), and gradient boosting machine (GBM). None of these methods makes any strong assumption about the functional form of the relationship between the dependent variable and the explanatory variables, neither about the data distribution. They can thus capture nonlinear relationships between the inputs (regional production or yield changes) and the output (global price change). We shortly present our implementation of CART here, while RF and GBM are presented in the next sections.

The purpose of CART is to build a binary decision tree. Let $p_{m,y}$ be a dependent variable and $x_{1,y}, x_{k,y}, ..., x_{K,y}$ a series of explanatory variables. The tree is constructed by repeatedly distributing the observations into homogeneous groups relative to $p_{m,y}$. The partitioning criteria is monotonous in the explanatory variable, x_k, which defines a cross-section of x_k. In contrast, higher valued observations belong to the right and lower-valued to the left branches. Additional partitions based on the same variable can be made, but one cut-off point is determined at each stage. The subgroups that define the tree are called nodes. CART performs recursive partitioning, and searches for splits that minimise the test error rate in the chosen objective function. The choice of the objective function depends on whether the output is continuous ($p_{m,y}$) or categorical ($p^b_{m,y}$). In the former case, i.e. for predicting $p_{m,y}$, CART is implemented using the residual sum of squares (RSS). To predict $p^b_{m,y}$ (classification), the objective function is a purity index based on the Gini index. Here, CART was implemented with the package `rpart` of the R software (Therneau et al., 2019) (`rpart` function).

3.2.4 Random Forest and Gradient Boosting

Although simple to visualise and interpret, CART results are usually unstable and tend to be sensitive to small data changes. Their price predictions are not always accurate (Kuhn and Johnson, 2013). For these reasons, ensemble learning algorithms based on bagging (for "bootstrap aggregating") and boosting methods are frequently used instead of CART trees (Breiman, 2000). In this study, we use random forests (RF) (Liaw et al., 2002) as a bagging-based algorithm and gradient boosting machine (GBM) as a boosting-based method.

The RF algorithm builds an ensemble of trees, each relying on a small subset of inputs (i.e., a subset of all regional productions or yields). Each tree is fitted to a randomly chosen training set generated using a bootstrap procedure. This approach reduces the effects of correlations between variables while allowing different input variables to be selected. In RF, predictions are derived by computing the average of all trees. Here, we find that 500 trees lead to stable results. RF can rank the inputs according to their predictive powers and, here, the resulting ranking can be used to identify the regions whose maize productions (or yields) show the strongest influence on global maize price. In this study, RF is implemented with the `randomForest` function of the package `randomForest` (Breiman et al., 2018), both for quantitative predictions and for classification.

The method GBM is also based on an ensemble of trees (Efron and Hastie, 2016). At each iteration, GBM builds a simple tree (weak-learner), each of which is learning from the prediction errors of all the trees built so far. The final prediction is the sum of all the models calculated earlier. As RF, GBM is able to rank the inputs according to their predictive powers. In our case, we fit GBM using the `gbm` function of the `gbm` package (Friedman, 2001) both for regression and classification based predictions. Here, we find that the most accurate results are obtained with 100 trees for GBM.

Neither RF nor GBM has analytical expressions. However, standard methods can be used to rank their inputs according to their importance and visualise their effects on the output on price changes. Using these methods, we rank the model inputs $x_{k,y}$ from the most influential to the least by computing the mean decrease accuracy criterion (Calle and Urrea, 2010) for each input (i.e. each regional production or yield changes). This criterion measures the extent to which the accuracy of model predictions or classifications decreases when each input variable is set to a random value. Lastly, we use partial dependence plots (Greenwell, 2017) to visualise the response of the model outputs to the most influential inputs, averaging the overall values of the other inputs. These plots allow us to analyse the shapes of the responses and detect non-linearity. The same approaches were applied to LM and CART to compare the input rankings and the dependence plots of all methods on the same basis.

3.2.5 Models Evaluation

The accuracy of the quantitative price-estimation is assessed by root mean squared error (RMSE), which we estimate using a leave-one-out cross-validation (LOOCV). In each step, one year of price ($p_{m,y}$, m=10,11,12) and production/yield ($x_{k,y}$) is extracted from the original data set. Then, the four models (CART, RF, GBM, and GLM) are trained using the remaining 55 years, to estimate the removed value of $p_{m,y}$ using the trained models. For each year, the procedure is performed to obtain a set of 56 estimations for each tested model and each month (m=10,11,12). Finally, we calculate a value RMSE for each model and each predicted month. We repeat the whole procedure twice, using regional maize production and regional maize yields as inputs, successively.

To evaluate the accuracy of the classification models, we apply the same LOOCV procedure, this time to calculate the area under the ROC curve (AUC). This criterion is commonly used to evaluate the performance of classification algorithms (Hernández-Orallo et al., 2012). An AUC higher than 0.5 indicates better performance than random classification. An AUC equal to 1 reveals a perfect classification.

3.3 Results

3.3.1 Quantitative effects of regional productions on price changes

Table 3.1 – Comparison of RMSE values for the four types of models (lm: linear model; cart: regression tree; rf: random forest; gbm: gradient boosting model). RMSE values (expressed in the same unit as a relative price change, i.e. in relative change ratio compared to the same month the previous year), were computed by cross-validation for predicting yearly price changes in October, November, and December using two types of inputs: relative regional production (left) or yield (right) changes. The lowest values obtained for each month are in red

	Production				Yield			
	LM	CART	RF	GBM	LM	CART	RF	GBM
October	0.169	0.140	0.137	0.135	0.132	0.136	0.122	0.128
November	0.153	0.148	0.140	0.135	0.163	0.147	0.139	0.158
December	0.144	0.148	0.130	0.129	0.139	0.129	0.129	0.147

Table 3.1 shows that the best methods are either RF or GBM, depending on the considered month. For example, the most accurate predictions of global price changes in October ($p_{10,y}$) are obtained by RF with an RMSE equal to 0.12. The least accurate results (i.e., the highest RMSE) are obtained either with the linear model (LM) or with CART, depending on the month considered.

The importance ranking of the regional maize yields is shown in Fig 3.2 for the three months considered and the four different statistical and machine learning methods. The ranking obtained when using regional production changes as inputs is shown in the supplementary materials (Fig 3.15). We determine the contribution to the prediction accuracy (RMSE) of the price as the relative importance of each region in a given month. We consider a region to be influential if a random choice of its corresponding input value (i.e., a yield change or production change chosen at random) leads to a substantial increase of the RMSE of the price change predictions. On the other hand, a region would be non-influential if a random choice of its corresponding input value does not affect the RMSE. Results clearly show that Northern America is by far the most influential region according to the four methods, with both types of inputs (production or yield changes), and for the three months considered. The only exception is the linear model (with yield change inputs) in November, but this model has low predictive power compared to others in November (Table 3.1). Considering the most accurate methods (GBM and RF), yield and production changes in Northern America have the most substantial influence on global price changes. Moreover, according to the linear models, the effects of yield and production change in Northern America on global price change are statistically significant ($p<0.01$) in October, November, and December, with and without the price change in year y-1 included as an additional explanatory input. This result indicates a Granger causality of yield and production changes in Northern America on global maize price. Furthermore, it reveals that yield and production changes are helpful in forecasting price changes, even when previous price changes were taken into account.

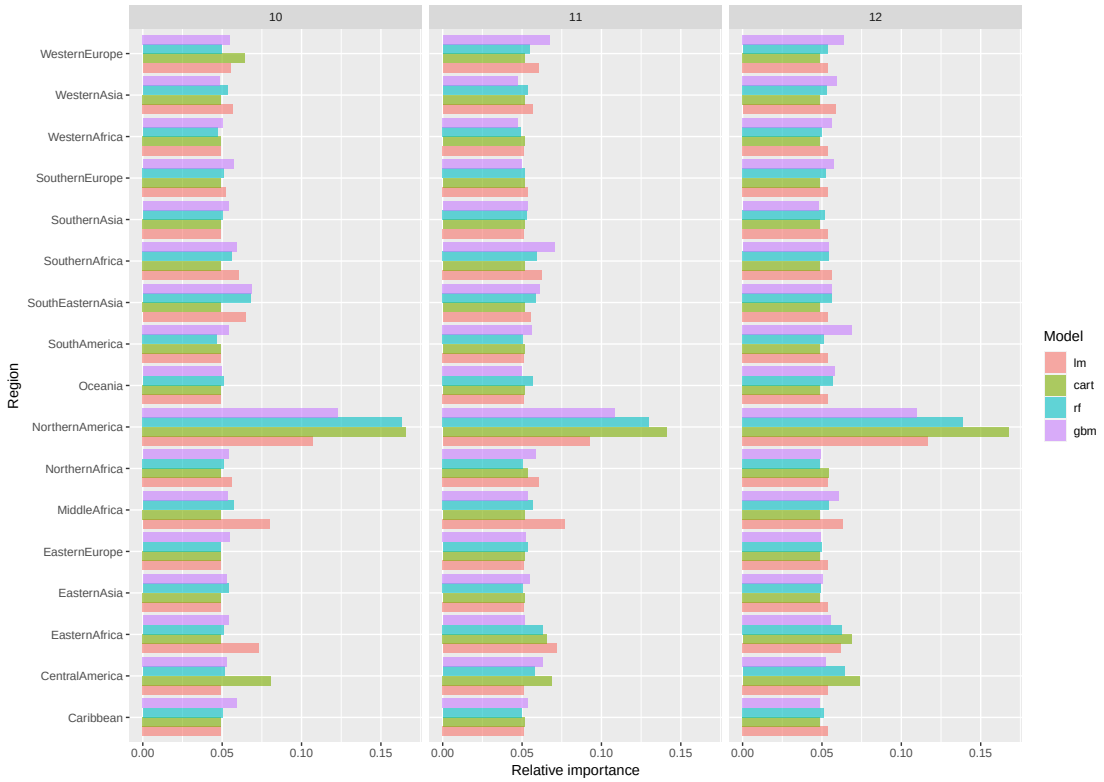


Figure 3.2 – Importance levels of regional yield changes for predicting the global maize price in October (10), November (11) and December (12). Importance levels are computed using the RMSE criterion and measure the extent to which the model accuracy decreases with a random permutation of each input.

The partial dependence plot (PDP) shown in Fig 3.3 presents the average response of price changes in October (10), November (11), and December (12) to variations of maize yield compared to the previous year in the most influential region, i.e., Northern America (similar PDPs are shown in the supplementary Fig 3.17 using production instead of yield). The PDPs obtained using the four models consistently show that an increase (decrease) of yield in Northern America leads to a decrease (increase) of global price. In October, for example, an 8% rise of relative maize yield in Northern America leads to a reduction of maize price of 7% according to the GBM model, while a 0.1% decrease of relative maize yield in Northern America is expected to increase the global price by 7% according to the same model. This result confirms the strong influence of Northern American yield on global maize prices. The PDPs obtained using the production and yield changes in other regions show much weaker trends and much flatter curves (see, for example, the PDPs obtained for the region Southern Africa, in supplementary Fig 3.16, Fig 3.18).

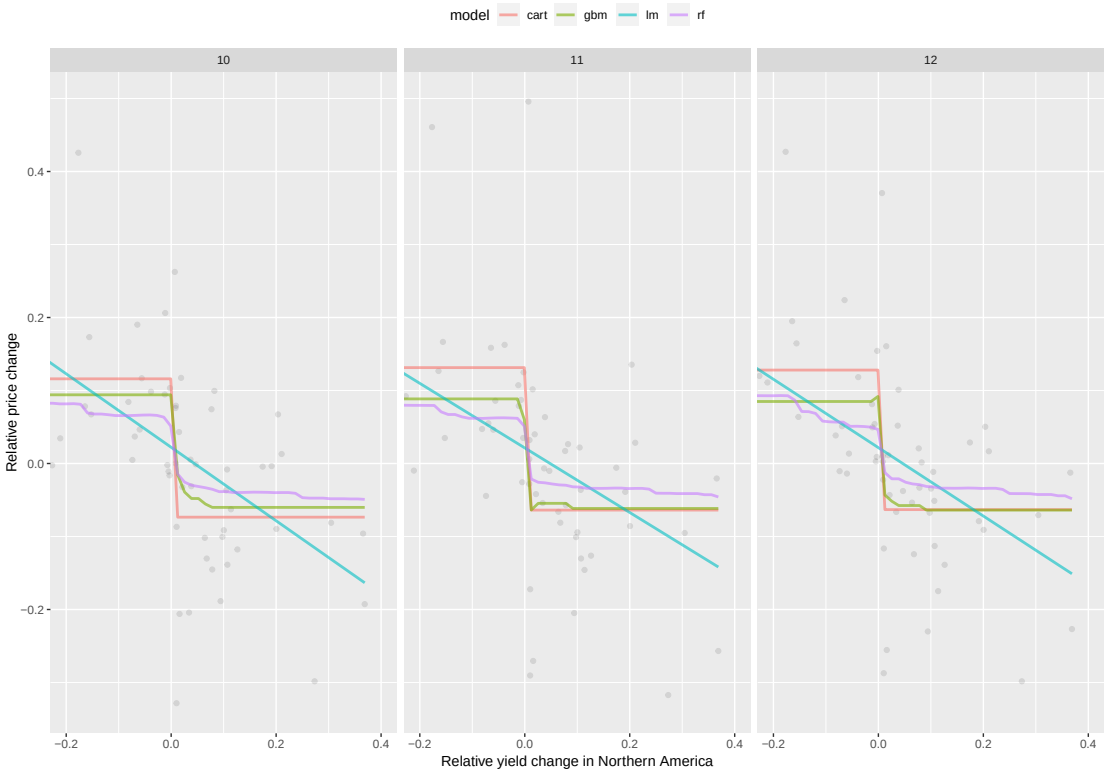


Figure 3.3 – Partial dependence plots obtained with LM, CART, rf and GBM showing the average response of relative price change in October (10), November (11) and December (12) to relative yield change in Northern America. The points indicate price variations as observed over the period of 1961-2019. The plot shows that, according to all models, any increase (decrease) of yield in Northern America compared to the previous year leads to a decrease (increase) of global price.

3.3.2 Classification of price increase vs. decrease

Fig 3.4 shows the results that ROC analyses for the classification models for the three months considered. The results are in favour of GBM and RF with AUC falling in the range of 0.7-0.8 for these methods in most cases. The 95%CI are relatively large, but those obtained with RF and GBM never include the benchmark value 0.5, characterising a random classification. On the contrary, the 95%CI of CART and the linear model sometimes include 0.5, revealing that these methods do not systematically perform better than a random classification. For a given month and a given type of input, the lowest AUC is obtained by the linear model or CART. The two types of inputs did not lead to any systematic difference in AUC values.

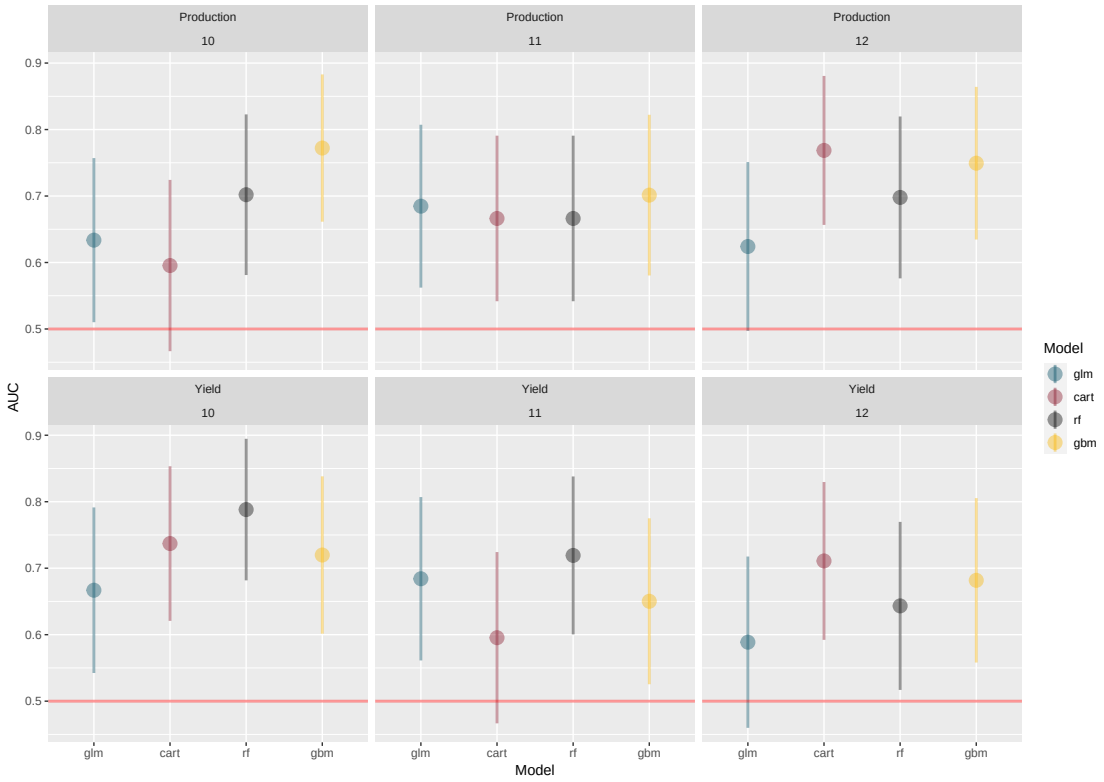


Figure 3.4 – AUC values obtained for the classification models predicting price increase vs. price decrease in October (10), November (11) and December (12). The horizontal red line indicates AUC=0.5, i.e random classification. Vertical bars indicate the 95% confidence intervals (CI). When these bars do not include 0.5, the AUC is significantly higher than 0.5 ($p<0.05$)

As already noticed in the case of regression, the importance ranking of the regional production and yield inputs of the classification models reveals that Northern America is the most influential region, in particular for the model GBM which has a good classification power. For more details, see figure 3.19 and figure 3.20 in the Appendix 3.A.

Fig 3.5 shows the PDPs of the classification models. These PDPs represent the average responses of the probability of price increase to relative yield changes in Northern America (PDPs obtained with regional production inputs are shown in Supplementary F). The probability of a global price increase strongly decreases below 0.5 as soon as the yield change is positive in Northern America compared to the previous year. In contrast, it increases above 0.5 when the yield change is negative. The effect is powerful with the model GBM. As already noticed for quantitative price changes, the PDPs obtained with the classification models show much weaker trends and much flatter curves for regions

other than Northern America (see, for example, the PDPs obtained for the region Southern Africa, in supplementary Fig 3.21 and Fig 3.23).

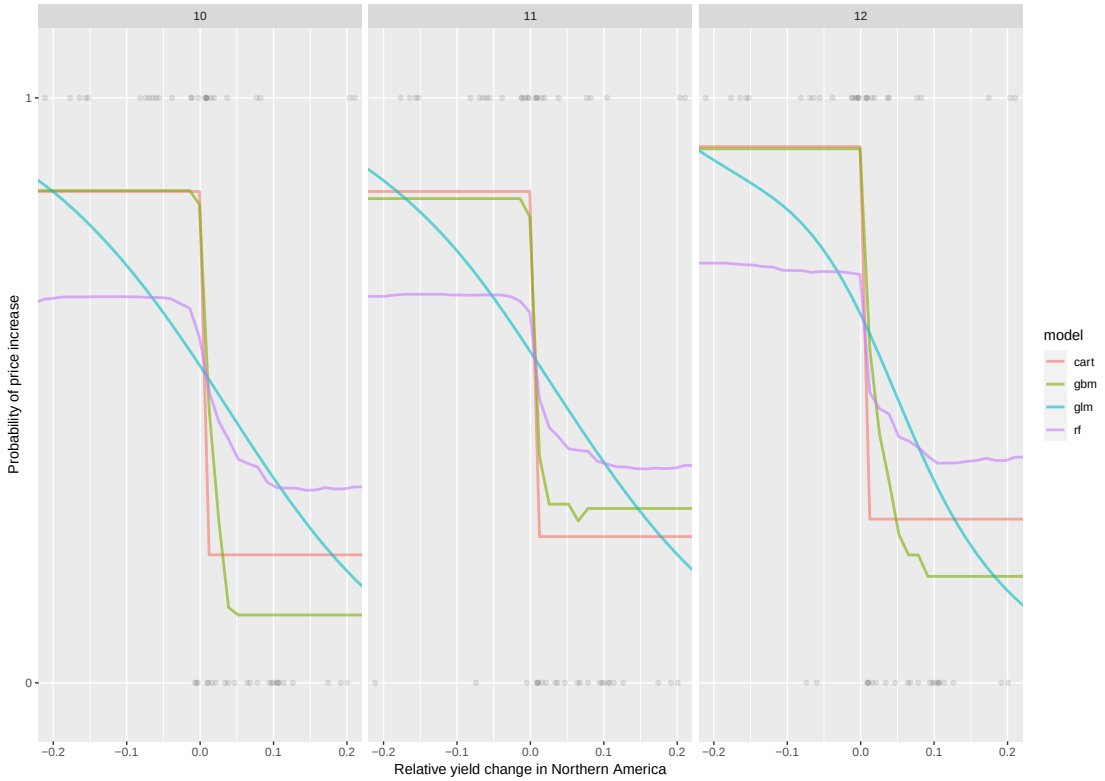


Figure 3.5 – Partial dependence plots showing the probability of price increase in October, November and December as a function of relative yield change in Northern America, for the four models considered. The points indicate price variations (on the y-axis, 1=price increase, 0=price decrease) as observed over the period of 1961-2019.

3.4 Discussion

Using regional maize production data and global maize prices, we assessed the effects of regional production and yield variations on late-season global maize prices. Because of the existing relationship between the global price and domestic prices, especially in the least developed countries (Caracciolo et al., 2014), the topic is essential to dealing with food security issues in vulnerable regions.

Our study is the first to address this question using various statistical and machine learning methods. Overall, all models consistently show that Northern

America is the most influential region, and both maize yields and maize productions seem to be equally influential. However, this result is somewhat trivial as Northern America (and, more specifically, the USA) is the leading maize producer and exporter at a global scale and as the USA is known to have a strong influence on the agricultural trade market (Chatzopoulos et al. (2019)). However, our models can provide data-driven quantitative information on the effect of regional production variations on global maize prices. Our analysis provides real added value because it allows us to quantify the effect of an increase or decrease in the annual production of maize in this region on the global price of this commodity. All methods reveal that a slight increase (decrease) of maize production or yield in Northern America would lead to a decrease (increase) of the global maize price by a few per cent compared to the previous year. When considering the most accurate methods, an increase of maize yield relative to the previous year of +8% in Northern America negatively affects the global maize price by about -7%, while a decrease of yield in Northern America as low as -0.1% will cause the global maize price to increase by more than 7%. The strong impact of maize production in Northern America is confirmed by the results obtained with the classification methods. Indeed, these methods indicate that the slight increase (decrease) in maize yield or production in Northern America has a strong negative (positive) effect on the probability of maize price increase compared to the previous year. Even a minimal decrease in maize production in Northern America can inflate the probability of a price increase.

Among all the considered modelling techniques, ensemble tree-based techniques (random forest and gradient boosting) show the lowest RMSE and highest AUC values, revealing that these methods were the best for both quantitative price prediction and classification. Indeed, in addition to predicting price changes quantitatively, the methods tested in this paper can be used to classify relative price increase vs decrease situations. The principle is to compute the probability of price change increase (or decrease) as a function of regional production (or yield) changes. The tree-based models tend to outperform the simpler GLM. Still, the rate of misclassification is approximately 25% with GBM and RF, which is relatively high but better than a random classification. As noticed for quantitative predictions, the production change in Northern America is, by far, the most meaningful input for classifying price increase vs price decrease situations. All these results concur in showing that maize production change in Northern America is a highly relevant indicator for assessing the risk of global maize price increase or decrease.

The nature of the inputs (i.e., production vs yield changes) has a marginal impact on the methods' performance. Thus, surprisingly, both GBM and RF do not perform better when regional production variations are used as inputs instead

of yield. However, production data combine two types of information, i.e., yields and cropping areas, whether yield variations alone do not account for possible variance in the regional maize cultivated areas.

Although the main purpose of our study is not to propose new forecasting tools, our models could potentially be used to predict global maize prices. Compared to other types of forecasting models, GBM and RF have several advantages but also a few disadvantages. Our models rely on public data and can be easily implemented using standard modelling open-source software. On the contrary, private forecasting techniques are usually unpublished, not freely available, and not transparent. Structural models constitute another category of models that can predict the prices of agricultural commodities. These models rely on theories describing economic systems and are developed by international organisations such as FAO, OECD, and IFPRI. They simulate price fluctuations using a series of functions describing partial or general market equilibrium. Although these models are used to predict product prices in the long run, they are not usually implemented to make short-term predictions. They are also complex and cannot be easily run by non-specialists. The WASDE model is another example of an operational tool for maize price predictions. Similarly to our models, WASDE can forecast maize price at a monthly time step. According to Hoffman et al. (2015), WASDE relies on a combination of nine different structural and non-structural sub-models while GBM and RF can be easily implemented using free R packages and publicly accessible data. They could be thus easily run by any interested stakeholder and updated every year based on the most recent data.

Our models could serve to predict price changes for other agricultural commodities from regional crop productions in the future. From a practical point of view, a disadvantage of the ML tree-based models is that they rely on yearly regional production input data. In principle, these data are only available after harvest, but relatively accurate values can be estimated shortly before harvest from local expert knowledge and model predictions. However, considering the maize growing season, it is not realistic to get reliable regional production data before the end of summer, especially regarding regions located in the Northern hemisphere, particularly Northern America, which is a key region for predicting global maize price. For this reason, all models were used here to predict global maize prices at the end of the year, more specifically in October, November, and December.

In this study, we analysed the effect of regional productions on global maize prices during the last three months of the year. We made this choice to be consistent with the harvest date for maize in the central maize-producing region - North America - which takes place in the very late summer and fall. Although we

did not carry out a detailed analysis for earlier months, we did perform a sensitivity analysis of the influence of North America depending on the month considered. As a result, we found that this region retained a significant but lesser influence in the months preceding the harvest, probably due to the influence of the harvest forecasts anticipated by the maize market players. In the future, however, it would be beneficial to deepen this analysis to identify more precisely the influence of the different producers on prices during the first months of the year.

Our approach could potentially be replicated for other crops whose production is less geographically concentrated. Such flexibility would allow us to assess the world food price sensitivity to production shocks or an export ban in a given country.

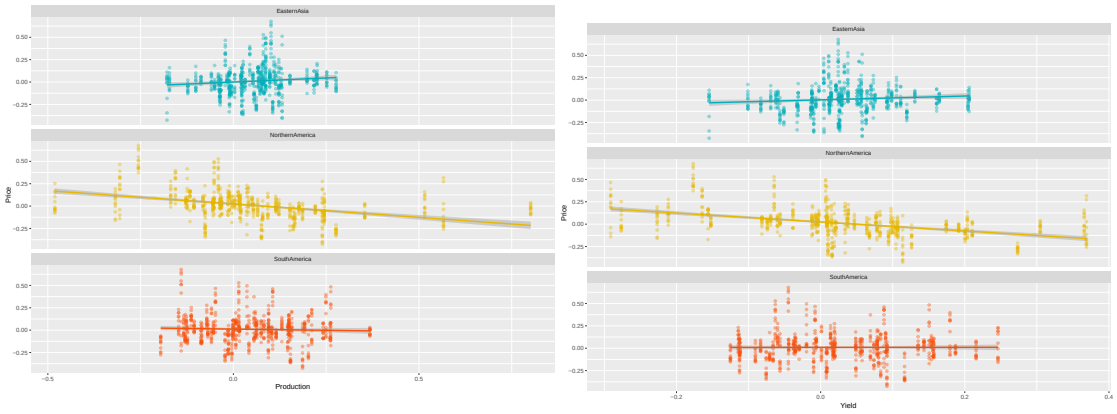
3.5 Conclusions

This study demonstrates that it is possible to assess the impact of regional maize production variations on the global price of maize using machine learning techniques on publicly available regional production and price data. As these methods can be easily implemented using only freely available packages and public information, our results contribute to forecasting the global price of maize more accessible. As such, our price prediction technique can be included food security management programs and policies and possibly serve as a price forecaster. Furthermore, the methods considered can rank regional producers according to their influence on global maize prices. Our results show that Northern America is the most influential out of all regions. More specifically, our results reveal that, for maize, small positive production changes relative to the previous year in Northern America have a strong and negative impact on global maize prices. Our study highlights the potential interest of ML for predicting global prices of major commodities from regional production and assessing price sensitivity to regional crop producers.

3.A Appendix

figures

Price data vs. Production and yield data



(a) Inputs=relative regional production changes

(b) Inputs=relative regional yield changes

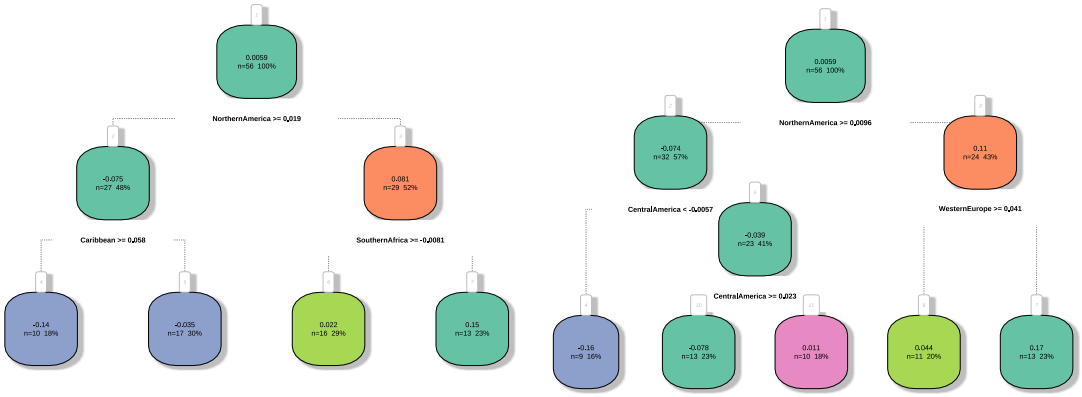
Figure 3.6 – Relative change in the global maize price versus relative regional production (a) and yield (b) changes, for three biggest maize producing regions

Regression based analysis

Fig 3.7[5] shows the CART tree fitted to predict $p_{m,y}$ in October as a function of the regional production (or yield) changes. This tree has four (fig 3.7a) or five (fig 3.7b) final nodes, defined by three or four inputs corresponding to different regions. The tree root (the upper rectangle in the diagram centre) includes 56 observations (i.e., the whole dataset) with an average $p_{10,y}$ of 0.59%. Referring to Fig 3.7a, after the algorithm examines all possible partitions according to the set of input variables, the optimisation function of CART finds that the maximum reduction of RSS. This result is achieved by splitting the 56 price data into two groups, defined by the maize production in Northern America, at a cut-off point of 1.9%. All regions with production change more significant than 1.9% are included in the right branch (no.2). On the contrary, when production change in Northern America is lower than 1.9%, the right branch of the tree (no.3) is used. The second partition is done based on the Caribbean (if x_{NA} $\geq$1.9%) or Southern Africa (if x_{NA} $<$1.9%). The final nodes at the bottom of the diagram include

[5]app 3.7 - fig 3.12 were implemented with the package `Rattle` of the R software (Williams, 2011) (`fancyRpartPlot` function)

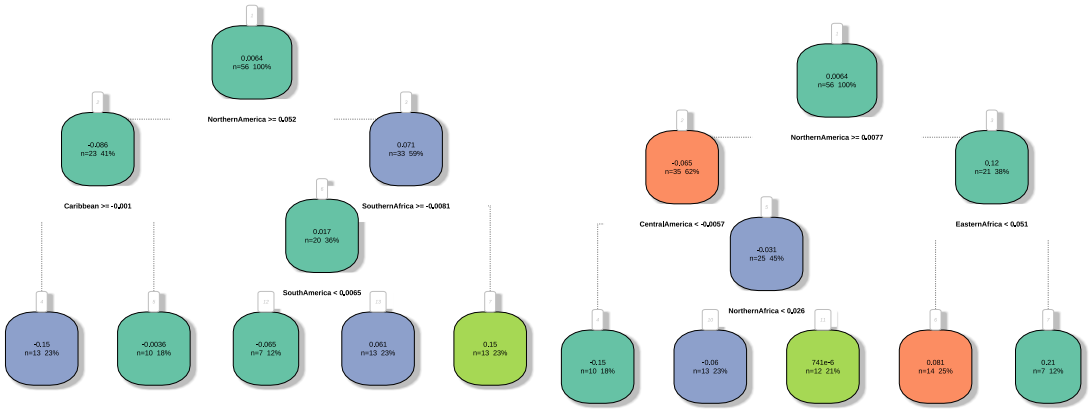
the average observed price change corresponding to five different production (or yield) situations. These results correspond to the average price changes reported in the final nodes. Here, the fitted tree produces four different price estimations determined by the values of three inputs.



(a) Inputs=relative regional production changes

(b) Inputs=relative regional yield changes

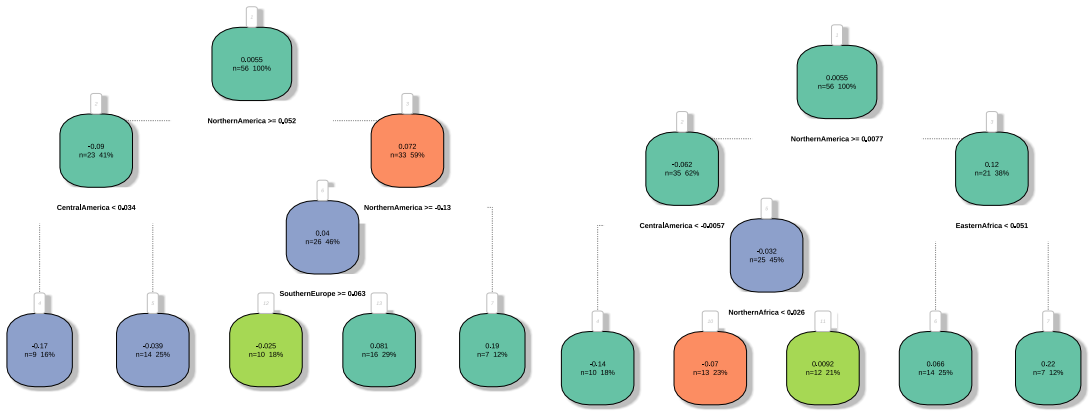
Figure 3.7 – CART models computing $p_{10,y}$ of maize (i.e., relative price change in October) as a function of relative regional production changes (a) and relative regional yield changes (b). All nodes of each tree include three numbers; the average relative price change value over all data falling in the considered node, the number of data in each node (n), the % of data in each node. The terminal nodes (at the bottom) report the relative price changes predicted by the CART models



(a) Inputs = relative regional production changes

(b) Inputs = relative regional yield changes

Figure 3.8 – CART models computing $p_{11,y}$ of maize (i.e., relative price change in November) as a function of relative regional production changes (a) and relative regional yield changes (b)



(a) Inputs=relative regional production changes

(b) Inputs=relative regional yield changes

Figure 3.9 – CART models computing $p_{12,y}$ of maize (i.e., relative price change in December) as a function of relative regional production changes (a) and relative regional yield changes (b)

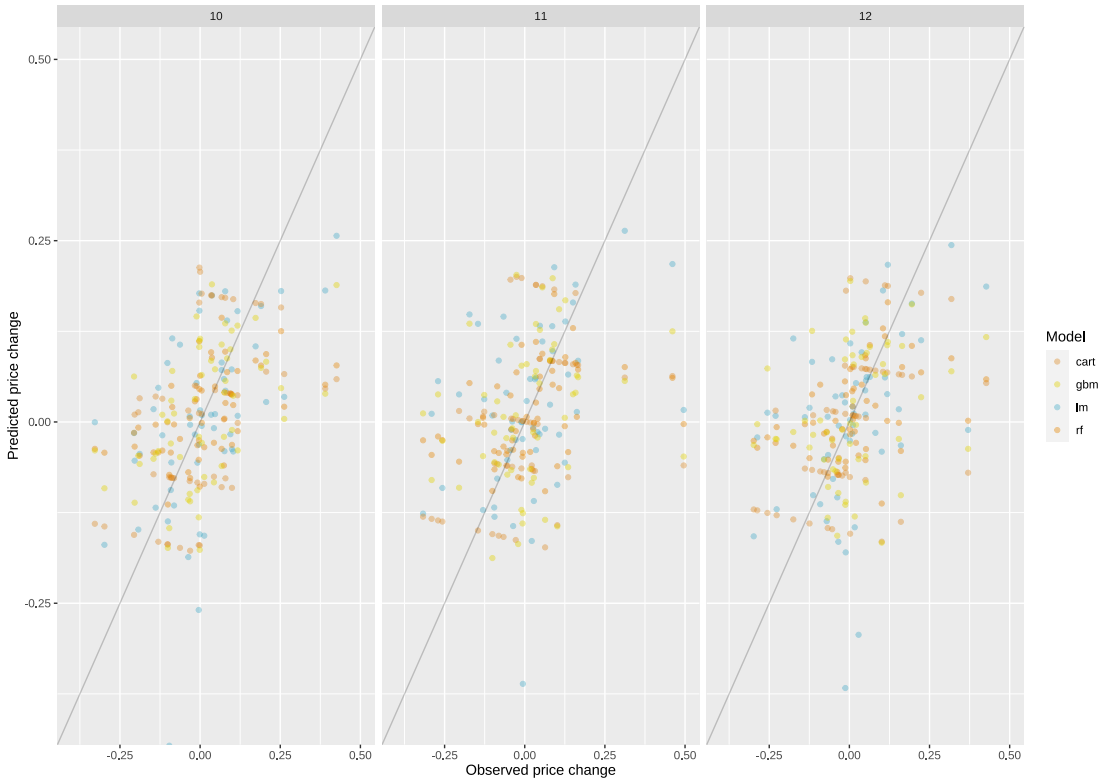


Figure 3.10 – Observed relative price change vs. Predicted relative price change, October (10) November (11) and December (12), with yield changes used as inputs.

