Instituto Tecnológico y de Estudios Superiores de Occidente

Reconocimiento de validez oficial de estudios de nivel superior según acuerdo secretarial 15018, publicado en el Diario Oficial de la Federación del 29 de noviembre de 1976.

Departamento de Matemáticas y Física

Master in Data Science

Markov Chain Monte Carlo approach to the analysis and forecast of grain prices and volatility monitoring

---

THESIS to obtain the **DEGREE** of
MASTER IN DATA SCIENCE

Presented by: **GABRIELA LOZANO OROZCO**

Thesis Advisor **ALVARO P. DÍAZ RUELAS, PH. D**

Tlaquepaque, Jalisco. November 2022.

Instituto Tecnológico y de Estudios Superiores de Occidente

Reconocimiento de validez oficial de estudios de nivel superior según acuerdo secretarial 15018, publicado en el Diario Oficial de la Federación del 29 de noviembre de 1976.

Departamento de Matemáticas y Física

Maestría en Ciencia de Datos

Markov Chain Monte Carlo approach to the analysis and forecast of grain prices and volatility monitoring

---

TESIS para la obtención de **GRADO** de
MAESTRÍA EN CIENCIA DE DATOS

Presenta: **GABRIELA LOZANO OROZCO**

Asesor: **ALVARO P. DÍAZ RUELAS, PH. D**

Tlaquepaque, Jalisco. Noviembre 2022.

ACKNOWLEDGMENTS

Thanks to my family; my husband who has been always the best teammate, my newborn girl who even being so little has taught me so much, my parents for all their support and my siblings. Thanks to all my teachers, and to Álvaro; my thesis advisor. Thanks to ITESO or the scholarship given.

AGRADECIMIENTOS

Gracias a mi familia; mi esposo quien siempre ha sido el mejor compañero, mi bebé que aún siendo tan pequeña me ha enseñado tanto, mis papás por su apoyo constante y mis hermanos. Gracias a todos mis maestros y a Álvaro, mi asesor de esta tesis. Gracias al ITESO por la beca otorgada.

DEDICATION

To my family, for their constant support.

DEDICATORIA

A mi familia, por su apoyo constante.

ABSTRACT

Public studies on the dynamics of food staples as important as cereals (grains) are relatively scarce. Here we undertake a preliminary analysis of the time series for corn, wheat, soybean, and oat prices first via classical ARIMA/GARCH models, and later complementing with the more complex Stochastic Volatility (SV) models. The goal is to improve upon the classical results by implementing a Bayesian analysis through the construction of a suitable Markov Chain Monte Carlo Model with improved volatility analysis and forecasting capabilities. The performance of the SV model is benchmarked against the classical ARMA/GARCH approach, and both are discussed as monitoring tools for the volatility prices.

RESUMEN

Estudios sobre la dinámica de alimentos básicos tan importantes como los cereales (granos) son relativamente escasos. En este trabajo llevamos a cabo un análisis preliminar de la serie de tiempo de los precios del maíz, el trigo, la soya y la avena, primero a través de los modelos clásicos ARIMA/GARCH, y con los modelos más complejos de volatilidad estocástica (SV). El objetivo es mejorar los resultados clásicos mediante la implementación de un análisis bayesiano a través de la construcción de un modelo Monte Carlo de cadena de Markov adecuado con capacidades mejoradas de análisis de volatilidad y pronóstico. El rendimiento del modelo SV se compara con el enfoque clásico ARMA/GARCH, y ambos se analizan como herramientas de seguimiento de la volatilidad de los precios.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

ACF	Auto Correlation Function
ADF	Augmented Dickey Fuller
AIC	Akaike Information Criterion
ARCH	Autoregressive Conditional Heteroskedasticity
ARIMA	Autoregressive Integrated Moving Average
ARMA	Autoregressive Moving Average
GARCH	Generalized Autoregressive Conditional Heteroskedasticity
MAD	Mean Absolute Deviation
MCMC	Markov Chain Monte Carlo
MLE	Maximum Likelihood Estimation
PACF	Partial Auto Correlation Function
SV	Stochastic Volatility

INTRODUCTION

The conflict in Ukraine has led to a raise in the prices of basic goods globally and an exacerbated increase in their variability. Food prices have been especially affected[1], worsening an already delicate economic situation resulting from the COVID-19 pandemic. According to a 2020 report by the International Monetary Fund (IMF), the recession caused by the pandemic has been the worst since the Great Depression[2]. Extraordinary fiscal support was extended to business and people by governments globally, reaching an amount of around $11.5 trillion as of September 2020[2].

It is then of high relevance to invest effort in detailed studies of the price evolution for the most relevant food sources, known as food staples. Grains (cereals) are by far the most important of staple foods, comprising an average of 48% of the total caloric intake for humans[3].

