Forecasting impacts of Agricultural Production on Global Maize Price

Rotem Zelingher, David Makowski, Thierry Brunelle

Forecasting impacts of Agricultural Production on Global Maize Price

Rotem Zelingher[1],David Makowski[2,3],Thierry Brunelle[3][1]

[1] *Université Paris-Saclay, INRAE, AgroParisTech, Economie Publique, 78850, Thiverval-Grignon, France*
[2] *UMR 318 INRAE AgroParisTech Université Paris-Saclay 75231 Paris France*
[3] *CIRAD, UMR CIRED, F-94736 Nogent-sur-Marne, France*

Abstract

Agricultural price shocks strongly affect farmers' income and food security. It is therefore important to understand the origin of these shocks and anticipate their occurrence. In this study, we explore the possibility of predicting global prices of one of the world main agricultural commodity - maize - based on variations in regional production. We examine the performances of several machine-learning (ML) methods and compare them with a powerful time series model (TBATS) trained with 56 years of price data. Our results show that, out of nineteen regions, global maize prices are mostly influenced by Northern America. More specifically, small positive production changes relative to the previous year in Northern America negatively impact the world price while production of other regions have weak or no influence. We find that TBATS is the most accurate method for a forecast horizon of three months or less. For longer forecasting horizons, ML techniques based on bagging and gradient boosting perform better but require yearly input data on regional maize productions. Our results highlight the interest of ML for predicting global prices of major commodities and reveal the strong sensitivity of global maize price to small variations of maize production in Northern America.

Keywords: Food-security, Maize, Agricultural commodity prices, Regional production, Machine learning

1 Introduction

In a context of population growth, evolving diet and climatic changes, the four components of food security - availability, stability, utilisation, and access - have become a vital matter. High levels of volatility in the food prices affect all aspects of food security, and pose a growing number of households population under uncertainty (Rosenzweig et al., 2001; Schmidhuber and Tubiello, 2007). At the turn of 2010, prices of main food crops in the international markets have shown high variability, sometimes doubling in the time frame of only a few years (Headey and Fan, 2010). For example, the price of maize increased by 75% from September 2007 to May 2008 (Headey, 2011). Poor harvest 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).

Reliable pre-market information concerning food-prices would enable decision makers to redesign their business and food-security strategies in a way that could mitigate the potential damage of price shocks or adapt to high levels of volatility (Barrett, 1997). Forecasts of price commodities can help governmental organisations and private institutions optimising purchase dates and help grain trading companies managing their stocks in an efficient way.

The economic and social risks associated with agricultural commodity price shocks have increased the need for crop price forecasting models that could be both accurate and accessible, especially for maize. Maize is indeed of great importance for global food security as it is one of the most consumed cereal both by human and livestock worldwide. Maize grains are also a major source of biofuel due to their high starch content

and relatively easy conversion to ethanol. Although maize is the most widely traded crop in the world, only few countries export their production, suggesting that maize price might be impacted by the production of a small number of regions. As some countries rely heavily on maize imports to ensure food security (Wu and Guclu, 2013; Rouf Shah et al., 2016), it is important to be able to anticipate price shocks for this commodity.

The objective of this study is to examine two main issues in the global maize market: which producing regions have the largest influence over the world price of maize through their production compared to others; and whether the global price of maize could be forecasted using regional production only. Large public institutions such as United States Department of Agriculture (USDA)[1], Food and Agriculture Organization (FAO)[2], Organisation for Economic Co-operation and Development (OECD) and the World-Bank[3]) as well as private trading companies currently forecast global prices of major crops at different time horizons. In public institutions, forecasting models are usually based on structural approaches and rely on economic-theory assumptions. FAO and OECD use a recursive-dynamic partial equilibrium model named Aglink-Cosimo (Gay, 2015). The great advantage of this model comes from its relatively simple structure of their easy-to-calculate equations. The World-Bank provides a general equilibrium model able to project annual prices for the horizon of between one year and up to 15 years ahead (Bank, 2019). In this model, crop-specific sub-models are run for all major producing countries and most of the consuming countries. Another price forecasting model is the World Agricultural Supply and Demand Estimates (WASDE) model used by USDA. This season-average price projection model is able to provide USDA staffs and policymakers with monthly projections over a period of between 1 to 16 months ahead (Hoffman et al., 2015).

The models of the World Bank and of USDA provide relatively short-term projections[4] relevant for many stakeholders. For example, the WASDE forecasts are used for risk calculation and design of the federal US crop insurance program (of Representatives, 2009). But these models were criticised because of their high complexity (Hoffman and Meyer, 2018) and, sometimes, because of their lack of accuracy (Hoffman, 2011; Warr, 1990; Hoffman et al., 2015; Lusk, 2016). Other forecasting models are run by private institutions, in particular by companies specialised in commodity trading. These tools are generally unpublished, not freely available, and are often considered as black boxes. It is thus difficult to know precisely their levels of performance. Yet, they were considered as not significantly more accurate than WASDE by Miftakhova and Pohl (2019).

Auto-regressive methods are widely used to forecast food price in the academic literature (Belke et al., 2013; Shively, 1996). These are well-established methods, but they could now be challenged by recent development in machine learning method. As long-time average price forecasts are not relevant for those who seek short-term predictions to handle price volatility, we focus here on monthly maize price projection over the course of the growing season. Such predictions are useful, in particular, for decision makers who need to optimise both their dates of commodity purchases and their stock usages (Bank, 2005). Monthly price forecasts are often considered as highly relevant for optimising this type of decision, especially as maize prices tend to change on a monthly basis (Ochieng and Baulch, 2019; Dorosh et al., 2004), partly due to weather-related supply shocks. Another advantage of this type of projections is that they can be derived from public price and production time-series. As these data are freely available, they can be used to develop free and transparent forecasting tools that can be implemented by a large range of stakeholders.

In this study, we predict the monthly average global price of maize using machine-learning (ML) techniques and statistical models trained on publicly available regional production and price data. Productions and yields directly inform on the level of commodity supply which is usually an unstable component of the market. These variables can thus be potentially useful for predicting crop prices. However, productions and yields are rarely used as predictors in econometric analysis due to a risk of endogeneity. In this paper,

[1]USDA Food Price Outlook is available and updated regularly at: `https://www.ers.usda.gov/data-products/food-price-outlook.aspx`

[2]The Market Monitor of the Agricultural Market Information System (AMIS) is available and updated regularly at: `http://www.amis-outlook.org/fileadmin/user_upload/amis/docs/Market_monitor/AMIS_Market_Monitor_current.pdf`

[3]World Bank Commodities Price Forecast is available and updated regularly at: `https://www.worldbank.org/en/research/commodity-markets`

[4]Popkin (1977) price forecasts covering three time horizons: (1) Near-term: up to three months; (2) Short-term: changes that occur between three to eighteen months ahead ; and (3) Long-term: any time horizon of more than a year and a half into the future.

we address this issue in two ways. We focus our analysis on end-of-season maize prices i.e., when all crop management decisions have been already made. Then we investigate the existence of a causal relationship between prices and production or yields using Granger causality tests.

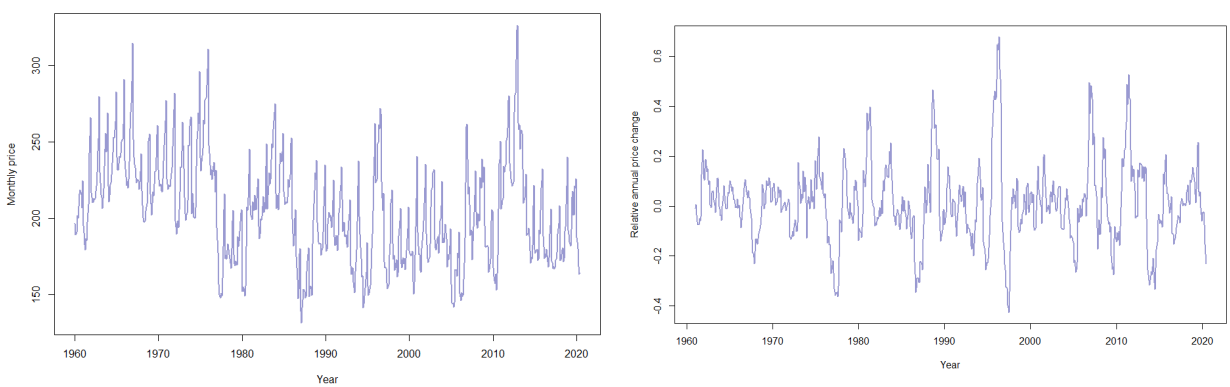
Regional productions determine the supply levels of the market and can thus potentially inform on price variations. However, prices depend on other factors as well, in particular on the demand levels, that are difficult to anticipate. So far, it has not been demonstrated that maize price variations could be predicted from maize production changes. Here, we analyse the relationships between regional production or yield and global prices and identify the most and least influential regional productions in the maize global market. We show that our ML tools are able to rank regional productions according to their influence on global maize prices and to predict maize prices a few months ahead. Our ML tools have an accuracy similar or higher than advanced time-series statistical models depending on forecast time horizon. Our results contribute in making the forecasting of global price of maize more accessible and, in this way, could help improve food security worldwide.

2 Materials and method

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.

Data on maize global monthly price were extracted from the World Bank's commodity markets database as a US No. 2 yellow free of board (FOB) Gulf of Mexico, U.S. nominal price. The time series of monthly price summarises over the counter (OTC) trading at settlement contracts from Chicago Mercantile Exchange (CME) from January 1960 to December 2019. We converted these prices into real 2010 USD, using the monthly agricultural index of the World-Bank[5] (Figure 1a).



(a) Real price, USD

(b) Relative annual changes

Figure 1: Time series of global maize price

The real prices are further denoted to as $q_{m,y}$, where m and y are the month and year indices, respectively. Maize crops are harvested once a year and levels of maize production can thus potentially have strong effects

[5]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 US CPI could be a biased deflater when dealing in a global market that includes both developed and developing countries. The second reason is a relatively smaller gap (RMSE) between the maize annual real prices as published by the World-Bank to the real maize global monthly price calculated for this study.

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

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

and their values are shown in Figure 1b. 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 predicted as a function of relative production (or yield) changes between the month m in year y and the same month in year $y-1$. Regional yield (grain weight per unit of cropping area, in tons per ha) and production (total regional grain weight, in tons) data were transformed in order to calculate the relative yield and production changes compared to the previous year, as follows:

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

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

We chose to forecast prices during the last quarter of each year, that is in October, November, and December (i.e., $m \in$10,11,12). These months were chosen because, in most regions, maize produced in a given calendar year is already harvested at this period. It is thus possible to obtain accurate maize yield and production statistics from October onwards, and to use them for price forecasting [6]. In the next sections, we present and compare several methods to predict $p_{m,y}$ 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.

2.2 Linear and generalised linear models

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

$$p_{m,y} = \alpha + \sum_{k=1}^{19} \beta_k x_{k,y} + \epsilon_{m,y} \tag{2}$$

where α and β_k are regression parameters and $\epsilon_{m,y}$ are residuals. We also define a variant of this model including the price change of year $y-1$ (i.e., $p_{m,y-1}$) as an additional input. This variant was used to investigate 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 $x_{k,y}$.

For classification, we use a generalised linear model (GLM) with a binomial family and logit link. This model computes the probability that $p^b_{m,y}$=1 (i.e., probability of 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. As done with the other methods, we fit linear models for each month (October, November, December) using successively production changes and yield changes as inputs.

2.3 CART

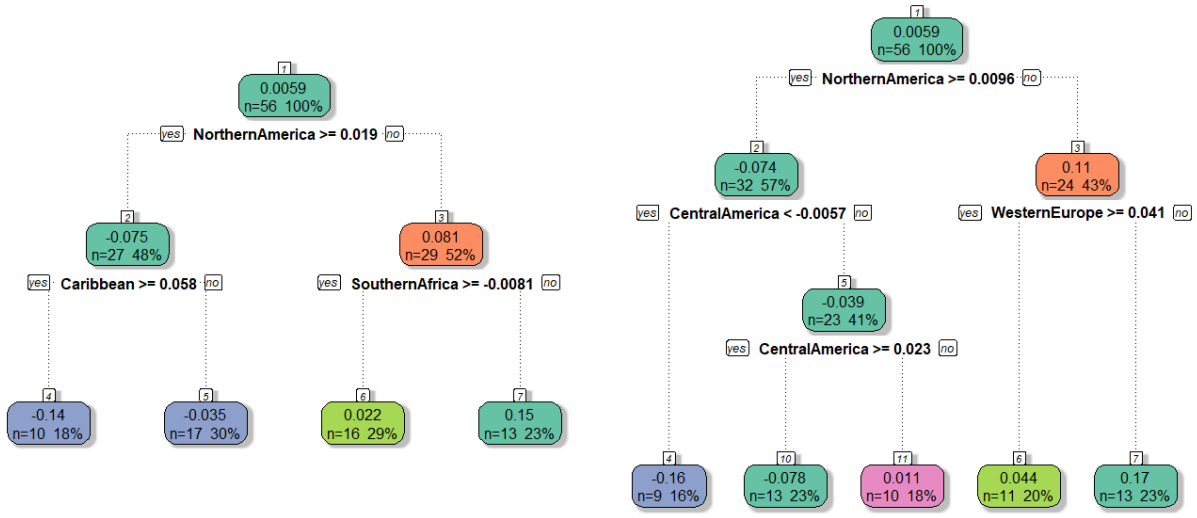
The three ML methods are decision-tree based algorithms: classification and regression trees (CART), Random-forest (RF), and gradient boosting machine (GBM). None of these methods make any strong assumption about the functional form of the relationship between the dependent variable and the explanatory

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

variables, neither about the data distribution. We shortly present our implementation of CART here. RF and GBM are presented in the next sections.

The purpose of CART is to build a binary decision tree as follows. Let $p_{m,y}$ be a dependent variable and $x_{1,y}, x_{2,y}, ..., x_{19,y}$ 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, whereas higher valued observations belong to the right branch and lower valued to the left branch. Additional distributions based on the same variable can be made, but at each stage one cut-off point is determined. The subgroups that define the tree are called nodes. Using a looping technique, CART performs recursive partitioning, or rather 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 this case, 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 is implemented with the package `rpart` of the R software (Therneau et al., 2019).

For illustration, Figure 2 shows an example of tree obtained for predicting $p_{m,y}$ in October as a function of the regional production (or yield) changes. This tree has five final nodes defined by three (2.a) or four (2.b) inputs corresponding to different regions. The tree root (the upper rectangle in the centre of the diagram) includes 56 observations (i.e., the whole dataset) with an average $p_{10,y}$ of 0.59%. Referring to figure 2.a, after the algorithm examined all possible partitions according to the set of input variables, it is found that the maximum reduction of RSS was 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 higher 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 the average observed price change corresponding to five different production (or yield) situations. These predictions correspond to the average price changes reported in the final nodes. Here, the fitted tree produces four different predictions determined by the values of three inputs.