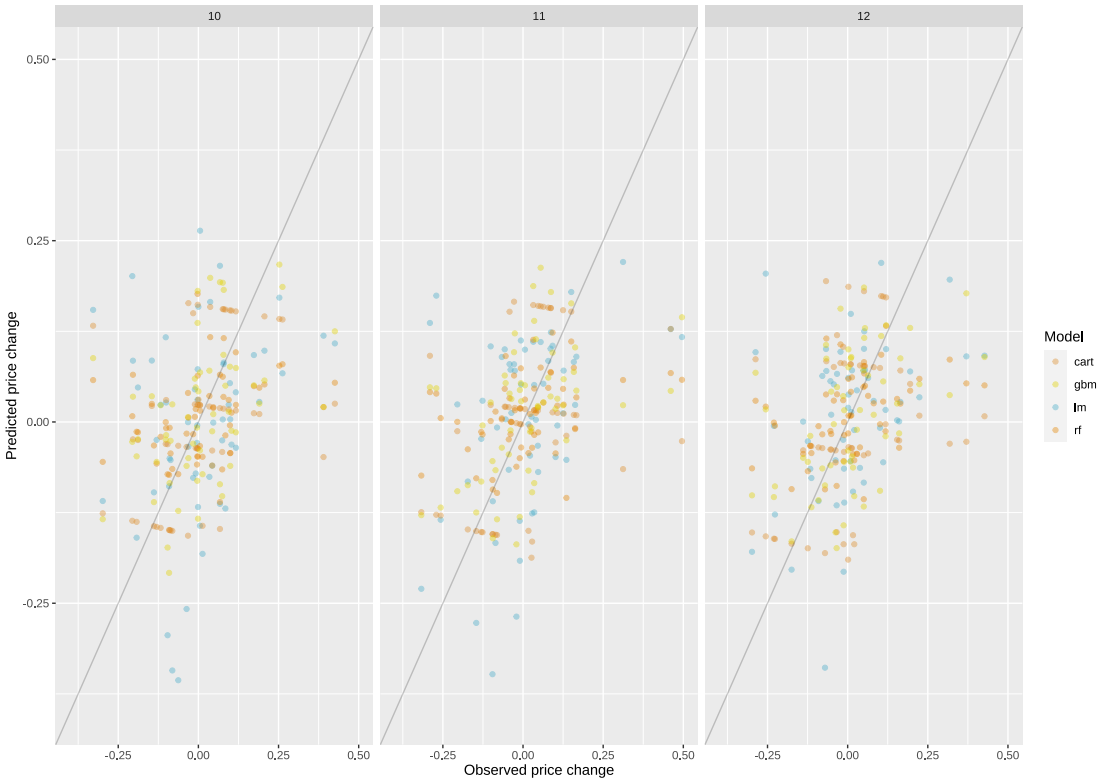


Figure 3.11 – Observed relative price change vs. Predicted relative price change, October (10) November (11) and December (12), with production changes used as inputs.

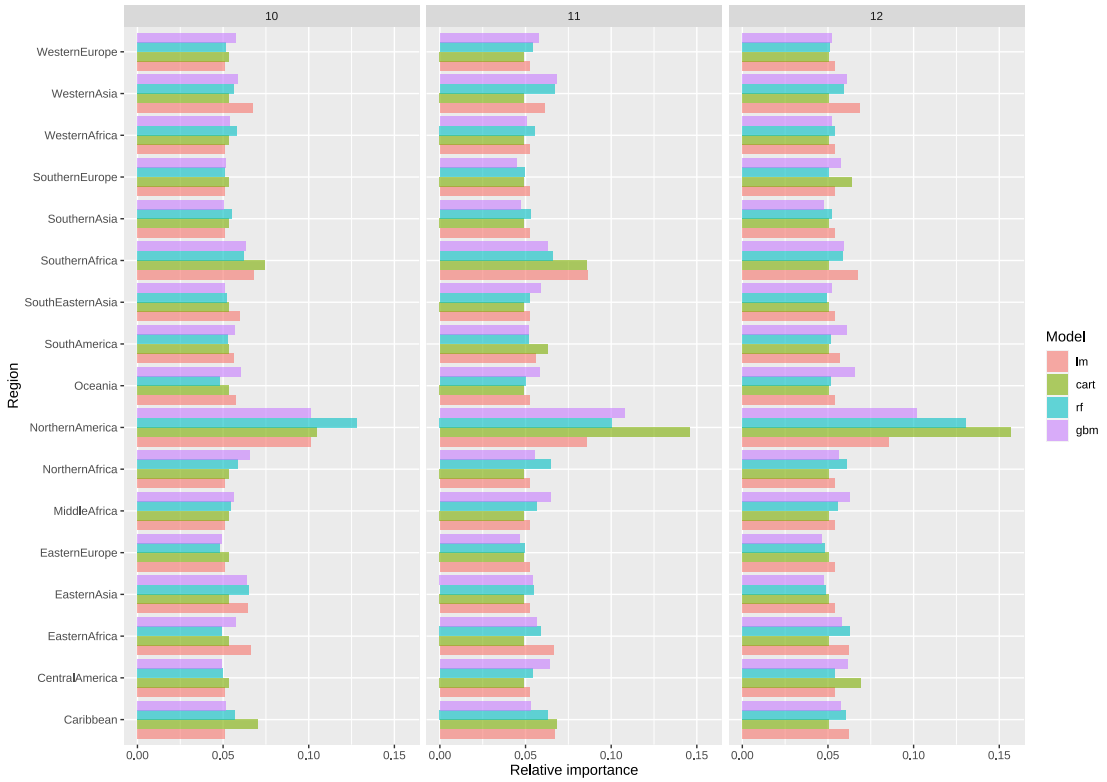


Figure 3.12 – Importance ranking of changes in production on the global maize October (10), November (11) and December (12) price. Importance levels are computed using the RMSE criterion and the permutation technique.

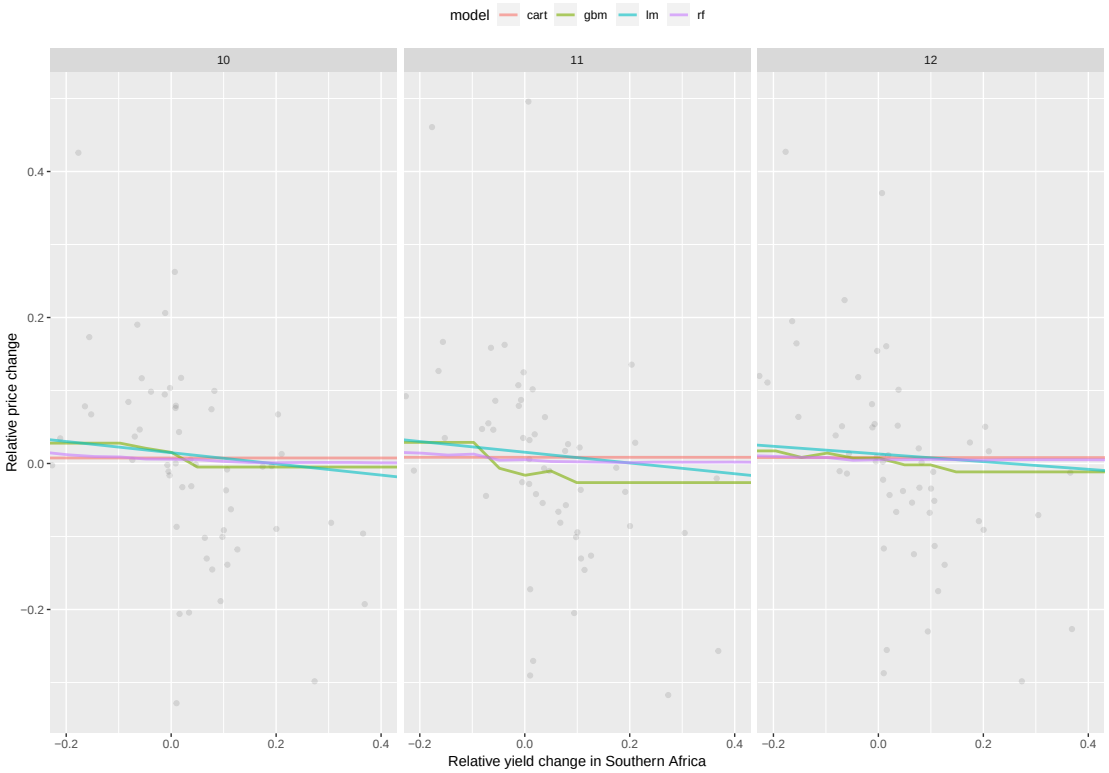


Figure 3.13 – Partial dependence plots obtained with LM, CART, RF and GBM showing the average response of relative price change in October (10), November (11) and December (12) to relative yield change in Southern Africa.

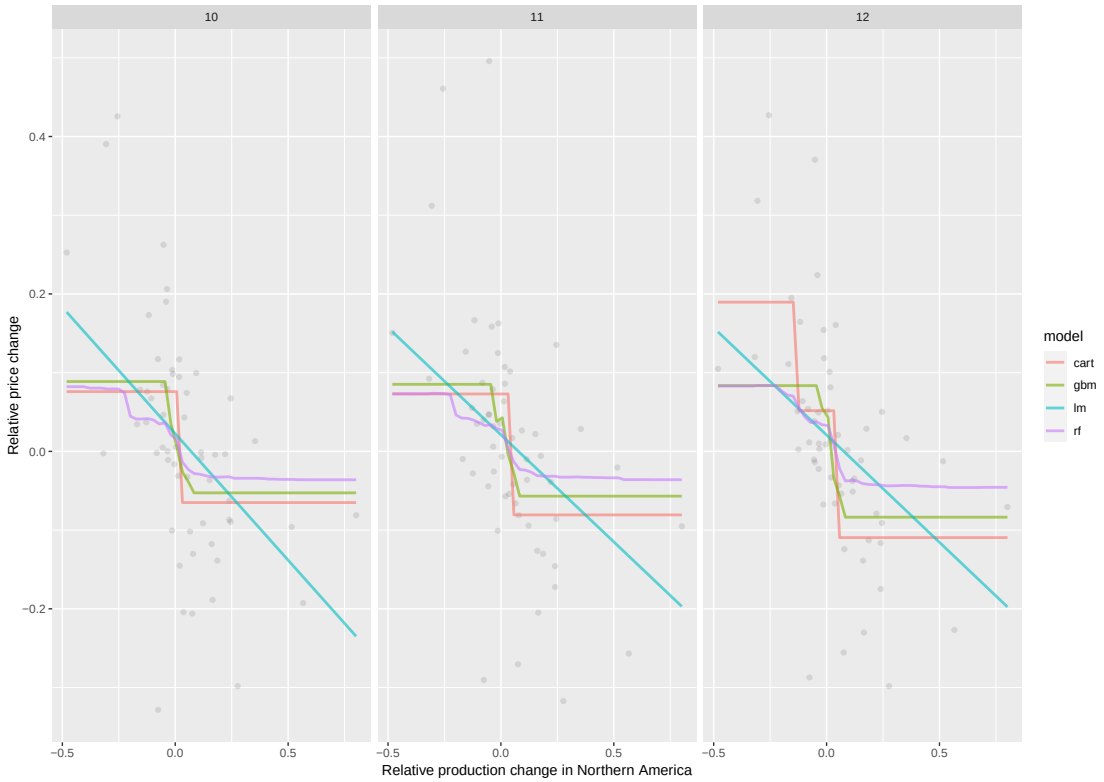


Figure 3.14 – Partial dependence plots obtained with LM, CART, RF and GBM showing the average response of relative price change in October (10), November (11) and December (12) to relative production change in Northern America.

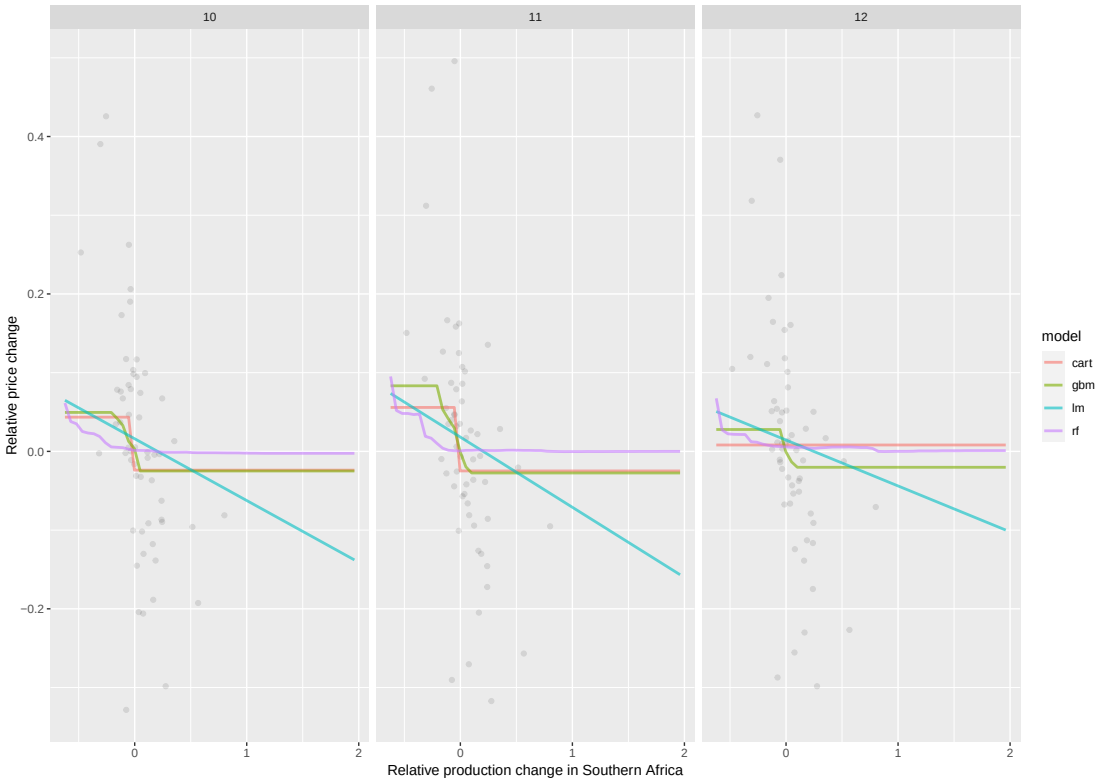
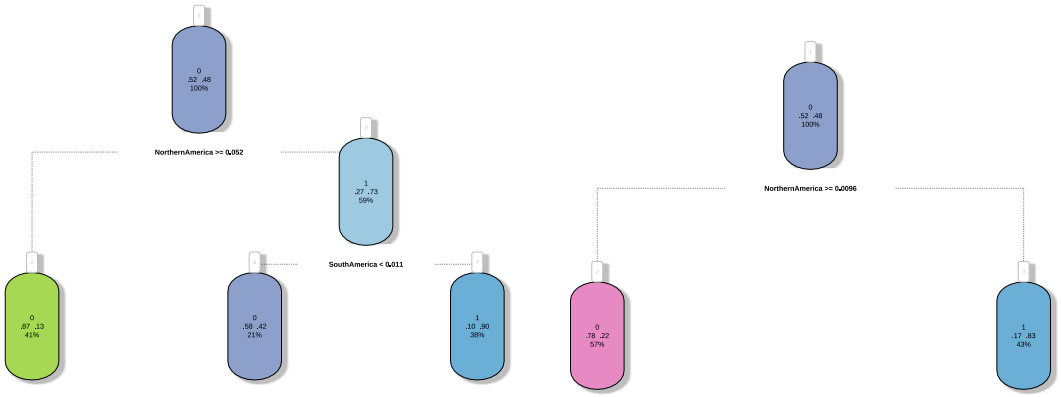


Figure 3.15 – Partial dependence plots obtained with LM, CART, RF and GBM showing the average response of relative price change in October (10), November (11) and December (12) to relative production change in Southern Africa.

Classification based analysis

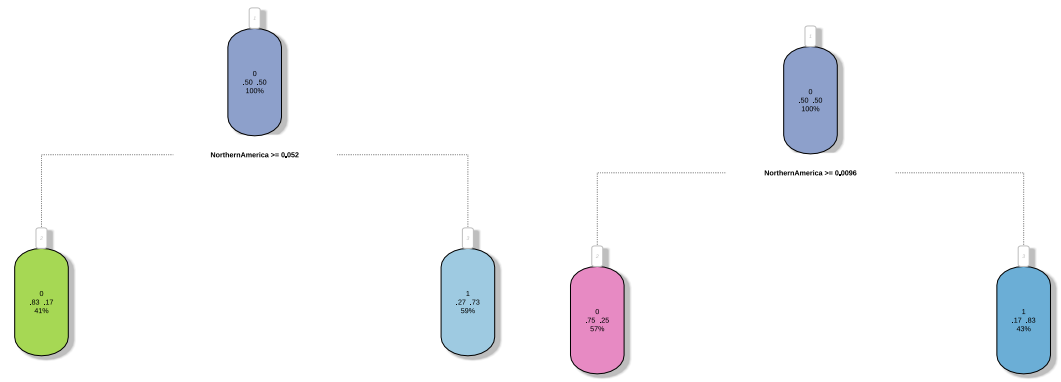
Fig 3.16 shows the tree obtained for classifying October price into two categories: price increase or decrease. Here, also, the most influential input is Northern America. According to the fitted tree, the highest chance of price decline in October occurs when the annual North-American production increases by more than 5.2%.



(a) Inputs=relative regional production changes

(b) Inputs=relative regional yield changes

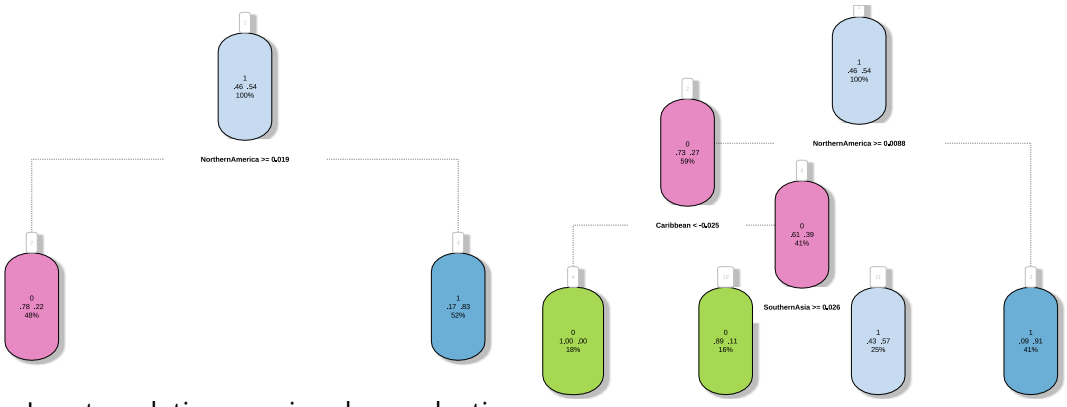
Figure 3.16 – CART models computing the probability of relative maize price increase in October as a function of relative regional production changes (a) and relative regional yield changes (b). Each node of each tree includes three numbers; the proportion of data showing a price increase among the data falling in the considered node, the number of data in each node (n), the % of data in each node. The terminal nodes (at the bottom) reports the probabilities of price increase computed by the CART models



(a) Inputs=relative regional production changes

(b) Inputs=relative regional yield changes

Figure 3.17 – CART models computing the probability of relative maize price increase in November as a function of relative regional production changes (a) and relative regional yield changes (b).



(a) Inputs=relative regional production changes

(b) Inputs=relative regional yield changes

Figure 3.18 – CART models computing the probability of relative maize price increase in December as a function of relative regional production changes (a) and relative regional yield changes (b).

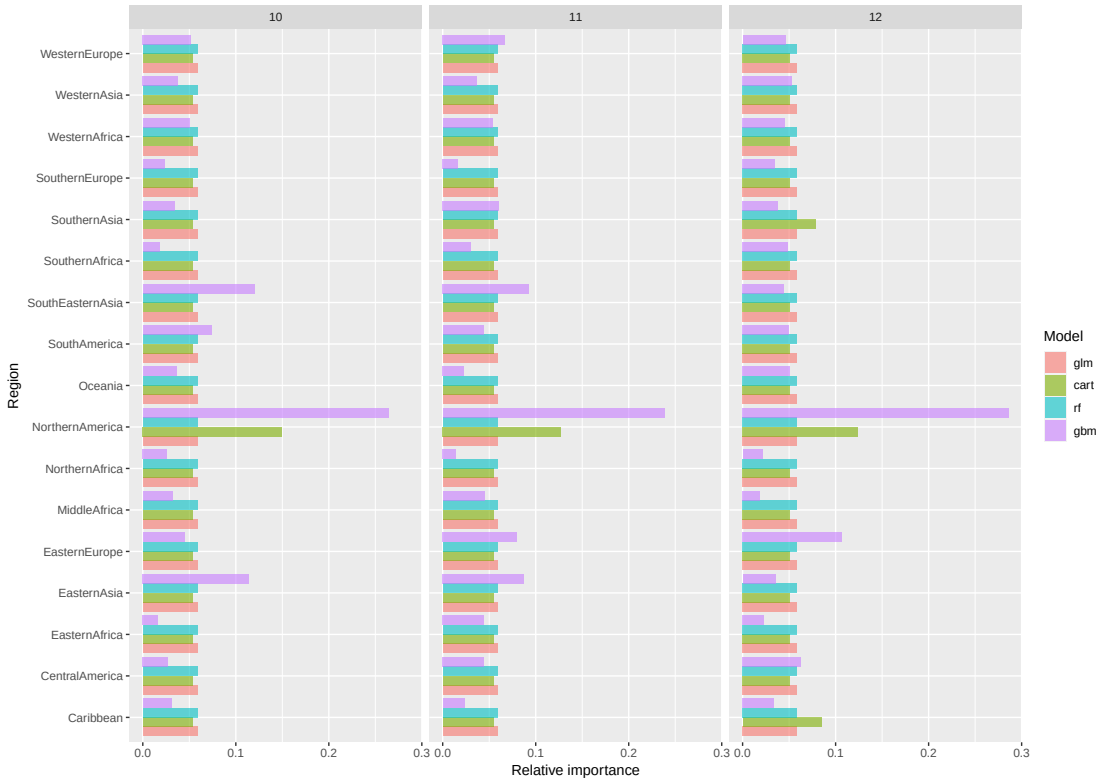


Figure 3.19 – Importance ranking of changes in yield on the global maize October (10), November (11) and December (12) price. Importance levels are computed using the cross-entropy loss (CE) with the permutation technique.

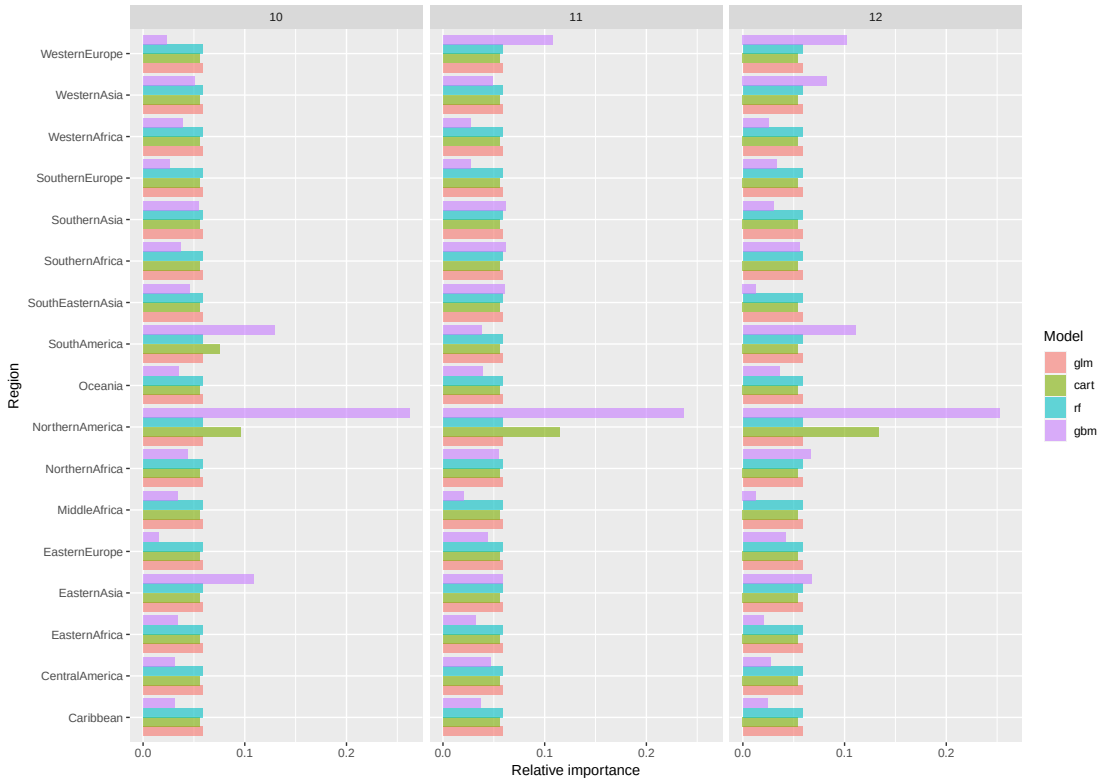


Figure 3.20 – Importance ranking of changes in production on the global maize October (10), November (11) and December (12) price. Importance levels are computed using the cross-entropy loss (CE) indicator with the permutation technique.

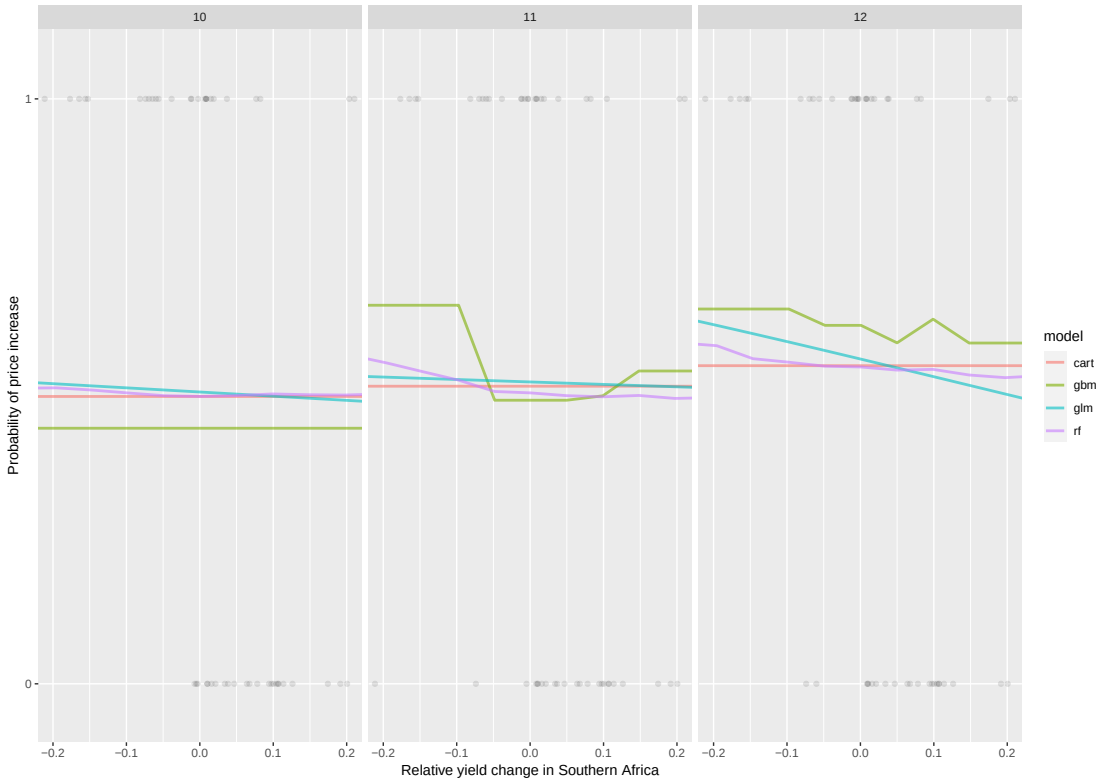


Figure 3.21 – Partial dependence plots showing the probability of price increase in October, November and December as a function of relative yield change in Southern Africa.

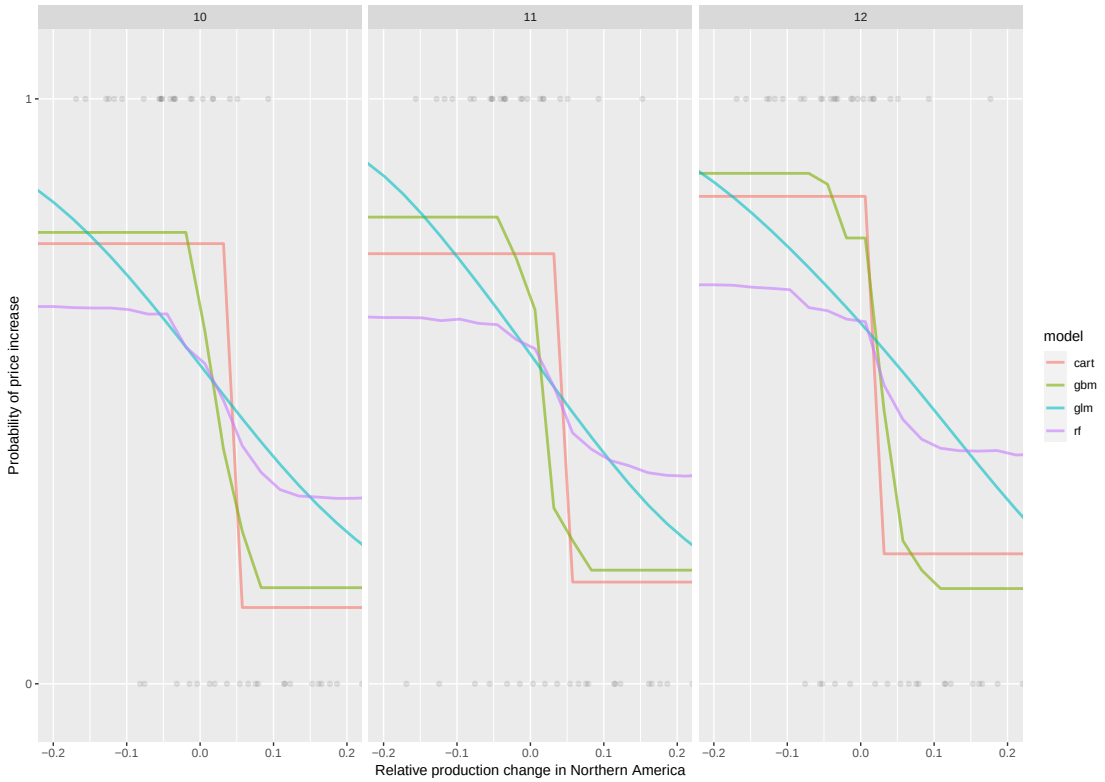


Figure 3.22 – Partial dependence plots showing the probability of price increase in October, November and December as a function of relative production change in Northern America.

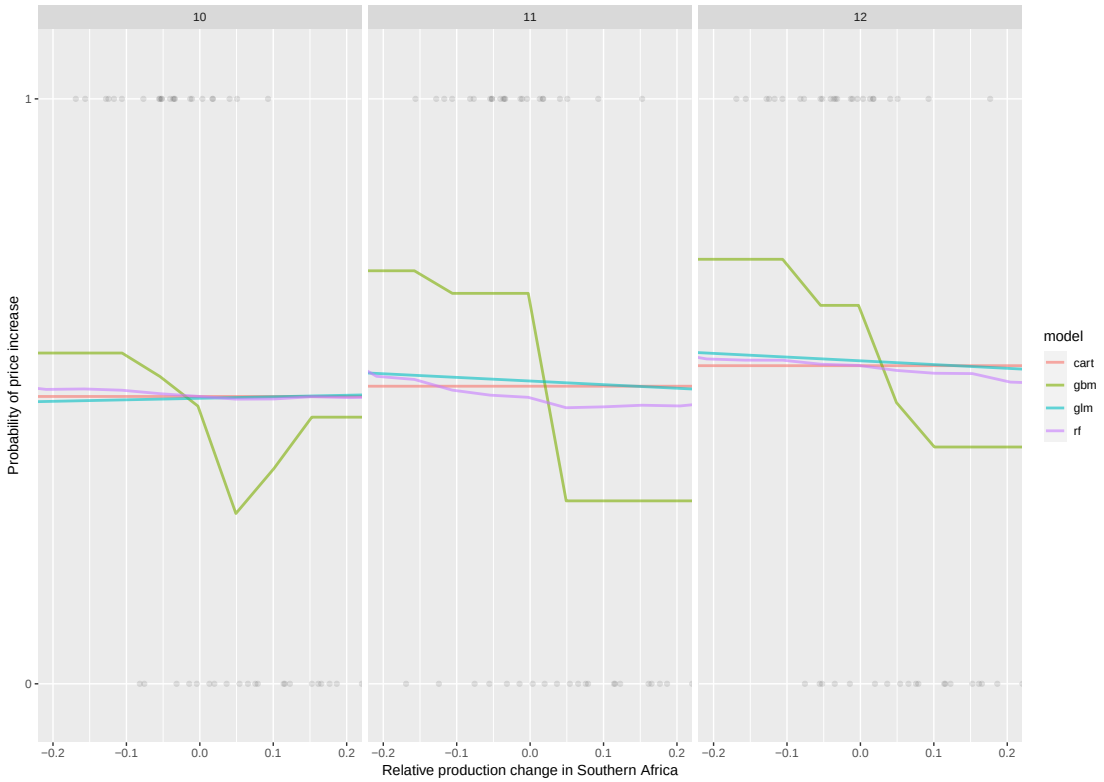


Figure 3.23 – Partial dependence plots obtained with LM, CART, RF and GBM showing the average response of relative price change in October (10), November (11) and December (12) to relative production change in Southern Africa.

Tables

Regression based analysis

Tables 3.4 and 3.5 show the summary statistics of multivariate linear regression models predicting relative price changes as a function of relative regional production changes (table 3.4) and relative regional yield changes (table 3.4). In the first row of each region are the estimated coefficients, β_k, namely, relative change in $p_{10,y}$, $p_{11,y}$ and $p_{12,y}$ induced by a one percent increase in regional production, $x_{k,y}$, where all other variables are fixed. The values in brackets show the levels of significance (p-value) of all estimated coefficients. The region with the strongest (and significant) impact is Northern-America.

Tables 3.6 and 3.7 show a summary statistics of the classification linear models, GLM, which compute the probability of relative maize price increase in October, November and December as a function of relative regional production

changes (tables 3.6) and relative regional yield changes (tables 3.7). The tables show the change in the logit of the probability of global maize price increase induced by each regional input, and the significance of the estimated coefficients (between brackets).

Table 3.2 – Summary statistics for regional data over 1961-2018.

Production data (FAOSTAT)		
	(1000 tonnes)	
Regions Caribbean	%/total 0.08%	Average 460
Central America	3.27%	17,825
Central Asia*	0.11%	1,251
Eastern Africa	2.64%	14,395
Eastern Asia	19.03%	103,592
Eastern Europe	6.23%	33,925
Middle Africa	0.47%	2,563
Northern Africa	0.91%	4,974
Northern America	40.59%	220,940
Northern Europe*	0.00%	38
Oceania	0.08%	424
South America	9.75%	53,066
South-eastern Asia	3.52%	19,186
Southern Africa	1.68%	9,170
Southern Asia	2.95%	16,037
Southern Europe	3.83%	20,845
Western Africa	1.47%	7,975
Western Asia	0.56%	3,069
Western Europe	2.80%	15,264

Table 3.3 – Summary statistics for regional data over 1961-2018.