Despite the clear importance of grains for humans, specialized literature on detailed studies of their price dynamics is not abundant, and there are few efforts to monitor the behavior of the price of these staple foods whose results are public. The most notable example of a public monitor is the online tool called Excessive Food Price Variability Early Warning System, available at the Food Security Portal[4] of the International Food Policy Research Institute (IFPRI)[5]. Although this monitor is well-implemented and based on a non-parametric quantile estimation regression model[6], there is still room for model testing and selection for the study of grain prices.

To this end, and to guarantee a worldwide perspective, we take the price time series for the following grains:[1]

- **Corn**. It represents the highest production of all the cereals with 817 million tons being produced in 2009[8]. The largest producer of corn is USA[8].
- **Wheat**. In 2007 it was the third most produced cereal after maize and rice with a world production of over 600 million tons[8] China has the largest land area devoted to wheat production, followed closely by the United States, India, and the Russian Federation[9].
- **Soybean**. It is a useful oil and protein source and can be used to improve the nutritional value of traditional foods. The main producers are Argentina and Brazil[8].
- **Oat**. This grain ranks around sixth in world cereal production statistics[10]. In 2020, Russia was the second leading global oat producer, after the European Union[11].

This work is divided in two main parts. In the first part we present a descriptive overview of the series and perform a detailed analysis with classic autoregressive models whose

[1] Data was taken from MacroTrends https://www.macrotrends.net/

parameters are estimated via log-likelihood maximization. In the second part we undertake a Bayesian approach to the problem via Markov Chain Monte Carlo (MCMC) sampling.

The power of MCMC allows us to estimate the posterior of the Autoregressive Moving Average (ARMA) and Generalized Autoregressive Conditional Heteroskedasticity (GARCH) parameters, obtaining a deeper picture of the parameter space of the models that were fitted on the first part.

Moreover, MCMC also enables us to fit a Stochastic Volatility (SV) model. Finally, we compare the performance of ARCH/GARCH and SV models to explain the price series and their volatility and close by discussing which model would be better to implement for monitoring the volatility of the grain prices.

PART 1. AUTOREGRESSIVE MODELING USING ARIMA/GARCH MODELS

1.1. Descriptive statistics of the series

In Table 1 we include a summary of the simplest most important descriptive statistical features for the price series considered in the rest of this study. There we can appreciate in red the record maximum values of the prices for wheat and oat, both crops whose main producers are also affected by the Russia-Ukraine conflict (e.g., EU, Russia, and Ukraine itself).

Table 1. Main descriptive statistics of the price (USD/bushel) series

Series	Corn	Wheat	Soybean	Oat
Date start Date end	1959-07-01 2022-09-16	1959-07-01 2022-09-16	1968-12-05 2022-09-16	1970-01-05 2022-09-16
Length, days	15,935	15,936	13,551	13,283
Date Min	1960-11-21 1.007	1968-08-12 1.171	1969-03-18 2.375	1970-02-06 0.582
Date Max	2012-08-21 8.3125	2022-03-07 12.94	2022-06-09 17.69	2022-04-12 8.07

In Figure 1, we show the rolling mean (simple moving average) superimposed to the price time series of the four crops in Table 1. Due to the daily frequency of the data, we chose a 130-day rolling window, representing half a business year (260 days per business year in the US).

The rolling mean provides a very simple and useful way to visualize the trend of the time series at different time scales. From Figure 1, we see that there are many changes of level at many time scales. There is a tendency of this price level to increase in the long run, and no obvious seasonality or cyclicality, despite the cyclic nature of the crops. We can also see how the four series appear to move together in a very similar fashion.