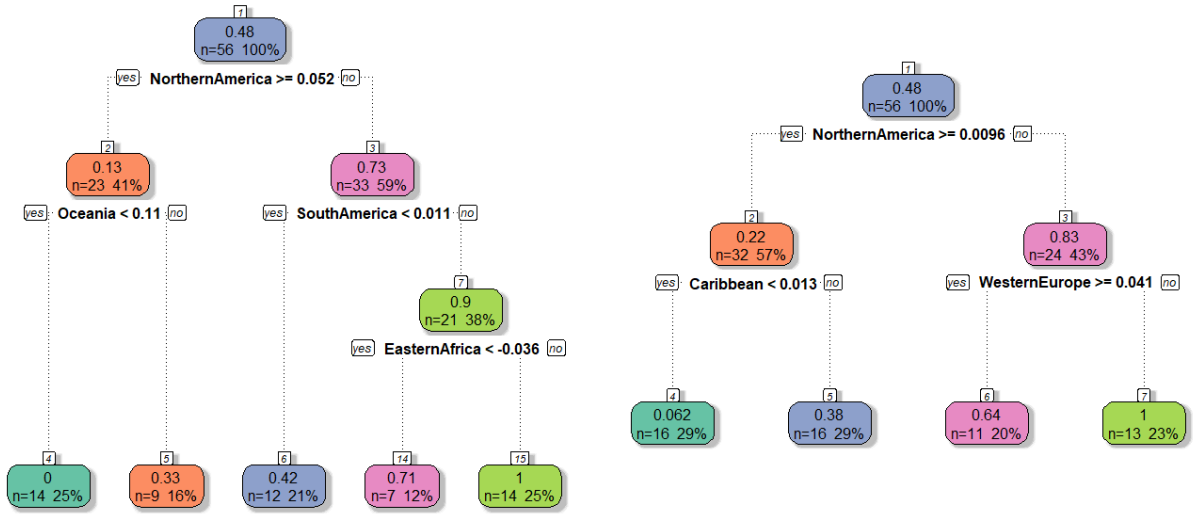


(a) Inputs=relative regional production changes

(b) Inputs=relative regional yield changes

Figure 2: CART models for $p_{10,y}$ 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

Figure 3 shows the tree obtained for classifying price in October into two categories, i.e., price increase or price 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 North-American production increases by more than 5.2% and that of Oceania grows by less than 11%. In this case, the probability of price decrease is estimated at 25%.



(a) Inputs=relative regional production changes

(b) Inputs=relative regional yield changes

Figure 3: CART models for 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

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 changes in data. Their predictions are not always very 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 chose to use Random-forest (RF) as a bagging-based algorithm, and gradient boosting machine (GBM) as a boosting-based method.

The RF algorithm builds an ensemble of trees, each of them relying on a small subset of inputs (i.e., a subset of the 19 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 giving the opportunity for 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 is able to rank the inputs according to their predictive powers and, here, the resulting ranking can be used to identify the regions whose maize productions show the strongest influence on maize global price. In this study, RF is implemented with the package `randomForest` (Leo Breiman and Wiener, 2018), both for quantitative predictions and for classification.

The method GBM is also based on an ensemble of trees (Bradley Efron, 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 expressed as 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, GBM is implemented with the `gbm` package Friedman (2001) both for regression and classification based predictions. As for RF, we find that the most accurate results are obtained with 500 trees for GBM.

Neither RF or GBM have analytical expressions, but standard methods can be used to rank their inputs according to their importance and visualise their effects on the output, here 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 of the input variables 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 over all values of the other inputs. These plots allow us to analyse the shapes of the responses and to detect non-linearity.

2.5 TBATS

The Trigonometric Seasonal Box Transformation with ARMA residuals Trend and Seasonal Components (TBATS) model (De Livera et al., 2011) is a recent and sophisticated time-series model able to deal with trends, seasonality and auto-correlations. This method automatically determine 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 auto-regressive and moving average components for predicting the target response variable.

We use TBATS for predicting price changes from past price change monthly data, without using inputs $x_{k,y}$ related to productions or yields. Indeed, TBATS allows us to predict price changes directly from the past series of observed price changes. We consider several time horizons for price change predictions, from one month ahead to one year ahead. Note that TBATS can only be used for quantitative prediction, not for classification. This method is implemented with the R package `forecast` (Hyndman et al., 2020).

2.6 Models Evaluation

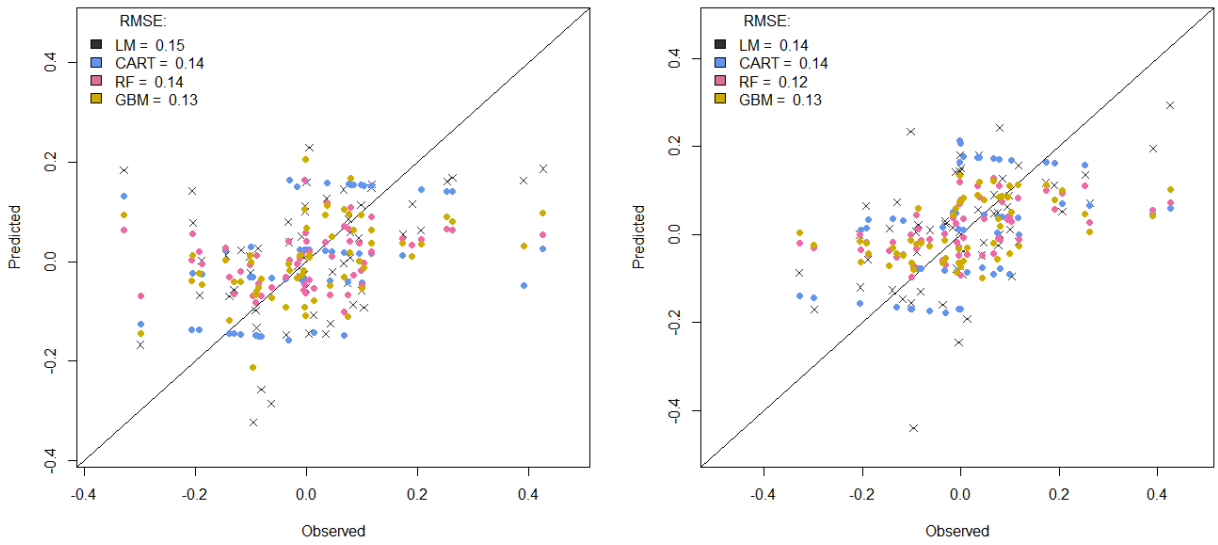
The accuracy of the quantitative predictions of relative price changes is assessed by computing root mean squared error (RMSE). For CART, RF, GBM, and GLM, the values of RMSE are estimated by leave-one-out cross-validation (LOOCV). One year of price ($p_{m,y}$, m=10,11,12) and of production/yield ($x_{k,y}$) is extracted from the original data set. Then, the models CART, RF, GBM, and GLM are trained using the remaining 55 years and the removed value of $p_{m,y}$ is predicted using the trained models. The procedure is performed 56 times - once for each year - to obtain a set of 56 predictions for each tested model and each month (m=10,11,12). Finally, a value RMSE is calculated for each model and each predicted month. The whole procedure is repeated twice, using regional maize production and regional maize yields as inputs, successively.

We assess the accuracy of the classifications performed by the same four models using another criterion of a similar procedure, namely 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. In addition to the AUC, several additional criteria were considered for assessing the classification models, namely sensitivity (true positive rate), specificity (true negative rate) and accuracy (proportion of correctly classified price change data). These three criteria were estimated for a classification rule base of a decision threshold of 0.5 (i.e., a price change data was classified in the category "increase" when the probability computed by the models was higher than 0.5).

The accuracy of TBATS is evaluated by computing the RMSE criterion for different types of forecasts corresponding to 12 different time horizons, i.e. h=1,2,...,12 months ahead. For a given year, a given month (m=10,11,12), 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 $p_{m,y}$. This procedure is repeated for every year, every months (m=10,11,12), and every time horizon. Then, a specific value of RMSE is computed for each month m and time horizon h combination by averaging the prediction errors among the 56 years.

3 Results

3.1 Quantitative predictions of price changes



(a) Inputs = relative regional production changes

(b) Inputs = relative regional yield changes

Figure 4: Observed relative price change vs. Predicted relative price change in October. Values of RMSE are reported for the different models.

Figure 4.b shows that CART, RF and GBM are better predictors of global price change in October ($p_{10,y}$) compared to the linear regression model. RF and GBM show a lower RMSE (0.12 and 0.13 accordingly) than LM, in particular when the predictions are derived from relative yield changes. Performances of all methods are closer when prices are predicted from relative production changes (Figure 4.a). The assessment results of the model predictions in November and December are shown in Figures 24, 25 (Supplementary E), and these results confirm that the tree-based methods tend to be slightly more accurate than LM.

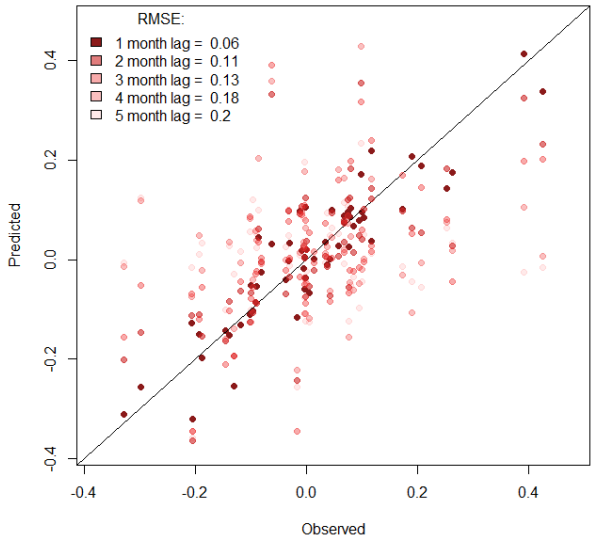
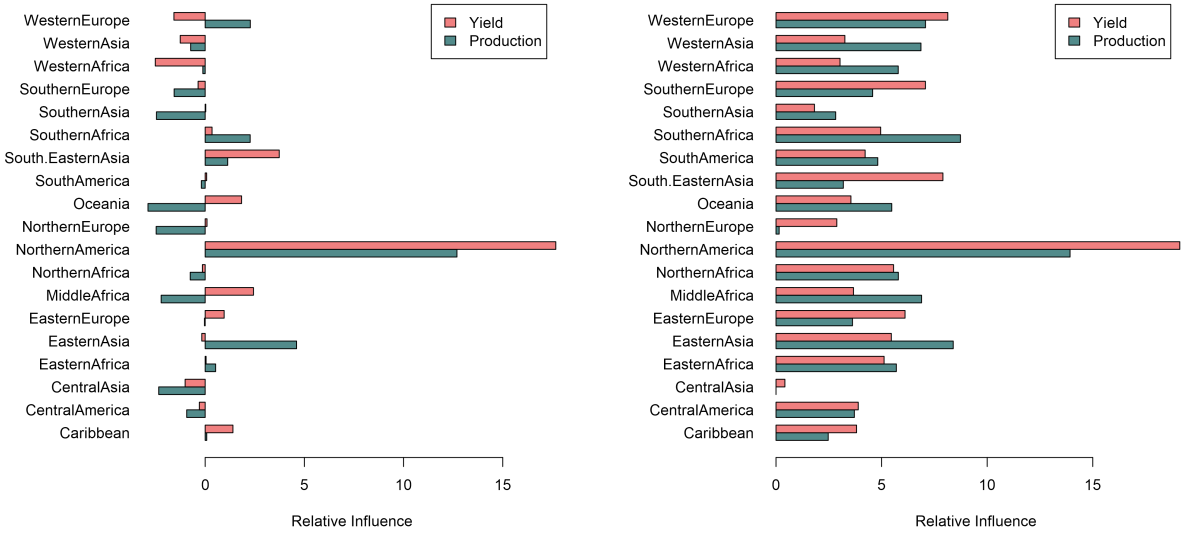


Figure 5: Observed relative price change vs. Predicted relative price change in October. Predictions were obtained with TBATS for different time lags. Values of RMSE are reported for time lags ranging from one to seven months.

Figure 5 compares observed relative price changes in October to TBATS predictions, for time lags ranging from one to five months. Clearly, the predictions of TBATS are more accurate when they are derived considering a short time lag. More specifically, the RMSE of TBATS is as low as 0.06 for a time lag of one month but reaches 0.2 for a time lag of five months. This forecasting error increment is logical as TBATS predicts price changes as a function of previous price change data observed up to the considered time lag. Predictions derived for a time lag of five months are thus computed using the data observed five month before the predicted date, while those derived for a time lag of one month are computed using the data observed just one month before the predicted date (i.e., price changes in October are predicted from price changes in September when the time lag is of one month). Compared to the tree-based models, TBATS is more accurate with a time lag of one or two months (RMSE ranging from 0.06 to 0.11) but becomes less accurate as soon as the time lag exceeds two months (Figure 5). Similar results were obtained for price changes in November and December; TBATS was more accurate than the other forecasting methods when the time lag was not bigger than 3 months but less accurate otherwise.

As RF and GBM show slightly better performances, we analyse in detail the importance ranking of their inputs (Figure 6). Clearly, the production (or yield) changes of maize in Northern America is by far the most influential input with both methods and the two types of inputs considered. The second most influential region is Oceania, Western Europe, Southern Europe, or South-Eastern Asia, depending on the method and the type of inputs considered (Figure 6). With RF, the importance measures obtained for some regions tend to be slightly negative revealing than these regions do not bring any useful information for predicting global price changes.

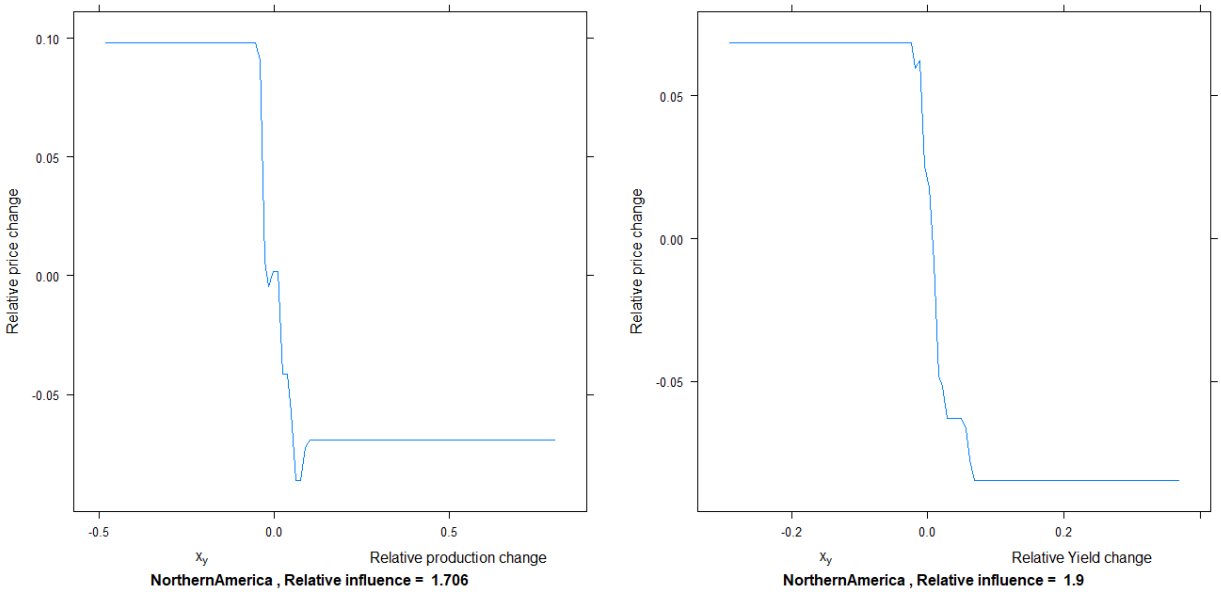


(a) Input ranking for RF

(b) Input ranking for GBM

Figure 6: Importance of the inputs of RF (a) and GBM (b) models predicting relative price change in October as a function of relative regional production changes and relative regional yield changes. Relative influences are computed using the Mean Decrease Accuracy indicator which measures the extent to which the model accuracy decrease with a random permutation of each input.