Production data (FAOSTAT)			
	Yield (1000 hg/ha)		
Regions	Average	Min. (year)	Max. (year)
Caribbean	11.13	8.54 (1993)	14.78 (2004)
Central America	20.13	9.74 (1961)	34.41 (2018)
Central Asia*	47.74	25.63 (1997)	65.72 (2018)
Eastern Africa	13.54	9.52 (1965)	19.44 (2014)
Eastern Asia	39.01	12.28 (1961)	60.89 (2017)
Eastern Europe	36.91	18.4 (1963)	70.02 (2018)
Middle Africa	8.5	6.74 (1979)	10.9 (2015)
Northern Africa	41.2	15.91 (1961)	69.12 (2012)
Northern America	73.34	39.23 (1961)	118.01 (2017)
Northern Europe*	34.36	10 (1985)	71.74 (2016)
Oceania	50.47	17.33 (1966)	87.81 (2015)
South America	27.72	12.95 (1964)	59.26 (2017)
South-eastern Asia	21.83	9.02 (1961)	46.14 (2018)
Southern Africa	23.89	7.88 (1992)	58.13 (2017)
Southern Asia	17.28	10.02 (1971)	34.29 (2017)
Southern Europe	54.31	21.13 (1961)	90.58 (2018)
Western Africa	12.02	6.96 (1972)	19.54 (2018)
Western Asia	34.87	11.4 (1962)	79.18 (2018)
Western Europe	69.51	22.56 (1962)	103.05 (2011)

	October	November	December
(Intercept)	-0.008	0.025	0.023
	(0.704)	(0.180)	(0.211)
EasternAfrica	0.241	0.369	0.290
	(0.091)	(0.013)	(0.040)
EasternAsia	0.334		
	(0.057)		
NorthernAmerica	-0.372	-0.269	-0.293
	(0.000)	(0.001)	(0.000)
Oceania	0.168		
	(0.075)		
SouthAmerica	0.244	0.209	0.232
	(0.062)	(0.119)	(0.071)
SouthEasternAsia	0.182		
	(0.147)		
SouthernAfrica	-0.075	-0.116	-0.074
	(0.049)	(0.004)	(0.050)
WesternAsia	-0.366	-0.298	-0.342
	(0.022)	(0.051)	(0.020)
Caribbean		-0.422	-0.295
		(0.011)	(0.061)
Num.Obs.	56	56	56
R2	0.495	0.483	0.466
R2 Adj.	0.409	0.420	0.400
AIC	-76.0	-74.0	-79.1
BIC	-55.8	-57.8	-62.9
Log.Lik.	48.013	45.004	47.549

Table 3.4 – Linear regression, Inputs=relative regional production changes

	October	November	December
(Intercept)	0.016	0.017	0.031
	(0.409)	(0.461)	(0.067)
EasternAfrica	0.319	0.466	0.307
	(0.017)	(0.003)	(0.029)
MiddleAfrica	0.876	0.821	0.540
	(0.001)	(0.006)	(0.036)
NorthernAfrica	-0.350	-0.394	
	(0.075)	(0.080)	
NorthernAmerica	-0.539	-0.481	-0.577
	(0.000)	(0.000)	(0.000)
SouthEasternAsia	0.738	0.595	
	(0.027)	(0.115)	
SouthernAfrica	-0.069	-0.077	-0.056
	(0.077)	(0.087)	(0.163)
SouthernEurope	-0.181	-0.185	
	(0.153)	(0.202)	
WesternAsia	-0.356	-0.326	-0.364
	(0.053)	(0.120)	(0.058)
WesternEurope	0.136	0.176	
	(0.095)	(0.061)	
Num.Obs.	56	56	56
R2	0.606	0.525	0.485
R2 Adj.	0.529	0.432	0.433
AIC	-88.0	-72.7	-83.2
BIC	-65.7	-50.4	-69.0
Log.Lik.	54.999	47.354	48.580

Table 3.5 – Linear regression, Inputs=relative regional yield changes

	October	November	December
(Intercept)	-1.241	-0.849	-0.734
	(0.069)	(0.245)	(0.359)
EasternAsia	11.273		13.760
	(0.029)		(0.082)
MiddleAfrica	11.782	24.357	13.839
	(0.113)	(0.044)	(0.142)
NorthernAmerica	-12.679	-14.458	-14.367
	(0.002)	(0.007)	(0.006)
Oceania	7.747	11.244	11.549
	(0.027)	(0.012)	(0.026)
SouthAmerica	13.491	10.101	16.375
	(0.012)	(0.049)	(0.027)
SouthEasternAsia	7.799	22.156	10.735
	(0.077)	(0.022)	(0.096)
SouthernAsia	-8.310	-9.282	-10.242
	(0.071)	(0.075)	(0.064)
WesternAsia	-10.081	-11.730	-24.163
	(0.046)	(0.059)	(0.029)
Caribbean		-17.261	
		(0.034)	
EasternEurope		18.163	15.166
		(0.007)	(0.038)
NorthernAfrica		-17.639	
		(0.036)	
SouthernEurope		-18.862	-24.117
		(0.034)	(0.041)
WesternEurope			4.198
			(0.173)
Num.Obs.	56	56	56
AIC	59.5	56.5	57.2
BIC	77.7	80.8	81.5
Log.Lik.	-20.727	-16.236	-16.597

Table 3.6 – Summary statistics of the classification linear models, GLM, Inputs=relative regional production changes

	October	November	December
(Intercept)	0.242	-0.943	1.727
	(0.691)	(0.259)	(0.010)
Caribbean	9.602	12.766	12.620
	(0.132)	(0.113)	(0.122)
MiddleAfrica	21.082	47.144	26.697
	(0.012)	(0.014)	(0.023)
NorthernAfrica	-14.884	-45.965	-13.985
	(0.022)	(0.014)	(0.041)
NorthernAmerica	-12.637	-21.116	-18.034
	(0.001)	(0.007)	(0.003)
SouthEasternAsia	20.966	90.067	
	(0.155)	(0.014)	
SouthernAsia	-8.086	-20.623	-14.116
	(0.090)	(0.019)	(0.025)
WesternAsia	-9.653		-23.056
	(0.118)		(0.012)
EasternAfrica		12.587	7.477
		(0.032)	(0.138)
EasternEurope		10.131	7.601
		(0.086)	(0.115)
SouthernEurope		-12.183	-13.813
		(0.189)	(0.092)
WesternAfrica		-27.689	
		(0.016)	
WesternEurope		9.432	5.866
		(0.061)	(0.062)
Num.Obs.	56	56	56
AIC	58.7	56.0	59.2
BIC	74.9	80.3	81.4
Log.Lik.	-21.354	-15.989	-18.580

Table 3.7 – Summary statistics of the classification linear models, GLM, Inputs=relative regional yield changes

Chapter 4

Forecasting global maize prices from regional productions

Co-author: David Makowski

Abstract

This study analyses the quality of six regression algorithms in forecasting the monthly price of maize in its primary international trading market, using publicly available data of agricultural production at a regional scale. The forecasting process covers a period of between one and twelve months ahead, using six different forecasting techniques. Three of them (CART, RF, and GBM) are tree-based machine learning techniques able to capture the relative influence of maize-producing regions on global maize price variations. Additionally, we consider two types of linear models - standard multiple linear regression and vector autoregressive (VAR) model. Finally, TBATS serves as an advanced time-series model that holds the advantages of several commonly used time-series algorithms. Using cross-validation, we compare the predictive capabilities of these six methods. We find RF and GBM have superior forecasting abilities relative to the linear models and that TBATS is more accurate for short time forecasts when the time horizon is shorter than three months. On top of that, the models assess the marginal contribution of each of the producing regions to the most extreme price shocks that occurred through the past six centuries, in both positive and negative directions, using Shapley decompositions. Our results reveal a strong influence of North-American yield variation on the global price, except for the last months preceding the new-crop season.

4.1 Introduction

The prices of food and agricultural products are of interest to many stakeholders, including policymakers, traders, and consumers. Moreover, these prices have a high impact on businesses and people who depend on agricultural products. Therefore, predicting the prices of agricultural commodities is a highly strategic issue.

Price forecasters commonly use the prediction methods depending on the target time horizon. For example, Partial-equilibrium (PE)and General equilibrium models (GEM) are common (Valin et al., 2014) for long-term predictions because long-term price changes (i.e., over several years or decades). In such horizons, price changes are primarily the results of political or climatic changes and long-run market structures and demographic dynamics. Therefore, such predictions are relevant in the context of the need for ahead-of-time adaptation and long-term strategy, particularly for policymakers.

Short-time agricultural price changes are relevant for traders who sell or buy agricultural commodities hourly or daily. At this time frame, price fluctuations depending on the short-term balance between supply and demand and the commodity market dynamics (Piot-Lepetit and M'Barek, 2011). Therefore, short-term predictions usually use standard time series analysis techniques such as smoothing methods or ARIMA models.

This paper focuses on medium-time fluctuations, i.e. over periods of up to one year. Those fluctuations mainly affect domestic markets but sometimes spill over into the global market, depending on their level, the crop in question, and region which had been affected (Headey and Fan, 2010). The United States Department of Agriculture (USDA) (ERS-USDA, 2021) publishes monthly price forecasts based on a model named World Agricultural Supply and Demand Estimates (WASDE), to provide USDA staffs and policymakers with price forecasts monthly and for up to 16 months ahead (Hoffman et al., 2015). However, the methodology used in WASDE is considered as complex (Hoffman et al., 2018) and is not fully accessible. Furthermore, (Hoffman, 2011; Warr, 1990; Hoffman et al., 2015; Lusk, 2016) have criticised it for its lack of accuracy.

Here, we focus on maize, a major agricultural commodity used worldwide. Maize plays a crucial role in global food security (directly or through livestock feed) and energy crops. More specifically, our objective is to predict maize's monthly average global price. To do so, we test three machine learning (ML) algorithms based on regression trees, predicting the annual change in the monthly maize price from the annual changes in regional maize productions or yields. These techniques aim at capturing the effect of the regional supply level change on global maize prices. In addition to these three ML algorithms, we use two

time-series methods: vector autoregressive model (VAR), which had previously proven to capture the effects of shocks in exogenous variables on feed prices (Schaub and Finger, 2020), and Trigonometric Seasonal Box Transformation with ARMA residuals Trend and Seasonal Components (TBATS), a model that enables us to predict price changes based on the combined influence of trends, seasonality, and auto-correlations of monthly prices.

In this paper, we compare the performances of these five models for out-of-sample predictions to those of a benchmark model based on linear regression for time horizons of one to twelve months ahead. Besides, we show that the three ML algorithms tested here can be used to identify the most influential maize-producing regions and to identify the origins of price shocks.

4.2 Data

The relationship between commodity price shocks and annual supplies depends not only on how production changes at the global scale but also on regional productions (Hertel et al., 2016). For this reason, we used regional production and yield annual changes as dependent variables (see table 4.2 and table 4.3 in Appendix 4.A). These data were collected in 242 countries and are publicly available in FAOSTAT for 1961 to 2019 to aggregate 19 regions (FAO, 2020). As the harvest dates differed across these regions (according to their location in the northern or southern hemispheres), we assumed that the production (or yield) in a given region would have an impact on maize prices during one year starting from the harvest month of the biggest producer of this region. This period corresponds roughly to the market year of each region. For example, based on this approach, we assume that production (yield) in Northern America in year y starts impacting monthly maize price from October of that year until September year $y + 1$. In contrast, we assume that the production in Southern America (located mainly in the southern hemisphere) impacts maize prices from March year y until February year $y + 1$. All the periods considered are shown in Appendix 4.A.

We converted the nominal maize prices (US No. 2 yellow from the World Bank's commodity market database) into real 2010 USD. Then, we defined $q_{m,y}$ as a series of deflated monthly global maize prices, where m and y are the months and year indices, respectively, so that m=1,...,12 and y=1,...,Y. The second series $z_{k,y}$ describes the production (or yield) in a region k (k=1, ...,K) and a year y. Since these variables have different units, we express them in relative terms as follows:

$$p_{m,y} = \frac{q_{m,y} - q_{m,y-1}}{q_{m,y-1}} \tag{4.1}$$

$$x_{k,y} = \frac{z_{k,y} - z_{k,y-1}}{z_{k,y-1}} \tag{4.2}$$

Fig 4.1 provides a visual representation of the three types of time series used in this study. Note the significant differences between the levels of variability of production and yield.

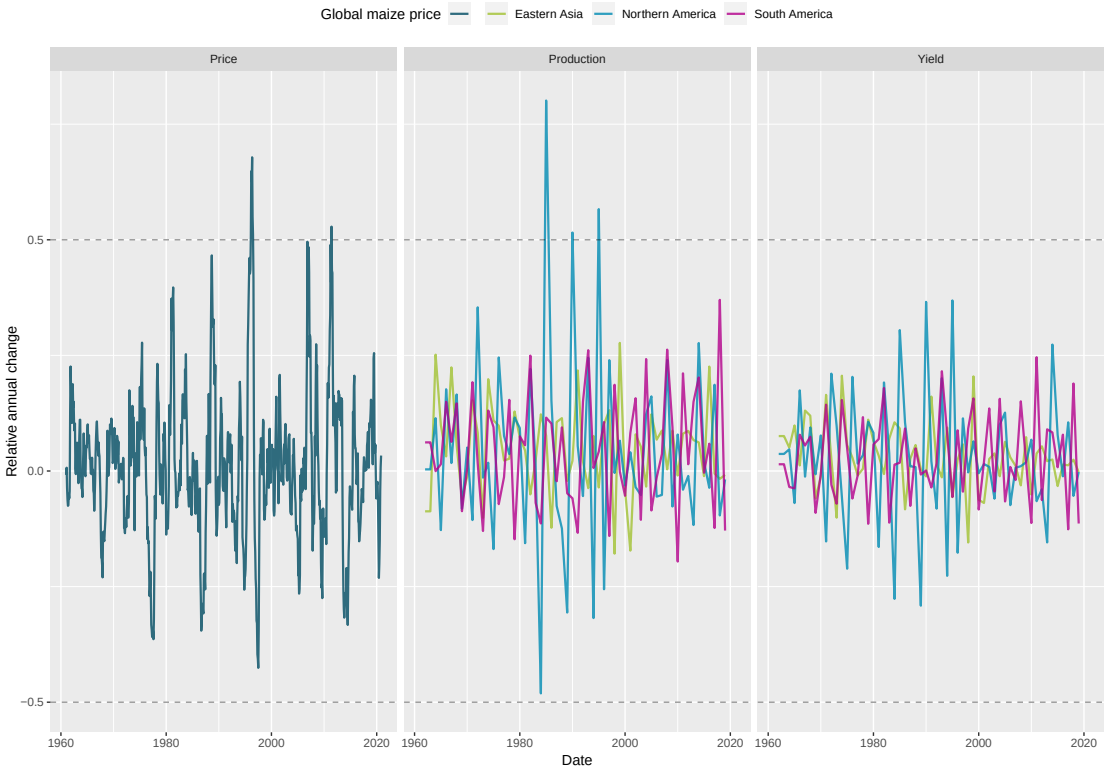


Figure 4.1 – Annual changes (%) in global price, regional production and regional yield (in the three biggest producing regions).

4.3 Methods

We consider two types of models, i.e. models predicting maize price changes as a function of yearly production (yield) changes and models predicting maize price changes from past monthly observations of price changes and yearly production (yield) changes. The first type of models can be expressed as

$$p_{m,y} = f(x_{1,y}, ..., x_{k,y}, ..., x_{19,y}) \tag{4.3}$$

and the second as:

$$p_{m,y} = f(p_{m,y-1}, x_{1,y}, ..., x_{k,y}, ..., x_{19,y}) \tag{4.4}$$

where k is the region index. We consider different types of function f, based on linear models and machine learning algorithms, as described below.

4.3.1 Models 1, 2, and 3 - Machine learning

The use of ML makes it possible to discover hidden patterns about the relationship between the direction and magnitude of changes in $p_{m,y}$ versus the variability in $x_{k,y}$. This way, we can detect non-linear relationships between variables without making any strong preliminary assumptions on the shapes of the relationships. More specifically, we use three different approaches, namely classification and regression trees (CART, model 1), Random Forest (RF, model 2), and gradient boosting (GBM, model 3).

Classification and regression trees (CART) is a recursive ML technique developed by Breiman et al. (1984). The algorithm receives all the observations that include information about the input variables ($x_{1,y}$,$x_{2,y}$,...,$x_{19,y}$), and build a regression tree to minimise the error rate in predicting $p_{m,y}$, measured here by the residual sum of squares (RSS). The partitioning process starts with a single leaf at the top of the tree (root). In each step, the algorithm splits the node into two, each defined by a different input (region), and stops when no further improvement is possible, i.e., when RSS cannot be any lower. We fit CART using the `rpart` package of R (Therneau et al., 2019). An illustration can be found in fig 4.8, Appendix 4.A.

CART models are usually easy to interpret but are considered weak learners (Luo et al., 2019), which might be highly biased. To overcome this problem, we apply two alternative methods based on the assembly of high numbers of individual trees, namely random forest (RF) and gradient boosting machine (GBM) (Liaw et al., 2002). RF takes a random subset of the original dataset and uses it to fit a basic decision tree to predict $p_{m,y}$. A bootstrapping process is implemented T times ($t = 1, ..., T$), and the T resulting trees are then averaged to produce the final predictions. Here, we find that RF leads to the most stable results with T = 500 trees. RF is applied here using the package `randomForest` (Breiman et al., 2018).

Similar to RF, GBM examines periods as a subsample of the data and uses them to fit a single tree. Nevertheless, unlike the latter, the selected sub-sample

is chosen according to the estimation error obtained in the analysis of the previous training set. In this study, we find that GBM returns the most accurate forecast when using T=100 trees. This method is implemented with the `gbm` R package Friedman (2001).

4.3.2 Model 4 - Multivariate linear regression

In linear model (LM), price change $p_{m,y}$ is related to $x_{k,y}$ as:

$$p_{m,y} = \alpha_m + \sum_{k=1}^{K^s} \beta_{k,m} x_{k,y} + \epsilon_{m,y} \tag{4.5}$$

where α_m is the intercept, $\beta_{k,m}$ are regression parameters, $\epsilon_{m,y}$ are the residuals, and K^s (< 19) is the number of selected regions. One model is fitted separately for each month m (with the function lm of the R software). To obtain a parsimonious model, we use a step-wise algorithm (based on AIC) to select the most influential K^s regions. Because of its simplicity and strong assumptions, this linear model serves as a benchmark model.

4.3.3 Model 5 - VAR

Model vector autoregressive (VAR) empirically examines the evolution and common effects that time series have on each other so that it describes the relationships over time between all the variables in question. In this case, the model includes several dynamic variables that affect each other and the effect of shocks in each explanatory variable on the global price. Unlike the models we have used so far, $p_{m,y}$ is not only a function of $x_{k,y}$ but also of the past price change values, $p_{m,y-1}$.

The basic purpose of VAR is to describe the interactions between all variables and try to predict future effects. Since firstly introduced by Sims (1980), VAR has been widely used and is considered a particularly effective tool in designing policy strategies (Bernanke et al., 2005; Jouchi et al., 2011). Here, we use this approach to predict $p_{m,y}$ as a function of $p_{m,y-1}$ and of $x_{k,y}$ as follows:

$$p_{m,y} = \alpha_m + \beta_{0,m} p_{m,y-1} + \sum_{k=1}^{K} \beta_{k,m} x_{k,y} + \epsilon_{m,y} \tag{4.6}$$

One separate model is fitted for each month m using the `vars` R package (Pfaff and Stigler, 2018).

4.3.4 Model 6 - TBATS

The Trigonometric Seasonal Box Transformation with ARMA residuals Trend and Seasonal Components (TBATS) model (De Livera et al., 2011) is an upgraded time-series model which can deal with trends, multiple-seasonality and auto-correlations. This method automatically determines whether a Box-Cox transformation of the data is required, whether seasonality needs to be accounted for (based on Fourier series), and whether a time trend should be included. It also automatically selects the optimal number of autoregressive and moving average components for predicting the target response variable.

Contrary to the models mentioned above, TBATS is fitted to the time series of the relative annual change in the monthly price of maize directly, without using the production data. TBATS aims at predicting price changes from the past series of observed price changes without taking regional productions into account. We consider several time horizons for price change predictions, from one month ahead to one year ahead. Here, this method is implemented with the R package `forecast` (Hyndman et al., 2020).

4.4 Model evaluation

The model prediction errors were assessed and compared using a cross-validation (CV) technique, implemented separately for each month and model. At each iteration of the CV, we select a sub-sample (training-set) containing observations from all the first $\tilde{Y}$ years plus the i following years (i is successively set equal to 1, 2, ...I, where I=13 or 14, depending on the month considered, and $\tilde{Y}$ is equal to 44 or 45). At each iteration, the training set trains the models, and the resulting trained models are used to predict the price change at year $\tilde{Y} + i + 1$. With this procedure, we ensure that at least $\tilde{Y} + 1$ years of data are available to train the models. Smaller datasets would lead to inaccurate predictions and a lack of identifiability.

We define the forecast error for the model in month m of the marketing year y as:

$$\epsilon_{m,y} = \tilde{p}_{m,y} - p_{m,y} \tag{4.7}$$

where $p_{m,y}$ is the observed price, and $\tilde{p}_{m,y}$ is the forecast made in m of the marketing year y by any of the models considered in this study. We then use these errors to compute an RMSE for each month and each model, as:

$$RMSE_m = \sqrt{\frac{\sum_{i=1}^{I}(\tilde{p}_{m,y} - p_{m,y})^2}{I}} \tag{4.8}$$

The accuracy of TBATS predictions is evaluated by computing the RMSE criterion for 12 different time horizons, i.e. h=1,2,...,12 months ahead. For a given year, a given month, and a given time horizon, TBATS is trained using all price data available before the month $m-h$, and the trained model is used to predict the value of $p_{m,y}$ ($\tilde{Y}$ = 28, I_{TBATS} = 690). This procedure is repeated relative to every year, every month, and time horizon. Then, a specific value of RMSE has computed for each month m and time horizon h combination by averaging the prediction errors among all years of data.

Finally, we assess and rank the influences of the producing regions using two different techniques. First, we use permutation ranking with RF and GBM to assess the importance of each region for predicting maize prices. This approach allows us to identify the most and least influential regions when forecasting maize price changes (Appendix 4.A). Second, using the Shapley decomposition technique (Shapley, 2016), we strive to identify the regional production variations responsible for specific extreme price change anomalies that occurred at some specific months and years in the past. Importance ranking and Shapley decomposition were implemented using the package `iml` of the R software.

4.5 Results

Fig 4.2 below presents the price change forecasts obtained by the different models considered.

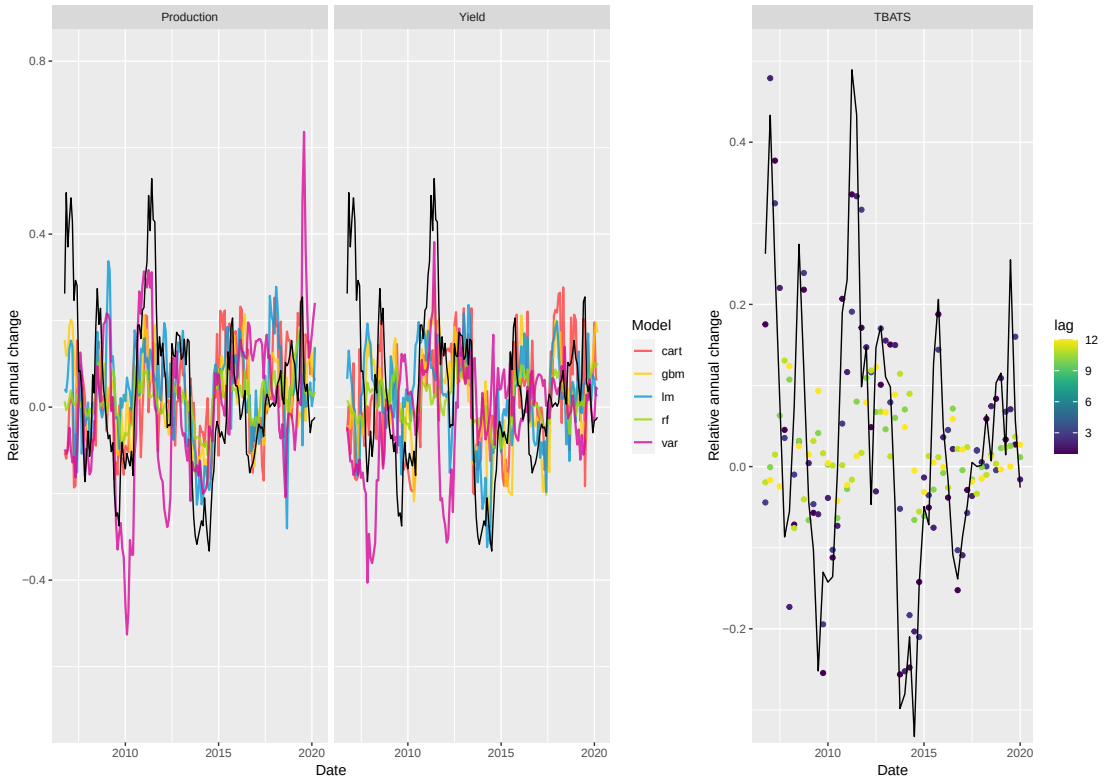


Figure 4.2 – Forecasted maize monthly price changes obtained with all models. CART, RF, GBM, LM, and VAR are shown on the left. The TBATS forecasts are displayed on the right for time lag ranging from 1 to 12 months. The black lines indicate the observed price changes

The left side of the figure (fig 4.2) presents the forecasts derived from the ML and linear models in the period between October 1990 and January 2020 (Segmentation by months is in Appendix 4.A). Generally, ML models tend to produce more accurate predictions than LM and VAR, as the latter two methods produce somewhat fluctuating predictions. Nonetheless, VAR seems to perform well in case of extreme price shocks.

TBATS predictions tend to diverge more from the observations when derived several months before the dates of forecast (right side of fig 4.2). For lag longer than three months, the predictions differ a lot from the observations.

Fig 4.3 below shows the relative advantage of using each model for forecasting $p_{m,y}$, with the reference value being the observed standard deviation of the price each month ($sd(p_{m,y})$). This measure corresponds to the difference between $sd(p_{m,y})$ and the RMSE of each model the same month, divided by $sd(p_{m,y})$, and expressed in percentages. A positive value indicates that the corresponding model is better than a constant prediction equal to zero. This way,

all the points below the black horizontal line indicate models showing ineffective price forecasts. In contrast, those above indicate models whose average forecast errors are lower than $sd(p_{m,y})$. Such models are better than a constant prediction equal to zero. The highest relative advantage values (located at the top of the graphic) indicate the most relevant models, which appear to be the tree-based methods in most cases (GBM, RF, and CART). The results are presented separately for TBATS to assess the influence of the time lags on the prediction accuracy. The relative advantage of TBATS compared to a constant prediction is high for a time horizon up to 3 months and became very low after six months.

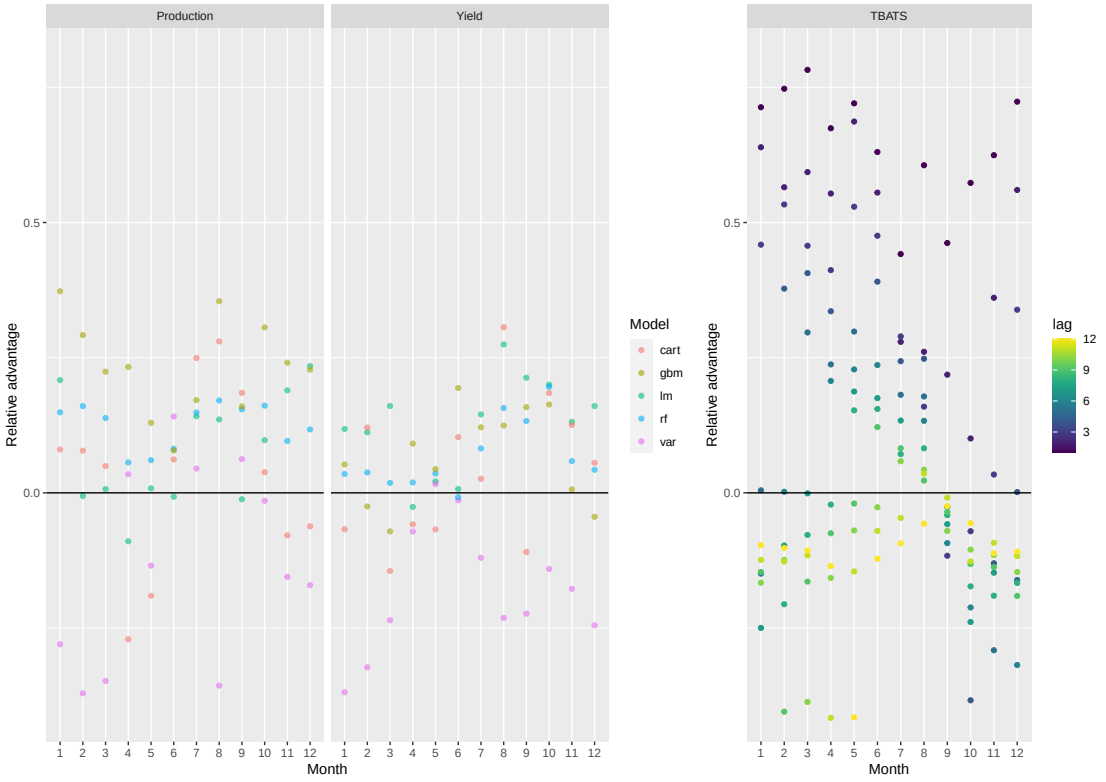


Figure 4.3 – Relative advantage in terms of prediction accuracy of the forecasting models, over 1990-2020. This measure corresponds to the difference between the standard deviation of the price changes in the whole dataset ($sd(p_{m,y})$) and the RMSE of each model the same month, divided by $sd(p_{m,y})$, and expressed in percentages. It indicates the relative benefit of using the models compared to a constant prediction equal to zero. ML methods, LM and VAR were used with production and yield inputs, successively.

Results show that several models are more accurate than constant predictions. The relative advantages of GBM tend to be higher when applying regional

productions as inputs rather than regional yields. However, the differences between the two types of inputs are not very high. The relative advantages of LM or VAR are often negative, revealing that these methods do not often perform better than constant predictions. Concerning TBATS (fig 4.3, right), price change predictions are more accurate than constant predictions, as long as the time-horizon for forecasting remains lower than 3 or 4 months. For such cases (dark points in Figure 4.3), the relative advantage of TBATS predictions can be higher by 78% higher than constant predictions. On the other hand, for longer time horizons, the accuracy of TBATS decreases rapidly and becomes not efficient at all for lag higher than six months.

We used the cross-validated values of RMSE to identify the most accurate models for each time horizon between one month and a year ahead, as shown in Table 4.1.

Table 4.1 – Best forecasting options for different months. The names reported for each month correspond to the models showing the lowest RMSE for predicting price change at this period. The numbers indicate the RMSE values of the best models. As TBATS tends to perform very well for short time lags, TBATS appears to be the best option for all months when the time lag is in the range of 1 to 5. For longer time lags, other models (in particular GBM) are more accurate. For CART and GBM, the name between brackets indicates whether the predictions were more accurate with regional yields or productions. The last column (sd p obs.) indicates the empirical standard deviations of the observed price changes, which can be considered as a benchmark

	Time lags (months)												sd p
m	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	obs.
Jan.	TBATS			GBM ($x_{k,y}$=Production)									
	0.05	0.06	0.09	0.11									0.17
Feb.	TBATS				GBM ($x_{k,y}$=Production)								
	0.05	0.08	0.09	0.11	0.13								0.19
Mar.	TBATS				GBM ($x_{k,y}$=Production)								
	0.04	0.08	0.11	0.12	0.14								0.18
Apr.	TBATS				GBM ($x_{k,y}$=Production)								
	0.06	0.08	0.11	0.13	0.14								0.18
May	TBATS				GBM ($x_{k,y}$=Production)								
	0.06	0.06	0.09	0.12	0.15								0.18
Jun.	TBATS						GBM ($x_{k,y}$ = Yield)						
	0.08	0.09	0.11	0.13	0.14	0.16	0.17						0.22
Jul.	TBATS			CART ($x_{k,y}$=Production)									
	0.11	0.15	0.14	0.15									0.20
Aug.	TBATS	GBM ($x_{k,y}$=Production)											
	0.07	0.12											0.19
Sept.	TBATS	CART ($x_{k,y}$=Production)											
	0.09	0.14											0.18
Oct.	TBATS	GBM ($x_{k,y}$=Production)											
	0.06	0.11											0.17
Nov.	TBATS		GBM ($x_{k,y}$=Production)										
	0.06	0.11	0.15										0.19
Dec.	TBATS			GBM ($x_{k,y}$=Production)									
	0.04	0.07	0.1	0.13									0.17

According to table 4.1, TBATS is the best model to predict $p_{m,y}$ in each of the 12 months of the year in a forecast range of two (September to November) to five months ahead (February March, and May to August). However, to predict a price for time horizons longer than three or four months, ML models are often more reliable and, in addition, offer the possibility to identify the most and least influential regions based on importance ranking and Shapley decomposition. For example, importance rankings (Supplementary fig 4.9 and fig 4.10 in Appendix 4.A) reveal a strong influence of Northern America for almost all months. Correspondingly, Western Asia, another central producing region, had strong relative influence substantially during the two months preceding the harvest season in Northern America (July and August).

The Shapley decompositions confirm the strong influence of Northern America. Two Shapley decompositions are shown in fig 4.4 for two extreme events corresponding to a substantial price increase and a firm price decrease over the period considered. Each regional Shapley value indicates the share of the price anomaly (either in December 1995 or in December 2013) explained by the corresponding region. According to these decompositions, the high maize price increase occurring in December 1995 appears to be mainly due to the changes in maize production obtained in Northern America and, to a lower extend, in Southern Africa. The maize productions in Northern America are also responsible for a significant share of the substantial price decrease in December 2013. Other examples confirming the significant role of Northern America are shown in Appendix 4.A.

4.6 Discussion

This research project analyses six decades of the global maize market. Maize is the highest produced crop worldwide and an essential energy source, especially in developing countries. Our study attempts to forecast the international monthly price of this commodity as a function of regional production. Although many have analysed and attempted to predict the price of maize accurately (see, for example, Hoffman et al. (2015), Xiaojie and Yun (2021), and Ahumada and Cornejo (2016)), very few have developed methods that are both easy to reproduce and interpret by users who are not necessarily specialists in price prediction. With regards to ML, our study offers a double contribution. First, on the academic side, it is the pioneer in performing Medium Term price forecasting of maize using ML, let alone detecting the main drivers for maize price changes through investigation of the ML algorithms. Second, it offers a practical, non-academic contribution - providing a range of price forecasting tools that stakeholders who do not have access to the best tools needed to trade in

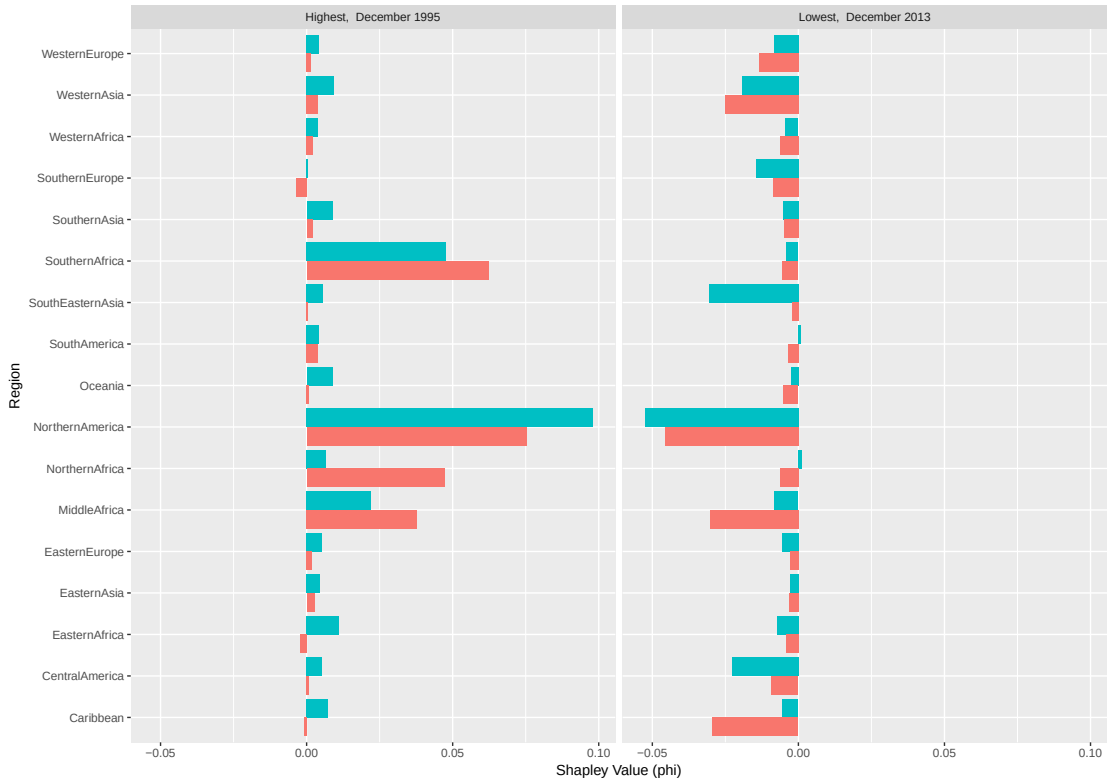


Figure 4.4 – Shapley values for December 1995 (strong price increase) and December 2013 (strong price decrease). The decompositions show the contributions of the producing regions to two extreme relative price changes (regional productions in red, and yields in blue). At a given date, the sum of the regional Shapley values is equal to the price change anomaly

global markets optimally can use easily.

Our study uses machine-learning algorithms and relies on publicly available data only. It is based on the use of annual regional yields and productions to enable the user to evaluate the sense behind the results, principally challenging the transparency of each model. The chosen models were those which had previously tested in relation with the global maize market and regional production (Zelingher et al., 2021), notable are CART (Breiman et al., 1984), RF (Hastie et al., 2009) and GBM (Friedman, 2001). To those are added two econometric models, each possess certain advantages: VAR (Sims, 1980), which can detect inter-and intra-effects of local productions shocks, and TBATS (De Livera et al., 2011), as a time-series based approach that has proved to achieve low forecasting errors (Lima and Laporta, 2020).

To understand the process behind the models' output and identify the forces

which drive these price change forecasts, we use two evaluation techniques: a relative importance examination and Shapley decomposition (Shapley, 2016). The integration of these two model-agnostic approaches guarantees an overall understanding of the different forces that act in the global commerce of maize. At first, the relative importance examination quantifies the impact of each producing region on annual changes in the monthly price (as a consequence of its contribution to the forecasting ability of the model). Next, Shapley provides a case-explicit examination, showing the nature (positive/negative correlation) and level (as a relation to the interaction of all regions with the dependent variable) regional production changes in a specific year affected price changes at some selected date. This measurement is especially critical for understanding the forces influencing extreme price changes, which might drive a global food crisis.

The paper emphasises the importance of conducting a constant comparison between the forecast values of several forecasting algorithms while looking at the marginal contribution of each factor to the output. Furthermore, this study highlights the importance of predicting global maize prices according to various scenarios using different models. This way, the impact of the various producing regions (input) can be examined and evaluated accordingly. That becomes crucial when a change in the production of a highly influential region is observed or projected.