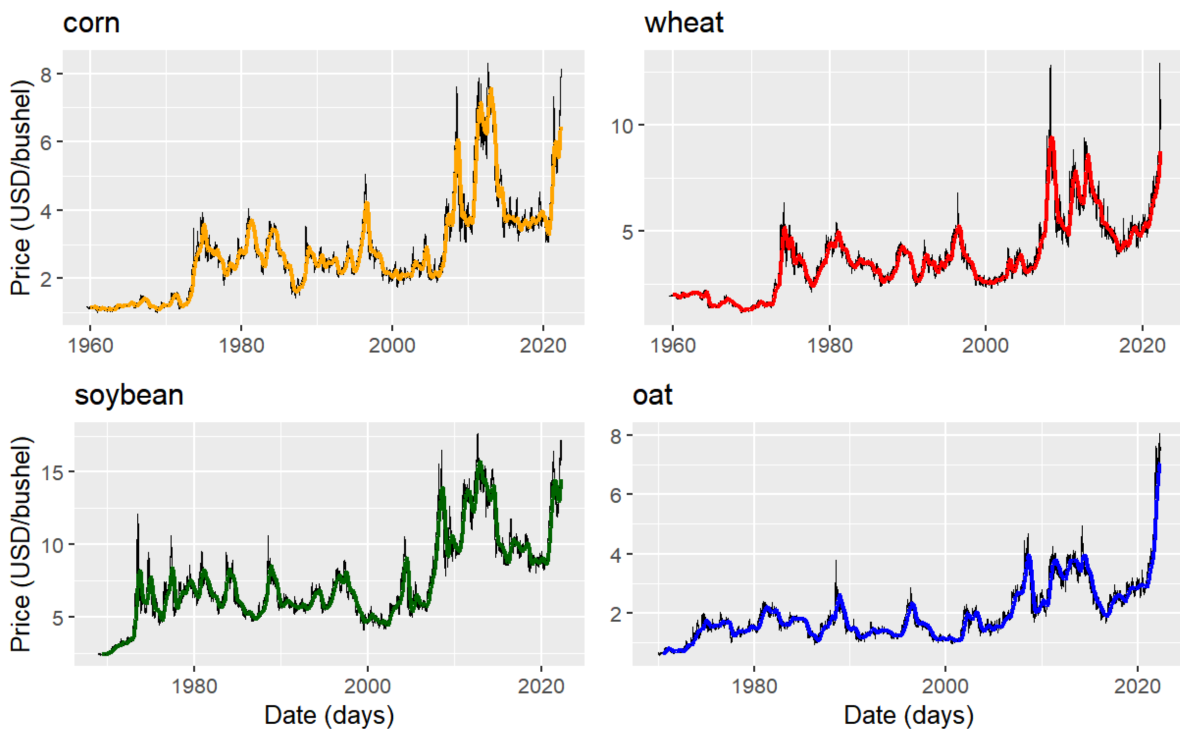

Figure 1. Price series and superimposed 130-day rolling mean

1.2. Autoregressive modeling

1.2.1. Preliminaries

Let us denote the price of the asset x at time t as x_t, and the corresponding time series as $\{x_t\}_{t\in I}^{i}$, with $i = corn,\ wheat,\ soybean,\ oats$ and I the index set corresponding to discrete $I \subset N$, or continuous time $I \subset R$. The plot below shows the time series for all four grains. It is evident that all series are non-stationary since a stepwise upwards trend is clearly visible.

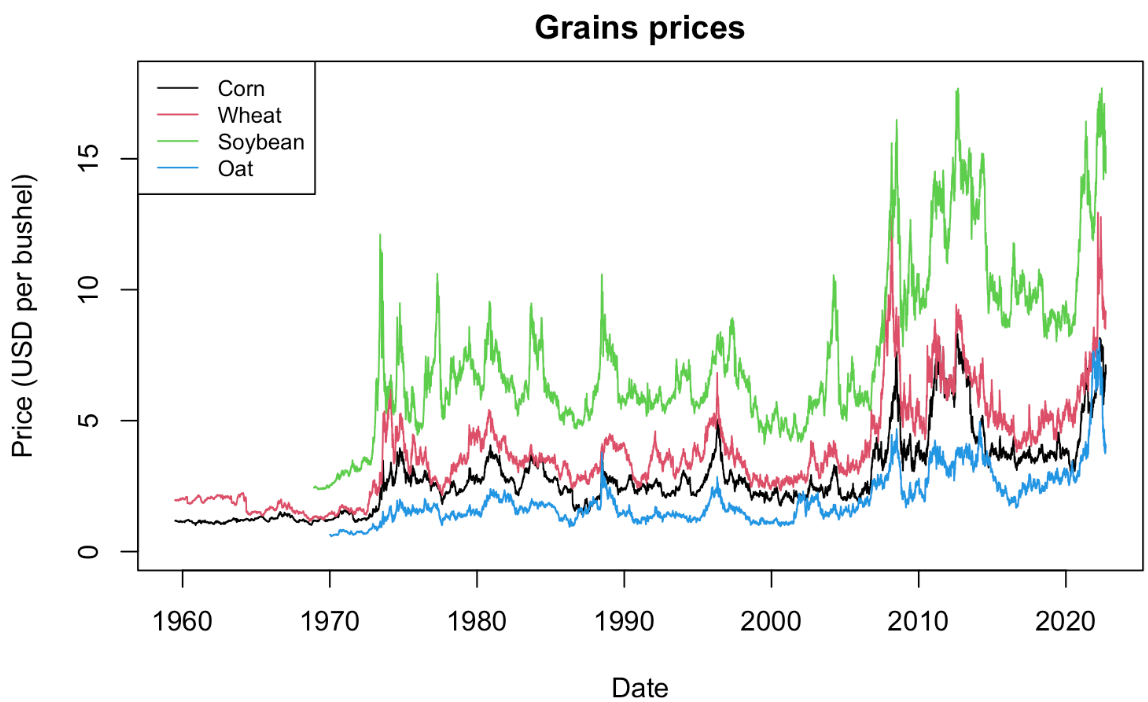

Figure 2. Corn, wheat, soybean, and oat prices, reported in USD per bushel[2]