The partial dependence plot (PDP) shown in Figure 7 shows the average response of price change in October as a function of changes of Maize production (7.a) and yield (7.b) in the most influential region, i.e., Northern America, both according to GBM model (RF based PDP's are in Supplementary F). The plot shows that any increase (decrease) of production or yield in Northern-America leads to a decrease (increase) of global price. More precisely, a 6% rise of relative maize production in Northern-America leads to a reduction of maize price of 8%, and an 5% decrease of maize production in that same region results in an 8% increase of maize price. Similarly, a positive yield change of 5% in Northern-America drives a drop of 6% in price change, whereas a decrease of merely 1% in the Northern American yield causes the global price to increase in more than 7%. The strong effect of Maize production change in Northern America is consistent with the results of the linear regression that showed that the effect of production change in Northern America is significant ($p<0.01$) in October, November, and December, even when the price change in year y-1 is included as an additional explanatory input. The latter result suggests a Granger causal relationship between global price change and production change in Northern America.



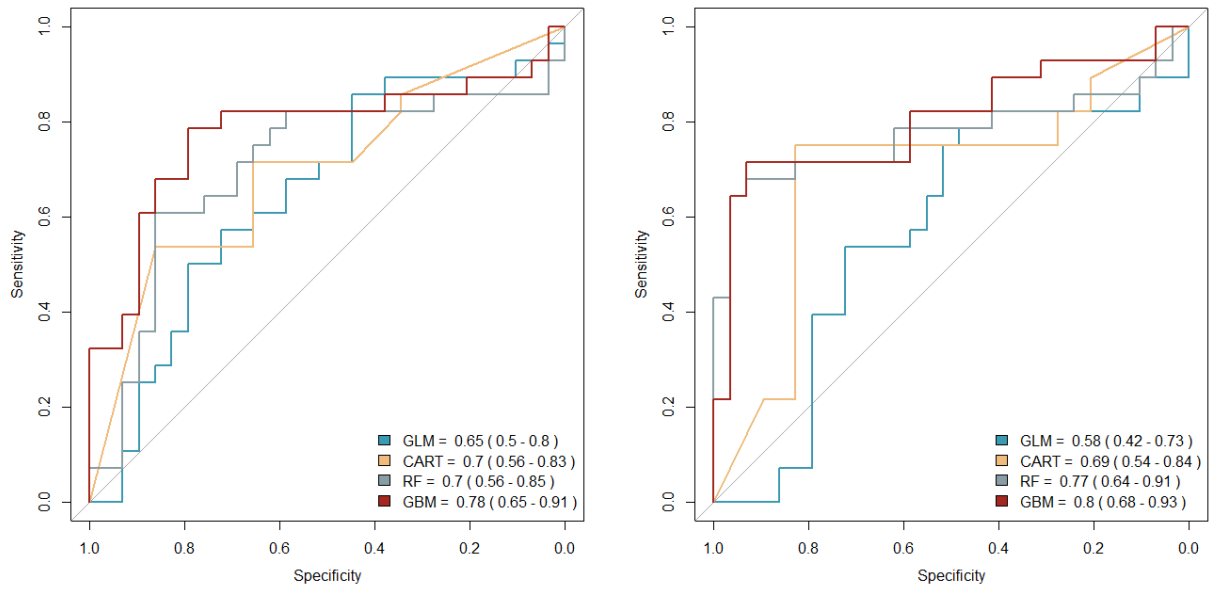
(a) Inputs = relative regional production changes

(b) Inputs = relative regional yield changes

Figure 7: Partial dependence plots obtained with GBM showing the average response of relative price change in October to relative production change in Northern America (a) and to relative yield change in Northern America (b).

3.2 Classification of price increase vs. decrease

Figure 8 shows the results that ROC analyses done for the models for classifying price increase vs. price decrease. Here also, the results are in favour of GBM, RF, and CART. For both input changes (production and yield) the best method is GBM with an AUC of 0.78 (relative to production) and 0.8 (relative to yield)). The 95%CI are relatively large but those obtained with RF and GBM never include the benchmark value 0.5 characterising random classifications. Based on these results, we conclude that RF and GBM consistently perform better than random classification. The additional classification criteria considered in Tables 1 and 2 indicate that tree-based methods (CART, GBM and RF) perform better than GLM in terms of sensitivity, specificity and accuracy, in particular when price change classifications are performed from relative production changes.



(a) Inputs = relative regional production changes

(b) Inputs = relative regional yield changes

Figure 8: ROC curves obtained for the classification models predicting price increase vs. price decrease in October. Each ROC curve relates "Sensitivity" (True Positive Rate) to "1-Specificity" (1-False Positive Rate). The area under the curve (AUC) is reported for all models. AUC=0.5 for a random classification. AUC=1 for a perfect classification. Values between brackets indicate the 95% confidence intervals

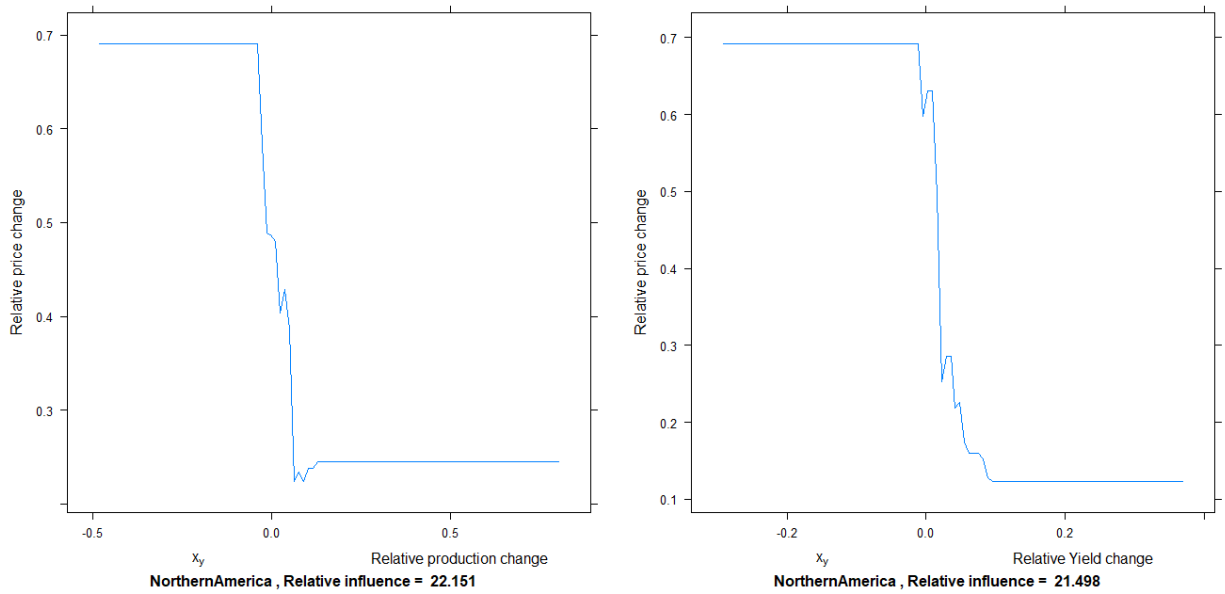
Table 1: Contingency tables, a (ACC), sensitivity (Sens) and specificity (Spec) of the classification models predicting global maize price increase vs. decrease in October. Inputs = relative regional production changes

	GLM		CART		RF		GBM	
	Predicted		Predicted		Predicted		Predicted	
Obs.	$p^b_{10,y}$=1	$p^b_{10,y}$=0	$p^b_{10,y}$=1	$p^b_{10,y}$=0	$p^b_{10,y}$=1	$p^b_{10,y}$=0	$p^b_{10,y}$=1	$p^b_{10,y}$=0
$p^b_{10,y}$=1	12	17	10	15	9	20	6	21
$p^b_{10,y}$=0	17	11	19	13	20	8	23	7
	Sens	59%	Sens	60%	Sens	69%	Sens	78%
	Spec	61%	Spec	59%	Spec	71%	Spec	77%
	ACC	60%	ACC	60%	ACC	70%	ACC	77%

Table 2: Contingency tables, accuracy (ACC), sensitivity (Sens) and specificity (Spec) of the classification models predicting global maize price increase vs. decrease in October. Inputs = relative regional yield changes

	GLM		CART		RF		GBM	
	Predicted		Predicted		Predicted		Predicted	
Obs.	$p^b_{10,y}$=1	$p^b_{10,y}$=0	$p^b_{10,y}$=1	$p^b_{10,y}$=0	$p^b_{10,y}$=1	$p^b_{10,y}$=0	$p^b_{10,y}$=1	$p^b_{10,y}$=0
$p^b_{10,y}$=1	12	15	8	21	5	20	7	20
$p^b_{10,y}$=0	17	13	21	7	24	8	22	8
	Sens	56%	Sens	72%	Sens	80%	Sens	74%
	Spec	57%	Spec	75%	Spec	75%	Spec	73%
	ACC	56%	ACC	74%	ACC	77%	ACC	74%

Figure 9 shows the average responses of the classification GBM model outputs (i.e., probability of price increase) to relative production and yield changes in Northern America (RF based PDP's are in Supplementary F). Clearly, the probability of global price increase shows a very strong decreasing trend as soon as the production (or yield) change is positive in Northern America. This probability falls below 0.3 when the production change exceeds +5% and falls below 0.2 when the yield change exceeds +10%.



(a) Inputs = relative regional production changes

(b) Inputs = relative regional yield changes

Figure 9: Partial dependence plots obtained with GBM showing the probability of price increase in October as a function of relative production change in Northern America (a) and relative yield change in Northern America (b).

4 Discussion

Our study is one of the first to compare such a wide variety of methods for predicting late-season maize price changes. Our analysis shows that price prediction accuracy strongly depends on the chosen algorithm. Among all the considered modelling techniques, ML tree-based techniques show a better root mean squared error than TBATS when the forecasting horizon is longer than 3 months. However, machine-learning techniques require more information to perform the prediction TBATS. The nature of the inputs used by these two methods is very different. With TBATS, price changes are predicted from past monthly data of price changes using sophisticated time series techniques accounting for trends, seasonality, and auto-correlation. With LM, CART, GBM and RF, price changes are predicted from regional production or yield changes using an ensemble of regression or classification trees. The performances of these methods are only marginally impacted by the nature of their inputs (i.e., production vs. yield changes).

Compared to TBATS, an advantage of GBM and RF is that these algorithms allow one to perform a sensitivity analysis of the price change predictions to the 19 regional production inputs through the computation of importance measures. Our results indicate that the input variable that has the most influence on maize prices is, by far, maize production in Northern America. Clearly, a small increase (decrease) of maize production in this region can lead to a substantial decrease (increase) of the global price. This result is somewhat expected as Northern America (and, more specifically, USA) is the main maize producer and exporter at the global scale and as USA is known to have strong influence on the agricultural trade market (Chatzopoulos et al. (2019)). However, our models are able to provide data-driven quantitative information on the effect of regional production variations on global maize prices. Surprisingly, both GBM and RF do not perform better when regional production variations are used as inputs instead of yield variations. This is despite the fact that productions data combine two types of information, i.e., yields and cropping areas, whether yield variations alone do not account for possible variations in the regional maize cultivated areas.

On a practical point of view, a disadvantage of the ML tree-based models compared to TBATS is that all require 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. Considering the maize growing season, it is not realistic to get reliable regional production data before the end of summer, at least in regions located in the Northern hemisphere, in particular in Northern America, a key region for predicting global maize price. For this reason, all models, apart from TBATS, were used here to predict maize prices at the end of year, more specifically in October, November, and December. There is no such restriction for TBATS because this method does not use input production data. This method can thus be used to predict global maize prices at any month during the growing season. However, our results show that the accuracy of TBATS predictions is highly dependent on the time lag, i.e. on the period of time between the data price change data used to train TBATS and the date at which price is predicted. In particular, our results show that the accuracy of TBATS predictions become substantially lower for time lags longer than three months.

Compared to other model types, GBM, RF and TBATS have several advantages but, also, a few disadvantages. Private forecasting models are typically updated every few minutes according to real-time trading data. Although they are usually based on simple models, they are unpublished, not freely available and not transparent. Structural models constitute another category of models able to predict 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 operational tool for maize price predictions. Similarly to our models, WASDE is able to forecast maize price at a monthly time step. According to Hoffman et al. (2015), WASDE projections for December have a RMSE of 0.19, which is larger than our RMSE of 0.13. The differences are even larger for predictions in October and November (RMSE of WASDE equal to 0.26 and 0.34, respectively, for WASDE vs. about 0.12-0.14 for RF and GBM, respectively). In addition, WASDE relies on the combination of nine different structural and non-structural sub-models while TBATS, 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. Moreover, in the future, these models could be adapted to predict price changes for other agricultural commodities.

In addition to being able to quantitatively predict price changes, 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. Here also, the tree-based models tend to outperform the simpler GLM model, at least according to some of the considered classification criteria. Still, the rate of misclassification is approximately 25% with GBM and RF, which is relatively high but better than a random classifications. As already noticed for quantitative predictions, the production change in Northern America is, by far, the most influential input for classifying price increase vs. price decrease situations. All these results concur to show that maize production change in Northern America is a highly relevant indicator for assessing the risk of global maize price increase or decrease.

The methods developed in this paper could be replicated for other crops whose production is less geographically concentrated. This would allow us to assess the world food price sensitivity to production shocks or to an export ban in a given country. Here, we focused on the relationship between world maize price and regional production, but other variables could clearly have an influence on the world maize price. In particular, the demand for biofuel (which is itself spur by oil price) can be an important driver, as maize grains are widely used to produce ethanol.