Our results demonstrate significant dissimilarities between the impact levels of the different regions, with monthly variance. Indeed, the relationship between maize prices and production changes in major producing regions are apparent, as Headey and Fan (2010) had already claimed. However, the "New-crop" period also plays a critical role. It is not by hazard that the impact of Northern America is evident throughout the entire year except for July and August. As it happens, the primary harvest season in this region begins next month, i.e., in September, so it is clear that the previous year's crop is no longer traded. However, it is not yet possible to predict with certainty the amount of crop harvested in the coming months. Therefore, the impact of Northern America in these months is low, despite being a big maize producer and exporter. Similarly, these two months are when the relative impact of Western Asia becomes high, as they present the main harvest season in this region.

This study proposes a significant contribution to the price forecasting literature of agricultural commodities. First and foremost, it is constructed to be replicated. Whereas to date, many have been obliged to base their food security strategy on paid data obtained from private companies or based on final results published as obscure numbers (see WASDE, World Bank Commodities Price Forecast or FAO-AMIS Market Database); our research offers an available

high-quality alternative. Indeed, activating the code through all stages, including those leading to the "black box" opening, will provide the user with predicted maize price values and an understanding of the processes and effects leading to these forecasted price values. Another contribution derives from the division of the forecasting period simultaneously to months and time horizon, giving the users the unique opportunity to adapt their strategy in case of possible changes in the maize market, principally in high influential regions. Lastly, the paper enables analysis of specific events through the Shapley-algorithm, while taking the opportunity to understand the existing gap between average marginal regional impacts and those that occur in times of extreme price changes.

Although this project deals with maize, the tested methodologies can be applied to other agricultural commodities. In future work, we will examine this assumption on several different internationally traded crops. There, we will strive to capture inter-and intra-sectoral differences and detect the factors impacting price volatilities in each of them. Indeed, the broader this open-price-forecasting will get, the higher will become its contribution to global food security.

4.A Appendix

4.A.1 Data information

Table 4.2 – Variable description and sources.

Final data				
Data Production	Unites % change/year	Time-range 1962 - 2019	Indices k = Region, y = Year	Sign $x_{k,y}$
Yield	% change/year	1962 - 2019	k = Region, y = Year	$x_{k,y}$
Price	% change /year	01/1961 - 11/2020	m = Month, y = Year	$p_{m,y}$
Initial information				
Data	Unites	Time-range	Source	Sign
Price	Nominal USD, m/tonne	01/1960 - 11/2020	World Bank, Pink Sheet (2020)	
Price index	USD (2010 = 100)	01/1960 - 11/2020	World Bank, Pink Sheet (2020)	
Production	tonnes / year	1961 - 2019	FAO STAT (2020)	$z_{k,y}$
Yield	hg / ha	1961 - 2019	FAO STAT (2020)	$z_{k,y}$
Real price	Real USD (2010)	01/1960 - 11/2020		$q_{m,y}$

Table 4.3 – Data composition (19 regions) and summary statistics of inputs.

Production data (FAOSTAT)					
	Production (1000 tonnes)		Yield (1000 hg/ha)		
Regions	%/total	Average	Average	Min. (year)	Max. (year)
Caribbean	0.08%	460	11.13	8.54 (1993)	14.78 (2004)
Central America	3.27%	17,825	20.13	9.74 (1961)	34.41 (2018)
Central Asia*	0.11%	1,251	47.74	25.63 (1997)	65.72 (2018)
Eastern Africa	2.64%	14,395	13.54	9.52 (1965)	19.44 (2014)
Eastern Asia	19.03%	103,592	39.01	12.28 (1961)	60.89 (2017)
Eastern Europe	6.23%	33,925	36.91	18.4 (1963)	70.02 (2018)
Middle Africa	0.47%	2,563	8.5	6.74 (1979)	10.9 (2015)
Northern Africa	0.91%	4,974	41.2	15.91 (1961)	69.12 (2012)
Northern America	40.59%	220,940	73.34	39.23 (1961)	118.01 (2017)
Northern Europe*	0.00%	38	34.36	10 (1985)	71.74 (2016)
Oceania	0.08%	424	50.47	17.33 (1966)	87.81 (2015)
South America	9.75%	53,066	27.72	12.95 (1964)	59.26 (2017)
South-eastern Asia	3.52%	19,186	21.83	9.02 (1961)	46.14 (2018)
Southern Africa	1.68%	9,170	23.89	7.88 (1992)	58.13 (2017)
Southern Asia	2.95%	16,037	17.28	10.02 (1971)	34.29 (2017)
Southern Europe	3.83%	20,845	54.31	21.13 (1961)	90.58 (2018)
Western Africa	1.47%	7,975	12.02	6.96 (1972)	19.54 (2018)
Western Asia	0.56%	3,069	34.87	11.4 (1962)	79.18 (2018)
Western Europe	2.80%	15,264	69.51	22.56 (1962)	103.05 (2011)

* Central Asia and Northern Europe are excluded from analysis due to lack of data

Table 4.4 – Market year of maize, relative to the majority of production in each region.

Regions	**Market Year**
Caribbean	July - June
CentralAmerica	October - September
CentralAsia*	July - June
EasternAfrica	July - June
EasternAsia	April - March
EasternEurope	October - September
MiddleAfrica	July - June
NorthernAfrica	July - June
NorthernAmerica	October - September**
NorthernEurope*	July - June
Oceania	April - March
SouthAmerica	March - February
SoutheasternAsia	January - December
SouthernAfrica	May - April
SouthernAsia	July - June
SouthernEurope	October - September
WesternAfrica	July - June
WesternAsia	September - August
WesternEurope	July - June

*Central Asia and Northern Europe are excluded from analysis due to lack of data.
**Agricultural year in the USA had changes in 1986. Starting from this year, any year y refers to September y to August $y + 1$

4.A.2 General presentation

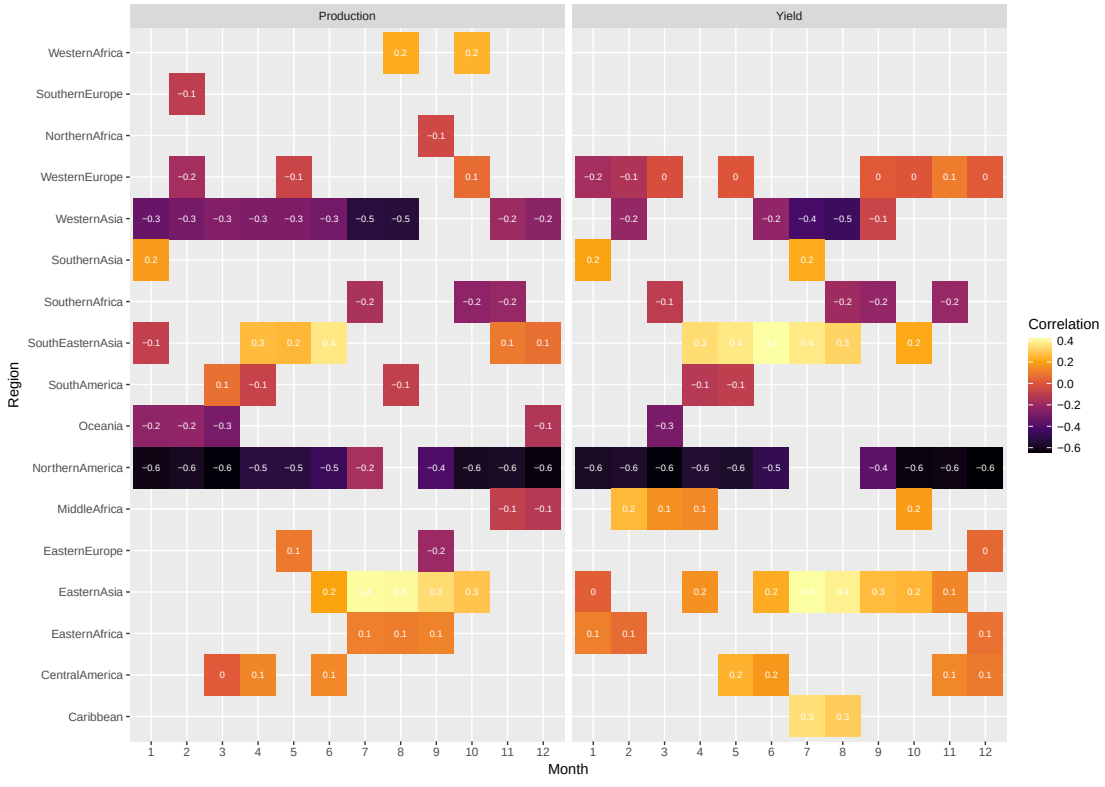


Figure 4.5 – Correlations between global maize price changes and regional maize productions (left) and yields (right)

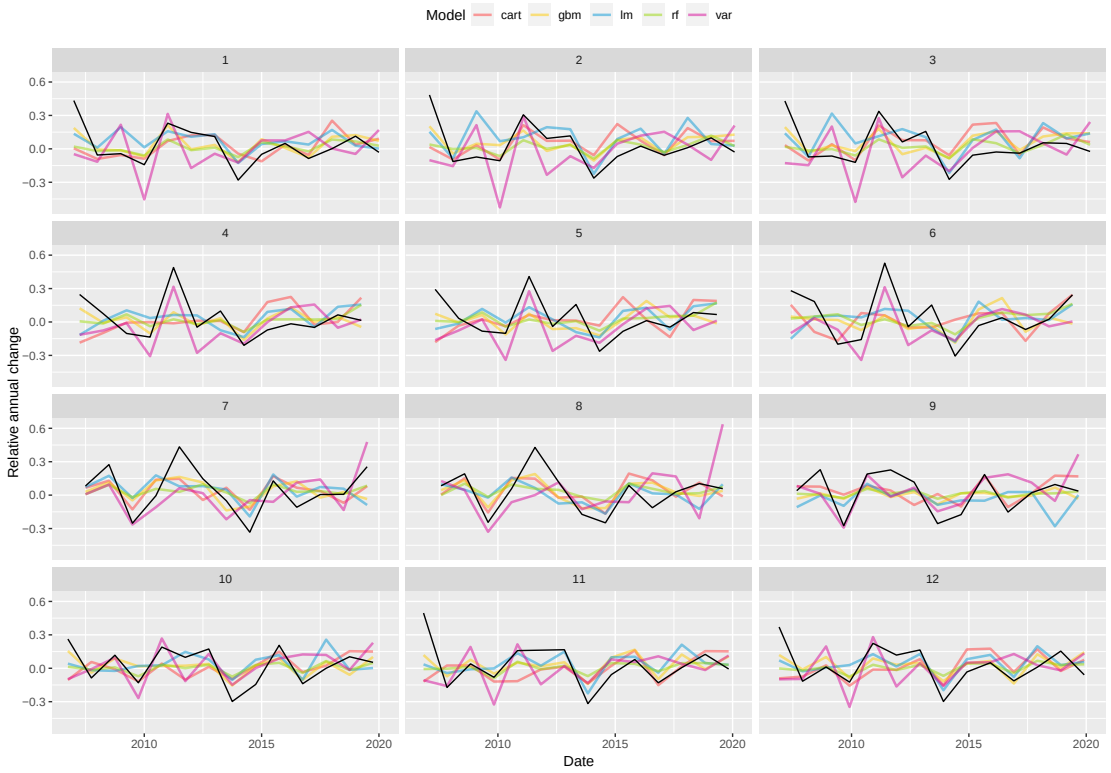


Figure 4.6 – Observed yearly price changes (in black) and model predictions (colours) using regional productions

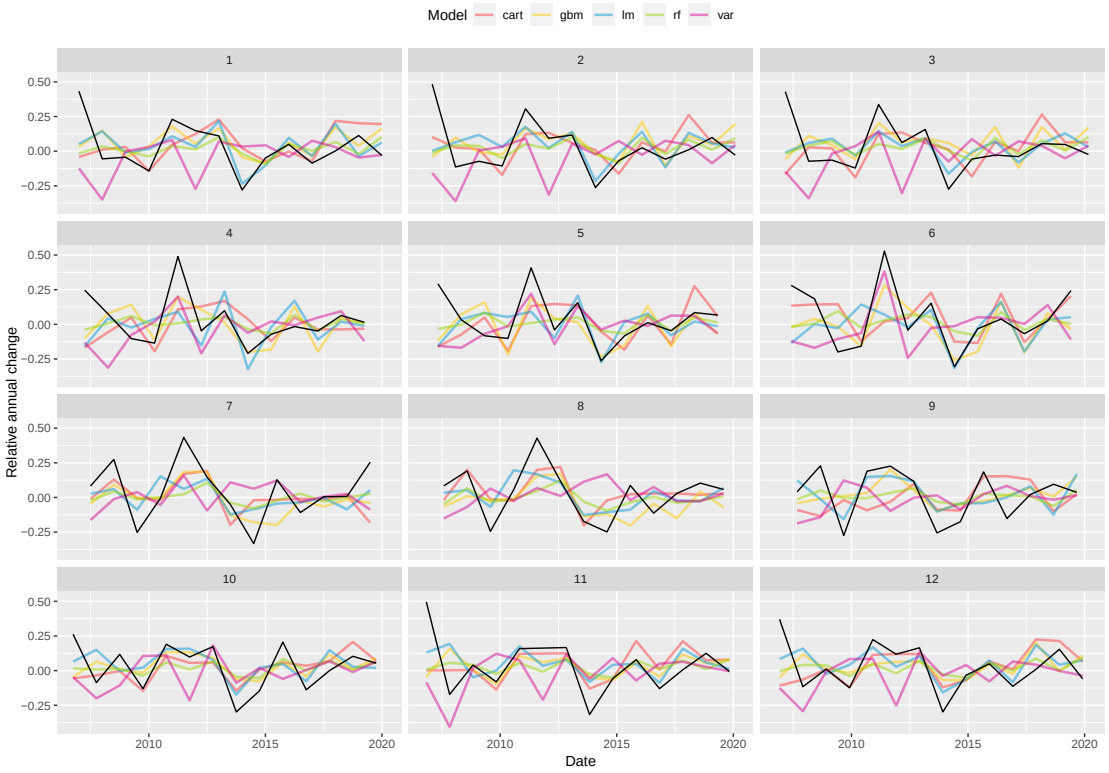


Figure 4.7 – Observed yearly price changes (in black) and model predictions (colours) using regional yields

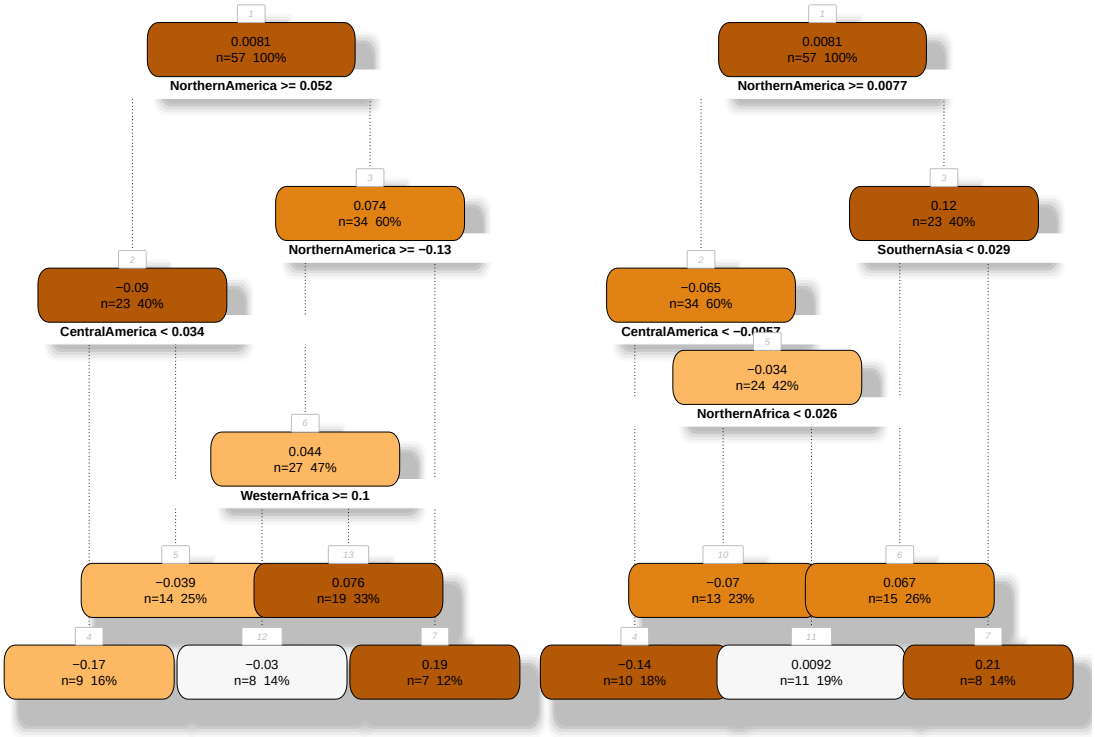


Figure 4.8 – CART models compute the relative price change of maize in December ($p_{12,y}$) as a function of relative regional production changes (right) and relative regional yield changes (left). The nodes of each tree include three numbers; the average relative price change value over all data falling in the considered node, the number of data in each node (n), the % of data in each node. The terminal nodes (at the bottom) report the relative price changes predicted by the CART models

Fig 4.8 was implemented with the package `Rattle` of the R software (Williams, 2011) `fancyRpartPlot` function.

4.A.3 Breakdown of the price change by inputs and regions

The importance-ranking maize regional output has been obtained for the most accurate forecasting model - GBM, as shown in fig 4.9 and fig 4.10 below. The contribution to the prediction accuracy of $p_{m,y}$ determines K relative importance values, as returned by these two models, separately. A region is considered as influential if a random choice of its corresponding input value ($x_{k,y}$) leads to a substantial increase of the mean squared error (MSE) of the price change pre-

dictions.

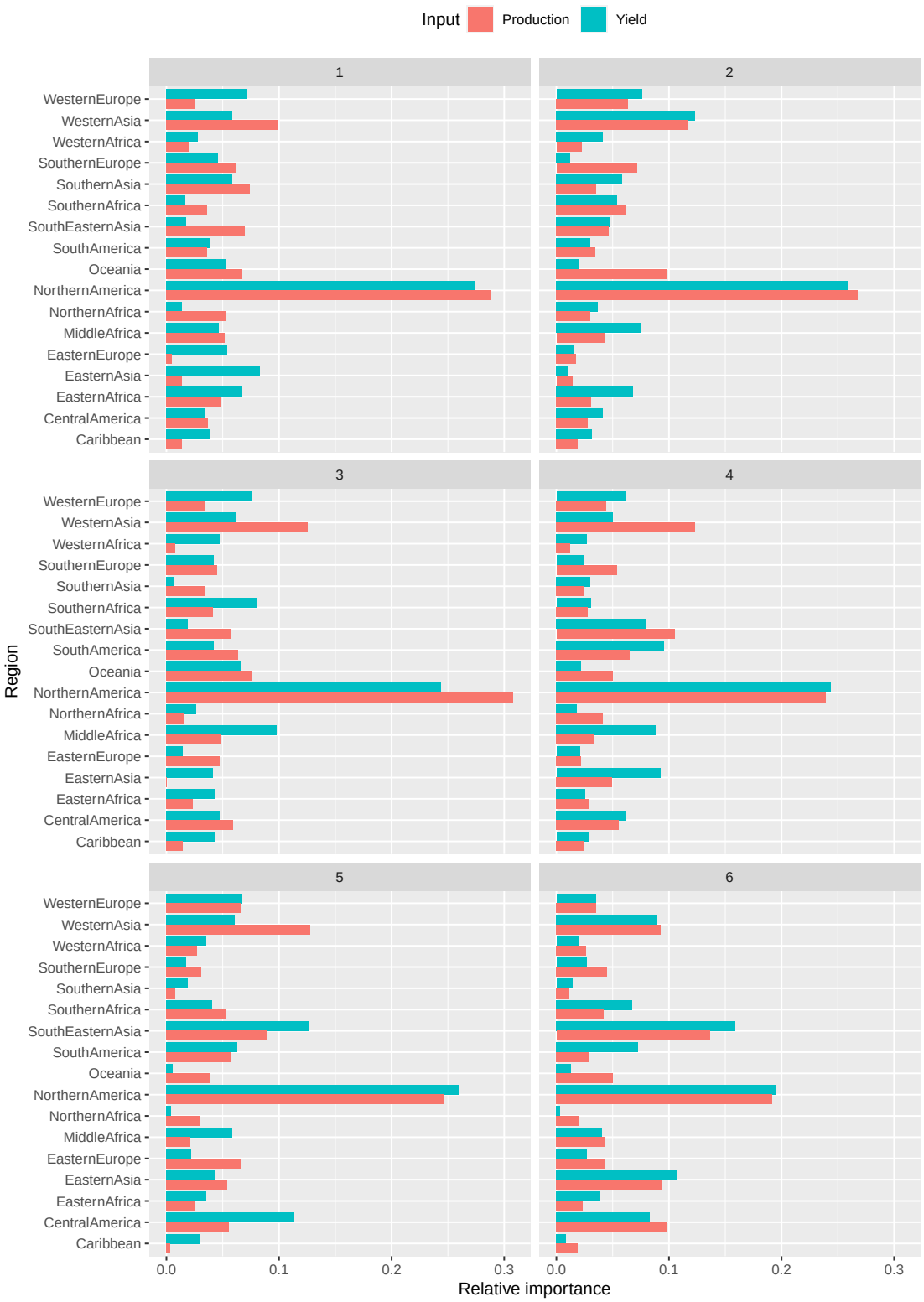


Figure 4.9 – Importance levels of $x_{k,y}$ for predicting the global maize price. Importance levels are computed using MSE and measure the extent to which the model accuracy decreases with a random permutation of each input relative to GBM. The figure shows the marginal importance of each region in the first semester of the year, i.e January - June. As such, the higher the bar reaches on the X-axis, the higher the marginal effect of that area on the maize price. The figure is divided into two parts: production in red and yield in blue.

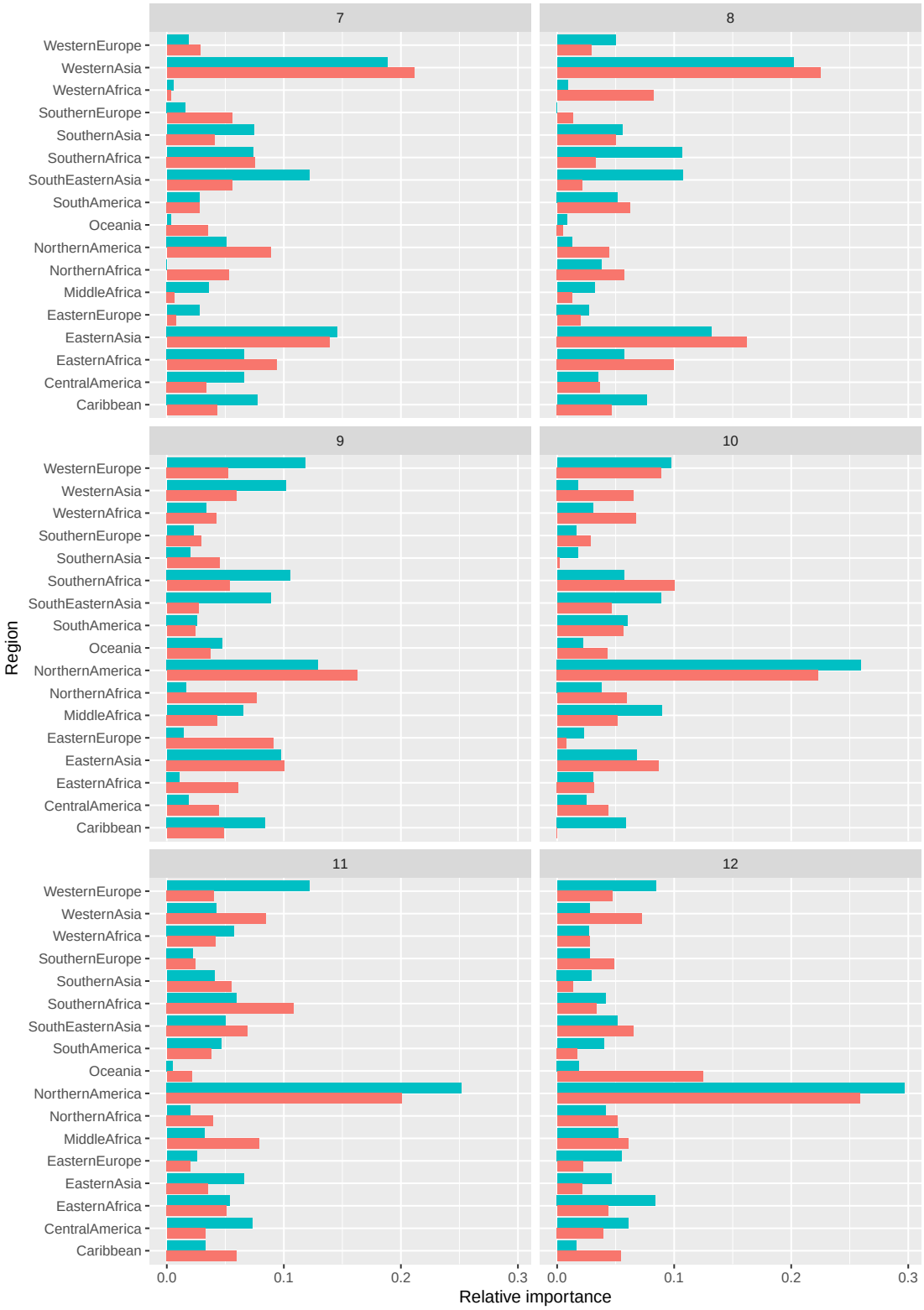


Figure 4.10 – Importance levels of $x_{k,y}$ for predicting the global maize price. Importance levels are computed using MSE and measure the extent to which the model accuracy decreases with a random permutation of each input relative to GBM. The figure shows the marginal importance of each region in the second semester of the year, i.e July - December. Inputs are presented in two colours: production in red and yield in blue.

The results, which are somehow similar across the two models, propose very big differences between the marginal impact of each region. Whereas most regions hold an average monthly influence of about 3% to 4%, Northern America's weight in the market is over 20%, with both types of inputs (production or yield changes), and it is followed by Western Asia. We note that while the relative influence of Northern America is greater through yield changes, in the case of Western Asia the strongest impact is rather through shifts in its production. We also note the variability of this regional importance across months. As for Northern America, its greatest influence throughout all year long, with exception of July August, which are the last months before its harvest, and where it has a minimal marginal impact is not considerably different from those of most regions. Not surprisingly, these are the months in which the relative importance of Western Asia is at its highest with an average of 24% concerning production and 27% through yield.

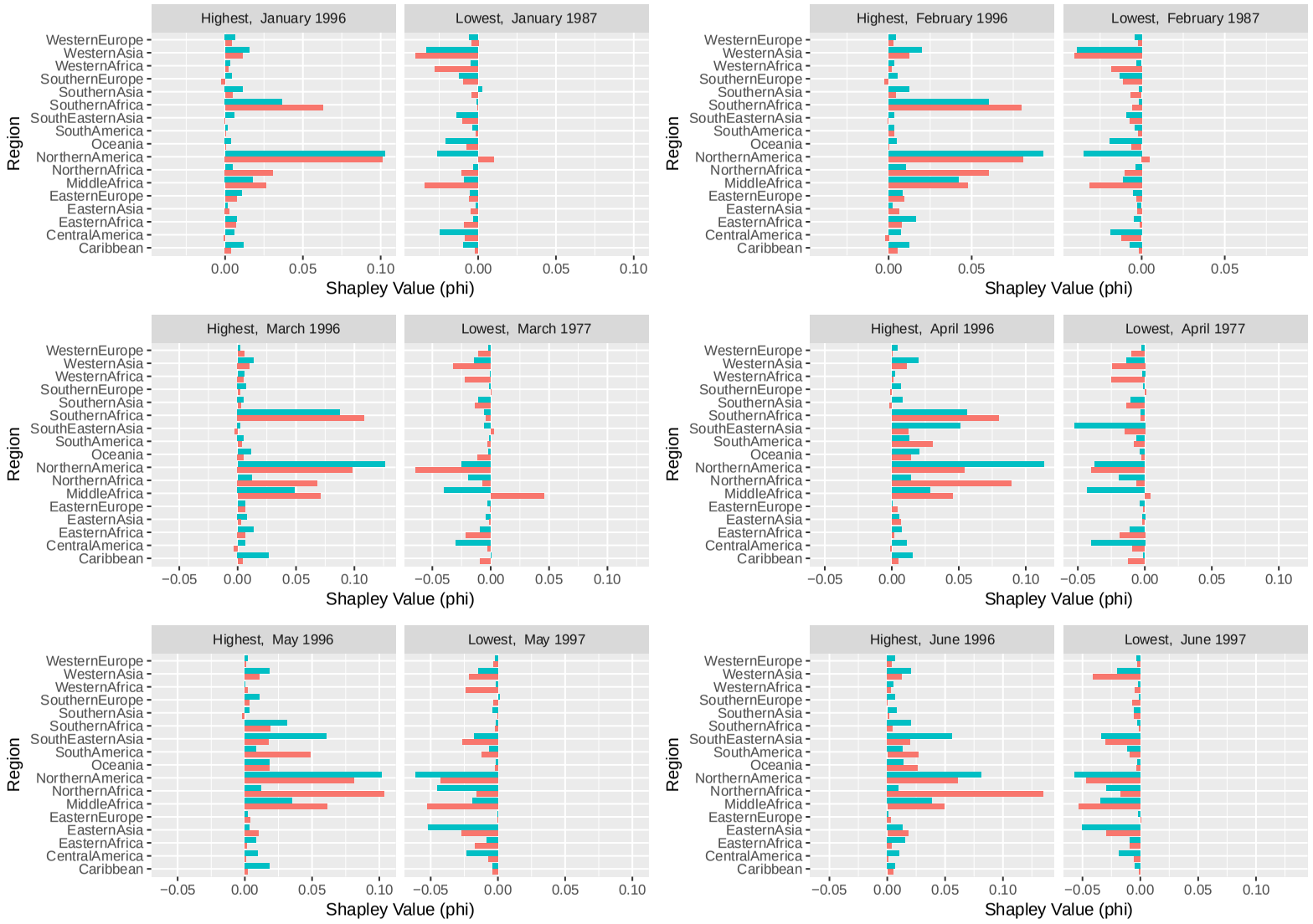


Figure 4.11 – Shapley values for the strongest price increase and strongest price decrease, Relative to January - June. The decomposition show the contributions of the producing regions to two extreme relative price changes (regional productions in red, and yields in blue). At a given date, the sum of the regional Shapley values is equal to the price change anomaly

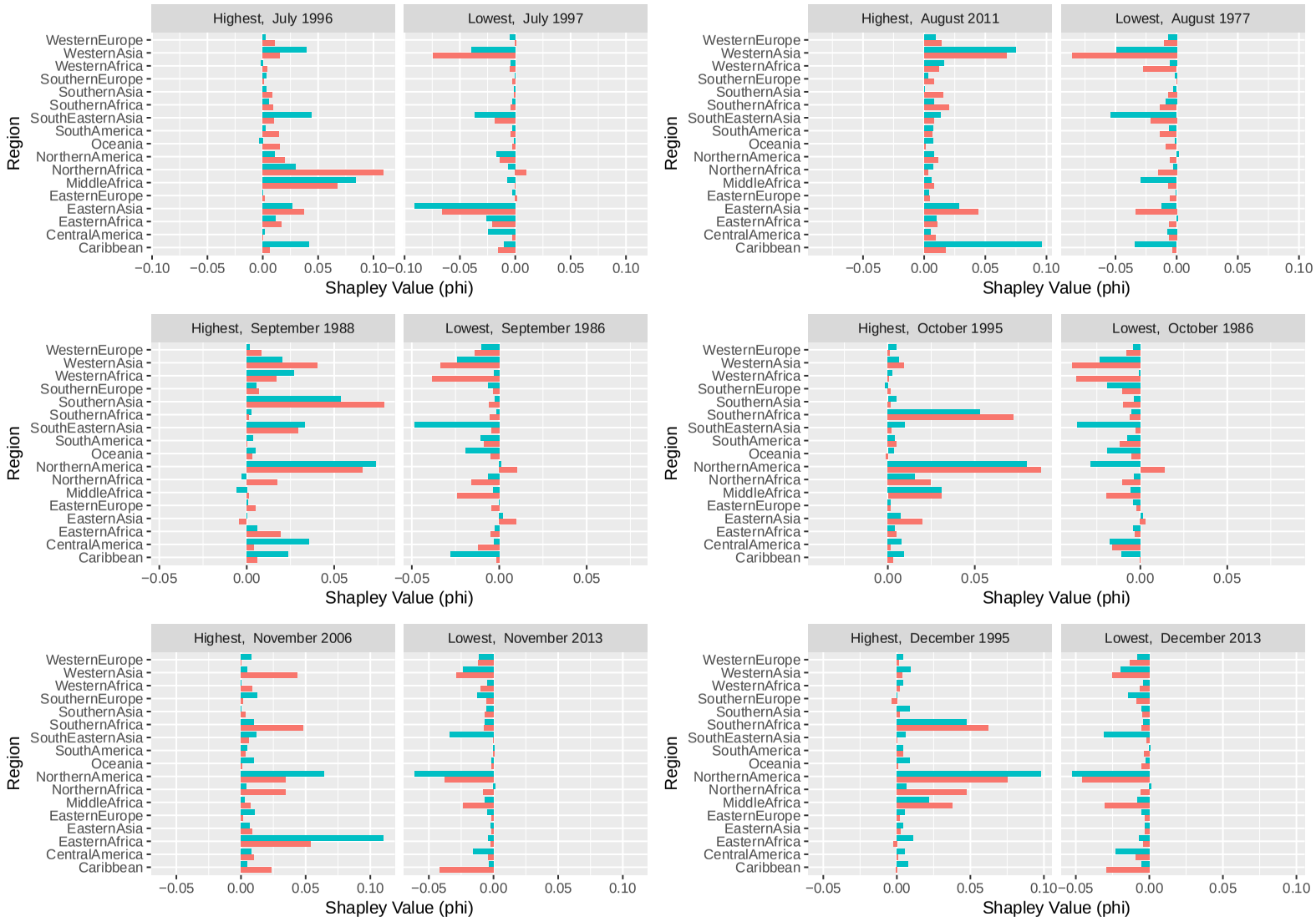


Figure 4.12 – Shapley values for the strongest price increase and strongest price decrease, Relative to July-December. The decomposition show the contributions of the producing regions to two extreme relative price changes (regional productions in red, and yields in blue). At a given date, the sum of the regional Shapley values is equal to the price change anomaly

Table 4.5 – Decomposition of Shapley values results:monthly actual prediction and monthly average prediction, relative to the highest/lowest extreme price changes, input = regional yield. For example, the forecasted -0.17 relative price change for January 1987 (Extreme low) is 0.18 lower than the average price forecast of 0.01. For this specific event, the sum of Shapley values yields the difference of actual and average prediction of -0.18.

Month	Direction	Actual Prediction	Average Prediction
1	Lowest, January 1987	-0.17	0.01
2	Lowest, February 1987	-0.16	0.01
3	Lowest, March 1977	-0.15	0.01
4	Lowest, April 1977	-0.18	0.01
5	Lowest, May 1997	-0.24	0.01
6	Lowest, June 1997	-0.27	0.02
7	Lowest, July 1997	-0.21	0.02
8	Lowest, August 1977	-0.24	0.01
9	Lowest, September 1986	-0.15	0.01
10	Lowest, October 1986	-0.17	0.01
11	Lowest, November 2013	-0.20	0.01
12	Lowest, December 2013	-0.20	0.01
1	Highest, January 1996	0.27	0.01
2	Highest, February 1996	0.33	0.01
3	Highest, March 1996	0.38	0.01
4	Highest, April 1996	0.38	0.01
5	Highest, May 1996	0.39	0.02
6	Highest, June 1996	0.38	0.02
7	Highest, July 1996	0.35	0.02
8	Highest, August 2011	0.26	0.01
9	Highest, September 1988	0.29	0.01
10	Highest, October 1995	0.27	0.01
11	Highest, November 2006	0.28	0.01
12	Highest, December 1995	0.25	0.01

Table 4.6 – Decomposition of Shapley values results:monthly actual prediction and monthly average prediction, relative to the highest/lowest extreme price changes, input = regional production. For example, the forecasted -0.17 relative price change for January 1987 (Extreme low) is 0.18 lower than the average price forecast of 0.01. For this specific event, the sum of Shapley values yields the difference of actual and average prediction of -0.18.

month	Direction	Actual Prediction	Average Prediction
1	Lowest, January 1987	-0.18	0.01
2	Lowest, February 1987	-0.18	0.01
3	Lowest, March 1977	-0.18	0.01
4	Lowest, April 1977	-0.22	0.01
5	Lowest, May 1997	-0.23	0.02
6	Lowest, June 1997	-0.24	0.02
7	Lowest, July 1997	-0.24	0.02
8	Lowest, August 1977	-0.21	0.02
9	Lowest, September 1986	-0.17	0.01
10	Lowest, October 1986	-0.19	0.01
11	Lowest, November 2013	-0.19	0.01
12	Lowest, December 2013	-0.18	0.01
1	Highest, January 1996	0.26	0.01
2	Highest, February 1996	0.31	0.01
3	Highest, March 1996	0.38	0.01
4	Highest, April 1996	0.37	0.01
5	Highest, May 1996	0.37	0.02
6	Highest, June 1996	0.34	0.02
7	Highest, July 1996	0.31	0.01
8	Highest, August 2011	0.27	0.01
9	Highest, September 1988	0.29	0.01
10	Highest, October 1995	0.26	0.01
11	Highest, November 2006	0.27	0.01
12	Highest, December 1995	0.26	0.01

Chapter 5

Data-driven assessment of the impacts of regional productions on the global prices of maize, soybean and cocoa

Co-author: David Makowski

Abstract

Prices of agricultural commodities (AC) have a crucial impact on food security worldwide. In order to anticipate their fluctuations, it is necessary to develop reliable predictive models. To facilitate their use by a large range of stakeholders - including those with few resources - it is necessary that these models are based on public data and on algorithms that are easy to implement. This study compares several simple econometric and Machine learning (ML) techniques to forecast price fluctuations of three globally traded AC - maize, soybean, and cocoa - with contrasted growing areas and market characteristics. For each AC and month, the most accurate model is selected using a cross validation procedure. Results reveal that the Gradient Boosting ML model is more accurate than other models in most cases. However, at a time horizon shorter than three months, the time series statistical method TBATS shows very good performance. We detect strong influence of Northern America over the global price of maize and soybean, except for the last months preceding the new-crop season. In the cocoa market, variations of production in Côte d'Ivoire, in Brazil and in Ghana have a substantial influence on cocoa prices. All the proposed models can be

easily trained from publicly available data of global monthly prices and local productions, at different geographical scales. Our approach is very accessible and requires few resources. It can therefore be implemented to anticipate and analyse agricultural price shocks by many stakeholders, including those with limited resources in developing countries.