5 Conclusions

This study demonstrates that it is possible to predict the monthly average global price of maize using machine-learning (ML) techniques and advanced time series models trained on publicly available regional production and price data. As these methods can be easily implemented using freely available packages, our results contribute in making the forecasting of global price of maize more accessible and, in this way, could help improve food security worldwide. In addition to their relatively good predictive and classification performances, several of the methods considered are able to rank regional producers according to their influence on global maize prices and our results show that, out of nineteen regions, Northern America is by far the most influential. 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 maize global price while production of other regions have weak influence. Our study highlights the potential interest of ML for predicting global prices of major commodities from regional production and assessing price sensitivity to crop producers.

References

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

Bank, W. (2019, Jan). Global economic prospects : Darkening skies (english). Technical report, World Bank, Washington, D.C.

Barrett, C. B. (1997). Heteroscedastic price forecasting for food security management in developing countries. *Oxford Development Studies 25*(2), 225–236.

Belke, A., I. G. Bordon, and U. Volz (2013). Effects of global liquidity on commodity and food prices. *World Development 44*, 31 – 43.

Bradley Efron, T. H. (2016, 7). *Computer Age Statistical Inference Algorithms, Evidence, and Data Science* (1 ed.). The address: Cambridge University Press.

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

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.

Chatzopoulos, T., I. P. 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.

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.

Dorosh, P. A., K. Subbarao, and C. Del Ninno (2004, Nov). 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.

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

Gay, H. (2015). Aglink-cosimo model documentation a partial equilibrium model of world agricultural markets. Technical report, FAO, OECD.

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

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

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

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.

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.

Hoffman, L. and L. Meyer (2018). Forecasting the us season-average farm price of upland cotton: Derivation of a futures price forecasting model.

Hoffman, L. A. (2011). *Using Futures Prices to Forecast US Corn Prices: Model Performance with Increased Price Volatility*. 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.

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. accessed on 5 April 2020.

Kuhn, M. and K. Johnson (2013). *Applied predictive modeling*, Volume 26. Springer.

Kuhn, S., B. Egert, S. Neumann, and C. Steinbeck (2008). Building blocks for automated elucidation of metabolites: Machine learning methods for nmr prediction. *BMC bioinformatics 9*(1), 400.

Leo Breiman, Adele Cutler, A. L. and M. Wiener (2018). Classification and regression based on a forest of trees using random inputs. accessed on 5 April 2020.

Lusk, J. L. (2016). From farm income to food consumption: Valuing usda data products. Technical report.

Miftakhova, A. and W. Pohl (2019). Financial markets and climate models: An empirical study on corn futures. *Available at SSRN 3330535*.

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

of Representatives, U. S. H. (2009, June). Hearing to review the federal crop insurance program : hearing before the subcommittee on general farm commodities and risk management of the committee on agriculture, house of representatives.

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

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

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.

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

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

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.

Therneau, T., B. Atkinson, B. Ripley, and M. B. Ripley (2019). Package 'rpart'. accessed on 23 March 2020.

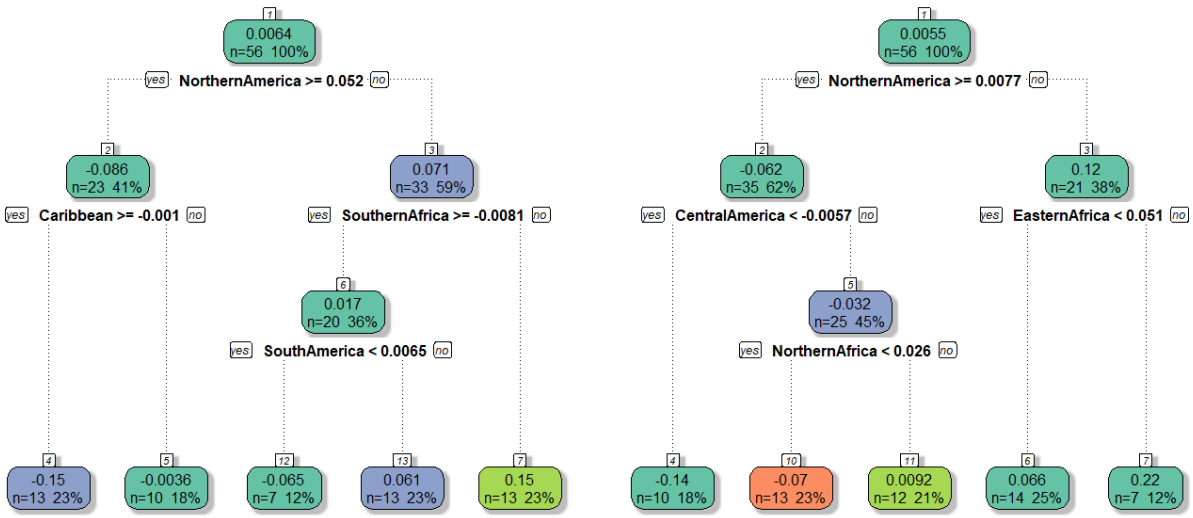
Warr, P. G. (1990). Predictive performance of the world bank's commodity price projections. *Agricultural Economics 4*(3), 365 – 379.

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

Appendices

A CART

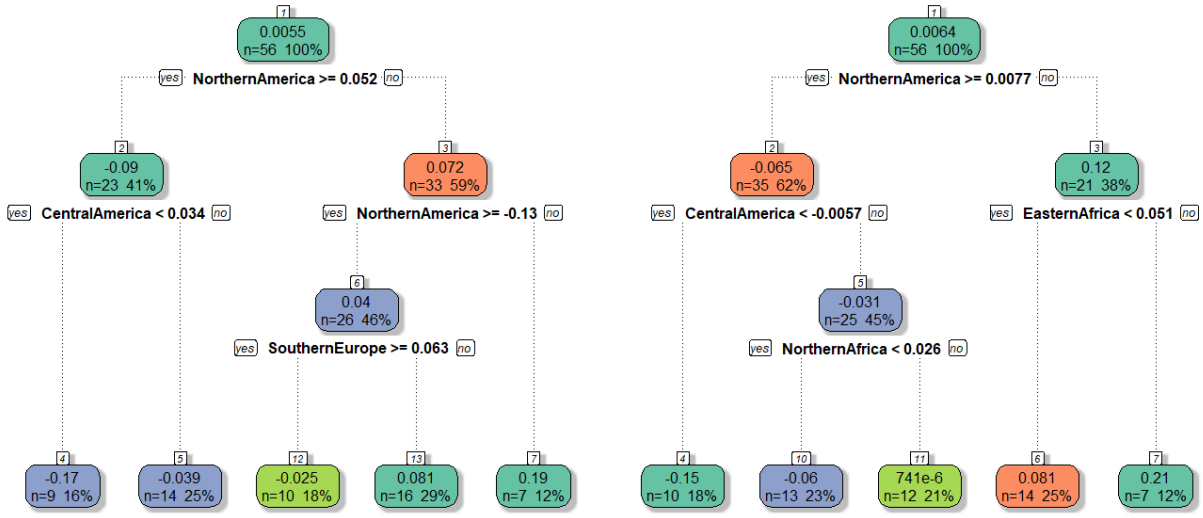
A.A Regression based forecast



(a) Inputs = relative regional production changes

(b) Inputs = relative regional yield changes

Figure 10: CART models for $p_{11,y}$ of maize (i.e., relative price change in November) as a function of relative regional production changes (a) and relative regional yield changes (b)

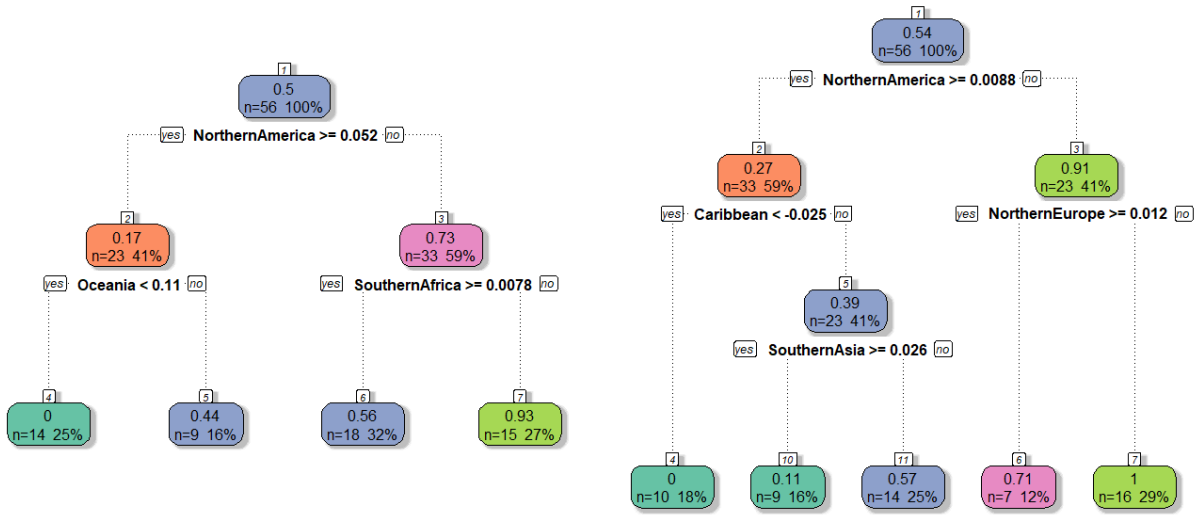


(a) Inputs=relative regional production changes

(b) Inputs=relative regional yield changes

Figure 11: CART models for $p_{12,y}$ of maize (i.e., relative price change in December) as a function of relative regional production changes (a) and relative regional yield changes (b)

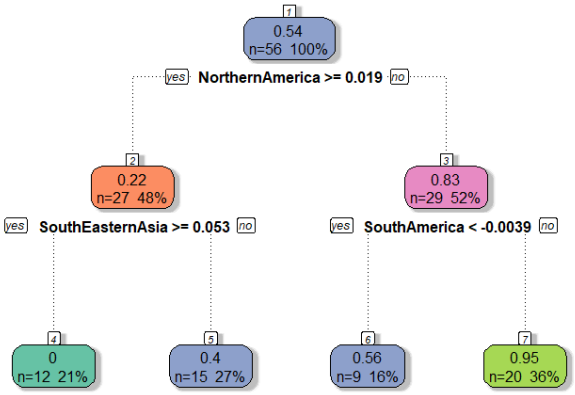
A.B Classification based forecast



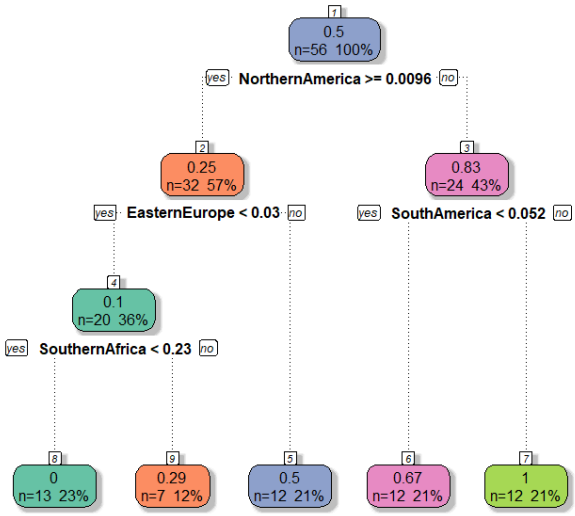
(a) Inputs=relative regional production changes

(b) Inputs=relative regional yield changes

Figure 12: CART models for 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

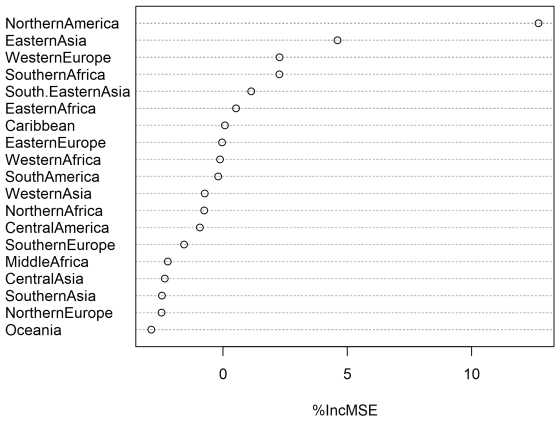


(b) Inputs=relative regional yield changes

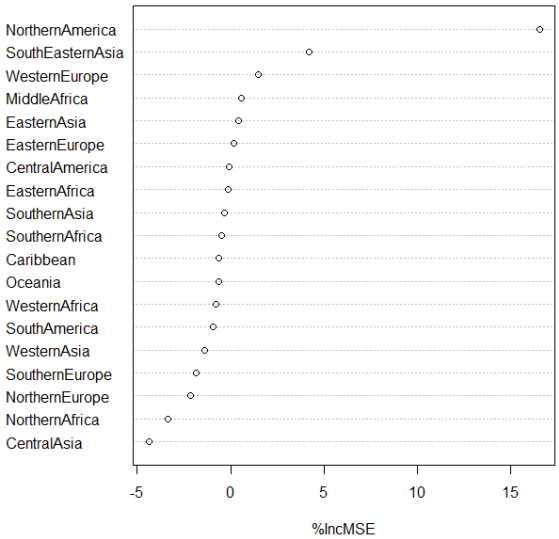
Figure 13: CART models for the probability of relative maize price increase in December as a function of relative regional production changes (a) and relative regional yield changes (b).

B Random Forest

B.A Regression based forecast

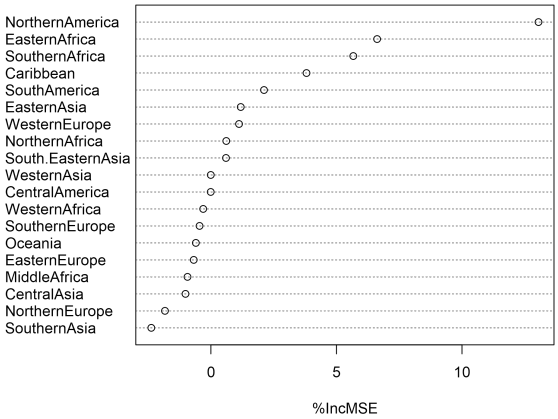


(a) Inputs = relative regional production changes

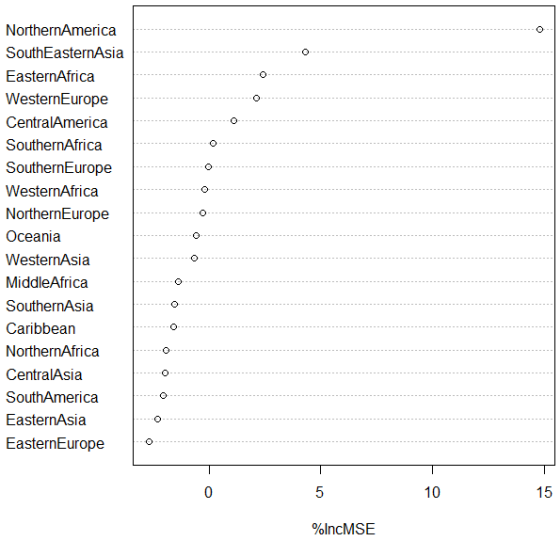


(b) Inputs = relative regional yield changes

Figure 14: Importance of the inputs of the Random-forest models predicting relative price change in October as a function of relative regional production changes (a) and relative regional yield changes (b). The Mean Decrease Accuracy (%IncMSE) indicator examines the extent to which the model performs less well without any specific x_k, so that a significant decrease is precisely expected when removing highly influential regions. The figure shows that one region is of almost exclusive importance in predicting $p_{10,y}$ (in terms of accuracy) - Northern-America, in both terms of production and yield. According to Random-forest based rank, other regions are also of importance here, although relatively less so. An interesting statistic is the existence of negative values for several regions. In other words, the inclusion of these areas in the prediction model may reduce its effectiveness.