Keywords: Food-security, Agricultural commodities, Price forecasting, Agricultural production, Machine learning

5.1 Introduction

Prices of agricultural commodities (AC) depend on many factors impacting the supply and demand sides of the food and feed balances. Therefore, international trade can serve as a tool for reducing price fluctuation. In principle, constant trade flows allow surpluses from high-productivity areas to be made available to those in short supply. However, AC prices sometimes suffer significant shocks in case of extreme events impacting crop productions, substantial shifts of food and feed demands or disruptions in storage and transportation chains. The extent of these shocks depends on the type of AC, time of occurrence, and the impacted areas (Abbott et al., 2009) (World-Bank, 2020a).

Historical evidence shows that local changes in production levels and export restrictions can sometimes have significant impacts on AC prices Headey and Fan (2008). For example, in 2008, rice prices increased by almost 300% in only four months due to export restrictions of major rice exporters with substantial social and economic impacts in importing countries (FAO, 2008). Furthermore, in 2020, COVID-19 caused some severe shortages in agricultural productions of several regions highly dependent on human-based labour, leading to an increase in prices of some AC and food products (Schmidhuber et al., 2020).

When not anticipated, global agricultural price shocks can have substantial impacts on food security (Laborde et al., 2021) because global price variations are often strongly correlated with local prices (Headey and Fan, 2010). In their study, Mundlak and Larson (1992) show that most of the changes in world prices are passed on to household (consumer) prices, with direct consequences on consumers. It is, therefore, crucial to be able to anticipate these shocks. In addition, price forecasting tools could be helpful to trigger mitigation strategies sufficiently in advance to reduce the risks for consumers and food security.

A diversity of methods have been proposed for AC price forecasting. However, accelerating technological advances, combined with improved access to local and global data, are opening up the possibility of using machine learning

(ML) techniques to forecast AC prices (Xiaojie and Yun, 2021; Ticlavilca and Feuz, 2010; Zhang et al., 2018). In particular, it is now becoming possible to train machine learning methods on open-access databases to predict the price of AC several times per year for various types of products. If adopted, this approach would be widely accessible, require few resources, and could be implemented by diverse actors in many world regions to anticipate agricultural price shocks. Moreover, while theoretical or econometric based approaches require making strong assumptions on the relationships between prices and influential factors (Storm et al., 2019), ML algorithms allow for the inclusion of a large number of input variables with minimal preliminary assumptions concerning their relationships with price variations.

However, the performance of machine learning methods for price forecasting can potentially depend on several factors, particularly on the nature of the input variables, the chosen algorithm, the type of commodity, and the forecasting time horizon. It is thus essential to assess and compare different types of ML techniques rigorously under contrasting conditions to determine their values. For this purpose, we choose three commodities with contrasted market structures - maize, soybean, cacao - to assess the performances of ML techniques under very different conditions. Maize belongs to the *grains* World Bank group and is the most produced crop worldwide. Maize is an essential energy source, especially in developing countries, where the total calories consumed from maize only (excluding its indirect contribution as farm animal food) is more than 10%.The USA is the world biggest maize producer, with about 30% of the global supply. In the world's southern hemisphere, Argentina's and Brazil's global market shares are lower. So is Eastern Asia's. However, they gradually increased during the past 60 years (from 9% to 23% and from 7% to 14%, respectively). Soybean is part of the *Oils & Meals* World Bank group and is the world primary protein source for livestock and play an essential role in the daily human diet (Thrane et al., 2017). The soybean market has experienced a substantial growth rate; its global production was multiplied by more than 13 in less than six decades. However, in terms of the market, the share of the USA decreased from 70% in 1961 to less than 30%. On the opposite, Argentina's and Brazil's shares have increased, and these two countries now grow more than half of the world production. Unlike maize and soybean, the cocoa market (included in the *beverages* World Bank group) has started being competitive only with the establishment of The International Cocoa Organization (ICCO) in 1973. Due to its significant vulnerability to external changes (mainly since it is primarily produced in small farms), the cocoa market is characterised by high and frequent price variations. All three commodities are of high global importance, and each is associated with a different group of AC, according to the official World Bank division

(World-Bank, 2021a).

This paper assesses the performance of six statistical and machine learning tools that can be easily implemented from open access data with freely available packages. We challenge these methods by considering the three crop species mentioned above and a wide range of forecasting time horizons (from 1 and up to 12 months), covering the needs of most of the decision-makers involved in food security management. We also show how these tools can identify the most influential producing regions and provide insights into significant factors influencing global trades.

The rest of the paper is structured as follows. The following section describes the data and explains their tailor to our specific research needs. Next, all models are presented, including the packages used for their implementation and the method used for assessing and comparing their performances. Finally, the results are presented and discussed.

5.1.1 Data

Our data set includes annual crop yields and crop productions, used as explanatory variables, and monthly prices, used as a dependent variable (full description of the data-sets is supplied in table 5.1 in Appendix 5.A). Annual crop yields and crop productions were extracted from the FAO STAT public database at two geographical scales, national and regional (the regions were those defined by FAO, and each region includes several countries). In addition, global monthly price data were extracted for maize, soybean, and cocoa from the World Bank's commodity market database between January 1960 and December 2020 (732 values). Finally, all three price time series were converted into real 2010 USD, using the Agricultural Price Index of the corresponding period, before being transformed into relative change from the corresponding month of the previous year (Fig 5.1).

Fig 5.1 (left) presents the global monthly prices of the three AC's over the past six decades, in real 2010 USD.

National and regional crop productions and yields were associated with some specific monthly prices, depending on local harvest seasons (FAS-USDA, 2021; ITC and UNCTAD/WTO., 2001). This procedure could predict monthly prices using production data available before the predicted months.

Fig 5.2 shows the production data for the leading producers of each of the considered commodities. The production shares of the different countries have changed substantially during the considered periods, particularly for cocoa, but for soybean and maize.

We now introduce the notations used to define our models. Let us define $q^{ca}_{m,y}$ as a series of real monthly global prices of a commodity ca, where m and y

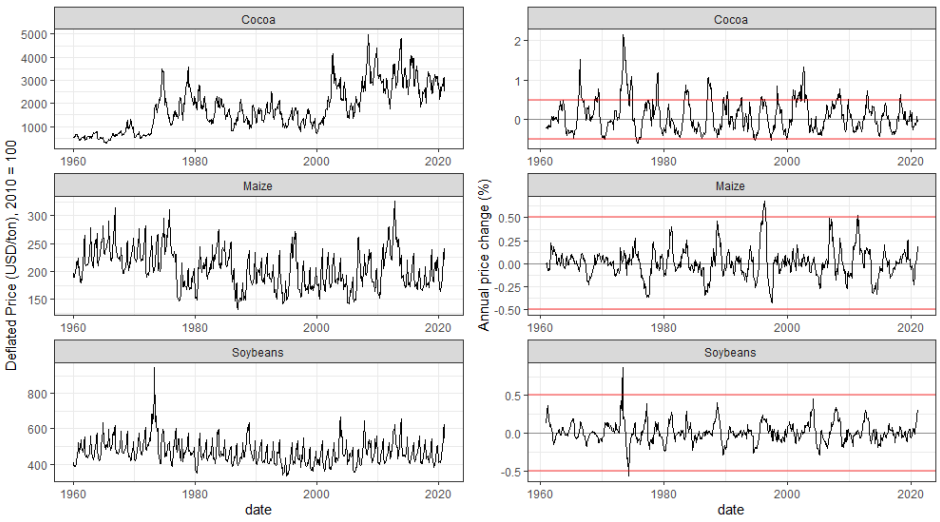


Figure 5.1 – Prices of cocoa, maize, and soybean from 1961 to 2020, in real USD (left) and relative price change compared to the same month of the previous year (ratio)

are the month and year indices, respectively, so that m=1,...,12 and y=1,...,Y, and ca is a crop index, ca=1, 2, 3, for maize, soybean and cocoa. A second time series $z^{ca}_{k,y}$ describes the production (or yield) in area k (k=1, ...,K) during year y for commodity ca. Since the two-time series $q^{ca}_{m,y}$ and $z^{ca}_{k,y}$ have very different units, we transform them in order to compute the relative annual changes of prices and productions (yields) as follows:

$$p^{ca}_{m,y} = \frac{q^{ca}_{m,y} - q^{ca}_{m,y-1}}{q^{ca}_{m,y-1}} \tag{5.1}$$

$$x^{ca}_{k,y} = \frac{z^{ca}_{k,y} - z^{ca}_{k,y-1}}{z^{ca}_{k,y-1}} \tag{5.2}$$

5.1.2 Models

The first model is a time-series based model, in which the price variation $p^{ca}_{m,y}$ is predicted solely based on a learning process of all the price changes that have occurred since the beginning of the observation period, without taking any external factor into account. With the other models, price variations are predicted either as a function of production (or yield) variations ($p^{ca}_{m,y} = f(x^{ca}_{k,y})$) or as a function of production variations and price variations observed the previous year ($p^{ca}_{m,y} = f(p^{ca}_{m,y-1}, x^{ca}_{k,y})$).

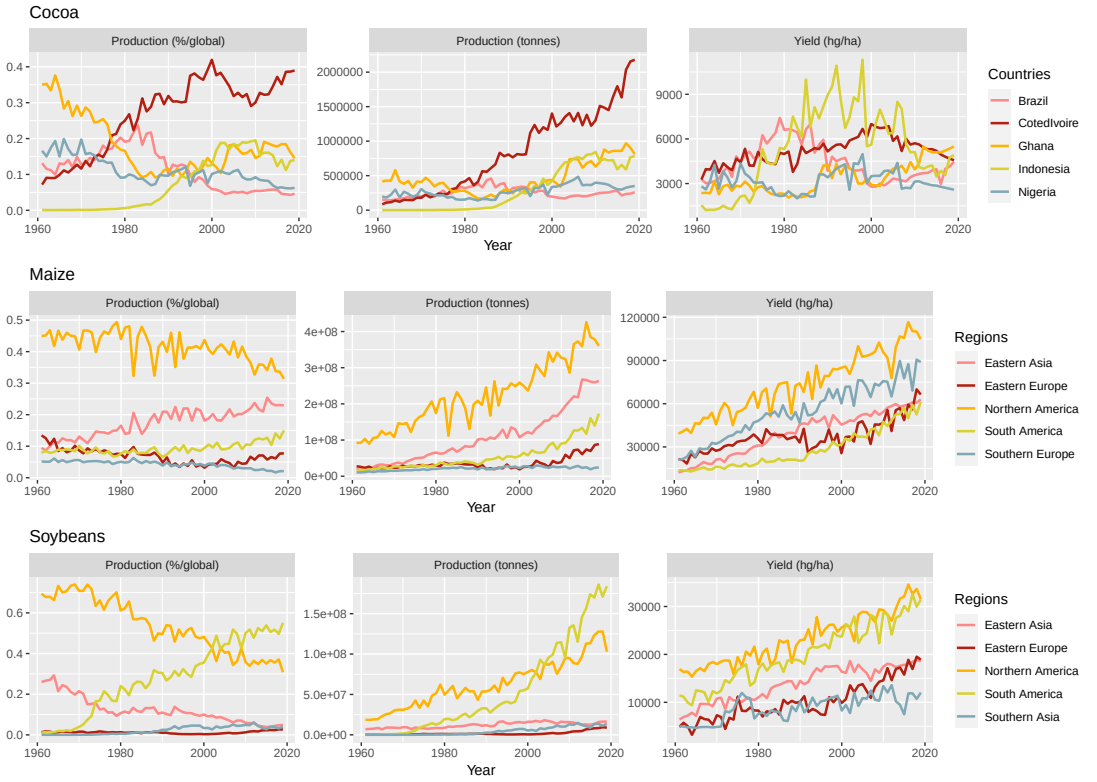


Figure 5.2 – Cocoa, maize, and soybean production data from 1961 to 2020: production as percentage of the global amount produced (left); production quantity, in tonnes (middle); and yield in hectograms per hectare (right).

Time series model

We use TBATS to forecast $p^{ca}_{m,y}$ using previously observed prices solely. Trigonometric Seasonal Box Transformation with ARMA residuals Trend and Seasonal Components (TBATS) (De Livera et al., 2011) is an innovative time series (TS) model. TBATS automatically handles non-linearity by using Box-Cox transformation of the data, recognising multiple seasonal components and determining possible time trends. For making the final forecast, TBATS chooses the optimal number of autoregressive and moving average components to minimise the forecasting error.

We apply TBATS using the R package `forecast` of Hyndman et al. (2020) considering forecasting time horizons from 1 to 12 months ahead after the last observed monthly price variation, for each month from January to December. Thus, TBATS is the only algorithm with no additional information apart from the historical prices. On the other hand, as TBATS ignores the effects of external factors,

it may lead to distorted results, especially when lags become larger (Gos et al., 2020) and cannot be used to analyse the effects of critical factors on price variation.

With that in mind, we integrate TBATS in our research as an additional tool for relatively short-term forecasts, assuming that it may sometimes lead to more accurate predictions than other models. To implement TBATS forecasts in R, we use the `forecast` package (Hyndman et al., 2020).

Linear models

Linear models describe the impact of annual regional outputs on the monthly global price through linear relationships. We define two versions. The first one relate price variations to production variations as follows:

$$p_{m,y}^{ca} = \alpha_{0,m}^{ca} + \sum_{k=1}^{K} \beta_{k,m}^{ca} x_{k,y}^{ca} + \epsilon_{m,y}^{ca} \tag{5.3}$$

The second model takes into account a possible dependence of monthly prices on the price the same month of the previous year:

$$p_{m,y}^{ca} = \alpha_{0,m}^{ca} + \alpha_{1,m}^{ca} p_{m,y-1}^{ca} + \sum_{k=1}^{K} \beta_{k,m}^{ca} x_{k,y}^{ca} + \epsilon_{m,y}^{ca} \tag{5.4}$$

In both equations $\alpha_{0,m}^{ca}$ is the intercept, $\beta_{k,m}^{ca}$ are regression parameters, $\epsilon_{m,y}^{ca}$ are the residuals, and K is the total number of producing areas included in the model (either countries or FAO regions, see Appendix tables 2-4). $\alpha_{1,m}^{ca}$, which appears exclusively in Eq 5.4, represents the marginal influence that $p_{m,y-1}^{ca}$ has over $p_{m,y}^{ca}$. The parameters were estimated using the software (R-Core-Team, 2020) base functions `lm` for each crop separately. Four different series of inputs were considered in turn: production changes per country, yield changes per country, production changes per region, yield changes per region. These four input sets led to four different models per crop (i.e., eight models when considering the two types of the linear model defined above). For each model, the number of inputs was reduced using a stepwise selection procedure based on AIC to select the most influential among the K regions or countries considered.

Machine-learning (ML) models

This study examines the forecasting accuracy of three ML models, namely CART, random forest, and gradient boosting.

Classification and regression trees (CART) was developed by Breiman et al. (1984) almost three decades ago, although it has only become popular in recent years. Its main advantage over traditional regression techniques is that CART does not require forcing any pre-assumptions onto the model while keeping a high level of interpretability. The principle of this method is to define a series of splitting rules based on the explanatory variables (here, the production or yield changes) able to minimise the prediction errors. The resulting set of splitting rules defines a tree that can be used to predict the target variable $p^{ca}_{m,y}$. Here, CART is fitted using the `rpart` package of the R software (Therneau et al., 2019).

CART is sometimes considered as a "weak learner" (Luo et al., 2019; Westreich et al., 2010). The resulting tree generated by CART is prone to instability (a different tree is often generated as soon as the data-set is slightly changed). Random forest (Hastie et al., 2009) was developed to reduce this instability by generating an ensemble of trees based on bootstrap samples drawn from the original data set. (Rokach, 2010). The ensemble approach assumes that if one model has not detected an important feature, it will be reflected in another model. Here, this method is implemented with the `randomForest` R package of Breiman et al. (2018). Another ensemble approach, which has proved very useful during the last few years, is Gradient Boosting (GBM). This method is implemented using the `gbm` R package (Friedman, 2001). GBM combines the results of several simple trees by selecting sub-samples according to the estimation errors of the previous trees. Both RF and GBM are implemented with 500 trees, as stable results were obtained with this number.

Vector autoregressive model (VAR)

Our last model is based on vector autoregressive model (VAR), fitted using the `vars` R package (Pfaff and Stigler, 2018) for each month and crop species, separately. VAR analyses several dynamic variables that simultaneously influence each other (Sims, 1980). In this study, VAR is used not only detects the effects of local productions shocks, $x^{ca}_{k,y}$, on global price variations but also the relationships between the different $x^{ca}_{k,y}$ across the regions considered. Moreover, VAR deals with the effect of the previous prices $p^{ca}_{m,y-1}$ on $p^{ca}_{m,y}$.

5.1.3 Model evaluation

We evaluate the accuracy of the price change predictions by implementing a Rolling Cross-Validation (RCV) with each modelling technique. We define twelve separate training sets (one per month m) each including the first $\tilde{Y}^{ca}$ = 44 years of observations. At each RCV iteration, we add one year of data i (the year im-

mediately following the last year of the training set), fit the regression model to the resulting expanded data-set, and use the trained model to forecast the price change of the next year $\tilde{Y}^{ca} + i + 1$. This procedure results in I predicted prices, where I=13 or 14, depending on the crop and month [1]. The choice of a minimal training set of 44 years enables a sufficiently high number of data to estimate the model parameters. The accuracy of the predictions is then evaluated by computing the root mean squared error (RMSE). We replicate the whole process for each regression model and each predicted month.

For TBATS, the calculation of RMSE is done in a slightly different manner because TBATS is used to predict the price at 12 different time horizons, h = 1,2,...12 months after the last observation. We train TBATS on a minimal set of $\tilde{Y}$ = 28 months. We then add each monthly data one by one, train TBATS each time, and predict the next 12 months of price changes using each trained TBATS model. A value of RMSE is finally computed for each time horizon h, each predicted month, and each crop species.

In order to measure the added values of the models compared to a constant prediction, the RMSE values are compared to the standard deviation of the observed values of $p^{ca}_{m,y}$ ($SD(p^{ca}_{m,y})$) for each crop species and each month. The standard deviation can be seen as an upper bound above which the model is useless compared to a constant prediction. To facilitate the comparison, we compute an index - Relative Advantage (RA) - defined as:

$$RA = 1 - \frac{RMSE^{ca}_{m}}{SD(p^{ca}_{m,y})} \tag{5.5}$$

A specific value of RA is computed for each model, month, and crop species. The whole set of values are presented in Appendix 5.A, Table 5.5, 5.6 and 5.7, for maize, soybean and cocoa, accordingly.

5.1.4 Ranking of the producing regions

In most studies, the relative importance of each producing region is usually assessed by calculating the increase in mean prediction errors (measured by the mean squared error (MSE)), resulting in a random choice of the value of each input $x^{ca}_{k,y}$ (Friedman, 2001; Jeung et al., 2019). A high increase of MSE reveals that the input is influential, whereas a low or zero MSE increase reveals that the corresponding input is non-influential. Although practical and very popular, this approach is not able to assess the direction of the effect of each feature on $\tilde{p}^{ca}_{m,y}$.

[1] Cocoa: I = 14 for all months; Maize: I = 13 for the months April to September, and 14 for the other six months; Soybean: I = 13 for the first nine months of the year, and 14 for the final three months.

This importance ranking technique cannot be used to determine whether each $x_{k,y}^{ca}$ has a positive or negative effect on prices. To do so, we implement here a more recent approach based on the method Shapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017). Shapley values measure the contributions of the model inputs to the model predictions. More specifically, considering a given prediction and the set of inputs used to produce this prediction, the Shapley values measure the contribution of the input values (here, the values of $x_{k,y}^{ca}$) on the deviation of the considered predicted value to the mean predicted value. The set of Shapley values computed for all predictions provides a global picture of the contributions of the different inputs to the predicted values. Here, Shapley values are significant because they describe the contribution of the regional productions to predicted price variations, in particular to the most extreme predicted price variations corresponding to major price shocks. Shapley values provide information on both the directions and magnitude of the effects of the inputs. When plotted as a function of the values of $x_{k,y}^{ca}$, they provide a visualisation tool to assess the risk of price shocks as a function of the levels of variation in regional productions. Finally, we rank the producing regions according to their influence on predicted price variations by taking the average of the absolute Shapley values by region.

5.2 Results

5.2.1 Maize prices

RA values are shown in Fig 5.3 for each model type and each predicted month, separately. Apart from TBATS, all the models' performances are reported considering different inputs, namely regional productions, regional yields, and regional productions/yields plus previous prices. For TBATS, RA values are shown for a forecasting time horizon of one to 12 months ahead.

Short-term predictions obtained with TBATS (one month ahead) tend to be more accurate than those obtained with machine learning tools and VAR. The RA reaches 37% with the best machine learning tool, while the best RA of TBATS is higher and stands at 78% in March when considering one-month ahead forecasts. For this type of forecast, the RA of TBATS is usually higher than 50% for most of the predicted months, revealing that this method can reduce the prediction errors by at least 50% compared to a constant prediction.

However, these good performances are not maintained when attempting more extended time-horizon forecasts with TBATS. RA levels decrease rapidly for any increment of time horizon. More precisely, while the average RA value for a month-ahead forecast is 64%, it drops to 45% for two-month ahead fore-

casts and decreases further for predictions at longer time horizons, becoming quickly close to zero or even negative (Fig 5.3).

Concerning machine learning tools, RA tends to be slightly higher when yield changes are used as predictors, and the GBM technique appears to perform better, especially compared to LM and VAR, in most cases. In addition, GBM tends to perform better than TBATS for time-horizon larger than six months.

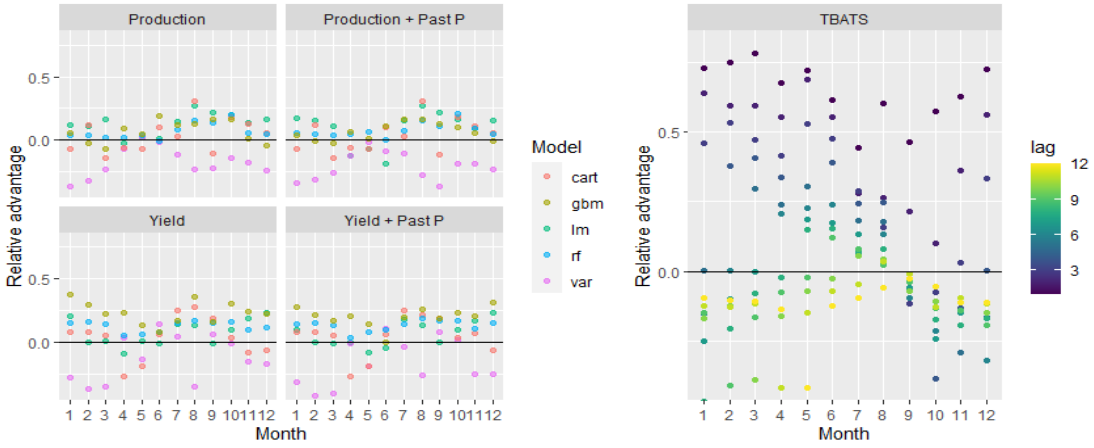


Figure 5.3 – Relative advantage (RA) in prediction accuracy of the forecasting models compared to constant prediction, over 1990-2020, for maize. RA is equal to 1 minus the ratio of the RMSE of each model to the standard deviation of the price changes in the whole data-set the same month, and expressed in percentages. It indicates the relative benefit of using the models compared to a constant prediction. ML methods, LM and VAR were used with either production or yield inputs, and with/without taking the price changes of the previous year (PastP) into account.

5.2.2 Soybean prices

In a similar way to maize, the best model to predict soybean price variation is TBATS for one, two or three-month ahead forecasts (Fig 5.4). However, for longer-time lags, the accuracy of TBATS declines as lag values become larger. Moreover, the relative advantage of TBATS varies between months. For example, whereas the RA exceeds 70% for one-month ahead forecasts in January-March, it only reaches 29% in May and 36% July.

Relative to the other techniques, the best results are obtained during the first few months (January-March) and between August and November. During these periods, RA can reach levels close to 50%. Predictions obtained in April-July tend to be less accurate, with RA often close to zero. The nature of the inputs does not

strongly influence the accuracy of the forecasting methods. Similar RA values are obtained with production or yield-related predictors, or whether previous prices are taken into account or not. GBM tends to be the most accurate among the tested algorithms and is ranked first in five months of the year. RF is ranked first in four months of the year, particularly with production inputs. LM provides relatively good results for January and September but is less accurate in other months. Compared to maize, ML models tend to perform better at forecasting soybean prices, while TBATS is slightly less accurate.

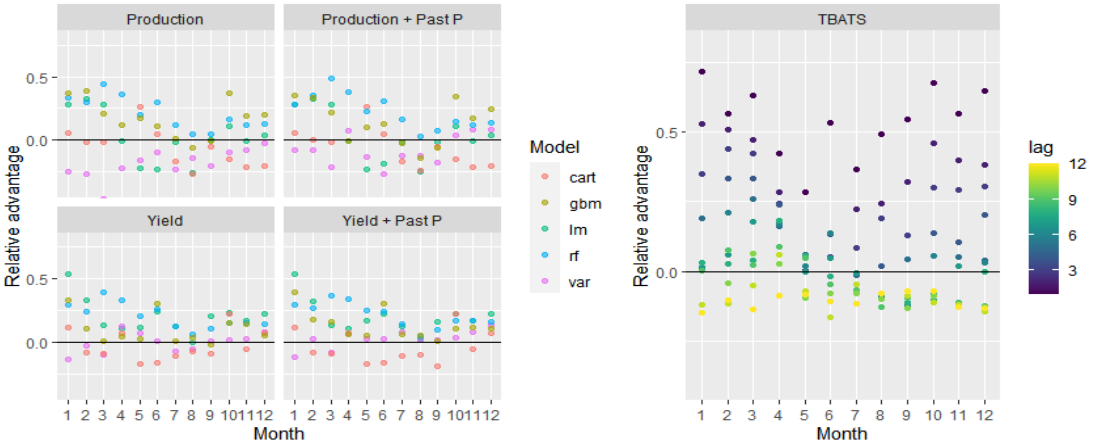


Figure 5.4 – Relative advantage (RA) in prediction accuracy of the forecasting models compared to constant prediction, over 1990-2020, for soybean. RA is equal to 1 minus the ratio of the RMSE of each model to the standard deviation of the price changes in the whole data-set the same month, and expressed in percentages. It indicates the relative benefit of using the models compared to a constant prediction. ML methods, LM and VAR were used with either production or yield inputs, and with/without taking the price changes of the previous year (PastP) into account.

5.2.3 Cocoa prices

For cocoa, TBATS gives relatively good results for short time horizons, with RA values exceeding 50% for most of the year. However, we note that the forecast accuracy decrease for a more extended time horizon is moderate, compared to maize and soybean, especially in August and September. Other predictive methods tend to perform poorly in most cases, with a few exceptions obtained with the GBM method in March (RA = 31%), April (RA = 43%) and May (RA = 48%) using national yield changes as predictors. The high RA values obtained between March and May may be since March is the last month of the main-crop har-

vest season in several major producing countries - Brazil, Côte d'Ivoire, Ghana and even Cameroon (ITC and UNCTAD/WTO., 2001). During the rest of the year, production data do not substantially influence cocoa's price, according to our results.

Figure 5.5 – Relative advantage (RA) in prediction accuracy of the forecasting models compared to constant prediction, over 1990-2020, for cocoa. RA is equal to 1 minus the ratio of the RMSE of each model to the standard deviation of the price changes in the whole dataset the same month, and expressed in percentages. It indicates the relative benefit of using the models compared to a constant prediction. ML methods, LM and VAR were used with either production or yield inputs, and with/without taking the price changes of the previous year (PastP) into account.

5.2.4 Most influential producing regions

We analyse the contributions of each producing region to price changes by computing SHapley Additive exPlanations (SHAP) with the most accurate ML techniques identified above (i.e., GBM). This approach allows us to express the difference between a specific price change prediction to the mean prediction as a sum of contributions (Shapley values) of the input values (yield or production changes). Producing regions of high relative importance are expected to have high absolute Shapley values, while those of lower influence are most likely to have Shapley values close to zero. Results are shown in Fig 5.6 to Fig 5.11.

For maize, Shapley values are reported for the forecasting technique GBM with regional yield changes and for January, which provides the most accurate results (Fig 5.6).

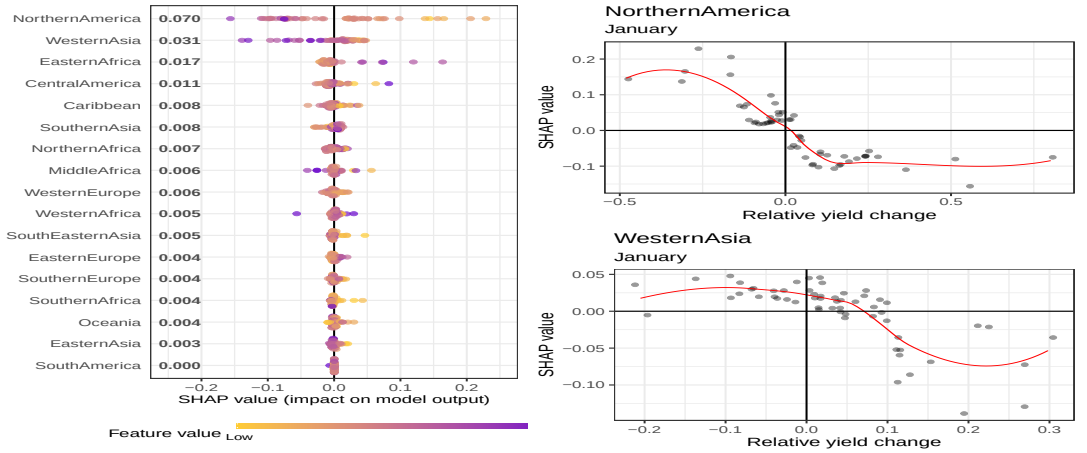


Figure 5.6 – Maize Shapley values for each regions (left) and for two most influential regions as a function of relative yield changes (right) computed with the forecasting model (GBM) for January. **Left**: Regions are ranked from the most influential (top) to the least (bottom). Each point correspond to one forecast of price change, where deep purple represents a high value of the considered feature (yield change input, here) and orange a low value. The points are located along the X-axis according to the level of the impact of regional yield change on price changes (the Shapley value), in a way that extreme negative impact on price change is at the most right, and vice versa. All points are centred around the black vertical line, which presents no impact on model predictions compared to mean prediction. The bold number to the right of the Y-axis is the mean absolute value of all the Shapley values by region, summarising the average impacts of the regions. **Right**: SHAP dependence plots, for the most influential region (on top) and the second-most influential. Here, the Shapley values of the two most influential producing regions are presented as a function of the relative yield changes of these regions, together with a smooth regression curve.

Shapley values confirm the strong impact of Northern American maize yield on the global maize price. For Northern America, Shapley values tend to take highly positive values when yield changes are low and negative (i.e., yield decrease compared to the previous year). At the same time, they are more likely to take negative values for positive yield changes (i.e., yield increase compared to the previous year) (Fig 5.6). This result indicates that a yield decrease (increase) in Northern America tends to be associated with a predicted global price increase (decrease). Thus, the importance of Northern America is considerably higher than that of all the other regions. This conclusion is relevant not only for January but for all other months as well, except for July and August (see Fig 5.12 in Appendix 5.B).

Another exciting finding stems from the second most important region, Western Asia. According to Shapley values, the impact of this region is minor compared to Northern America. Although the Shapley values of Western Asia show a similar declining trend with yield variations as in Northern America, the range of Shapley values is narrower in Western Asia. The average absolute value of the Shapley's is more than two times smaller in this region (0.031 versus 0.07) (Fig 5.6).

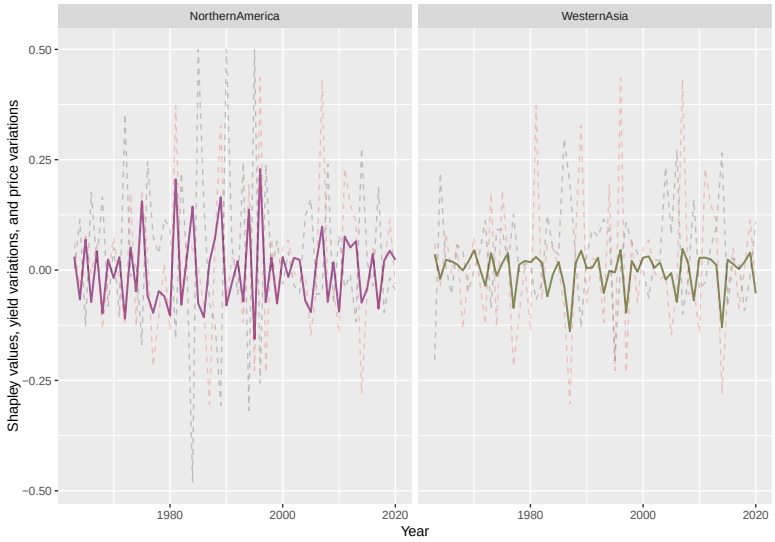


Figure 5.7 – Time series of Shapley values for the two most influential maize producing regions. The bold lines indicate Shapley values. The dashed red lines indicate price variations (in January). The dashed blue lines indicate yield variations. For facilitating the visualisation, variations higher than 0.5 in absolute values are rounded down to 0.5

Fig 5.7 presents the Shapley values as a function of years, where they are compared to the relative yield changes (in blue) and the predicted price variations in January (in bright red). Again, the high impact of Northern America on global price prediction is evident, as peaks of Shapley values appear throughout the period, in particular during years when Northern American productions (or yield) reached very high or low levels. The variability of the Shapley values of the second most influential region (Western Asia) is much lower. Interestingly, the Shapley values obtained for Western Asia remain close to zero most of the time but drop to relatively low levels during years characterised by strongly positive yield variations.

Results obtained for soybean are similar to those obtained for maize. However, we note weaker domination by lead producers (compared to the maize

Shapley values, the mean are smaller and their ranges of variations narrower), suggesting a more competitive global market for soybean than maize. Similar to maize, we found a declining trend between Shapley values and regional production variations when considering the two most influential regions (Fig 5.8 and Fig 5.9).

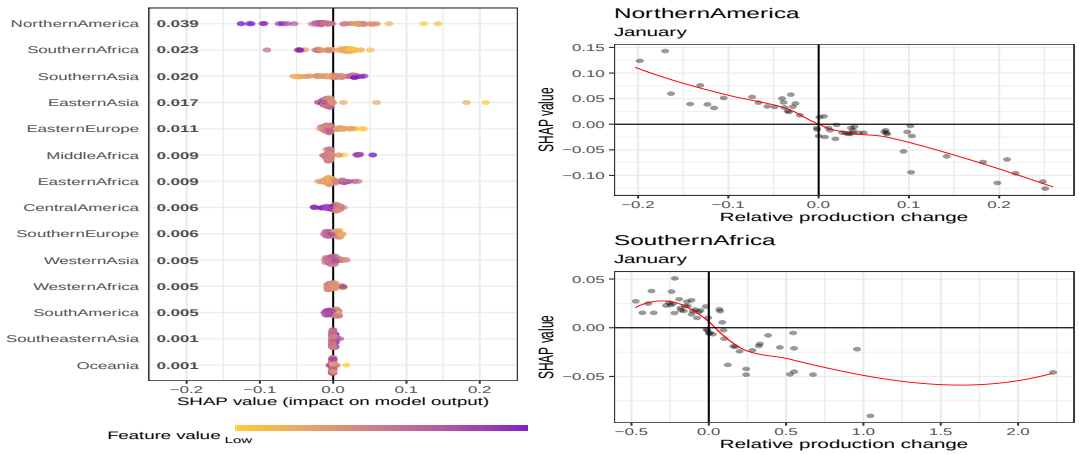


Figure 5.8 – Soybean Shapley values for each regions (left) and for two most influential regions as a function of relative production changes (right) computed with the forecasting model (GBM) for the month of January. **Left**: Regions are ranked from the most influential (top) to the least (bottom). Each point correspond to one forecast of price change, where deep purple represents a high value of the considered feature (production change input, here) and orange a low value. The points are located along the X-axis according to the level of the impact of regional production change on price changes (the Shapley value), in a way that extreme negative impact on price change is at the most right, and vice versa. All points are centred around the black vertical line, which presents no impact on model predictions compared to mean prediction. The bold number to the right of the Y-axis is the mean absolute value of all the Shapley values, summarising the average impact of the corresponding region. **Right**: SHAP dependence plots, for the two most influential regions. Here, the Shapley values of the two most influential producing regions are presented as a function of the relative production changes of these regions, together with a smooth regression curve.