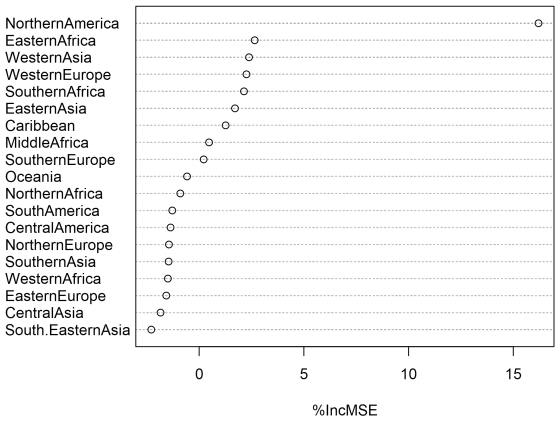


(a) Inputs = relative regional production changes

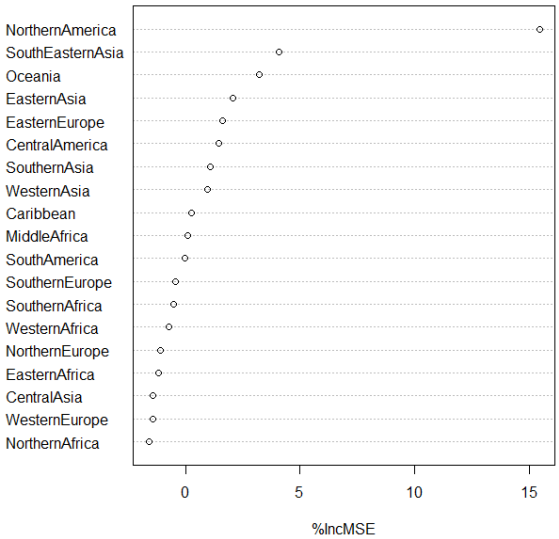


(b) Inputs = relative regional yield changes

Figure 15: Importance of the inputs of the Random-forest models predicting relative price change in November as a function of relative regional production changes (a) and relative regional yield changes (b).



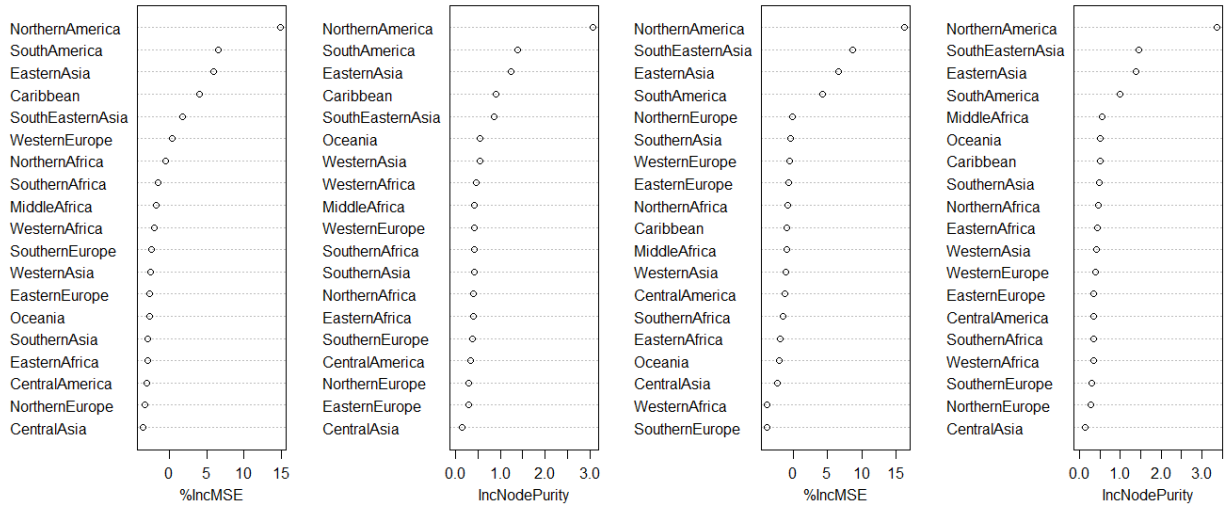
(a) Inputs = relative regional production changes



(b) Inputs = relative regional yield changes

Figure 16: Importance of the inputs of the Random-forest models predicting relative price change in December as a function of relative regional production changes (a) and relative regional yield changes (b).

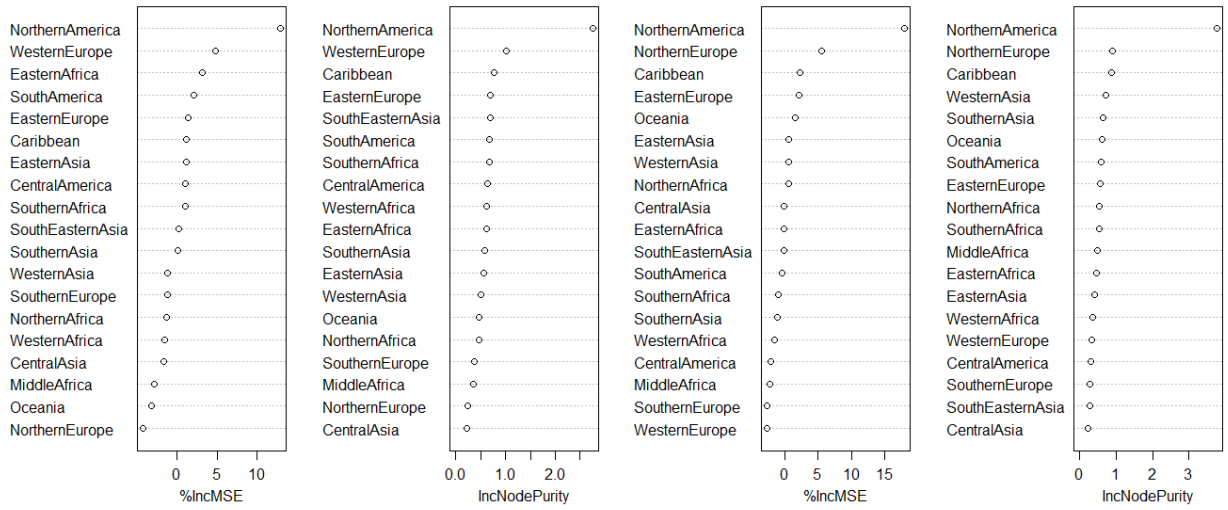
B.B Classification based forecast



(a) Inputs=relative regional production changes

(b) Inputs=relative regional yield changes

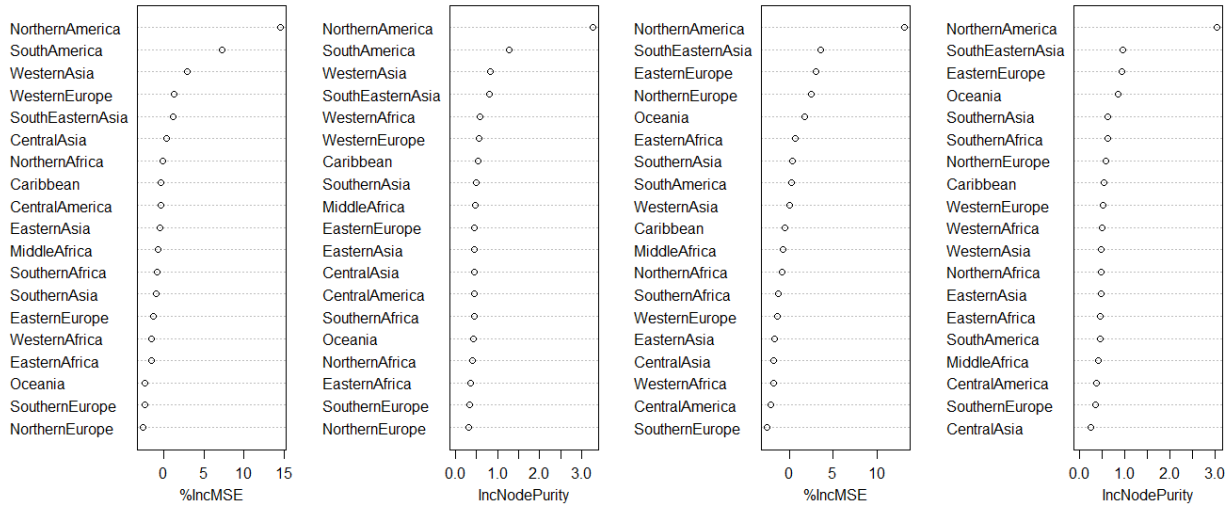
Figure 17: Importance of the inputs of the Random-forest models predicting probability of relative maize price increase in October as a function of relative regional production changes (a) and relative regional yield changes (b). The RIA results of RF point out one variable with an almost exclusive importance in predicting the global maize price change rate (%IncMSE) - the Northern-America production. The Mean Decrease Gini (IncNodePurity) measures how different the splits at the bottom of the trees are, even when high scores express relative importance. Here, too, high score signifies strong influence (Kuhn et al., 2008)



(a) Inputs=relative regional production changes

(b) Inputs=relative regional yield changes

Figure 18: Importance of the inputs of the Random-forest models predicting probability of relative maize price increase in November as a function of relative regional production changes (a) and relative regional yield changes (b).



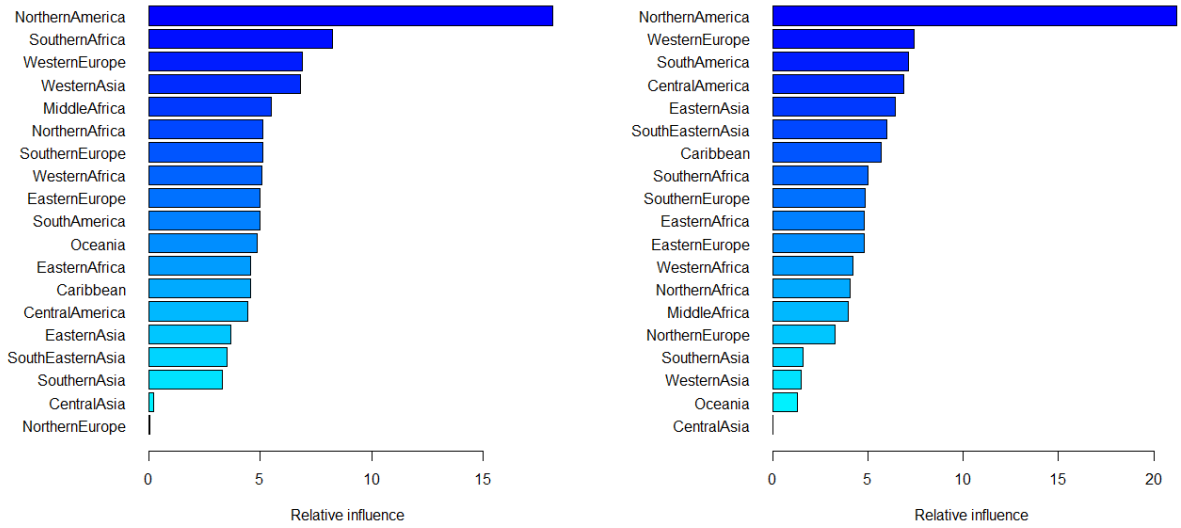
(a) Inputs=relative regional production changes

(b) Inputs=relative regional yield changes

Figure 19: Importance of the inputs of the Random-forest models predicting probability of relative maize price increase in December as a function of relative regional production changes (a) and relative regional yield changes (b).

C GBM

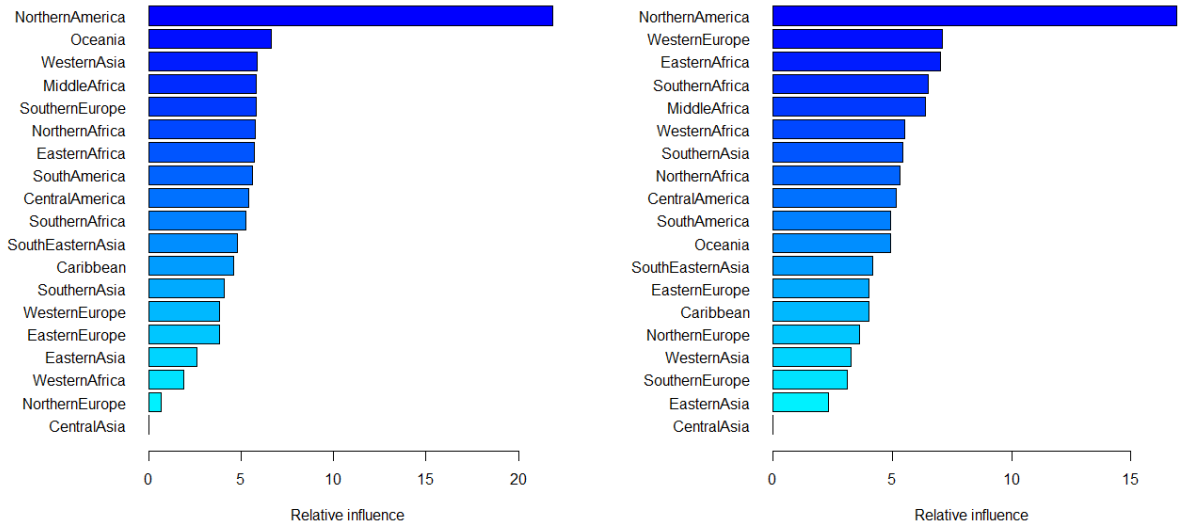
C.A Regression based forecast



(a) Inputs = relative regional production changes

(b) Inputs = relative regional yield changes

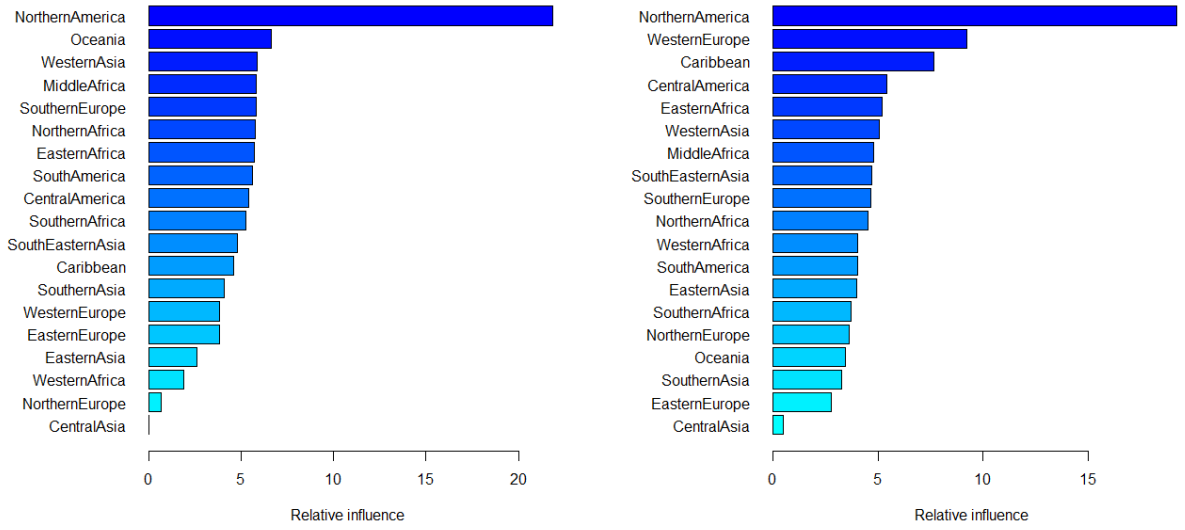
Figure 20: Importance of the inputs of the GBM models predicting relative price change in October as a function of relative regional production changes (a) and relative regional yield changes (b).