The results obtained for cocoa show a different pattern. The cocoa production is highly geographically concentrated in only three regions (see Table 5.4 in Appendix 5.A). Interestingly, this is the sole of the three commodities considered for which prices are predicted more accurately with national than regional productions. Although Côte d'Ivoire - the first world producer - is ranked first ac-

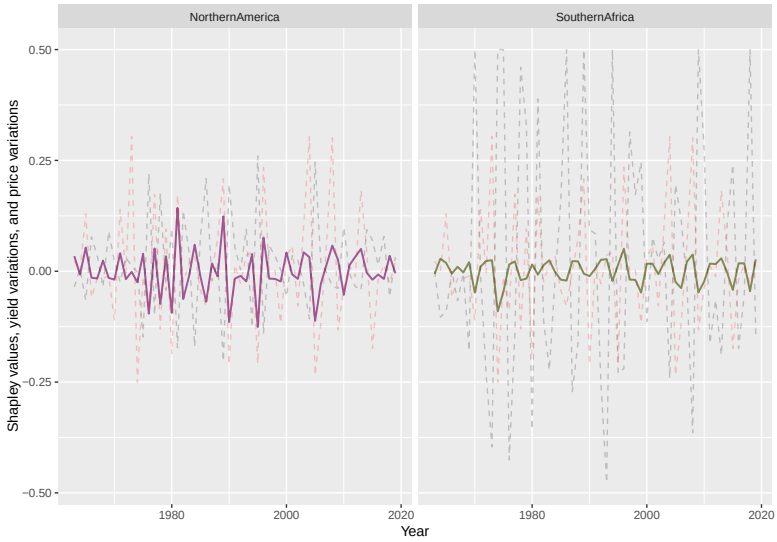


Figure 5.9 – Time series of Shapley values for the two most influential soybean producing regions. Shapley values are indicated by the bold lines. Price variations (in January) are indicated by the dashed red lines. Production variations are indicated by the dashed blue lines. For facilitating the visualisation, variations higher than 0.5 in absolute values are rounded down to 0.5.

cording to mean Shapley values, Brazil and Ghana also show a strong influence (Fig 5.10). Nevertheless, the mean absolute Shapley value difference between the first two regions is relatively small compared to maize and soybean.

For Côte d'Ivoire, the negative relationship between Shapley values and yield changes is not very strong (Fig 5.10). However, Brazil's decreasing trend is apparent - the second most influential region.

Shapley time series confirm the similar influence of Côte d'Ivoire and Brazil on cocoa prices. However, Shapley values of Brazil tend to be opposite of yield variations in this region (Fig 5.11).

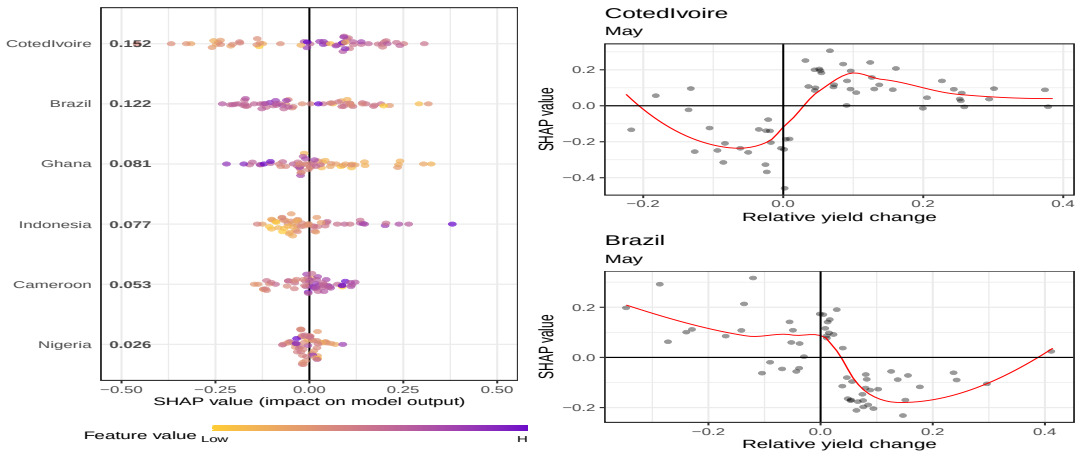


Figure 5.10 – Cocoa Shapley values for each regions (left) and for two most influential regions as a function of relative yield changes (right) computed with the forecasting model (GBM) for the month of May. **Left**: Regions are ranked from the most influential (top) to the least (bottom). Each point correspond to one forecast of price change, where deep purple represents a high value of the considered feature (yield change input, here) and orange a low value. The points are located along the X-axis according to the level of the impact of regional production change on price changes (the Shapley value), in a way that extreme negative impact on price change is at the most right, and vice versa. All points are centred around the black vertical line, which presents no impact on model predictions compared to mean prediction. The bold number to the right of the Y-axis is the mean absolute value of all the Shapley values, summarising the average impact of the corresponding region. **Right**: SHAP dependence plots, for the most influential region (on top) and the second-most influential. Here, the Shapley values of the two most influential producing regions are presented as a function of the relative yield changes of these regions, together with a smooth regression curve.

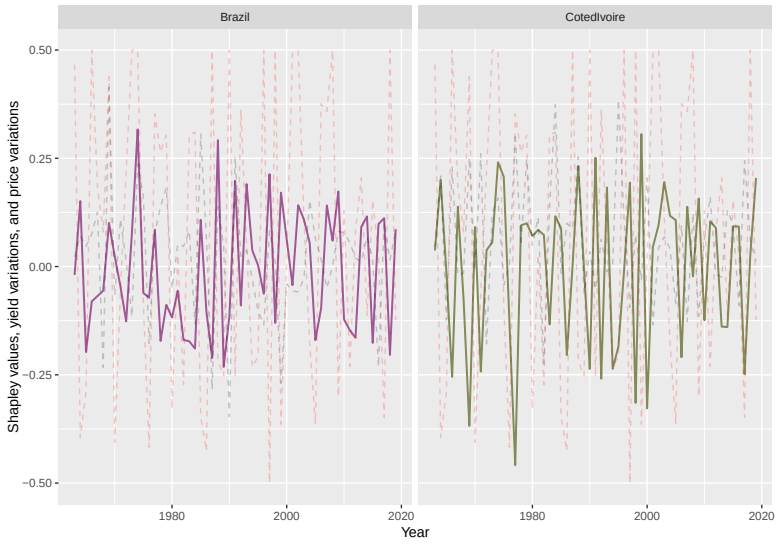


Figure 5.11 – Time series of Shapley values for the two most influential cocoa producing countries. Shapley values are indicated by the bold lines. Price variations (in May) are indicated by the dashed red lines. Yield variations are indicated by the dashed blue lines. For facilitating the visualisation, variations higher than 0.5 in absolute values are rounded down to 0.5

5.3 Discussion

This study examines the effectiveness of several machine learning and econometric methods for forecasting the international price of three globally traded crops: maize, soybean and cocoa. The selected methods rely on open-source data and software. Through a comparative analysis, it explores the genericity of the proposed approach and captures the uniqueness of AC from three different categories, as determined by the World Bank: *grains*, for maize; *oils & meals*, for soybean; and *beverages*, for cocoa. The robustness of the model performances is assessed by conducting an in-depth sensitivity analysis across three geographic scales (regional, continental and national), two types of production metrics (production or yield variations) and the inclusion or not of the relative annual change of last year's price. All in all, each forecasted monthly price is the result of the best performing model, selected out of 60 (5 algorithms $\times$ 3 geographic scales $\times$ 4 versions, excluding TBATS). The analysis of three AC market categories and the comparison of three geographic scales reveal the significance of the economic structure of the market, in particular the utmost importance that market structures have on how crops production influences AC prices globally. The study shows that regional changes in maize production in Northern America (the crop's lead world producer and exporter) have undeniably high impacts on its global price. The market is thus clearly dominated by the world's leading producer of maize, but this is more minor the case for the other crops. The other extreme is the cocoa market, which is concentrated for the most part in two regions (Western Africa and South America), and whose production is typically the work of smallholder farmers in family farms in relatively poor areas. Contrary to maize and soybean, which are traded mainly in the international market located and managed in the country of the biggest producer,[2] cocoa is mainly traded in the importer side, New York and London, i.e., far from its country of origin. This market structuring contributes to a lack of market information among cocoa farmers and prevents them from controlling the price they will receive for their crop or the preferred date to sell it.

Beyond evaluating the accuracy of the predictions, many techniques have been applied to interpret the results of the trained models in this study: relative importance analysis (Greenwell et al., 2020), Shapley values (Molnar et al., 2018; Tianqi et al., 2021), (Greenwell, 2017; Liu and Just, 2020) and standard correlation analysis. For cocoa, none of these methods indicated a strong relationship between the production volume of the leading producer (Côte d'Ivoire) and price

[2]The mainstay of maize and soybean's trade is at the CME in US dollar. However, for soybean, since December 2007, the price listed on the World Bank website comes from the Construction Industry Federation (CIF) Rotterdam

changes. However, the results did show a lack of absolute market power concentrated in a particular area and a relatively uniform distribution of monthly impact per country over the year. Moreover, a comprehensive examination pointed out a rather complex relationship between Shapley values and crop yield variations in Côte d'Ivoire. Finally, focusing on specific extreme price shocks indicated a high contribution of cocoa yield in Indonesia to events of exceptionally high price increases.

The results did show a lack of absolute market power concentrated in a particular area and reasonably uniform distribution of monthly impact per country over the year. Moreover, our comprehensive analysis pointed out a rather complex relationship between Shapley values and crop yield variations in Côte d'Ivoire. Focusing on specific extreme price shocks, we showed a high contribution of cocoa yield in Indonesia to events of exceptionally high price increases. The results seemed surprising at first, as Indonesia's market share in the global cocoa market is significantly lower than that of Côte d'Ivoire. However, an in-depth study of the cocoa literature has revealed a multiplex system in which some factors undermine the natural equilibrium of the market. ForCôte d'Ivoire (as well as Cameroon and Nigeria), the vast majority of production is collected from small farmers, most of whom have no access to market information. Alternatively, they received a price set by the local government at the beginning of the season, depending on future prices on significant stock exchanges (ITC and UNCTAD/WTO., 2001). Over the years, the cocoa export market in Côte d'Ivoire has been privatised, and private export companies are now responsible for collecting cocoa production, thus reducing farmers' room for action during and after harvest (Abbott et al., 2019). In Côte d'Ivoire, farmers take critical decisions right at the beginning of the growing season, trying to increase their production when the price they receive from their government increases and vice versa.

The soybean market has grown significantly over the past six decades and is today the most important legume in the world for producing oils and proteins. Like maize, here too, the influence of Northern America is evident above that of the other provinces. However, Northern America's market share has shrunk significantly over the years, and now South America (mainly Argentina and Brazil) is a leader in global soybean production. Nevertheless, Northern America is still the first to impact price changes in the global soybean market for most months of the year, other than those prior to the harvest season (similar to the maize market).

In terms of forecasting results, ML methods (RF and GBM) usually perform better than the other models for medium to long time horizons, as shown in Appendix 5.B. For a short-time horizon (one to three months), TBATS was generally more accurate, revealing that information about production changes is

probably already partly integrated into the market prices one to three months in advances. GBM provides a noticeable higher forecasting accuracy for Northern America's new-crop (the beginning of the standardised trading year) months concerning maize and soybean. For cocoa, March, April and May, are the only months in which ML models are substantially better than the other models, including TBATS. As a result, one might conclude that the benefit of taking regional production into account tends to be more decisive for competitive markets with fewer price distortions.

5.4 Conclusion

As the proposed forecasting tools rely on public data and open-source software, they can be easily implemented by many stakeholders, even with limited resources. Furthermore, we demonstrated that our framework is relevant for different crop types, namely maize, soybean and cocoa. Therefore, we believe that, in the future, it could cover other agricultural commodities.

Our analysis shows that, for short-term predictions, time series forecasting techniques such as TBATS provided accurate predictions of price variations. However, for longer forecasting horizons, the accuracy level of this technique declines rapidly, and it becomes more relevant to use alternative methods based on regression models and machine learning tools, including production/yield variations as predictors. We found that machine learning techniques based on ensembles of trees, such as random forest and gradient boosting, were potent. An added value of these machine learning models stays in their ability to rank producing units according to their influence on price changes and to quantify the contributions of major producing regions on the occurrence of significant price shocks. Moreover, they can help analyse the relationships between price and production changes in major producing regions.

Thanks to its transparency and ease of application, the proposed framework can help improve the analysis of price variations, especially in developing countries with limited resources for price modelling projects.

5.A Appendix

Table 5.1 – Variable description and data sources.

Final data				
Data Production	Unites % change/year	Time-range 1962 - 2019	Indices k = Region, y = Year	Sign $x_{k,y}$
Yield	% change/year	1962 - 2019	k = Region, y = Year	$x_{k,y}$
Price	% change/year	01/1961 - 11/2020	m = Month, y = Year	$p_{m,y}$
Initial information				
Data	Unites	Time-range	Source	Sign
Price	Nominal USD/mt*	01/1960 - 11/2020	World Bank, Pink Sheet (2020)	
Ag. Price index	USD (2010 = 100)	01/1960 - 11/2020	World Bank, Pink Sheet (2020)	
Production	tonnes/year	1961 - 2019	FAO STAT (2020)	$z_{k,y}$
Yield	hg/ha	1961 - 2019	FAO STAT (2020)	$z_{k,y}$
Real price	Real USD (2010)	01/1960 - 11/2020		$q_{m,y}$
Additional information				
Data	Unites	Time-range	Source	Sign
Production	1000 mt/year	1960 - 2020	PSD, USDA	$z_{k,y}$
Production	% change/year	1961 - 2020	k = Region, y = Year	$x_{k,y}^{USDA}$

* For cocoa, prices are given by units of kg, and were manually converted to metric tonnes units

5.A.1 Maize

Table 5.2 – Production data of maize, relative to region

	Production (1000 tonnes)		Yield (1000 hg/ha)		
Regions	%/total	Average	Average	Min. (year)	Max. (year)
Caribbean	0.10%	462	11.16	8.54 (1993)	14.78 (2004)
Central America	3.30%	18,050	20.4	9.74 (1961)	36.31 (2019)
Central Asia*	0.10%	1,288	48.34	25.63 (1997)	65.72 (2018)
Eastern Africa	2.60%	14,673	13.64	9.52 (1965)	19.49 (2019)
Eastern Asia	19.20%	106,304	39.41	12.28 (1961)	62.94 (2019)
Eastern Europe	6.30%	34,856	37.4	18.4 (1963)	70.02 (2018)
Middle Africa	0.50%	2,652	8.54	6.74 (1979)	10.9 (2015)
Northern Africa	0.90%	5,017	41.68	15.91 (1961)	69.32 (2019)
Northern America	40.30%	223,304	73.87	39.23 (1961)	118.01 (2017)
Northern Europe*	0.00%	41	35.43	10 (1985)	75.84 (2019)
Oceania	0.10%	426	50.76	17.33 (1966)	87.81 (2015)
South America	9.90%	55,095	28.29	12.95 (1964)	61.38 (2019)
South-eastern Asia	3.60%	19,741	22.26	9.02 (1961)	46.9 (2019)
Southern Africa	1.70%	9,209	24.26	7.88 (1992)	58.13 (2017)
Southern Asia	3.00%	16,495	17.59	10.02 (1971)	35.28 (2019)
Southern Europe	3.80%	20,902	54.9	21.13 (1961)	90.58 (2018)
Western Africa	1.50%	8,252	12.11	6.96 (1972)	19.54 (2018)
Western Asia	0.60%	3,142	35.6	11.4 (1962)	79.18 (2018)
Western Europe	2.80%	15,336	69.82	22.56 (1962)	103.05 (2011)

*Excluded from analysis due to lack of data

Table 5.3 – Production data of maize, relative to country

	Production (1000 tonnes)		Yield (1000 hg/ha)		
Countries	%/total	Average	Average	Min. (year)	Max. (year)
Argentina	2.70%	14,813	43.13	16.48 (1963)	78.62 (2019)
Brazil	6.10%	33,777	26	11.61 (1964)	57.73 (2019)
China	18.80%	104,055	39.39	11.85 (1961)	63.17 (2019)
India	2.10%	11,472	16.17	9 (1971)	30.7 (2019)
Mexico	2.80%	15,409	21.53	9.87 (1963)	40.7 (2019)
USA	38.90%	215,455	73.83	39.18 (1961)	117.43 (2016)

Table 5.4 – Production data of maize, relative to continent

	Production (1000 tonnes)		Yield (1000 hg/ha)		
Continents	%/total	Average	Average	Min. (year)	Max. (year)
Africa	7.23%	2,361,926	15.47	9.8 (1964)	21.66 (2017)
Americas	53.45%	17,463,588	50.48	25.83 (1964)	82.19 (2016)
Asia	26.41%	8,628,804	31.39	11.37 (1961)	55.41 (2019)
Europe	12.84%	4,195,971	46.82	20.87 (1963)	75.28 (2018)
Oceania	0.08%	25,118	50.8	17.34 (1966)	87.98 (2015)

5.A.2 Soybean

Table 5.5 – Production data of soybean, relative to region

	Production (1000 tonnes)		Yield (1000 hg/ha)		
Regions	%/total	Average	Average	Min. (year)	Max. (year)
Central America	0.27%	372	18.2	12.85 (2010)	21.34 (1991)
Central Asia*	0.03%	96	12.46	2.5 (1972)	22.93 (2017)
Eastern Africa	0.16%	229	15.18	6.47 (1996)	21.68 (2016)
Eastern Asia	8.69%	12,087	13.19	5.54 (1961)	22.38 (1978)
Eastern Europe	1.33%	1,852	24.19	6.76 (1968)	36.85 (2014)
Middle Africa	0.02%	21	13.82	6.42 (1961)	18.82 (2018)
Northern Africa	0.04%	61	15.41	3.86 (1966)	23.29 (2008)
Northern America	45.11%	62,764	10.68	6.37 (1968)	14.85 (2015)
Northern Europe*	0.00%	1	5.75	2.29 (1983)	15.63 (2010)
Oceania	0.04%	51	10.19	3.22 (1964)	19.53 (2018)
South America	39.00%	54,263	7.08	4.61 (2012)	8.87 (1988)
South-eastern Asia	0.92%	1,279	25.79	9.49 (1973)	32.99 (2012)
Southern Africa	0.17%	235	20.08	9.39 (1964)	32.61 (2017)
Southern Asia	3.24%	4,505	23.27	15.35 (1964)	34.6 (2016)
Southern Europe	0.56%	783	22.69	7.83 (1964)	43.5 (2015)
Western Africa	0.25%	345	15.33	12.69 (2019)	17.97 (2018)
Western Asia	0.05%	69	8.84	4.66 (1965)	13.59 (2012)
Western Europe*	0.13%	220	23.58	12.51 (1974)	29.97 (2017)

*Excluded from analysis due to lack of data

Table 5.6 – Production data of soybean, relative to country

	Production (1000 tonnes)		Yield (1000 hg/ha)		
Countries	%/total	Average	Average	Min. (year)	Max. (year)
Argentina	12.90%	17,941	20.68	9.77 (1961)	33.34 (2019)
Brazil	23.10%	32,075	20.09	8.48 (1964)	33.9 (2018)
China	8.20%	11,375	13.9	6.26 (1961)	18.98 (2018)
India	3.10%	4,374	8.51	4.35 (1965)	13.53 (2012)
USA	43.60%	60,660	23.3	15.31 (1964)	34.94 (2016)

Table 5.7 – Production data of soybean, relative to continent

	Production (1000 tonnes)		Yield (1000 hg/ha)		
Continents	%/total	Average	Average	Min. (year)	Max. (year)
Africa	1.00%	880	8.36	3.48 (1965)	14.59 (2010)
Americas	84.00%	117,399	22.34	15.17 (1964)	32.57 (2017)
Asia	13.00%	17,986	12.06	6.44 (1961)	15.35 (2010)
Europe	2.00%	2,811	13.38	3.31 (1964)	23.16 (1997)
Oceania	0.00%	51	15.41	3.86 (1966)	23.29 (2008)

5.A.3 Cocoa

Table 5.8 – Production data of cocoa, relative to region

	Production (1000 tonnes)		Yield (1000 hg/ha)		
Regions	%/total	Average	Average	Min. (year)	Max. (year)
Caribbean	2.20%	60	3.6	2.21 (2005)	5.37 (2019)
Central America	1.80%	49	4.89	2.83 (1965)	6.65 (2007)
Eastern Africa	0.50%	14	3.96	1.72 (1963)	5.84 (2015)
Eastern Asia*	0.00%	0			
Middle Africa	6.20%	172	2.98	1.92 (1976)	3.94 (2019)
Oceania	1.40%	40	4.14	3.24 (1994)	5.11 (1971)
South America	15.90%	437	4	2.66 (2000)	5.68 (1985)
South-eastern Asia	12.90%	355	6.13	2.65 (1962)	10.32 (1998)
Southern Asia	0.30%	9	3.15	1.01 (1972)	5.74 (1994)
Western Africa	58.80%	1,620	3.99	2.55 (1965)	5.07 (1996)

*Excluded from analysis due to lack of data

Table 5.9 – Production data of cocoa, relative to country

	Production (1000 tonnes)		Yield (1000 hg/ha)		
Countries	%/total	Average	Average	Min. (year)	Max. (year)
Brazil	9.23%	254	78.61	2.79 (2000)	7.42 (1979)
Cameroon	5.30%	146	57.73	1.98 (1961)	4.16 (2019)
Côte d'Ivoire	30.50%	840	63.17	3.27 (1961)	7.01 (2000)
Ghana	16.70%	460	30.70	2.05 (1981)	5.5 (2012)
Indonesia	10.60%	292	40.70	1.22 (1962)	11.32 (1998)
Nigeria	9.90%	273	117.43	2 (1962)	4.98 (1998)

Table 5.10 – Production data of cocoa, relative to continent

	Production (1000 tonnes)		Yield (1000 hg/ha)		
Continents	%/total	Average	Average	Min. (year)	Max. (year)
Africa	65.54%	106,525	3.86	2.54 (1961)	4.93 (2006)
Americas	19.82%	32,210	4.00	2.69 (2000)	5.57 (2019)
Asia	13.20%	21,445	5.73	2.64 (1961)	10.16 (1998)
Oceania	1.44%	2,344	4.14	3.24 (1998)	5.11 (1971)

5.B Detailed results and interpretation

5.B.1 Maize

Table 5.11 – Best forecasting options for different months, relative to maize price. The names reported for each month correspond to the models showing the highest RA for predicting price change at this period. As TBATS tends to perform very well for short time lags, TBATS appears to be the best option for all months when the time lag is relatively small. For longer time horizons, other models (in particular GBM) are more accurate. The name between brackets indicates whether the predictions were more accurate with regional yields or productions and whether historical prices were found to have a substantial impact on price.

	Time lags (months)											
Month	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
January	TBATS			GBM (Yield)								
	0.73	0.64	0.46	0.37								
February	TBATS				GBM (Yield)							
	0.75	0.59	0.53	0.38	0.29							
March	TBATS					GBM (Yield)						
	0.78	0.59	0.47	0.41	0.3	0.22						
April	TBATS					GBM (Yield + $p_{m,y-1}$)						
	0.67	0.55	0.41	0.34	0.24	0.18						
May	TBATS						GBM (Yield + $p_{m,y-1}$)					
	0.68	0.64	0.47	0.30	0.21	0.19	0.18					
June	TBATS						GBM (Production)					
	0.61	0.55	0.48	0.39	0.31	0.24	0.19					
July	TBATS				CART (Yield)							
	0.44	0.28	0.29	0.24	0.23							
August	TBATS	CART (Production)										
	0.61	0.31										
September	TBATS	LM (Production)										
	0.46	0.21										
October	TBATS	GBM (Yield)										
	0.57	0.31										
November	TBATS		GBM (Yield)									
	0.63	0.36	0.24									
December	TBATS			GBM (Yield + $p_{m,y-1}$)								
	0.72	0.56	0.33	0.31								

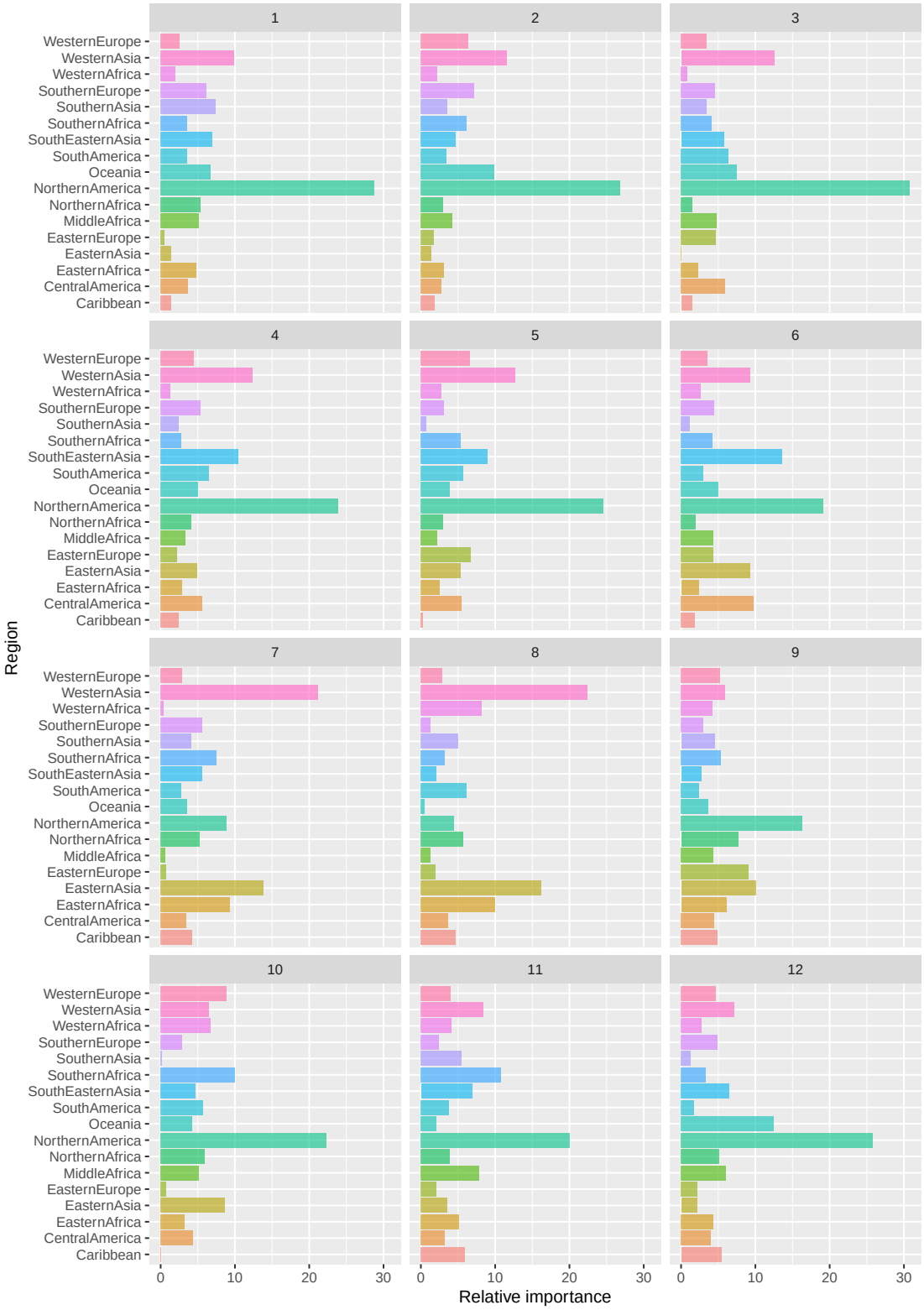


Figure 5.12 – Relative importance, Maize

5.B.2 Soybean

Table 5.12 – Best forecasting options for different months, relative to soybean price. The names reported for each month correspond to the models showing the highest RA for predicting price change at this period. As TBATS tends to perform very well for short time lags, TBATS appears to be the best option for all months when the time lag is relatively small. For longer time horizons, other models (in particular GBM) are more accurate. The name between brackets indicates whether the predictions were more accurate with regional yields or productions and whether historical prices were found to have a substantial impact on price.

	Time lags (months)											
Month	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
January	TBATS	LM (Yield)										
	0.72	0.54										
February	TBATS			GBM (Production)								
	0.56	0.51	0.44	0.39								
March	TBATS	RF (Production + $p_{m,y-1}$)										
	0.63	0.48										
April	TBATS	RF (Production + $p_{m,y-1}$)										
	0.42	0.38										
May	TBATS	CART (Production)										
	0.29	0.26										
June	TBATS	GBM (Yield)										
	0.53	0.31										
July	TBATS		RF (Production + $p_{m,y-1}$)									
	0.36	0.22	0.16									
August	TBATS			RF (Yield)								
	0.49	0.24	0.19	0.06								
September	TBATS		LM (Yield)									
	0.55	0.32	0.21									
October	TBATS		GBM (Production)									
	0.68	0.46	0.37									
November	TBATS			GBM (Production)								
	0.56	0.40	0.29	0.19								
December	TBATS			GBM (Production + $p_{m,y-1}$)								
	0.64	0.38	0.31	0.24								

Figure 5.13 – Relative importance, Soybean

5.B.3 Cocoa

Table 5.13 – Best forecasting options for different months, relative to cocoa price. The names reported for each month correspond to the models showing the highest RA for predicting price change at this period. As TBATS tends to perform very well for short time lags, TBATS appears to be the best option for all months when the time lag is relatively small. For longer time horizons, other models (in particular GBM) are more accurate. The name between brackets indicates whether the predictions were more accurate with national yields or productions and whether historical prices were found to have a substantial impact on price.

Month	Time lags (months)											
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
January	TBATS				CART (Production + $p_{m,y-1}$)							
	0.69	0.55	0.45	0.27	0.15							
February	TBATS					LM (Yield)						
	0.61	0.44	0.40	0.37	0.20	0.13						
March	TBATS		GBM (Yield + $p_{m,y-1}$)									
	0.51	0.40	0.31									
April	TBATS	GBM (Yield)										
	0.68	0.43										
May	TBATS	GBM (Yield)										
	0.59	0.48										
June	TBATS				RF (Production)							
	0.62	0.48	0.34	0.30	0.20							
July	TBATS					RF (Production)						
	0.62	0.46	0.32	0.22	0.21	0.13						
August	TBATS						RF (Production)					
	0.71	0.54	0.39	0.28	0.21	0.22	0.13					
September	TBATS							GBM (Yield)				
	0.63	0.62	0.43	0.29	0.21	0.13	0.16	0.03				
October	TBATS				GBM (Yield + $p_{m,y-1}$)							
	0.59	0.35	0.37	0.24	0.11							
November	TBATS		VAR (Yield)									
	0.49	0.22	0.12									
December	TBATS			VAR (yield)								
	0.59	0.41	0.24	0.12								

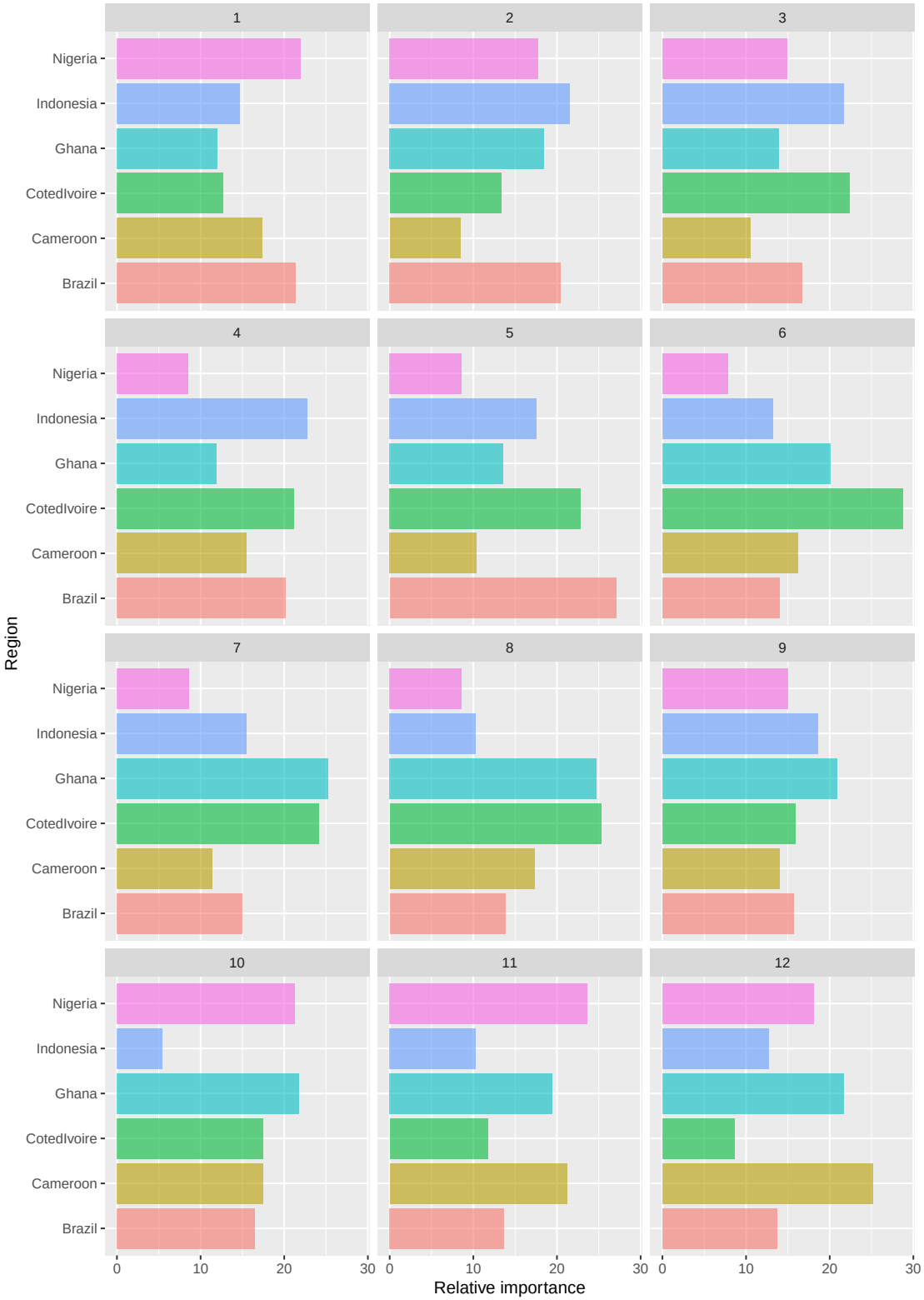


Figure 5.14 – Relative importance, Cocoa

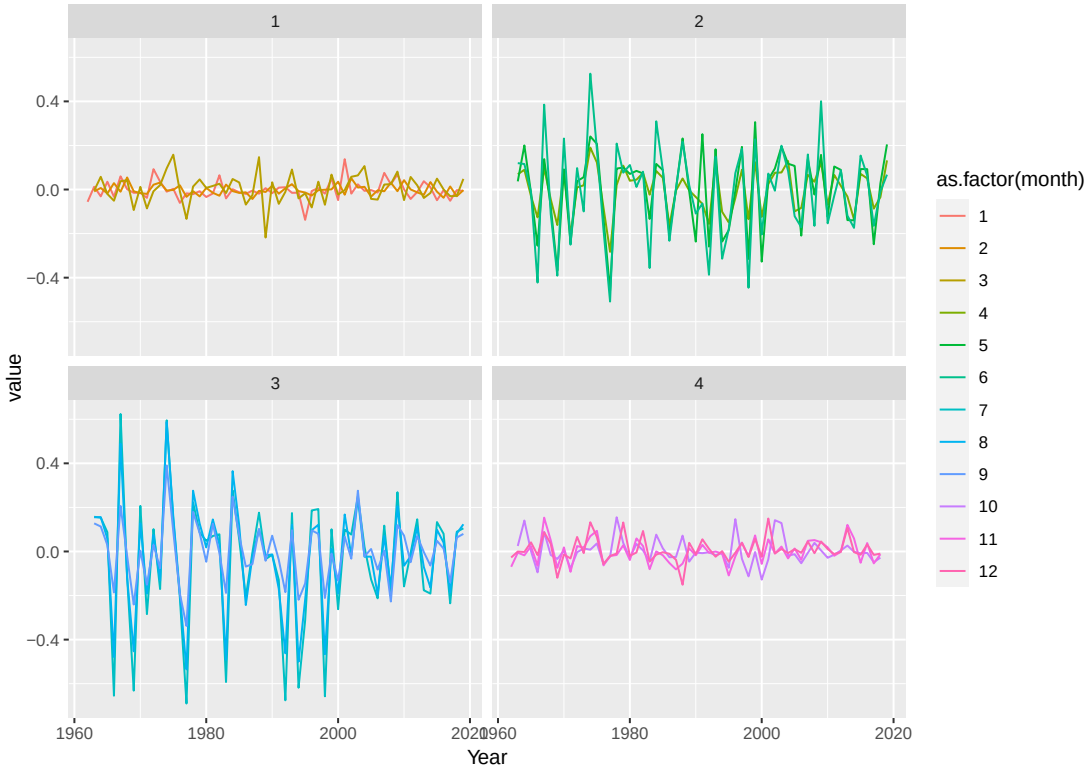


Figure 5.15 – Time-series of Shapley values, Cote D'Ivoire, Cocoa

Chapter 6

Discussion

```
The fear of population
   running ahead of food production
   has been regularly voiced.
   It is not my intention
   to dismiss these problems and fears.
```

– Amartya Sen, Poverty and Famines (p.150)

This research was motivated by a long-time desire to contribute to the worldwide efforts of improving food security and diets worldwide. While already familiar with the subjects of climate change and their effects on agriculture and the problem of food security from previous research papers, the topic of international trade was new to me.

Beginning this research, we recognised the scarcity of an AC price forecasting tool that would be accurate and bridge the gap of dis-information that strongly affects the most vulnerable producers and consumers around the globe. Completing this thesis, we achieved to provide a novel methodology to forecast agricultural-commodity prices. This method ensures high forecasting accuracy and is interpretable and technically accessible. We show that AC prices can be predicted for time-frames of one month to a year ahead while maintaining a high forecast quality and the principles of scientific transparency.

6.1 The basic idea behind the study

As already explained in the introduction chapter, turning the theory of AC price forecasting into an available and accessible tool would be socially beneficial, especially for those who currently cannot access it - those are mainly the residents

of low-income countries. Considering that every crop has unique nutritional values, consuming several crops from different groups can create a complete balanced diet. In the context of minimally processed food, such crop integration can build a low-priced and healthy diet. As an inexpensive source of energy and micro-nutrients, maize, combined with soybean as a cheap source of protein, fat, and other micro-nutrients, can promote food security for low-income consumers. Parallel, and in combination with cocoa, cultivated mainly by small farmers in developing countries, can help in terms of welfare as a tool for rescuing poor farmers from the cycle of poverty. That has often been the historical role: maize, as a relatively sustainable crop cultivated in varied climate zones, and soybean, which has been an important food component of the Chinese diet for thousands of years. However, in contrast to the tremendous potential of these crops and their clear potential to feed the entire world population today (Helms, 2004), their supply in certain areas is still limited.

Although food production has more than tripled over the past six decades, the growing use of the world's major crops as a source for energy or livestock feed has increased their consumption in high-income regions, which comes at the expense of the low-income ones. Furthermore, the global diet evolution leads to upward in meat consumption even in the least affluent regions, contributing to increased demand. The tendency to stock food as part of a food security (business) strategy to protect populations (of risks management) in times of food price fluctuations also causes relatively high vulnerability among residents of low-income countries who do not possess adequate food stocks.

Despite repeated requests from the World Trade Organisation (WTO) to refrain from government interventions, which could harm markets competitiveness (G20, 2020; WTO, 2020), the recent coronavirus crisis has proved, once again, that the first to suffer from price fluctuations are the low GDP per capita's countries. Furthermore, the ongoing health (and economic) crisis revealed the negative consequences of international tensions, especially between the major powers, and a lack of coordination between the trading countries (IFPRI et al., 2020). The most vulnerable are the poor, who could not stand the uncertainty, including the inability to sell their agricultural goods or purchase food. Given that not many countries have held sufficient food reserves for three months, millions have joined those who already suffered from food insecurity. Based on the existing potential in international food trade and the historical claims that it will balance the food distribution among all countries while ensuring a competitive market and balancing price fluctuations, this study focused on the international prices of agricultural goods. The entire study was done with an aspiration to promote the symmetry in information regarding global AC markets, including the drivers for price fluctuations, their likelihood to occur in the coming future

(month to a year ahead) and their quantification. I hope this study could advance social welfare, especially among the underprivileged in low-income countries.

6.2 Empirical Approach

This study is composed of three essays that give a good picture of the performance of the proposed empirical approach. As a whole, they create an accessible framework for analysing the international AC markets by unravelling the impacts of agricultural productions on the global price of the same crop.

6.2.1 Essay I: Assessing the sensitivity of global maize price to regional production using statistical and machine learning methods

The opening article of the doctoral dissertation attempts to trace two key points: The first is an assessment of the impact of the producing areas on maize price fluctuations; the second is an establishment of an empirical relationship between global maize price and production variation.

This article examines maize as a prototype market for an almost six decades period. The uniqueness of this study lies in its pioneer use of ML models to analyse AC prices in the Medium Term while providing a glimpse of what stands behind the results obtained. For each model, two-month specific versions are built: regression to quantify annual price changes and classification, to assess the chance of a price decrease or increase relative to the year before. All models were evaluated by a leave one out cross-validation. The study focuses on the fourth quarter of the year, i.e., the beginning of the period in which the North American (mainly USA) maize's is physically sold on the Chicago Merchandise Exchange as *new crop*. The results quantify the impact of maize production in Northern America on the global maize price in October, November and December, i.e. during and after the North-American harvest season. The results point out the potential of using machine learning models for price prediction but do not compare the predictive performance of this approach with standard forecasting tools.

6.2.2 Essay II: Forecasting global maize prices from regional productions

The second essay directly continues the first while focusing on forecasting. This article was motivated by two main objectives: The first was to forecast the monthly

global price of maize in a Medium Term time horizon, and the second was to find the most accurate forecasting method for different months and time-frames (lags).

We compare machine learning tools to two econometric models; both are ubiquitous tools in forecasting studies:

1. TBATS - an autoregressive tool that automatically handles non-linear features and multi-seasonality. TBATS has already demonstrated impressive predictive capabilities in relatively short ranges for a variety of topics, including daily electricity price (Karabiber and Xydis, 2019), gas consumption (Naim et al., 2018), and even rainfall (Farheen, 2021). As for AC price forecasting, however, this is the first time to test TBATS.

2. VAR - a multivariate autoregressive model. VAR is a widely used and relatively simple forecasting tool with great importance in building and analysing monetary policies. VAR models excel in detecting shocks within the data and combining their effects on the variability of the main variables or, in our case, maize prices. While VAR is an effective tool for forecasting variables such as inflation, GDP growth, currency exchange or interest rates (Bjørnland, 2008; Kapetanios et al., 2008), its effectiveness has not yet been tested in the context of Medium Term AC prices.

The second article compares the forecasting models relative to a benchmark corresponding to a naive constant prediction. As such, the model evaluation included a rolling cross-validation process, which yielded a forecasting error (RMSE) used to rank all the models; and a comparison with a naive prediction represented by a mean price change value. Beyond comparing the attractiveness of the models for forecasting maize prices, the study includes an analysis of the nature of the relationship between the level of change in the regional annual maize production and the change in its global price; Identification of the regions with the highest impact on the maize price, broken down by month. In addition, the study provides an accurate breakdown of the preferred method for forecasting monthly maize prices according to the forecasting horizon (by month, in one-month lags).

Relative importance analysis, which seeks to uncover the overall relevance of each region for price forecasting (König et al., 2021), confirmed the substantial relative influence of Northern American production on the global price during most of the year, starting from the beginning of the market year in October until May. However, Western Asia showed a more substantial influence on maize price changes in July and August.

Additionally, the Shapley values provided a glimpse into the main drivers of inevitable extreme price fluctuations. By taking into account specific extreme

cases, the results show that certain regions greatly impacted extreme price fluctuations observed at specific years. For example, Shapley brings into light the strong positive influence of the Eastern African maize yields of 2006 over the November price of that same year. Undeniably, 2006 was a year of extreme droughts in the region (Solomon et al., 2007), which harmed the agricultural sector (Gebrechorkos et al., 2020), and resulted in exceptionally high maize importation, notably from the United States.

6.2.3 Essay III: Data-driven assessment of the impacts of crop productions on the global prices of maize, soybean and cocoa

The final article applies the knowledge accumulated throughout the two former essays by examining the effectiveness of the forecasting methods for two additional crops: soybeans and cocoa. This analysis explores the genericity of the proposed approach and captures the uniqueness of AC from three different categories, as determined by the World Bank: *grains*, for maize; *oils & meals*, for soybean; and *beverages*, for cocoa. Additionally, this chapter assesses the sensitivity of the model performances to three geographic scales considered for the inputs, i.e., regional (as in the two first essays), continental and national. Finally and for all three commodities, we implemented each model with two sets of inputs: 1. regional production or yield variations; and 2. the same variables with the addition of the relative annual change of last year's price. All in all, each forecasted monthly price is the result of the best performing model, out of 60 (5 algorithms $\times$ 3 geographic scales $\times$ 4 versions, excluding TBATS), and relative to the most relevant geographic scale division. The specification of three AC market categories and three geographic scales brings out the significance of the economic structure of the market. The results reveal the utmost importance that market structures have on the level crop productions influence AC prices globally. Furthermore, the study shows that regional changes in maize production have undeniably high impacts on its price, especially when coming from Northern America - the crop's lead world producer and exporter, and by a considerable difference compared to other regions. The other extreme is the cocoa market. Western Africa and South America concentrate alone most of the cocoa production in their area, typically as the work of smallholder farmers in family farms in relatively poor areas. Contrary to maize and soybean, which are traded mainly in the international market located and managed in the country of the biggest producer, [1] cocoa is mainly traded in the importer side, New York and

[1] the mainstay of maize and soybean trade activity is at the CME in US dollar. However, for soybean, since December 2007, the price listed on the World Bank website comes from the

London, i.e., far from its country of origin. This fact contributes to the lack of market information among cocoa producers and detracts from their ability to control the price they will receive for their crop or the preferred date for selling it.

Numerous techniques were applied for interpreting the results of the fitted models: relative importance analysis (Greenwell et al., 2020), Shapley values (Molnar et al., 2018; Tianqi et al., 2021), (Greenwell, 2017; Liu and Just, 2020) and standard correlation analysis. For cocoa, none of these methods indicated a strong relationship between the production volume of the leading producer (Côte d'Ivoire) and price changes. The results did show a lack of absolute market power concentrated in a particular area and a fairly uniform distribution of monthly impact per country over the year. Moreover, a comprehensive examination pointed out a rather complex relationship between Shapley values and crop yield variations in Côte d'Ivoire. Focusing on specific extreme price shocks indicated a high contribution of cocoa yield in Indonesia to events of exceptionally high price increases. The results seemed surprising at first, as it is the most concentrated market among the three markets examined, not to mention that Indonesia's market share in the global cocoa market is significantly lower than that of Côte d'Ivoire. However, an in-depth study of the cocoa literature has revealed a complex system in which some factors undermine the natural equilibrium of the market. For Côte d'Ivoire (as well as Cameroon and Nigeria), one or very few local organisations collect the vast majority of production from the small farmers, most of whom have no access to market information. Alternatively, they received a price set by the local government at the beginning of the season, depending on future prices and stock exchanges (ITC and UNCTAD/WTO., 2001). This price, however, is set low enough to ensure a positive return to the paying body. As it comes, over the years, the cocoa export market in Ivory Coast has been privatised, so private export companies now collect the production. As a result, there has not yet been an improvement in the small farmers' situation (Abbott et al., 2019). In such conditions, the farmers of Côte d'Ivoire take the critical decision right at the beginning of the growing season: they aim at increasing their production when the price they receive from their government increases and vice versa.

In terms of forecasting results, ML methods (RF and GBM) usually perform better than the other models for any horizon that is longer than three months into the future (Depends on the crop and month. For detailed results see Appendix 5.B). Concerning maize and soybean, GBM provides noticeable higher forecasting accuracy for Northern America's new crop months. In cocoa, these months, namely March, April and May, are the only ones in which ML models

Construction Industry Federation (CIF) Rotterdam

are considerably favourable over the other models, including TBATS.

As a result, one might conclude that the effectiveness of the forecasting tool proposed in a research study increases for sectors with fewer market distortions. That is, the model is more suitable for competitive markets.

6.3 Contributions

This research project offers several contributions to the literature on price forecasting. Detecting the main drivers for maize prices changes using several ML and econometric techniques, Essay I (3rd chapter) mainly contributed by presenting novel analytical methods for forecasting AC prices. The first article, which was our first try to only accessible data and relatively simple models, revealed that ML algorithms are a legitimate tool to be used in the AC price forecasting science. Moreover, it proved that these ML models do not have to be of the "black-box" type and that their behaviour becomes interpretable when using powerful visualisation techniques.

The second essay (4th chapter) continued the path started in the one before and provided evidence, through several model interpretation techniques, that prices in the maize market react strongly to changes in crop output changes in Northern America, mainly of yield. The latter applies to 10 monthly prices per year, apart from the last two months of the North American trade year. Our importance ranking technique revealed a strong lead of Western Asia during those two months. However, when focusing on specific events, Shapley value presents relatively strong Western Asia and Northern Africa influences.[2]

The third and final essay (5th chapter) found, through expanding the second essay, substantial differences in the optimal forecasting approaches for each unique AC price. It showed that for forecasting maize prices with the highest accuracy, relative to the models tested in this thesis, regional yields is the most recommended input to use. For soybean, most of the impact comes from regional production, while cocoa prices are greatly affected by the local yield of the six biggest producing countries. The application of multiple interpretation techniques (PDP, relative importance, Shapley value; with SHAP based analysis: Shapley value and PDP) uncovered the remarkable impact each producing unit has over the global monthly price and thus supplied an original tool to preprepare to extreme price fluctuations.

[2]Together, Western Asia and Northern Africa (MENA) compose a whole region sharing some common characteristics. In both areas, July and August open the regional trading year of maize. With regards to Western Asia, these impacts are more of the negative direction, i.e., its main contribution pushes the global price down

This study followed three leading principles, i.e., concise comprehensibility, interpretability and accessibility. Generally speaking, it contributed to the efforts for promoting global food security.

6.3.1 Contribution I - Concise Comprehensible Forecasting Tool

As noted in the Approach and Methodologies chapter, an ideal model for this research can capture multiple drivers to AC price fluctuations while remaining relatively succinct. Whereas data collection could be an obstacle in the way of such a model, this study successfully restricted the input to publicly available data (annual crop production/yield), which the user can obtain in one simple click. The user can turn this raw information into a usable grouped variable by uploading the data into his personal computer and running the code. The dependent variable of the model (global prices) transformed into its "model-adapted" style in the same manner.

6.3.2 Contribution II - Interpretable Forecasting Tool

In their article, Coyle and Weller (2020) criticise researchers' choice of ML models as a tool to analyse and predict policy-related questions. They argue that ML models are often non-interpretable and thus prevent their users from understanding them and verifying the validity of their results. To overcome this challenge, we chose to construct all three articles based on interpretable models and then used several model-agnostic (Molnar, 2019) visualisation techniques.

More specifically, the first step of conducting this research was to investigate the causal relationship between the model's inputs and output. For this primal information, we used the Granger indicator of causality (Granger, 1969). After training the models, the contribution level to the prediction accuracy (RMSE) determined the relative regional importance separately for each algorithm for each month. Finally, the Partial Dependence Plots (PDP) visually described the average responses of the maize price to relative maize yield changes in the regions that ranked the highest by their contribution to the prediction accuracy of the models.

In the second paper, the relative importance technique showed, again, the relative influence of the features. Later, integration of the game-theory based Shapley value enabled us to assess the marginal contribution of each of the producing regions to specific events of the most extreme price shocks, in both positive and negative directions.

The third article combines several model interpretation techniques. Amongst them is Shapley Additive Explanations (SHAP) of Lundberg and Lee (2017). SHAP

is an ML specific interpretation method based on the traditional Shapley algorithm. Similar to the Shapley values, SHAP measures the contributions of each feature to the model predictions. However, the main advantage of this innovative algorithm derives from its ability to combine a Shapley adapted PDP for an interpretation that combines both quantification and visualisation measures.

6.3.3 Contribution III - Accessible Forecasting Tool

By being accessible, the model must be constantly ready for adaptation by the food security strategic designer, i.e., the policymaker who uses it. The policymaker must have regular access to the model's input to achieve this goal. Such a model provides its users with the ability to understand what stands behind it and a global comprehension regarding the market they face.

First, the forecasting program includes an alternative error valuation tool named "Relative Advantage". Using this tool, users can explore whether the forecasting model is sufficiently efficient relative to constant prediction and the other models. "Relative Advantage" provides a dynamic evaluation, dependent on the required month and the time remaining until the due date. Second, the model-agnostic techniques mentioned in the previous section indicate which actor should the model-user examine with caution. By using it, policymakers can design their strategy for up to one year before buying/selling the AC, then verifying its accuracy as the actual trade date approach.

We summarise the overall contribution with a hypothetical example showing how to profit from this forecasting tool.

Often, decision-makers involved in food security programmes have a limited annual budget, and the objective is to maximise the food security level of a population in need. The policymaker aims to purchase a sufficient amount of maize (source of energy for humans and livestock), wheat, rice (sources of energy for humans), soybean (source of protein for humans and livestock) for one year; while saving as much as possible for a local production of fruits and vegetables. Any budget left can potentially increase social welfare through steps such as investments in grain stock building or technologies. We assume that protein-rich foods (eggs and dairy, with a limited amount of meat) do not have additional costs besides feeding. Considering the costs and risks associated with storing AC increase by the day, our model will return, for each crop and each month, the forecasted costs of the actual crop purchase. The policymaker will then integrate this information into the local costs function (including any details regarding freight, storage, quality, or any other relevant factor) to detect the optimal month(s) to purchase some quantity of each AC.

6.4 Recommendations and future work

This doctorate thesis provides a comprehensible, interpretable and accessible AC price forecasting and analysis tool for analysis horizons of one to twelve months ahead. Moreover, it is accessible to whoever needs it as a ready to use R or python package, which uses freely available data only. As of today, the model examines three different internationally traded crops thoroughly. However, it involves the price forecasting of eight crops in total (final results are in Appendix A) and, thus, proves to be a substantially outperforming tool, which is applicable for other AC besides maize, soybean and cocoa. Price forecasts in this project are the output of one type of input (crop production or yield). Apart from presenting a final result, the model provides an error analysis that indicates the estimated risk, corresponding to the approximated forecasting error of the model. The overall recommendation is to consider both forecasted price and the model error and prefer acting in months where the risk of errors is small.

We use our model to highlight critical price change events, which should be detected correctly to enable the model to be transparent for its users. One challenge with forecasting AC prices is that while one might care most about forecasting events of extreme fluctuations, these events are relatively rare. From a food-security perspective, failing to identify extreme price change events might be a worse outcome than missing events of moderate price changes. Techniques such as the Shapley value algorithm can reflect these policy priorities in the model. Our analysis highlights the importance of understanding the trade-off between missing some production (yield) shocks in influential regions and mistakenly putting more attention on the total global supply or even on the production of regions with low-impact. Data on the costs of misinterpretation could assist in evaluating the potential damage of any agricultural output shocks on the food security level of vulnerable areas.

While doing the research, I first encountered the field of international AC trade, along with the subject of machine learning. Over the past few years, I have read numerous articles and listened to countless lectures, conferences and opinions of experts from numerous fields. The chapters included in this research paper do not show all the attempts to maximise our tool's contribution to the world of AC price forecasting. We examined different explanatory variables, either together or separately; analysed the predictive power based on data from various sources of information and even examined dependence according to harvest seasons versus local trading year dates. In addition, we experimented with models from a wide array of possibilities while performing different versions of the run in the models that we eventually included. To maximise the familiarity with the currently accepted methods and the overall op-

tions available, segmentation of explanatory variables was also done based on profound research work, which included the exploration of economic and agronomic databases parallel to an investigation of the existing literature.

This forecasting tool may not be imminently relevant to all countries. The international (World-Bank) prices examined in this study show the average monthly value paid in direct global trading markets. This price is not necessarily a good indicator of the consumer price level (part of income spent on food), which ultimately determines his/her level of food security. As discussed concerning cocoa, these prices do not always reflect the price paid for the farmer who produced it. Here comes the great importance of the nature of the state importing or exporting each AC. As explained in Section 1.2.3, while many high-income countries manage well-planned programmes to protect consumers and producers from price fluctuations, low-income countries cannot always do it efficiently. The bottom line of the existence or in-existence of such programmes is the level at which each country's domestic price will fluctuate by the global price.

Alternatively, by the time data will be available in a sufficiently reliable and rich manner, it could be beneficial to include annual grain-stocks change as one of the model's inputs. Indeed, by their original role, large food stocks can compensate for periods of poor harvest or high AC prices, and thus they function as a social safeguard. Unfortunately, food stocking is a costly matter that is not economically available for all nations. Furthermore, sufficient stocks can mitigate competition over food products. On the downside, overstocking can put global markets out of their natural stabilisation, as happened at the beginning of 2020, when China re-filled its grain stocks.

Undeniably, those who are less protected are also the most vulnerable ones. However, unfortunately, those are also the ones who lack the tools to analyse the global markets and forecast the optimal moment to purchase or sell agricultural commodities. In such an uncertain environment, excessive volatility in commodity prices negatively affects both producers and consumers. This lack of information generally impacts farmers' incomes and production and leads to worse input investment decisions. The repercussions of commodity market instability can also exacerbate poverty problems, particularly in rural areas. This way, lack of information adds to the negative impact on food security in the most vulnerable and import-dependent countries.

Another significant issue not considered here is the relationship between the prices of different commodities. Throughout the thesis work, a price forecasting model refers to each AC as independent of the others. However, in practice, as the Shapley value showed very visibly in the result interpretation of the maize, there is a strong relationship between the prices of substitute commodities. On the nutritional level, maize is a carbohydrate and, hence it is a substitute for

wheat, rice and sometimes soybean. Indeed, as a feed for livestock, price fluctuations of these AC also reflect price fluctuations of other AC on the international market and are used as a source of protein: meat and dairy products. In terms of local prices, egg prices will also shift eventually, in line with grain prices. In its aspect as an energy source, maize is used as a bio-fuel and therefore coordinated, along with other commodities, with the prices of energy commodities: coal, crude oils and natural gas.

As for cocoa and coffee, these AC are not essential in terms of nutritional value to the consumer but constitute a single or significant source of income for many small farmers, notably in developing countries in Western Africa. These AC are grown mainly in the tropics and imported in the vast majority by high-income countries. Producer prices fluctuate with the international price and are frequently determine their decision regarding land-allocation, changing between cocoa and coffee trees (Gilbert, 2016).

Beyond the discussed above, the results derived by model-opening techniques open the possibility for future studies that will address model improvement. In this context, it is possible exploring other options for adding an explanatory variable or converting to a different explanatory variable, examining the forecast quality of additional models or constructing a forecast based on running several models simultaneously. Another piece of advice is to analyse possibilities for combining the different algorithms to create a model covering several crops.

To conclude, this work offers a comprehensive and available tool for analysing and forecasting prices of agricultural commodities in time ranges of one month to one year ahead. If used correctly, the proposed mechanism may contribute to the food and economic security of households, farmers or other entities in need. However, as already written in the first paragraph of the work, this tool will bring maximum benefit if incorporated as part of a multidisciplinary food security plan. Undoubtedly, food prices are a critical component but investing in education (including education for proper nutrition and maintaining a healthy lifestyle), health, employment and security are also necessary, as fully detailed in the SDG website.

Bibliography

Abbott, P. C., C. Hurt, and W. E. Tyner (2009, 03). What's driving food prices? march 2009 update. Issue Reports 48495, Farm Foundation.

Abbott, P. C., C. Hurt, and W. E. Tyner (2019). Cote d'ivoire economic update (vol. 2) (french). Technical Report 138517, World Bank, Washington, D.C.

Abdoulkarim, E. and S. Zainab (2011). Assessing the effect of oil price on world food prices: Application of principal component analysis. *Energy Policy 39*(2), 1022–1025. Special Section on Offshore wind power planning, economics and environment.

Adjemian, M. K. and S. H. Irwin (2018). Usda announcement effects in real-time. *American Journal of Agricultural Economics 100*(4), 1151–1171.

Ahumada, H. and M. Cornejo (2016). Forecasting food prices: The case of corn, soybeans and wheat. *International Journal of Forecasting 32*(3), 838–848.

Allen, P. G. (1994). Economic forecasting in agriculture. *International Journal of Forecasting 10*(1), 81–135.

Baquedano, F. G. and W. M. Liefert (2014). Market integration and price transmission in consumer markets of developing countries. *Food Policy 44*, 103–114.

Beillouin, D., B. Schauberger, A. Bastos, P. Ciais, and D. Makowski (2020). Impact of extreme weather conditions on european crop production in 2018. *Philosophical Transactions of the Royal Society B: Biological Sciences 375*(1810), 20190510.

Bernanke, B. S., J. Boivin, and P. Eliasz (2005, 02). Measuring the effects of monetary policy: A factor-augmented vector autoregressive (favar) approach*. *The Quarterly Journal of Economics 120*(1), 387–422.

Bjørnland, H. C. (2008). Monetary policy and exchange rate interactions in a small open economy. *The Scandinavian Journal of Economics 110*(1), 197–221.

Blanchard, O. J. and D. Quah (1988, October). The dynamic effects of aggregate demand and supply disturbances. Working Paper 2737, National Bureau of Economic Research.

Blumenstock, J., G. Cadamuro, and R. On (2015). Predicting poverty and wealth from mobile phone metadata. *Science 350*(6264), 1073–1076.

Brandt, J. A. and D. A. Bessler (1983). Price forecasting and evaluation: An application in agriculture. *Journal of Forecasting 2*(3), 237–248.

Breiman, L. (2000). Randomizing outputs to increase prediction accuracy. *Machine Learning 40*(3), 229–242.

Breiman, L., A. Cutler, A. Liaw, and W. Matthew (2018). *Breiman and Cutler's Random Forests for Classification and Regression*. UC Berkeley. R package version 4.6-14.

Breiman, L., J. Friedman, C. J. Stone, and R. A. Olshen (1984). *Classification and regression trees*. CRC press.

Bren d'Amour, C., L. Wenz, M. Kalkuhl, J. C. Steckel, and F. Creutzig (2016). Teleconnected food supply shocks. *Environmental research letters 11*(3), 035007.

Brorsen, B. W. and S. H. Irwin (1996). Improving the relevance of research on price forecasting and marketing strategies. *Agricultural and Resource Economics Review 25*(1), 68–75.

Bukenya, J. O. and W. C. Labys (2007). Do fluctuations in wine stocks affect wine prices? Technical Report 386-2016-22741, American Association of Wine Economists, 2007-10.

Burfisher, M. E. (2021). *Introduction to Computable General Equilibrium Models* (3 ed.). New York, NY: Cambridge University Press.

Calle, M. L. and V. Urrea (2010, 03). Letter to the editor: Stability of random forest importance measures. *Briefings in Bioinformatics 12*(1), 86–89.

Caracciolo, F., L. Cembalo, A. Lombardi, and G. Thompson (2014). Distributional effects of maize price increases in malawi. *The Journal of Development Studies 50*(2), 258–275.

Chatzopoulos, T., I. Pérez Domínguez, M. Zampieri, and A. Toreti (2019). Climate extremes and agricultural commodity markets: A global economic analysis of regionally simulated events. *Weather and Climate Extremes In press*, 100193.

Childs, N. W., J. Kiawu, et al. (2009). *Factors behind the rise in global rice prices in 2008*. US Department of Agriculture, Economic Research Service.

Christian, R. L. and K. Marco (2006). Tomorrow's hunger: A framework for analysing vulnerability to food security. WIDER Research Paper 2006/119, NU-WIDER, Helsinki.

Clapp, J. and S. R. Isakson (2018). Risky returns: The implications of financialization in the food system. *Development and Change 49*(2), 437–460.

Cleveland, R. B., W. S. Cleveland, J. E. McRae, and I. Terpenning (1990). Stl: A seasonal-trend decomposition. *J. Off. Stat 6*(1), 3–73.

CME-Group (2021). *Introduction to Grains and Oilseeds*. CME.

Coleman-Jensen, A., M. P. Rabbitt, C. A. Gregory, and A. Singh (2021). Household food security in the united states in 2020. Report, ERR 298, U.S. Department of Agriculture, Economic Research Service.

Colling, P. L., S. H. Irwin, and C. R. Zulauf (1996). Reaction of wheat, corn, and soybean futures prices to usda "export inspections" reports. *Review of Agricultural Economics 18*(1), 127–136.

Corong, E., T. Hertel, R. McDougall, M. Tsigas, and D. van der Mensbrugghe (2017). The standard gtap model, version 7. *Journal of Global Economic Analysis 2*(1), 1–119.

Costinot, A. and D. Donaldson (2016, December). How large are the gains from economic integration? theory and evidence from u.s. agriculture, 1880-1997. Working Paper 22946, National Bureau of Economic Research.

Costinot, A. and A. Rodríguez-Clare (2018). The US gains from trade: Valuation using the demand for foreign factor services. *Journal of Economic Perspectives 32*(2), 3–24.

Coyle, D. and A. Weller (2020). "explaining" machine learning reveals policy challenges. *Science 368*(6498), 1433–1434.

Daily, G., P. Dasgupta, B. Bolin, P. Crosson, J. du Guerny, P. Ehrlich, C. Folke, A. M. Jansson, B.-O. Jansson, N. Kautsky, A. Kinzig, S. Levin, K.-G. Mäler, P. Pinstrup-Andersen, D. Siniscalco, and B. Walker (1998). Food production, population growth, and the environment. *Science 281*(5381), 1291–1292.

De Livera, A. M., R. J. Hyndman, and R. D. Snyder (2011). Forecasting time series with complex seasonal patterns using exponential smoothing. *Journal of the American statistical association 106*(496), 1513–1527.

Deaton, A. and G. Laroque (1992, 01). On the behaviour of commodity prices. *The Review of Economic Studies 59*(1), 1–23.

Deng, H. L. and Q. S. Yǔ (2019, 01). Soybean price pattern discovery via toeplitz inverse covariance-based clustering. *International Journal of Agricultural and Environmental Information Systems (IJAEIS) 10*(4), 1–17.

Dorosh, P. A., K. Subbarao, and C. Del Ninno (2004, 11). Food aid and food security in the short and long run: country experience from asia and sub-saharan africa (english). Working Paper 538, World Bank, Washington, D.C.

Efron, B. and T. Hastie (2016, 07). *Computer Age Statistical Inference Algorithms, Evidence, and Data Science* (1 ed.). The address: Cambridge University Press.

ERS-USDA (2021). *Food Price Outlook*. Washington, D.C.: USDA.

FAO (2002). Etudes de la FAO sur des aspects sélectionnés des négociations de l'OMC sur l'agriculture. Technical Report 21, FAO, Rome.

FAO (2008). Food outlook global market analysis. Technical report, Markets and Trade Division, FAO, Rome. Rice.

FAO (2012). Price volatility from a global perspective. In *Food price volatility and the role of speculation*, pp. 1–7. Rome: FAO.

FAO (2018, 08). Trade and nutrition technical note. Technical Report 21, FAO, Rome.

FAO (2020). *FAOSTAT Statistical Database*. Rome: FAO.

FAO (2021). *GIEWS FPMA Tool monitoring and analysis of food prices*. FAO.

FAO (2021). *New Food Balances: Description of utilization variables*. Rome: FAO.

FAO, IFAD, UNICEF, WFP, and WHO (2020). *Transforming food systems for affordable healthy diets.* Number 2020 in The State of Food Security and Nutrition in the World (SOFI). Rome: FAO, IFAD, UNICEF, WFP and WHO.

FAO and WFP (2010). Addressing food insecurity in protracted crises. Technical report, FAO, Rome.

Farheen, N. (2021, 05). Rainfall prediction and suitable crop suggestion using machine learning prediction algorithms. In A. Hassanien, S. Bhattacharyya, S. Chakrabarti, A. Bhattacharya, and S. Dutta (Eds.), *Emerging Technologies in Data Mining and Information Security*, Volume 2 of *Proceedings of IEMIS 2020*, pp. 497–513. Springer Singapore.

FAS-USDA (2021). *The Foreign Agricultural Service (FAS) Data & Analysis*. Washington, D.C.: FAO.

Frederic, A. V. and A. B. Gerald (1999). Understanding crop statistics. Technical Report 1554, USDA ERS.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics 29*(5), 1189–1232.

Fuss, S., P. Havlík, J. Szolgayová, E. Schmid, W. H. Reuter, N. Khabarov, M. Obersteiner, Y. Ermoliev, T. Ermolieva, and F. Kraxner (2015). Global food security and adaptation under crop yield volatility. *Technological Forecasting and Social Change 98*, 223–233.

G20 (2020, 03). Extraordinary G20 leaders' summit, statement on COVID-19. Report, WTO. accessed 2021-08-31.

Gebrechorkos, S. H., S. Hülsmann, and C. Bernhofer (2020). Analysis of climate variability and droughts in east africa using high-resolution climate data products. *Global and Planetary Change 186*, 103130.

Ghorbani, A. and J. Zou (Eds.) (1996, 11). *Report of the World Food Summit*, Rome. FAO: FAO.

Gilbert, C. L. (1989, 09). The impact of exchange rates and developing country debt on commodity prices. *The Economic Journal 99*(397), 773–784.

Gilbert, C. L. (2016). The dynamics of the world cocoa price. In *The economics of chocolate*, Chapter 16, pp. 307–338. Oxford University Press.

Gilbert, C. L. and P. Varangis (2003, May). Globalization and international commodity trade with specific reference to the west african cocoa producers. Working Paper 9668, National Bureau of Economic Research.

Gos, M., J. Krzyszczak, P. Baranowski, M. Murat, and I. Malinowska (2020). Combined tbats and svm model of minimum and maximum air temperatures applied to wheat yield prediction at different locations in europe. *Agricultural and Forest Meteorology 281*, 107827.

Gouel, C. (2011). *Instabilité des prix agricoles et politiques optimales de stabilisation*. Ph. D. thesis, Ecole Polytechnique X.

Gouel, C. (2012). Agricultural price instability: a survey of competing explanations and remedies. *Journal of economic surveys 26*(1), 129–156.

Gouel, C., M. Gautam, and W. J. Martin (2016, 02). Managing food price volatility in a large open country: the case of wheat in India. *Oxford Economic Papers 68*(3), 811–835.

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica 37*(3), 424–438.

Greenwell, B., B. Boehmke, and J. Cunningham (2020). *Package 'gbm'*. Greg Ridgeway. R package version 2.1.8.

Greenwell, B. M. (2017). pdp: An r package for constructing partial dependence plots. *The R Journal 9*(1), 421–436. R package version 0.7.0.

Haile, M. G., M. Kalkuhl, and J. von Braun (2016). Worldwide acreage and yield response to international price change and volatility: A dynamic panel data analysis for wheat, rice, corn, and soybeans. *American Journal of Agricultural Economics 98*(1), 172–190.

Haile, M. G., T. Wossen, K. Tesfaye, and J. von Braun (2017). Impact of climate change, weather extremes, and price risk on global food supply. *Economics of Disasters and Climate Change 1*(1), 55–75.

Hamzei, J. and M. Seyyedi (2016). Energy use and input–output costs for sunflower production in sole and intercropping with soybean under different tillage systems. *Soil and Tillage Research 157*, 73–82.

Hao, N., P. Pedroni, G. Colson, and M. Wetzstein (2017). The linkage between the u.s. ethanol market and developing countries' maize prices: a panel svar analysis. *Agricultural Economics 48*(5), 629–638.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *Random Forests*, pp. 587–604. New York, NY: Springer New York.

Hativ, A. and A. Mazouz (2021). Useing statistical methods to learn foreign exchange market transactions datasets. In *Statistics at glance 2020*, Number 2 in Statistics at glance 2020, Chapter 2, pp. 56–65. Jerusalem: Bank of Israel.

Havlík, P., H. Valin, A. Mosnier, S. Frank, P. Lauri, D. Leclère, A. Palazzo, M. Batka, E. Boere, A. Brouwer, et al. (2018, 06). Globiom documentation. draft.

Havlík, P., U. Schneider, E. Schmid, H. Böttcher, S. Fritz, R. Skalský, K. Aoki, S. De Cara, G. Kindermann, F. Kraxner, S. Leduc, I. McCallum, A. Mosnier, T. Sauer, and M. Obersteiner (2011). Global land-use implications of first and second generation biofuel targets. *Energy Policy 39*(10), 5690–5702. Sustainability of biofuels.

Headey, D. (2011). Rethinking the global food crisis: The role of trade shocks. *Food Policy 36*(2), 136–146.

Headey, D. and S. Fan (2008). Anatomy of a crisis: the causes and consequences of surging food prices. *Agricultural Economics 39*(s1), 375–391.

Headey, D. and S. Fan (2010). *Reflections on the global food crisis: how did it happen? how has it hurt? and how can we prevent the next one?*, Volume 165. Intl Food Policy Res Inst.

Headey, D. D. and W. J. Martin (2016). The impact of food prices on poverty and food security. *Annual Review of Resource Economics 8*(1), 329–351.

Helms, M. (2004). Food sustainability, food security and the environment. *British Food Journal 106*(5), 380–387.

Hernández-Orallo, J., P. Flach, and C. Ferri (2012). A unified view of performance metrics: Translating threshold choice into expected classification loss. *Journal of Machine Learning Research 13*(91), 2813–2869.

Hertel, T. W. (1997). *Global trade analysis: modeling and applications*. Number 7685 in GTAP Books. Cambridge university press.

Hertel, T. W., U. L. C. Baldos, and D. van der Mensbrugghe (2016). Predicting long-term food demand, cropland use, and prices. *Annual Review of Resource Economics 8*(1), 417–441.

HLPE (2013). Biofuels and food security. Report, High Level Panel of Experts on Food Security and Nutrition (HLPE). A report by the High Level Panel of Experts on Food Security and Nutrition of the Committee on World Food Security.

Hobbs, J. E. (2020). Food supply chains during the covid-19 pandemic. *Canadian Journal of Agricultural Economics/Revue canadienne d'agroeconomie 68*(2), 171–176.

Hoffman, L., L. Meyer, et al. (2018). Forecasting the us season-average farm price of upland cotton: Derivation of a futures price forecasting model. Technical Report CWS-18I-01, USDA ERS.

Hoffman, L. A. (2011). *Using Futures Prices to Forecast US Corn Prices: Model Performance with Increased Price Volatility*, Chapter 7, pp. 107–132. New York, NY: Springer New York.