(a) Inputs = relative regional production changes

(b) Inputs = relative regional yield changes

Figure 21: Importance of the inputs of the GBM models predicting relative price change in November as a function of relative regional production changes (a) and relative regional yield changes (b).

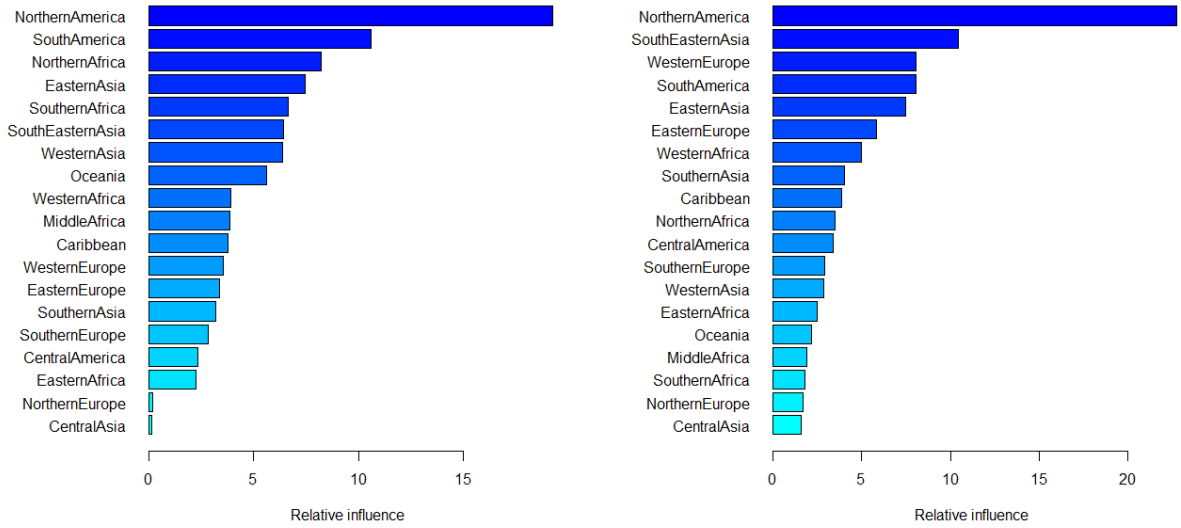


(a) Inputs = relative regional production changes

(b) Inputs = relative regional yield changes

Figure 22: Importance of the inputs of the GBM models predicting relative price change in December as a function of relative regional production changes (a) and relative regional yield changes (b).

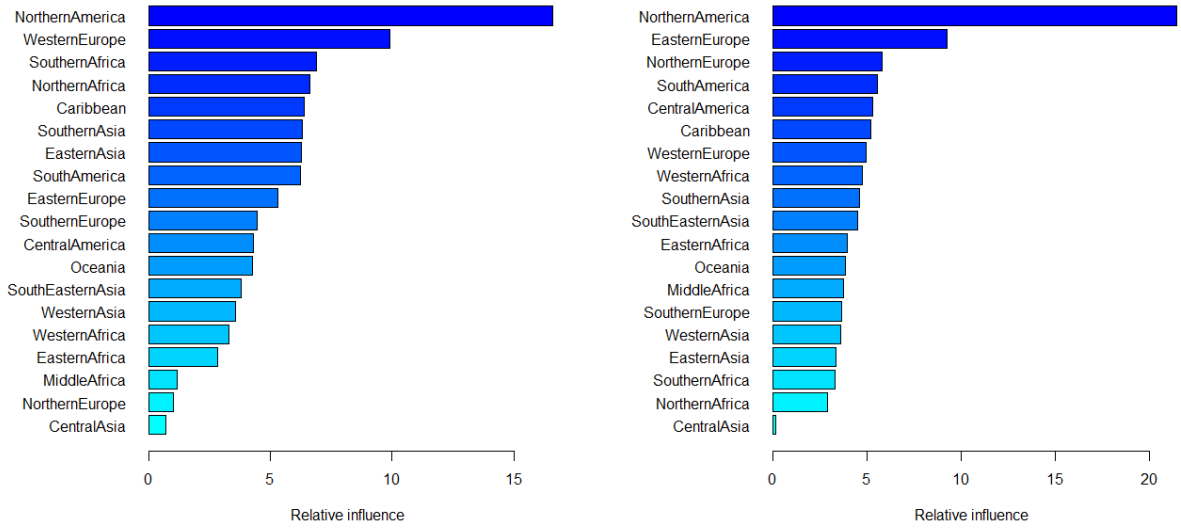
C.B Classification based forecast



(a) Inputs=relative regional production changes

(b) Inputs = relative regional yield changes

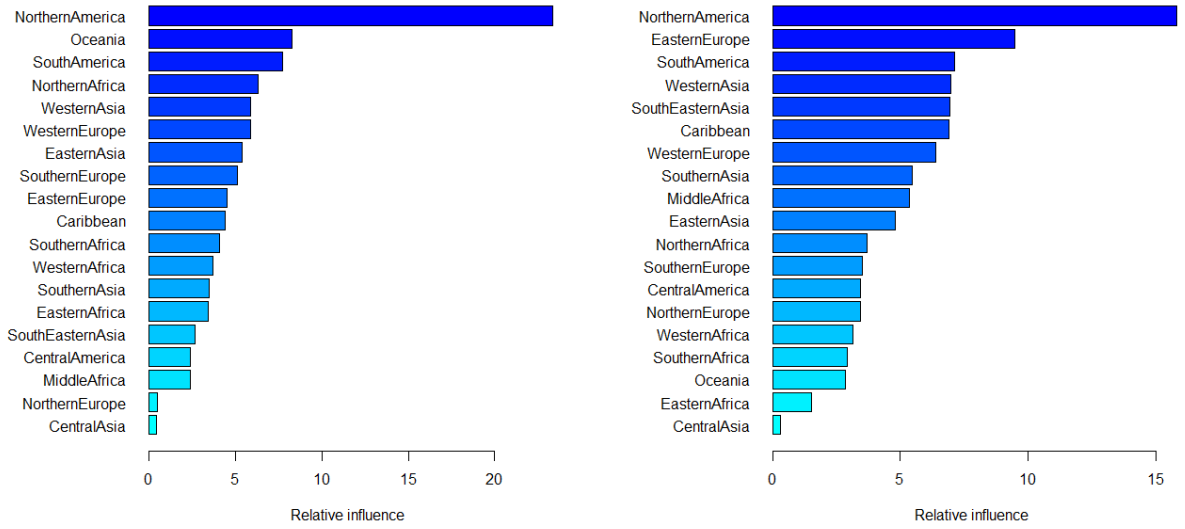
Figure 23: Importance of the inputs of the GBM models predicting probability of relative maize price increase in October as a function of relative regional production changes (a) and relative regional yield changes (b).



(a) Inputs=relative regional production changes

(b) Inputs=relative regional yield changes

Figure 24: Importance of the inputs of the GBM models predicting 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 25: Importance of the inputs of the GBM models predicting probability of relative maize price increase in December as a function of relative regional production changes (a) and relative regional yield changes (b).

D Linear Models

D.A Regression based forecast, LM

Tables 3 and 4 show the summary statistics of multivariate linear regression models predicting relative price changes as a function of relative regional production changes (3) and relative regional yield changes (4). In the first row of each region are the statistics coefficients, β_k, namely, relative change in $p_{10,y}$, $p_{11,y}$ and $p_{12,y}$ given one percent variance in regional production, $x_{k,y}$, where all other variables are fixed. The values in brackets show the level of significance (p-value) of each coefficient. As can be seen, most of the coefficients are insufficiently significant (p-value¿5%) in both tables but especially so when $p_{m,y}$'s are explained by regional production.

In both cases, the region with the highest negative (and significant) impact is Northern-America, with a coefficient of at least -0.350 (3) and -0.275 (4). This means that a 1% reduction in annual Northern American production will cause an estimated increase of at least 0.35% in global maize price at the end of that year. Similarly, a 1% negative change in Northern-America annual yield is expected to induce a mean rise of at least 0.275% in global maize price.

	October	November	December
(Intercept)	-0.013	0.009	0.016
	(0.646)	(0.758)	(0.576)
Caribbean	-0.113	-0.334	-0.266
	(0.580)	(0.120)	(0.204)
CentralAmerica	0.028	0.120	0.081
	(0.856)	(0.462)	(0.611)
CentralAsia	-0.262	-0.289	-0.252
	(0.146)	(0.123)	(0.169)
EasternAfrica	0.259	0.355	0.264
	(0.125)	(0.045)	(0.124)
EasternAsia	0.363	0.089	0.057
	(0.088)	(0.680)	(0.788)
EasternEurope	0.151	0.181	0.143
	(0.237)	(0.173)	(0.270)
MiddleAfrica	-0.163	-0.117	-0.276
	(0.531)	(0.665)	(0.300)
NorthernAfrica	0.176	-0.062	0.089
	(0.461)	(0.802)	(0.712)
NorthernAmerica	-0.331	-0.275	-0.282
	(0.001)	(0.009)	(0.006)
NorthernEurope	0.013	0.001	0.012
	(0.587)	(0.967)	(0.633)
Oceania	0.119	0.085	0.006
	(0.266)	(0.444)	(0.956)
SouthAmerica	0.215	0.113	0.142
	(0.186)	(0.498)	(0.385)
SouthEasternAsia	0.124	0.095	-0.016
	(0.423)	(0.552)	(0.920)
SouthernAfrica	-0.095	-0.106	-0.074
	(0.043)	(0.028)	(0.114)
SouthernAsia	-0.024	-0.003	0.066
	(0.877)	(0.983)	(0.672)
SouthernEurope	-0.281	-0.307	-0.260
	(0.205)	(0.181)	(0.248)
WesternAfrica	0.145	0.161	0.147
	(0.319)	(0.286)	(0.321)
WesternAsia	-0.354	-0.364	-0.328
	(0.058)	(0.059)	(0.082)
WesternEurope	0.067	0.106	0.060
	(0.408)	(0.209)	(0.461)
Num.Obs.	56	56	56
R2	0.579	0.584	0.548
R2 Adj.	0.357	0.365	0.310
AIC	-64.3	-60.2	-62.5
BIC	-21.7	-17.7	-20.0
Log.Lik.	53.140	51.108	52.255

Table 3: linear regression, Inputs=relative regional production changes

	October	November	December
(Intercept)	0.025	0.028	0.034
	(0.342)	(0.340)	(0.221)
Caribbean	0.240	0.149	0.241
	(0.282)	(0.543)	(0.302)
CentralAmerica	0.089	0.458	0.374
	(0.750)	(0.144)	(0.206)
CentralAsia	-0.225	-0.054	-0.092
	(0.414)	(0.858)	(0.748)
EasternAfrica	0.373	0.515	0.394
	(0.011)	(0.002)	(0.010)
EasternAsia	-0.018	-0.146	-0.154
	(0.936)	(0.553)	(0.510)
EasternEurope	0.137	0.083	0.044
	(0.313)	(0.577)	(0.755)
MiddleAfrica	0.904	0.934	0.768
	(0.003)	(0.005)	(0.013)
NorthernAfrica	-0.479	-0.633	-0.422
	(0.067)	(0.031)	(0.121)
NorthernAmerica	-0.421	-0.350	-0.378
	(0.003)	(0.023)	(0.010)
NorthernEurope	-0.100	-0.154	-0.129
	(0.095)	(0.023)	(0.042)
Oceania	-0.186	-0.250	-0.290
	(0.256)	(0.169)	(0.094)
SouthAmerica	-0.009	-0.217	-0.112
	(0.966)	(0.366)	(0.620)
SouthEasternAsia	0.804	0.779	0.366
	(0.064)	(0.103)	(0.411)
SouthernAfrica	-0.081	-0.069	-0.049
	(0.078)	(0.175)	(0.305)
SouthernAsia	-0.075	-0.075	0.004
	(0.649)	(0.679)	(0.980)
SouthernEurope	-0.340	-0.301	-0.251
	(0.120)	(0.211)	(0.270)
WesternAfrica	-0.005	-0.073	0.036
	(0.979)	(0.743)	(0.865)
WesternAsia	-0.395	-0.399	-0.403
	(0.056)	(0.080)	(0.062)
WesternEurope	0.192	0.281	0.182
	(0.062)	(0.015)	(0.090)
Num.Obs.	56	56	56
R2	0.678	0.637	0.632
R2 Adj.	0.508	0.446	0.439
AIC	-79.3	-67.9	-74.1
BIC	-36.7	-25.3	-31.5
Log.Lik.	60.631	54.939	58.036

Table 4: linear regression, Inputs=relative regional yield changes

D.B Classification based forecast, GLM

Tables 5 and 6 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 (5) and relative regional yield changes (6). The tables show the change in the probability of the global maize price to increase or decrease, given a change in regional input with all other variables fixed; and the significance of the result.