Hoffman, L. A., X. L. Etienne, S. H. Irwin, E. V. Colino, and J. I. Toasa (2015). Forecast performance of wasde price projections for us corn. *Agricultural economics 46*(S1), 157–171.

Horne, C., K. Malden, L. Nelson, and N. Salidjanova (2020, 04). The U.S.-china "phase one" deal: A backgrounder. Technical report, U.S.-CHINA ECONOMIC AND SECURITY REVIEW COMMISSION.

Huang, G.-B., Q.-Y. Zhu, and C.-K. Siew (2006). Extreme learning machine: Theory and applications. *Neurocomputing 70*(1), 489–501. Neural Networks.

Hyndman, R., G. Athanasopoulos, G. Caceres, L. Chhay, M. O'Hara-Wild, F. Petropoulos, S. Razbash, E. Wang, F. Yasmeen, R. Ihaka, D. Reid, D. Shaub, Y. Tang, and Z. Zhou (2020). *Forecasting Functions for Time Series and Linear Models*. NA. R package version 8.12.

IFPRI (2018). Ifpri strategy refresh 2018-2020. Washington, DC.

IFPRI, Akademiya2063, and IISD (Eds.) (2020, 11). *Évènement virtuel - Commerce alimentaire et agricole dans le nouvel environnement politique : Comment les membres de l'OMC peuvent-ils appuyer la reprise économique et la résilience de l'Afrique?* IFPRI, Akademiya2063 and IISD: IFPRI.

ITC (2021). *COVID-19 Temporary Trade Measures*. Geneva: The International Trade Centre (ITC).

ITC and UNCTAD/WTO. (2001). *Cocoa: a guide to trade practices*. Geneva, Switzerland: International Trade Centre UNCTAD/WTO.

Jeung, M., S. Baek, J. Beom, K. Hwa Cho, Y. Her, and Y. Kwangsik (2019). Evaluation of random forest and regression tree methods for estimation of mass first flush ratio in urban catchments. *Journal of Hydrology 575*, 1099–1110.

Jichlinski, M. (1983). The price policy of cocoa in ivory coast. Master's thesis, The Hebrew University of Jerusalem, Rehovot.

Jouchi, N., K. Munehisa, and W. Toshiaki (2011). Bayesian analysis of time-varying parameter vector autoregressive model for the japanese economy and monetary policy. *Journal of the Japanese and International Economies 25*(3), 225–245.

Kalkuhl, M. (2016). *How Strong Do Global Commodity Prices Influence Domestic Food Prices in Developing Countries? A Global Price Transmission and Vulnerability Mapping Analysis*, pp. 269–301. Cham: Springer International Publishing.

Kan, I., A. Reznik, A. Kimhi, and J. Kaminski (2018). The impacts of climate change on cropland allocation, crop production, output prices and social welfare in israel: A structural econometric framework. Technical Report 888-2019-2205, The Hebrew University of Jerusalem, Rehovot, Israel.

Kapetanios, G., V. Labhard, and S. Price (2008). Forecast combination and the bank of england's suite of statistical forecasting models. *Economic Modelling 25*(4), 772–792.

Karabiber, O. A. and G. Xydis (2019). Electricity price forecasting in the danish day-ahead market using the tbats, ann and arima methods. *Energies 12*(5), 1–29.

Karali, B., O. Isengildina-Massa, S. H. Irwin, M. K. Adjemian, and R. Johansson (2019). Are usda reports still news to changing crop markets? *Food Policy 84*, 66–76.

Kim, Y. and S. Dharmasena (2018). Price discovery and integration in u.s. pecan markets. *Journal of Food Distribution Research 49*(856-2018-3108), 39–47.

Kuhn, M. and K. Johnson (2013). *Applied predictive modeling*, Volume 26, Chapter 8, pp. 173–190. Springer.

König, G., C. Molnar, B. Bischl, and M. Grosse-Wentrup (2021). Relative feature importance. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 9318–9325.

Laborde, D., A. Herforth, and D. e. a. Headey (2021). Covid-19 pandemic leads to greater depth of unaffordability of healthy and nutrient-adequate diets in low- and middle-income countries. *Nature Food 2*(7), 473–475.

Laborde, D., W. Martin, J. Swinnen, and J. Vos (2020). Covid-19 risks to global food security. *Science 369*(6503), 500–502.

Labys, W. (2003). New directions in the modeling and forecasting of commodity markets. *Mondes en developpement 122*(2), 3–19.

Latet (2021). The situation of food insecurity in israel. Report, Latet. In Hebrew.

Laudien, R., B. Schauberger, D. Makowski, and C. Gornott (2020). Robustly forecasting maize yields in tanzania based on climatic predictors. *Scientific Reports 19650*(10), 1–12.

Lentz, E., H. Michelson, K. Baylis, and Y. Zhou (2019). A data-driven approach improves food insecurity crisis prediction. *World Development 122*, 399–409.

Li, G.-q., S.-w. Xu, and Z.-m. Li (2010). Short-term price forecasting for agro-products using artificial neural networks. *Agriculture and Agricultural Science Procedia 1*, 278–287.

Liangyue, C. and J. Greenville (2020, 06). Understanding how china's tariff on australian barley exports will affect the agricultural sector. Technical Report 20.14, Australian Bureau of Agricultural and Resource Economics and Sciences (ABARES), Canberra.

Liaw, A., M. Wiener, et al. (2002). Classification and regression by randomforest. *R news 2*(3), 18–22.

Lima, M. V. M. d. and G. Z. Laporta (2020). Evaluation of the models for forecasting dengue in brazil from 2000 to 2017: An ecological time-series study. *Insects 11*(11), 1–14.

Liu, Y. and A. Just (2020). *SHAPforxgboost: SHAP Plots for 'XGBoost'*. dmlc XGBoost. R package version 0.1.0.

Lobell, D. B. and M. B. Burke (2010). On the use of statistical models to predict crop yield responses to climate change. *Agricultural and Forest Meteorology 150*(11), 1443–1452.

Lundberg, S. and S.-I. Lee (2017, 5). A unified approach to interpreting model predictions. Technical report, Harvard.

Luo, Z., X.-N. Bui, H. Nguyen, and H. Moayedi (2019). A novel artificial intelligence technique for analyzing slope stability using pso-ca model. *Engineering with Computers 37*(1435-5663), 1–12.

Lusk, J. L. (2016). From farm income to food consumption: Valuing usda data products. Technical Report 643-2018-161, Council on Food, Agricultural and Resource Economics (C-FARE).

Mallory, M. l. (2021). How to find information. In *Price Analysis: A Fundamental Approach to the Study of Commodity Prices*, Chapter 3, pp. 16–23. University of Illinois.

Malthus, T. (1789). Na essay on the principle of population, murray.

Mantzura, A. (2016). Seasonal adjustment from economic and financial series at the bank of israel - demonstration of the series "mortgage performance". In *Bank of Israel, Statistical Perspective 2016*, pp. 56–64. Jerusalem: Bank of Israel.

Mayzel, Y. (2021). The food security system in israel. Report, research and information center, The Knesset. In Hebrew.

Mitchell, D. (2008). A note on rising food prices. In *World bank policy research working paper*, Number 4682 in World bank policy research working papers. Washington, DC: World Bank.

Molnar, C. (2019). *Interpretable Machine Learning*. Bavarian State Ministry of Science and the Arts.

Molnar, C., G. Casalicchio, and B. Bischl (2018). iml: An r package for interpretable machine learning. *Journal of Open Source Software 3*(26), 786.

Mundlak, Y. and D. F. Larson (1992, 09). On the transmission of world agricultural prices. *The World Bank Economic Review 6*(3), 399–422.

Naim, I., T. Mahara, and A. R. Idrisi (2018). Effective short-term forecasting for daily time series with complex seasonal patterns. *Procedia Computer Science 132*, 1832–1841. International Conference on Computational Intelligence and Data Science.

Nidhiprabha, B. (2019, 06). Commodity price cycles, the agricultural trap, and thailand's incessant subsidies. *Asian Economic Papers 18*(2), 49–69.

Nwachukwu, M. U. and H. Chike (2011). Fuel subsidy in nigeria: Fact or fallacy. *Energy 36*(5), 2796–2801.

Ochieng, Dennis O.; Botha, R. and B. Baulch (2019). Structure, conduct and performance of maize markets in malawi. Technical Report 29, IFPRI, Washington, DC.

O'Connor, D. and M. Keane (2011). *Empirical Issues Relating to Dairy Commodity Price Volatility*, pp. 63–83. New York, NY: Springer New York.

Palatnik, R. R. and M. Shechter (2008). Can climate change mitigation policy benefit the israeli economy? a computable general equilibrium analysis. In *Environmental Economcis Abstarcts-WPS -Vo. 13, No.31; Presented at the 11th Annual Conference on Global Economic Analysis, Helsinki,*, Department of Agricultural Economics, Purdue University, West Lafayette, IN.

Pearl, J. (2009). A theory of inferred causation. In *Causality: Models, Reasoning, and Inference* (2 ed.)., Chapter 2, pp. 41–64. Cambridge University Press.

Pfaff, B. and M. Stigler (2018). *Package 'vars'*. Kronberg im Taunus. R package version 1.5-3.

Piot-Lepetit, I. and R. M'Barek (2011). *Methods to Analyse Agricultural Commodity Price Volatility*, pp. 1–11. New York, NY: Springer New York.

Popkin, J. (1977). Price forecasting. *Business Economics 12*(1), 33–37.

Puma, M. J., S. Bose, S. Y. Chon, and B. I. Cook (2015). Assessing the evolving fragility of the global food system. *Environmental Research Letters 10*(2), 024007.

R-Core-Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. version 3.6.3.

Ranum, P., J. P. Peña-Rosas, and M. N. Garcia-Casal (2014). Global maize production, utilization, and consumption. *Annals of the New York Academy of Sciences 1312*(1), 105–112.

Ratnayaka, R., D. Seneviratne, W. Jianguo, and H. Arumawadu (2015). A hybrid statistical approach for stock market forecasting based on artificial neural network and arima time series models. In *2015 International Conference on Behavioral, Economic and Socio-Cultural Computing, BESC 2015*, pp. 54–60. cited By 17.

Rokach, L. (2010). Ensemble-based classifiers. *Artificial Intelligence Review 33*, 1–39.

Rosenzweig, C., A. Iglesias, X. Yang, P. R. Epstein, and E. Chivian (2001, 12). Climate change and extreme weather events; implications for food production, plant diseases, and pests. *Global Change and Human Health 2*(2), 90–104.

Rosenzweig, C., J. W. Jones, J. L. Hatfield, A. C. Ruane, K. J. Boote, P. Thorburn, J. M. Antle, G. C. Nelson, C. Porter, S. Janssen, et al. (2013). The agricultural model intercomparison and improvement project (agmip): protocols and pilot studies. *Agricultural and Forest Meteorology 170*, 166–182.

Rouf Shah, T., K. Prasad, and P. Kumar (2016). Maize—a potential source of human nutrition and health: A review. *Cogent Food & Agriculture 2*(1), 1166995.

Schaub, S. and R. Finger (2020, 02). Effects of drought on hay and feed grain prices. *Environmental Research Letters 15*(3), 034014.

Schmidhuber, J., J. Pound, and B. Qiao (2020). Covid-19: Channels of transmission to food and agriculture. Technical report, FAO.

Schmidhuber, J. and F. N. Tubiello (2007). Global food security under climate change. *Proceedings of the National Academy of Sciences 104*(50), 19703–19708.

Shapley, L. S. (2016). *17. A Value for n-Person Games*, pp. 307–318. Princeton University Press.

Shively, G. E. (1996). Food price variability and economic reform: An arch approach for ghana. *American Journal of Agricultural Economics 78*(1), 126–136.

Sims, C. A. (1980). Macroeconomics and reality. *Econometrica 48*(1), 1–48.

Smith, L. C., O. Dupriez, and N. Troubat (2014). Assessment of the reliability and relevance of the food data collected in national household consumption and expenditure surveys. Report, FAO, World Bank. International Household Survey Network.

Smith, L. C. and A. Subandoro (2007). *Measuring food security using household expenditure surveys*. Number 3 in Food Security in Practice Technical Guide Series. Washington, D.C.: International Food Policy Research Institute (IFPRI).

Solomon, S., D. Qin, M. Manning, Z. Chen, M. Marquis, K. Averyt, M. Tignor, and H. M. (eds.) (2007). *Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, 2007*. Assessment report (Intergovernmental Panel on Climate Change).: Working Group. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press.

Storm, H., K. Baylis, and T. Heckelei (2019, 08). Machine learning in agricultural and applied economics. *European Review of Agricultural Economics 47*(3), 849–892.

Swinnen, Johan, e. and e. McDermott, John (2020, 07). *COVID-19 and global food security* (1 ed.). Washington, DC: International Food Policy Research Institute (IFPRI).

Tadasse, G., B. Algieri, M. Kalkuhl, and J. Von Braun (2016). Drivers and triggers of international food price spikes and volatility. In *Food price volatility and its implications for food security and policy*, pp. 59–82. Springer, Cham.

Taylor, H. C. (1919). *Agricultural economics.* New York,Macmillan,. https://www.biodiversitylibrary.org/bibliography/23695.

Therneau, T., B. Atkinson, B. Ripley, and M. B. Ripley (2019). *Package 'rpart'*. CRAN. R package version 4.1-15.

Thomsen, M. R. (2021, 1). Partial vs. General Equilibrium Models. [Online; accessed 2021-08-11].

Thrane, M., P. Paulsen, M. Orcutt, and T. Krieger (2017). Soy protein: Impacts, production, and applications. In S. R. Nadathur, J. P. Wanasundara, and L. Scanlin (Eds.), *Sustainable Protein Sources*, Chapter 2, pp. 23–45. San Diego: Academic Press.

Tianqi, C., H. Tong, B. Michael, K. Vadim, T. Yuan, C. Hyunsu, C. Kailong, M. Rory, C. Ignacio, Z. Tianyi, L. Mu, X. Junyuan, L. Min, G. Yifeng, and L. Yutian (2021). *Extreme Gradient Boosting*. dmlc XGBoost. version 1.4.1.1.

Ticlavilca, A. M. and D. M. Feuz (2010, 4). Forecasting agricultural commodity prices using multivariate bayesian machine. In *NCCC-134 Applied Commodity Price Analysis, Forecasting, and Market Risk Management*, NCCC-134 Applied Commodity Price Analysis, Forecasting, and Market Risk Management, St. Louis, MO.

UN (2019). *State of Commodity Dependence 2019*. State of Commodity Dependence. United Nations.

UN (2021). *Sustainable Development Goals*. UN.

US-HR (2009, 06). Hearing to review the federal crop insurance program : hearing before the subcommittee on general farm commodities and risk management of the committee on agriculture. Technical Report 110, United States House of Representatives, U.S. Government, IDCC, Washington, DC.

USDA (2021). Biden-harris administration's actions to reduce food insecurity amid the covid-19 crisis. Washington, DC.

Valenzuela, E., T. W. Hertel, R. Keeney, and J. J. Reimer (2007). Assessing global computable general equilibrium model validity using agricultural price volatility. *American Journal of Agricultural Economics 89*(2), 383–397.

Valin, H., R. D. Sands, D. van der Mensbrugghe, G. C. Nelson, H. Ahammad, E. Blanc, B. Bodirsky, S. Fujimori, T. Hasegawa, P. Havlik, E. Heyhoe, P. Kyle, D. Mason-D'Croz, S. Paltsev, S. Rolinski, A. Tabeau, H. van Meijl, M. von Lampe, and D. Willenbockel (2014). The future of food demand: understanding differences in global economic models. *Agricultural Economics 45*(1), 51–67.

van Meijl, H. and G. Woltjer (2012). The development of the magnet strategy. In *Presented at the 15th Annual Conference on Global Economic Analysis, Geneva, Switzerland*, Department of Agricultural Economics, Purdue University, West Lafayette, IN.

van Meijl, J., P. Havlík, H. Lotze-Campen, E. Stehfest, P. Witzke, I. P. Domínguez, B. Bodirsky, M. van Dijk, J. Doelman, T. Fellmann, et al. (2017). Challenges of global agriculture in a climate change context by 2050: Agclim50. Technical report, JRC.

Vries, J. d. (1980a). Forecasting world banana trade flows. Document de travail sur les produits de base 5, World Bank.

Vries, J. d. (1980b). The world sugar economy: an econometric analysis of long term developments. Document de travail sur les produits de base 5, World Bank.

Warr, P. G. (1990). Predictive performance of the world bank's commodity price projections. *Agricultural Economics 4*(3), 365–379.

Wegren, S. K. (2011). Food security and russia's 2010 drought. *Eurasian Geography and Economics 52*(1), 140–156.

Westreich, D., J. Lessler, and M. Jonsson Funk (2010). Propensity score estimation: neural networks, support vector machines, decision trees (cart), and meta-classifiers as alternatives to logistic regression. *Journal of Clinical Epidemiology 63*(8), 826–833.

Weymar, F. H. (1965). *The dynamics of the world cocoa market*. Ph. D. thesis, Massachusetts Institute of Technology.

Williams, G. (2011). *Package 'Rattle'*. Ise R! R package version 5.4.0.

World-Bank (2005, 05). Managing food price risks and instability in an environment of market liberalization (english). Technical report, World Bank, Washington, D.C.

World-Bank (2014, 10). Focus: The role of income growth in commodities. Technical report, World Bank, Washington, D.C.

World-Bank (2016, 07). Focus: From energy prices to food prices: Moving in tandem? Technical report, World Bank, Washington, D.C.

World-Bank (2020a, 10). Commodity markets outlook—persistence of commodity shocks. Technical report, World Bank, Washington, D.C. License: Creative Commons Attribution CC BY 3.0 IGO.

World-Bank (2020b, 04). A shock like no other: The impact of covid-19 on commodity markets. Technical report, World Bank, Washington, D.C. License: Creative Commons Attribution CC BY 3.0 IGO.

World-Bank (2021a). *Commodity Markets Outlook*. Washington, D.C.: World-Bank.

World-Bank (2021b). *Food Security and COVID-19*. Washington, D.C.: World Bank.

Wright, B. (2008, 01). Speculators, storage and the price of rice. Technical Report 2, Giannini Foundation of Agricultural Economics.

WTO (2020, 09). How wto members have used trade measures to expedite access to covid-19 critical medical goods and services. Report, WTO. accessed 2021-08-31.

Wu, F. and H. Guclu (2013). Global maize trade and food security: Implications from a social network model. *Risk Analysis 33*(12), 2168–2178.

Xiaojie, X. and Z. Yun (2021). Corn cash price forecasting with neural networks. *Computers and Electronics in Agriculture 184*, 106120.

Xiong, T., C. Li, and Y. Bao (2018). Seasonal forecasting of agricultural commodity price using a hybrid stl and elm method: Evidence from the vegetable market in china. *Neurocomputing 275*, 2831–2844.

Zeevi, D., T. Korem, N. Zmora, D. Israeli, D. Rothschild, A. Weinberger, O. Ben-Yacov, D. Lador, T. Avnit-Sagi, M. Lotan-Pompan, J. Suez, J. A. Mahdi, E. Matot, G. Malka, N. Kosower, M. Rein, G. Zilberman-Schapira, L. Dohnalová, M. Pevsner-Fischer, R. Bikovsky, Z. Halpern, E. Elinav, and E. Segal (2015). Personalized nutrition by prediction of glycemic responses. *Cell 163*(5), 1079–1094.

Zelingher, R., A. Ghermandi, E. De Cian, M. Mistry, and I. Kan (2019, 10). Economic Impacts of Climate Change on Vegetative Agriculture Markets in Israel. *ENVIRONMENTAL & RESOURCE ECONOMICS 74*(2), 679–696.

Zelingher, R., D. Makowski, and T. Brunelle (2021). Assessing the sensitivity of global maize price to regional productions using statistical and machine learning methods. *Frontiers in Sustainable Food Systems 5*, 171.

Zhang, D., G. Zang, J. Li, K. Ma, and H. Liu (2018). Prediction of soybean price in china using qr-rbf neural network model. *Computers and Electronics in Agriculture 154*, 10–17.

Zhang, G. (2003). Time series forecasting using a hybrid arima and neural network model. *Neurocomputing 50*, 159–175.

Appendix A

Forecasting cycle for monthly price of eight agricultural commodities

Table A.1 – Forecasting cycle for monthly price of Arabica coffee

	Time lags (months)											
Month	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Jan	TBATS									GBM (Prod. Region)		
	0.68	0.65	0.54	0.44	0.40	0.36	0.40	0.23	0.08	0.05		
Feb	TBATS								RF (Prod. local + $p_{m,y-1}$)			
	0.54	0.33	0.30	0.21	0.10	0.09	0.08	0.12	0.08			
Mar	TBATS			RF (Prod. Local + $p_{m,y-1}$)								
	0.67	0.28	0.14	0.13								
Apr	TBATS			VAR (Prod. Local + $p_{m,y-1}$)								
	0.82	0.54	0.20	0.08								
May	TBATS			VAR (Prod. Local + $p_{m,y-1}$)								
	0.62	0.50	0.33	0.12								
Jun	TBATS				VAR (Prod. Local + $p_{m,y-1}$)							
	0.66	0.41	0.31	0.22	0.03							
Jul	TBATS				RF (Prod. Region)							
	0.64	0.44	0.20	0.12	0.12							
Aug	TBATS						VAR (Prod. Region)					
	0.60	0.67	0.48	0.26	0.18	0.17	0.07					
Sep	TBATS							RF (Prod. Region)				
	0.78	0.60	0.61	0.45	0.21	0.14	0.13	0.07				
Oct	TBATS							VAR (Prod. Region)				
	0.65	0.59	0.53	0.46	0.27	0.14	0.09	0.09				
Nov	TBATS						GBM (Prod. Region + $p_{m,y-1}$)					
	0.71	0.48	0.42	0.34	0.41	0.23	0.11					
Dec	TBATS							GBM (Prod. Region + $p_{m,y-1}$)				
	0.65	0.47	0.39	0.31	0.21	0.37	0.20	0.14				

Best models over annual production and different geographic scales (continental, regional, local), for horizons of 1 to 12 months, relative to month.

Table A.2 – Forecasting cycle for monthly price of cocoa

	Time lags (months)											
Month	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Jan	TBATS				VAR (Yield. Region)							
	0.69	0.55	0.45	0.27	0.18							
Feb	TBATS					LM (Yield Local)						
	0.61	0.44	0.40	0.37	0.20	0.13						
Mar	TBATS		GBM (Yield Local + $p_{m,y-1}$)									
	0.51	0.40	0.31									
Apr	TBATS	GBM (Yield Local)										
	0.68	0.43										
May	TBATS	GBM (Yield Local)										
	0.59	0.48										
Jun	TBATS				RF (Prod. Local)							
	0.62	0.48	0.34	0.30	0.20							
Jul	TBATS					RF (Prod. Local)						
	0.62	0.46	0.32	0.22	0.21	0.13						
Aug	TBATS						LM (Yield. Region)					
	0.71	0.54	0.39	0.28	0.21	0.22	0.16					
Sep	TBATS							LM (Yield. Region)				
	0.63	0.62	0.43	0.29	0.21	0.13	0.16	0.10				
Oct	TBATS				LM (Yield Local)							
	0.59	0.35	0.37	0.24	0.11							
Nov	TBATS		VAR (Yield. Region)									
	0.49	0.19	0.18									
Dec	TBATS			VAR (Yield Region + $p_{m,y-1}$)								
	0.59	0.41	0.24	0.19								

Best models over different variables (annual production or yield) and geographic scales (continental, regional, local), 1 to 12 months ahead, relative to month.

Table A.3 – Forecasting cycle for monthly price of maize

	Time lags (months)											
Month	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Jan	TBATS			GBM (Yield. Region)								
	0.73	0.64	0.46	0.37								
Feb	TBATS				GBM (Yield. Region)							
	0.75	0.59	0.53	0.38	0.29							
Mar	TBATS					GBM (Yield. Region)						
	0.78	0.59	0.47	0.41	0.3	0.22						
Apr	TBATS				VAR (Prod. Local + $p_{m,y-1}$)							
	0.67	0.55	0.41	0.34	0.25							
May	TBATS				VAR (Prod. Local + $p_{m,y-1}$)							
	0.68	0.64	0.47	0.30	0.23							
Jun	TBATS						GBM (Prod. Region)					
	0.61	0.55	0.48	0.39	0.31	0.24	0.19					
Jul	TBATS				CART (Yield. Region)							
	0.44	0.28	0.29	0.24	0.23							
Aug	TBATS	CART (Prod. Region)										
	0.61	0.31										
Sep	TBATS	LM (Prod. Region)										
	0.46	0.21										
Oct	TBATS	GBM (Yield Local + $p_{m,y-1}$)										
	0.57	0.37										
Nov	TBATS		GBM (Yield Local)									
	0.63	0.36	0.28									
Dec	TBATS			GBM (Yield Region + $p_{m,y-1}$)								
	0.72	0.56	0.33	0.31								

Best models over different variables (annual production or yield) and geographic scales (continental, regional, local), 1 to 12 months ahead, relative to month.

Table A.4 – Forecasting cycle for monthly price of palm-oil

	Time lags (months)											
Month	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Jan	TBATS			GBM (Yield Local)								
	0.71	0.53	0.35	0.37								
Feb	TBATS				LM (Prod. Region)							
	0.77	0.57	0.48	0.36	0.29							
Mar	TBATS					LM (Prod. Region + $p_{m,y-1}$)						
	0.72	0.61	0.45	0.41	0.29	0.22						
Apr	TBATS				VAR (Prod. Local + $p_{m,y-1}$)							
	0.57	0.44	0.41	0.28	0.25							
May	TBATS				VAR (Prod. Local + $p_{m,y-1}$)							
	0.67	0.43	0.31	0.33	0.23							
Jun	TBATS					VAR (Prod. Local + $p_{m,y-1}$)						
	0.72	0.46	0.33	0.18	0.22	0.13						
Jul	TBATS						VAR (Prod. Local + $p_{m,y-1}$)					
	0.54	0.44	0.23	0.37	0.09	0.11	0.08					
Aug	TBATS					VAR (Prod. Local + $p_{m,y-1}$)						
	0.51	0.33	0.29	0.14	0.16	0.09						
Sep	TBATS				VAR (Prod. Region + $p_{m,y-1}$)							
	0.77	0.34	0.33	0.3	0.18							
Oct	TBATS		GBM (Yield Local + $p_{m,y-1}$)									
	0.65	0.55	0.37									
Nov	TBATS			GBM (Yield Local)								
	0.66	0.42	0.43	0.28								
Dec	TBATS				GBM (Yield Local)							
	0.72	0.56	0.33	0.31	0.27							

Best models over different variables (annual production or yield) and geographic scales (continental, regional, local), 1 to 12 months ahead, relative to month.

Table A.5 – Forecasting cycle for monthly price of rice

	Time lags (months)											
Month	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Jan	TBATS		LM (Prod. Region + $p_{m,y-1}$)									
	0.71	0.53	0.42									
Feb	TBATS					VAR (Yield Local + $p_{m,y-1}$)						
	0.77	0.57	0.48	0.36	0.18	0.14						
Mar	TBATS					VAR (Prod. Region)						
	0.72	0.61	0.45	0.41	0.29	0.19						
Apr	TBATS					VAR (Prod. Local + $p_{m,y-1}$)						
	0.57	0.44	0.41	0.28	0.09	0.05						
May	TBATS					VAR (Prod. Local + $p_{m,y-1}$)						
	0.67	0.43	0.31	0.33	0.20	0.11						
Jun	TBATS					GBM (Yield Local)						
	0.72	0.46	0.33	0.18	0.22	0.16						
Jul	TBATS					RF (Yield Local)						
	0.54	0.44	0.23	0.37	0.09	0.09						
Aug	TBATS				GBM (Yield Local)							
	0.51	0.33	0.29	0.14	0.12							
Sep	TBATS						RF (Yield Local + $p_{m,y-1}$)					
	0.77	0.34	0.33	0.3	0.13	0.08	0.05					
Oct	TBATS		LM (Prod. Region + $p_{m,y-1}$)									
	0.65	0.55	0.44									
Nov	TBATS			LM (Prod. Region)								
	0.66	0.42	0.43	0.24								
Dec	TBATS				LM (Prod. Region + $p_{m,y-1}$)							
	0.73	0.45	0.30	0.36	0.19							

Best models over different variables (annual production or yield) and geographic scales (continental, regional, local), 1 to 12 months ahead, relative to month.

Table A.6 – Forecasting cycle for monthly price of Robusta coffee

	Time lags (months)											
Month	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Jan	TBATS											
	0.76	0.62	0.51	0.26	0.17	0.14	0.12	0.12	0.08			
Feb	TBATS									RF (Prod. Local)		
	0.76	0.64	0.54	0.45	0.23	0.14	0.12	0.10	0.09	0.05		
Mar	TBATS											
	0.73	0.53	0.45	0.42	0.35	0.16	0.10	0.11	0.09	0.07	0.04	
Apr	TBATS											
	0.76	0.57	0.39	0.35	0.35	0.31	0.12	0.08	0.08	0.05	0.03	0.02
May	TBATS											
	0.82	0.57	0.42	0.27	0.25	0.26	0.24	0.06	0.03	0.04	0.01	
Jun	TBATS											
	0.80	0.68	0.55	0.37	0.23	0.22	0.22	0.22	0.04	0.03	0.02	
Jul	TBATS											
	0.81	0.69	0.60	0.51	0.36	0.23	0.21	0.19	0.20	0.02	0.01	0.01
Aug	TBATS											
	0.71	0.65	0.64	0.55	0.44	0.36	0.20	0.21	0.16	0.16		
Sep	TBATS				CART (Prod. Local)							
	0.74	0.50	0.44	0.44	0.34							
Oct	TBATS											
	0.47	0.40	0.3	0.28	0.28	0.25	0.13	0.1				
Nov	TBATS											
	0.76	0.37	0.3	0.25	0.23	0.21	0.19	0.07	0.05			
Dec	TBATS											
	0.74	0.59	0.28	0.20	0.18	0.16	0.15	0.11	0.01			

Best models over annual production and different geographic scales (continental, regional, local), for horizons of 1 to 12 months, relative to month.

Table A.7 – Forecasting cycle for monthly price of soybean

	Time lags (months)											
Month	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Jan	TBATS	LM (Yield. Region)										
	0.72	0.53										
Feb	TBATS		VAR (Yield Local + $p_{m,y-1}$)									
	0.56	0.51	0.44	0.4								
Mar	TBATS		RF (Prod. Region + $p_{m,y-1}$)									
	0.63	0.59	0.48									
Apr	TBATS	RF (Prod. Region + $p_{m,y-1}$)										
	0.42	0.38										
May	TBATS	CART (Prod. Region)										
	0.29	0.26										
Jun	TBATS	GBM (Yield. Region)										
	0.53	0.31										
Jul	TBATS	VAR (Prod. Local + $p_{m,y-1}$)										
	0.36	0.28										
Aug	TBATS		GBM (Prod. Local)									
	0.49	0.24	0.19	0.12								
Sep	TBATS		GBM (Prod. Local)									
	0.55	0.32	0.22									
Oct	TBATS		GBM (Prod. Region)									
	0.68	0.46	0.37									
Nov	TBATS		GBM (Prod. Region)									
	0.56	0.40	0.29	0.19								
Dec	TBATS		GBM (Prod. Region + $p_{m,y-1}$)									
	0.64	0.38	0.31	0.24								

Best models over different variables (annual production or yield) and geographic scales (continental, regional, local), 1 to 12 months ahead, relative to month.

Table A.8 – Forecasting cycle for monthly price of wheat

	Time lags (months)											
Month	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Jan	TBATS				GBM (Yield Local)							
	0.79	0.55	0.36	0.32	0.28							
Feb	TBATS					LM (Prod. Region + $p_{m,y-1}$)						
	0.72	0.64	0.47	0.37	0.32	0.19						
Mar	TBATS		GBM (Yield Local + $p_{m,y-1}$)									
	0.51	0.40	0.19									
Apr	TBATS	VAR (Prod. Local + $p_{m,y-1}$)										
	0.68	0.25										
May	TBATS	VAR (Prod. Local + $p_{m,y-1}$)										
	0.59	0.23										
Jun	TBATS				VAR (Prod. Local + $p_{m,y-1}$)							
	0.62	0.48	0.34	0.30	0.13							
Jul	TBATS					LM (Prod. Region)						
	0.62	0.46	0.32	0.22	0.21	0.12						
Aug	TBATS						VAR (Prod. Local + $p_{m,y-1}$)					
	0.71	0.54	0.39	0.28	0.21	0.22	0.09					
Sep	TBATS							CART (Prod. Local)				
	0.63	0.62	0.43	0.29	0.21	0.13	0.16	0.11				
Oct	TBATS				GBM (Yield Local + $p_{m,y-1}$)							
	0.59	0.35	0.37	0.24	0.37							
Nov	TBATS		GBM (Yield Local)									
	0.49	0.22	0.28									
Dec	TBATS			GBM (Yield Local)								
	0.66	0.45	0.45	0.27								

Best models over different variables (annual production or yield) and geographic scales (continental, regional, local), for horizons of 1 to 12 months, relative to month.

Table A.9 – Data sources 1/2

Crop	Agricultural output	Monthly prices
Arabica, coffee	PSD Data Setes (2021)	PSD Data Setes (2021)
Cocoa	FAO STAT (2020)	FAO STAT (2020)
Maize	FAO STAT (2020)	World Bank, Pink Sheet (2020)
Palm-oil	FAO STAT (2020)	World Bank, Pink Sheet (2020)
Rice	FAO STAT (2020)	World Bank, Pink Sheet (2020)
Robusta, coffee	PSD Data Setes (2021)	World Bank, Pink Sheet (2020)
Soybean	FAO STAT (2020)	World Bank, Pink Sheet (2020)
Wheat	FAO STAT (2020)	World Bank, Pink Sheet (2020)

Table A.10 – Data sources 2/2

	References
PSD Data Setes	apps.fas.usda.gov/psdonline
FAO STAT	www.fao.org/faostat
World Bank, Pink Sheet	www.worldbank.org/en/research

université PARIS-SACLAY

ÉCOLE DOCTORALE
Agriculture, alimentation, biologie, environnement, santé (ABIES)

Titre : Prévision des prix des produits agricoles à l'aide de techniques d'apprentissage automatique

Mots clés : Matières premières agricoles, Commerce international, Prévision des prix, Apprentissage automatique, Sécurité alimentaire

Résumé : Serait-il possible de développer un outil de prévision des prix des produits agricoles de base qui soit à la fois précis, interprétable et accessible au plus grand nombre ? Un tel outil permettrait à ceux qui n'ont pas la capacité financière ou le bagage technique appropriés de prévoir les prix des produits agricoles de base, un ou plusieurs mois à l'avance. Ce doctorat explore la faisabilité de cette idée en trois parties : L'objectif de la première partie est de tester la capacité de plusieurs modèles statistiques et d'apprentissage automatique à simuler les variations du prix du maïs en fonction des variations annuelles de production et de rendement du maïs observées dans les principales régions productrices. Dans la deuxième partie de la thèse, les modèles développés dans la première partie sont adaptés pour effectuer des prévisions mensuelles de prix du maïs. Nous comparons les performances de ces modèles à celles de techniques prédictives souvent utilisées pour l'analyse des séries chronologiques. Enfin, dans la troisième partie, nous étendons le travail réalisé sur le maïs à deux autres cultures très différentes - le et le cacao. Nous analysons la capacité des techniques de prévision mises au point dans la partie précédente à prédire les variations de prix du soja et du cacao et nous analysons également l'effet de l'échelle géographique considérée pour calculer les variations de production. Dans cette partie également, nous montrons comment les méthodes d'apprentissage machine peuvent être utilisées pour identifier les chocs de production à l'origine des chocs de prix. Globalement, cette thèse montre que les méthodes d'apprentissage automatique sont des outils potentiellement utiles à la fois pour comprendre l'impact de la production agricole sur les variations de prix et pour prédire ces variations plusieurs mois à l'avance. Ces approches sont assez faciles à appliquer et peuvent être calibrées avec des données de prix et de production publiquement accessibles. Elles peuvent ainsi contribuer à démocratiser l'analyse et la prévision des variations de prix agricoles.

Title : Agricultural Commodity Price Forecasting Using Comprehensive Machine-Learning Techniques

Keywords : Agricultural commodities, International trade, rice Forecasting, Machine Learning, Food security

Abstract : Would it be possible to develop a forecasting tool for agricultural commodity (AC) prices that is both accurate and interpretable and publicly accessible? Such a tool could turn the forecasting and analysis of food prices into an implementable instrument used by whoever is concerned by food security. This PhD explores the feasibility of this idea in three parts: The first part aims to test the ability of several statistical and machine learning (ML) models to simulate changes in maize prices based on annual changes in maize production and yield observed in major producing regions. The second part of the thesis applies the models developed in the first part and adapt them to produce monthly forecasts of maize prices. We compare the performance of these models to that of forecasting techniques often used for time series analysis. Finally, the third part extends the model to consider two other different crops – soybeans and cocoa. We evaluate the forecasting ability of the techniques developed in the previous stages to predict price changes for soybeans and cocoa. Additionally, we test the sensitivity of the results relative to three geographic scales. Also is the application of ML methods to identify which production shocks drive price shocks. Overall, this thesis shows that ML methods are a potential tool for understanding and forecasting the impact of agricultural production on price variations. These approaches can be easily implemented since they rely on publicly available data, accessible via public website. These tools can thus contribute to democratising the analysis and forecasting of variation in AC prices.

Maison du doctorat de l'Université Paris-Saclay
2ème étage aile ouest, Ecole normale supérieure Paris-Saclay
4 avenue des Sciences,
91190 Gif sur Yvette, France