	October	November	December
(Intercept)	-1.486	-0.882	-1.202
	(0.213)	(0.540)	(0.440)
Caribbean	-14.074	-44.006	2.202
	(0.157)	(0.118)	(0.865)
CentralAmerica	-8.082	-2.265	-12.226
	(0.260)	(0.739)	(0.138)
CentralAsia	5.446	-17.845	-18.665
	(0.513)	(0.341)	(0.181)
EasternAfrica	-2.006	9.053	13.410
	(0.751)	(0.543)	(0.261)
EasternAsia	19.968	2.306	13.571
	(0.047)	(0.769)	(0.211)
EasternEurope	4.105	40.401	13.710
	(0.536)	(0.121)	(0.246)
MiddleAfrica	12.840	70.015	3.241
	(0.305)	(0.112)	(0.827)
NorthernAfrica	15.993	-47.516	39.674
	(0.321)	(0.207)	(0.096)
NorthernAmerica	-26.390	-36.043	-40.389
	(0.043)	(0.152)	(0.025)
NorthernEurope	3.556	2.592	2.714
	(0.098)	(0.278)	(0.148)
Oceania	18.759	24.220	27.053
	(0.038)	(0.120)	(0.026)
SouthAmerica	23.271	18.154	35.837
	(0.029)	(0.128)	(0.036)
SouthEasternAsia	15.011	58.183	18.130
	(0.098)	(0.136)	(0.180)
SouthernAfrica	0.937	-0.462	-3.007
	(0.629)	(0.917)	(0.251)
SouthernAsia	-22.518	-17.729	-21.024
	(0.060)	(0.164)	(0.041)
SouthernEurope	13.256	-48.775	-13.858
	(0.427)	(0.123)	(0.366)
WesternAfrica	-2.379	-15.136	7.631
	(0.700)	(0.273)	(0.342)
WesternAsia	-31.642	-23.678	-43.512
	(0.078)	(0.165)	(0.075)
WesternEurope	-15.510	-2.120	9.184
	(0.125)	(0.825)	(0.130)
Num.Obs.	56	56	56
AIC	70.3	64.6	62.5
BIC	110.8	105.1	103.0
Log.Lik.	-15.154	-12.321	-11.243

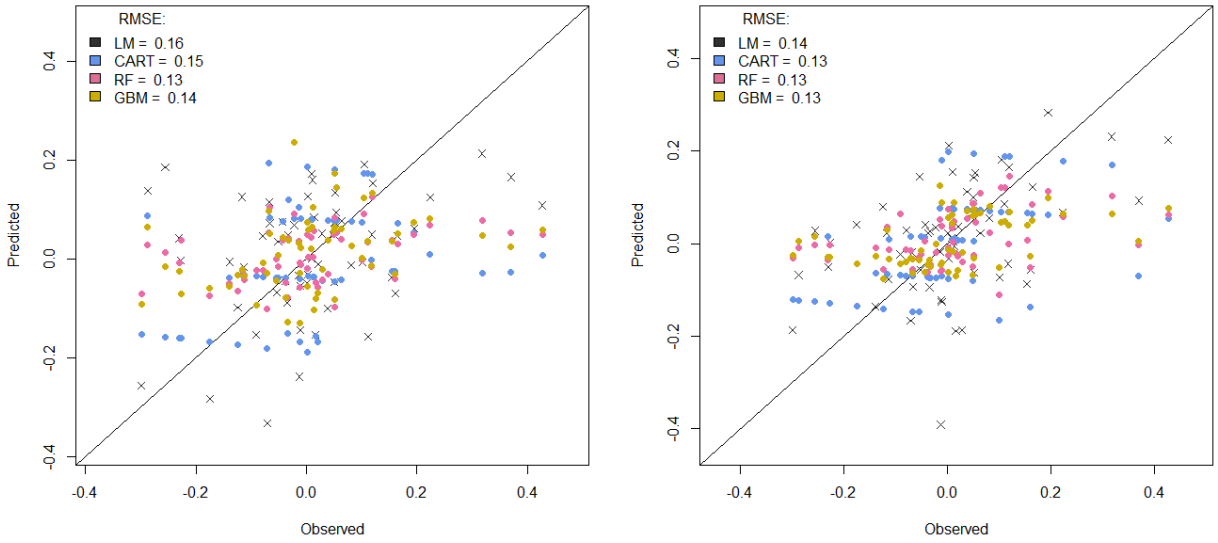
Table 5: Summary statistics of the classification linear models, GLM, Inputs=relative regional production changes

	October	November	December
(Intercept)	0.620	-73.499	4.782
	(0.514)	(0.998)	(0.083)
Caribbean	14.871	5498.778	38.476
	(0.125)	(0.989)	(0.093)
CentralAmerica	-6.937	-186.847	14.237
	(0.542)	(1.000)	(0.416)
CentralAsia	-15.512	-9295.218	-42.824
	(0.103)	(0.988)	(0.140)
EasternAfrica	8.493	8140.243	38.411
	(0.206)	(0.987)	(0.097)
EasternAsia	3.833	74.314	-0.781
	(0.590)	(1.000)	(0.936)
EasternEurope	0.884	3191.542	29.363
	(0.859)	(0.990)	(0.098)
MiddleAfrica	31.916	15318.640	32.934
	(0.053)	(0.987)	(0.031)
NorthernAfrica	-21.947	-18975.743	-43.565
	(0.072)	(0.988)	(0.078)
NorthernAmerica	-18.198	-9773.343	-38.749
	(0.036)	(0.987)	(0.044)
NorthernEurope	0.382	-1495.583	-9.467
	(0.840)	(0.988)	(0.115)
Oceania	-1.736	162.448	-6.470
	(0.743)	(0.999)	(0.577)
SouthAmerica	5.714	-2649.987	-24.827
	(0.531)	(0.992)	(0.244)
SouthEasternAsia	39.946	47256.879	33.770
	(0.184)	(0.987)	(0.220)
SouthernAfrica	-2.384	-718.440	-0.961
	(0.221)	(0.989)	(0.713)
SouthernAsia	-15.161	-6147.984	-10.216
	(0.090)	(0.987)	(0.235)
SouthernEurope	-2.640	-5178.372	-55.883
	(0.777)	(0.991)	(0.092)
WesternAfrica	-6.273	-13111.716	3.962
	(0.481)	(0.988)	(0.713)
WesternAsia	-10.955	-663.729	-65.221
	(0.183)	(0.995)	(0.049)
WesternEurope	-8.600	5017.556	23.672
	(0.145)	(0.988)	(0.072)
Num.Obs.	56	56	56
AIC	75.3	40.0	63.3
BIC	115.8	80.5	103.8
Log.Lik.	-17.660	-0.000	-11.650

Table 6: Summary statistics of the classification linear models, GLM, Inputs=relative regional yield changes

E Model Evaluation

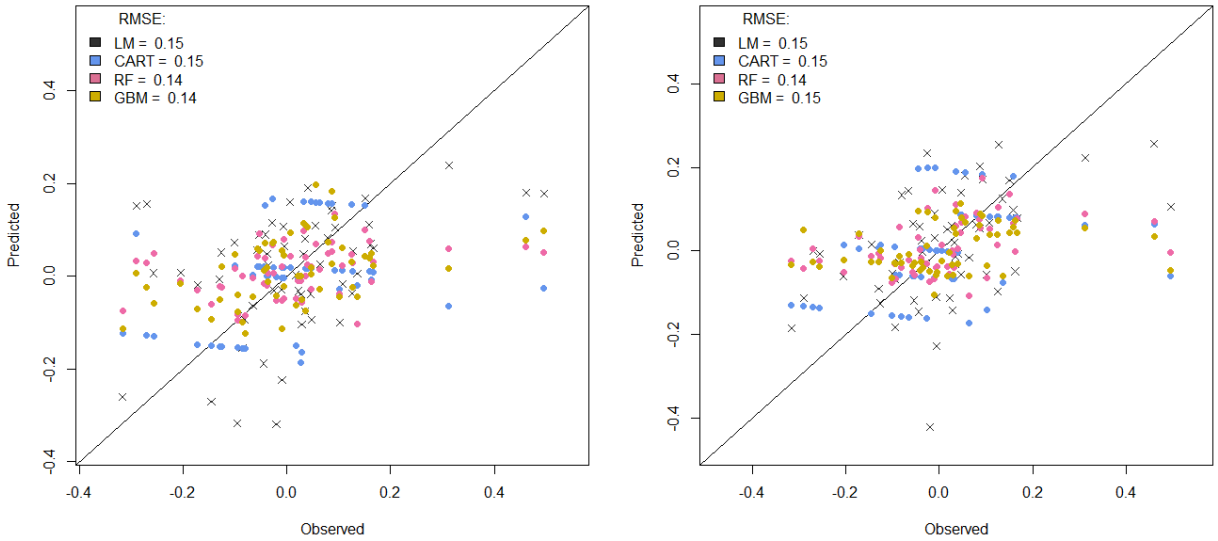
E.A Regression based forecasting models, RMSE



(a) Inputs = relative regional production changes

(b) Inputs = relative regional yield changes

Figure 26: Observed relative price change vs. Predicted relative price change, November. Values of RMSE are reported for the different models.



(a) Inputs = relative regional production changes

(b) Inputs = relative regional yield changes

Figure 27: Observed relative price change vs. Predicted relative price change, December. Values of RMSE are reported for the different models.

E.B Time-series based forecasting model, TBATS, RMSE

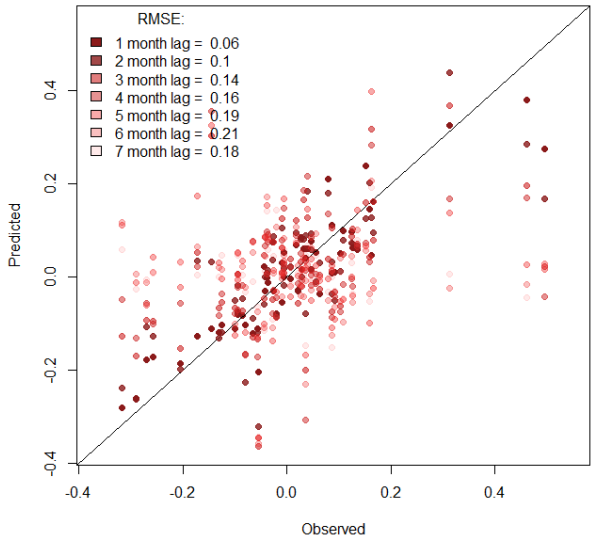


Figure 28: Observed relative price change vs. Predicted relative price change, November. Predictions were obtained with TBATS for different time lags (from one to five months ahead). Values of RMSE are reported for all time lags.

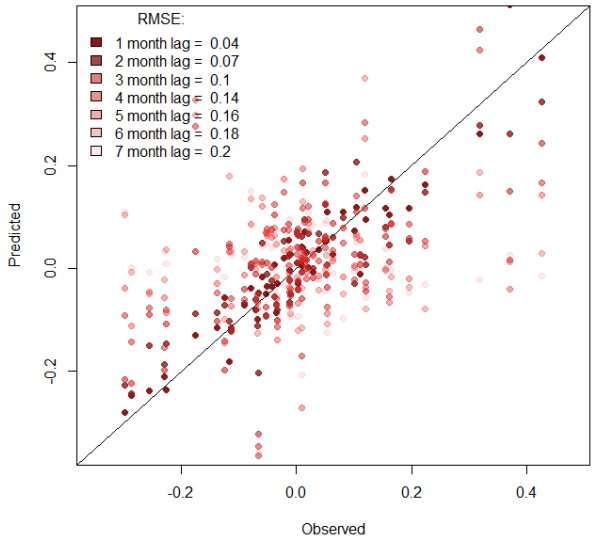
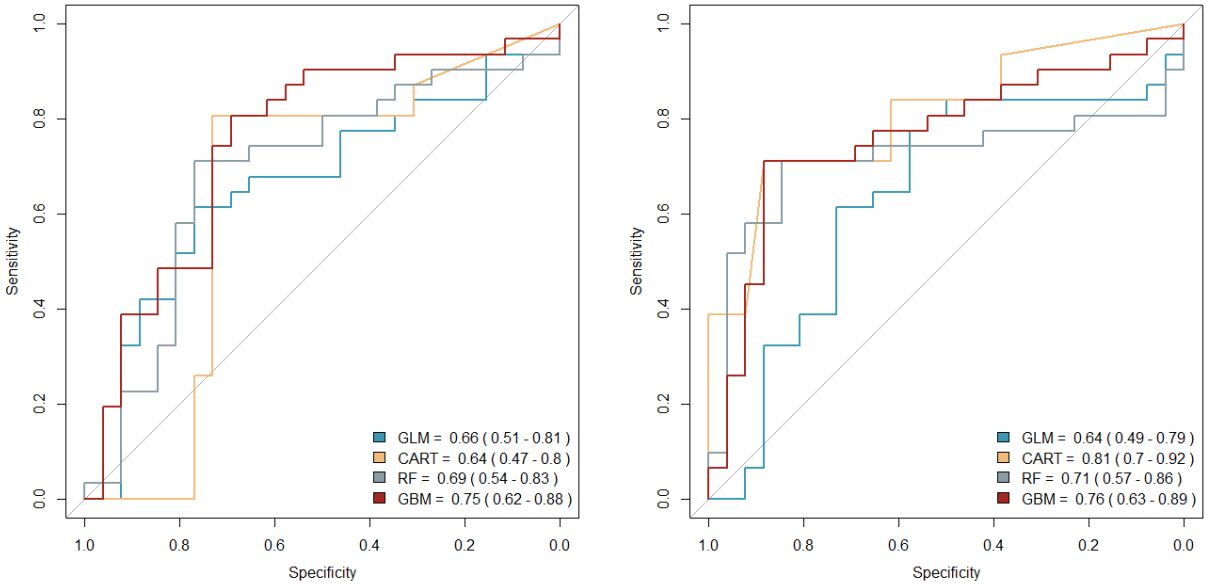


Figure 29: Observed relative price change vs. Predicted relative price change, December. Predictions were obtained with TBATS for different time lags (from one to six months ahead). Values of RMSE are reported for all time lags.

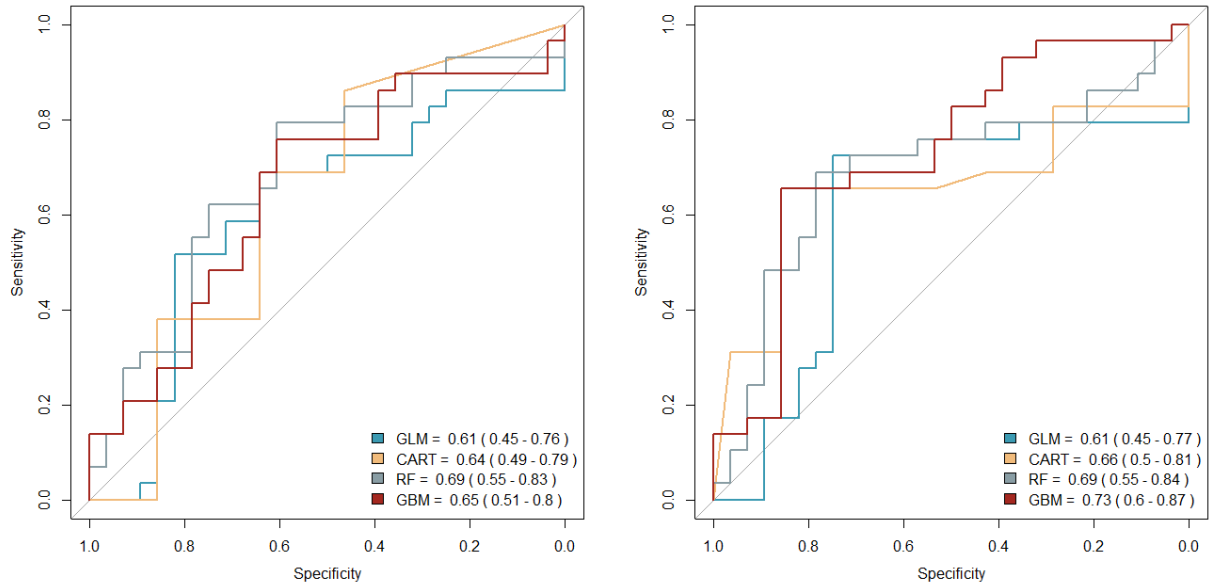
E.C Classification based forecasting models, AUC



(a) Inputs = relative regional production changes

(b) Inputs = relative regional yield changes

Figure 30: ROC curves obtained for the classification models predicting November price increase vs. price decrease (95% confidence interval).



(a) Inputs = relative regional production changes

(b) Inputs = relative regional yield changes

Figure 31: ROC curves obtained for the classification models predicting December price increase vs. price decrease (95% confidence interval).

E.D Accuracy of models, probability of maize price to increase or decrease

Table 7: Accuracy (ACC), sensitivity (Sens) and specificity (Spec) of the classification models predicting global maize increase vs. decrease in November. Inputs = relative regional production changes

	GLM		CART		RF		GBM	
	Predicted		Predicted		Predicted		Predicted	
Obs.	$p^b_{11,y}$=1	$p^b_{11,y}$=0	$p^b_{11,y}$=1	$p^b_{11,y}$=0	$p^b_{11,y}$=1	$p^b_{11,y}$=0	$p^b_{11,y}$=1	$p^b_{11,y}$=0
$p^b_{11,y}$=1	10	18	15	20	11	19	11	21
$p^b_{11,y}$=0	18	11	13	9	17	10	17	8
	Sens	64%	Sens	57%	Sens	63%	Sens	66%
	Spec	62%	Spec	59%	Spec	63%	Spec	68%
	ACC	63%	ACC	58%	ACC	63%	ACC	67%

Table 8: Accuracy (ACC), sensitivity (Sens) and specificity (Spec) of the classification models predicting global maize increase vs. decrease in November. Inputs = relative regional production changes

	GLM		CART		RF		GBM	
	Predicted		Predicted		Predicted		Predicted	
Obs.	$p^b_{11,y}$=1	$p^b_{11,y}$=0	Decrease	Increase	Decrease	Increase	Decrease	Increase
$p^b_{11,y}$=1	8	21	16	20	8	20	8	19
$p^b_{11,y}$=0	20	8	12	9	20	9	20	10

Sens	72%	Sens	56%	Sens	71%	Sens	70%
Spec	71%	Spec	57%	Spec	69%	Spec	67%
ACC	72%	ACC	56%	ACC	70%	ACC	68%

Table 9: Accuracy (ACC), sensitivity (Sens) and specificity (Spec) of the classification models predicting global maize increase vs. decrease in December. Inputs = relative regional production changes

	GLM		CART		RF		GBM	
	Predicted		Predicted		Predicted		Predicted	
Obs.	Decrease	Increase	Decrease	Increase	Decrease	Increase	Decrease	Increase
$p^b_{12,y}$=1	12	21	11	25	10	23	7	23
$p^b_{12,y}$=0	14	10	15	6	16	8	19	8

Sens	64%	Sens	69%	Sens	70%	Sens	77%
Spec	58%	Spec	71%	Spec	67%	Spec	70%
ACC	61%	ACC	70%	ACC	68%	ACC	74%

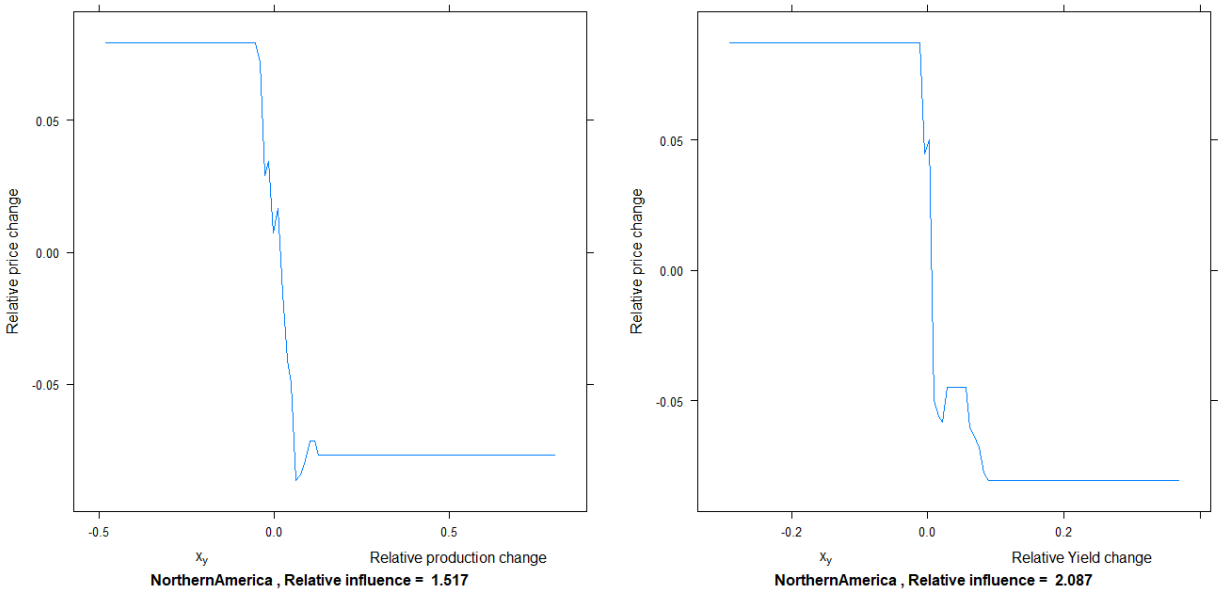
Table 10: Accuracy (ACC), sensitivity (Sens) and specificity (Spec) of the classification models predicting global maize increase vs. decrease in December. Inputs = relative regional yield changes

	GLM		CART		RF		GBM	
	Predicted		Predicted		Predicted		Predicted	
Obs.	Decrease	Increase	Decrease	Increase	Decrease	Increase	Decrease	Increase
$p^b_{12,y}$=1	10	20	10	25	9	22	9	23
$p^b_{12,y}$=0	16	11	16	6	17	9	17	8

Sens	67%	Sens	71%	Sens	71%	Sens	72%
Spec	59%	Spec	73%	Spec	65%	Spec	68%
ACC	63%	ACC	72%	ACC	68%	ACC	70%

F Partial dependence

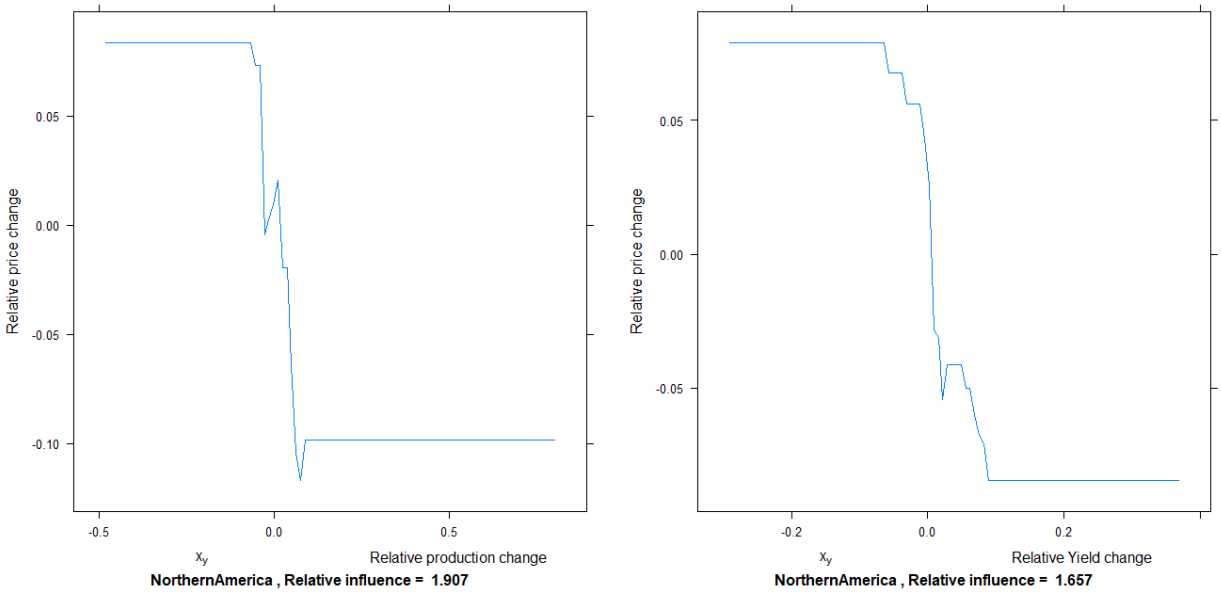
F.A Regression based partial dependence plots



(a) Inputs = relative regional production changes

(b) Inputs = relative regional yield changes

Figure 32: Partial dependence plots obtained with GBM showing the average response of relative price change in November to relative production change in Northern-America (a) and to relative yield change in Northern-America (b).

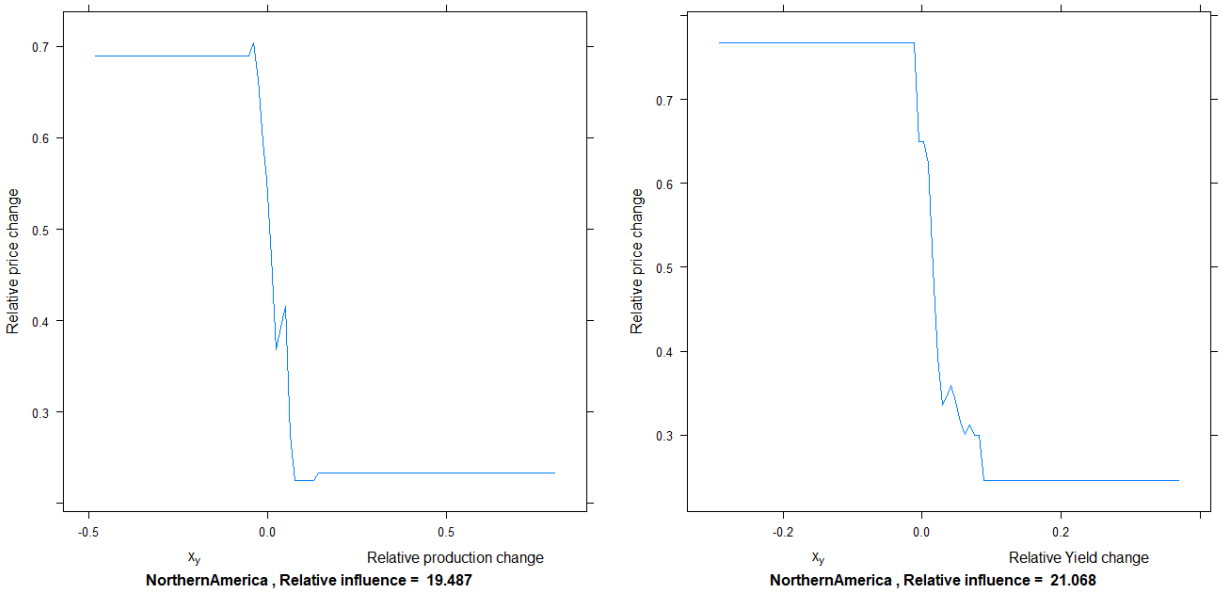


(a) Inputs = relative regional production changes

(b) Inputs = relative regional yield changes

Figure 33: Partial dependence plots obtained with GBM showing the average response of relative price change in December to relative production change in Northern-America (a) and to relative yield change in Northern-America (b).

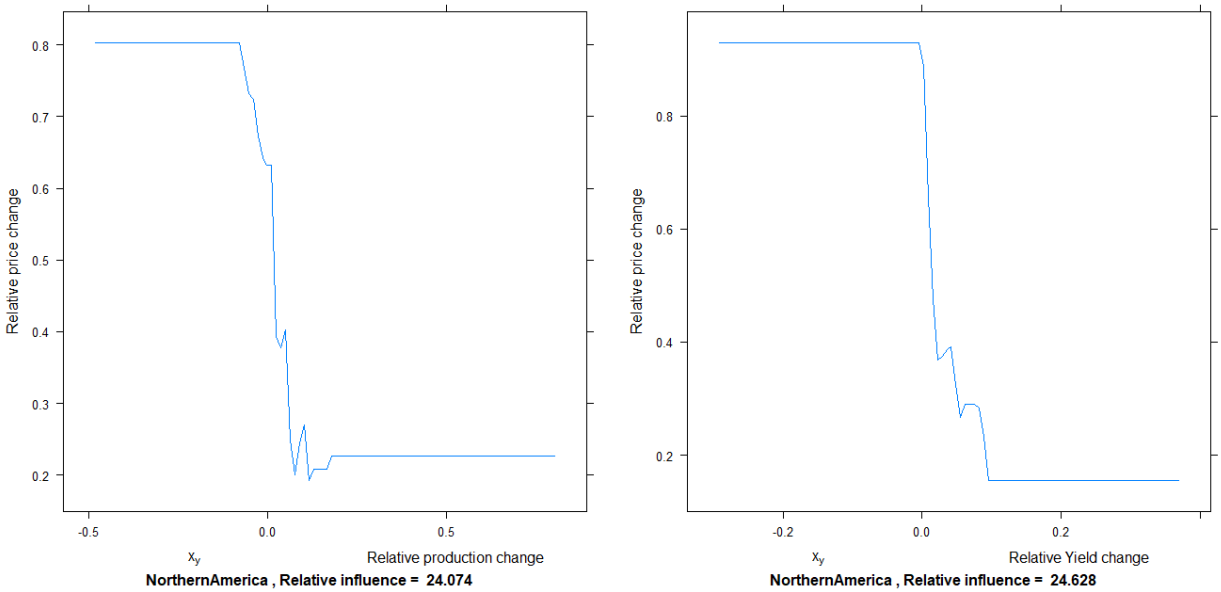
F.B Classification based partial dependence plots



(a) Inputs = relative regional production changes

(b) Inputs = relative regional yield changes

Figure 34: Partial dependence plots obtained with GBM showing the probability of price increase in November as a function of relative production change in Northern-America (a) and to relative yield change in Northern-America (b).



(a) Inputs = relative regional production changes

(b) Inputs = relative regional yield changes

Figure 35: Partial dependence plots obtained with GBM showing the probability of price increase in December as a function of relative production change in Northern-America (a) and to relative yield change in Northern-America (b).