However, to rigorously check for non-stationary behavior, we implemented an Augmented Dickey-Fuller (ADF) test. If the increments $\Delta x_t = x_t - x_{t-1}$ of the series $\{x_t\}_{t\in I}^{i}$ can be modeled by an AR(p) process[12],

$$\boldsymbol{x_t = \mu_t + \beta x_{t-1} + \sum_{i=1}^{p-1} \phi_i \Delta x_{t-i} + \epsilon_t} \quad (1)$$

where μ_t represents the trend (a deterministic function of time), and $\epsilon_t \sim N(0,1)$ are the error terms, then the ADF test works under the null hypothesis that there is a unit root, this is, the test contrasts the hypotheses

$$\boldsymbol{H_0\colon \beta = 1,\ vs\ H_a\colon |\beta| < 1}$$

by building a t-statistic for the parameter β. Then, we can conclude that a series possesses a unit root according to the ADF test if the null hypothesis cannot be rejected at the selected level of confidence. This is taken as a strong indication that the series is non-stationary.

[2] Bushel is a unit of measurement for grain created many years ago to facilitate grain trade, it corresponded to how much grain would fit in a bushel basket. Nowadays, the United States Department of Agriculture (USDA) created a weight equivalent for a bushel, different for every commodity. Corn was assigned a bushel weight of 56 pounds, while soybeans and wheat were assigned bushel weights of 60 pounds. Oat bushel weight is 32 pounds.

The plots below show the differenced series Δx_t, which clearly display volatility clustering (heteroscedasticity). The Augmented Dickey-Fuller test indicates that the differenced series are stationary, meaning that the integrated component of the ARIMA model is $d = 1$, as we explain in Sec. 1.2.2

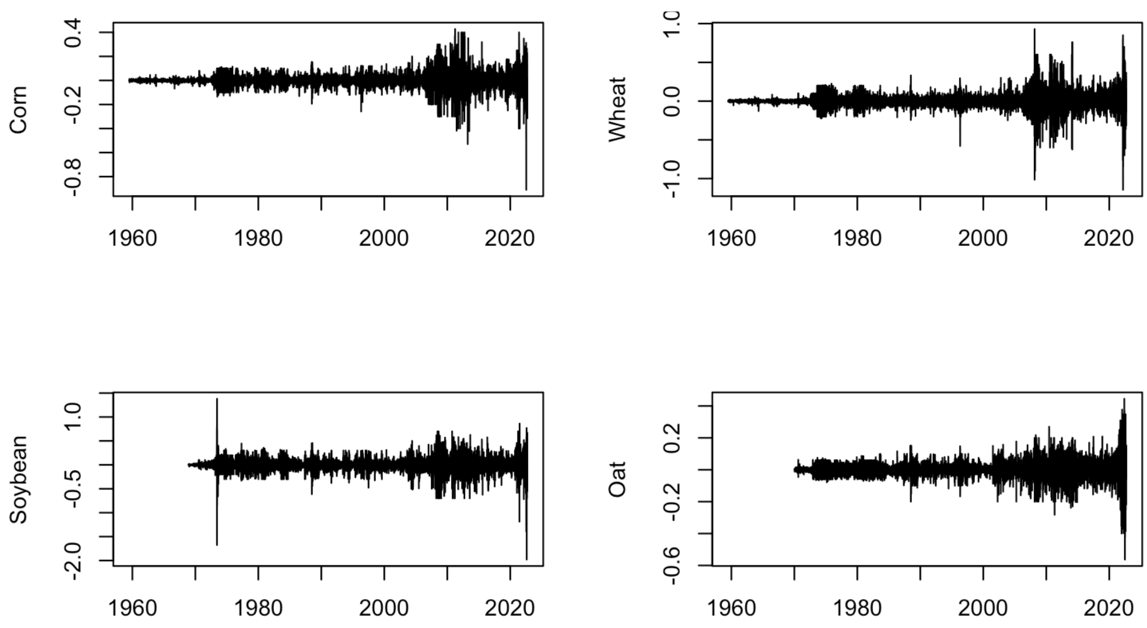

Figure 3. Variances of grain prices

Transformations such as logarithms, and then differencing of the log-transformed series can help stabilize the variance of a price series. The following figure shows the time plots for the differences of adjacent points in time for the price logarithms

$$r_t = log(x_t) - log(x_{t-1}) \quad (2)$$

for all series. The differenced logarithms in Eq. (2) above are known as log-returns of the prices, but we can also call them simply returns instead[13].

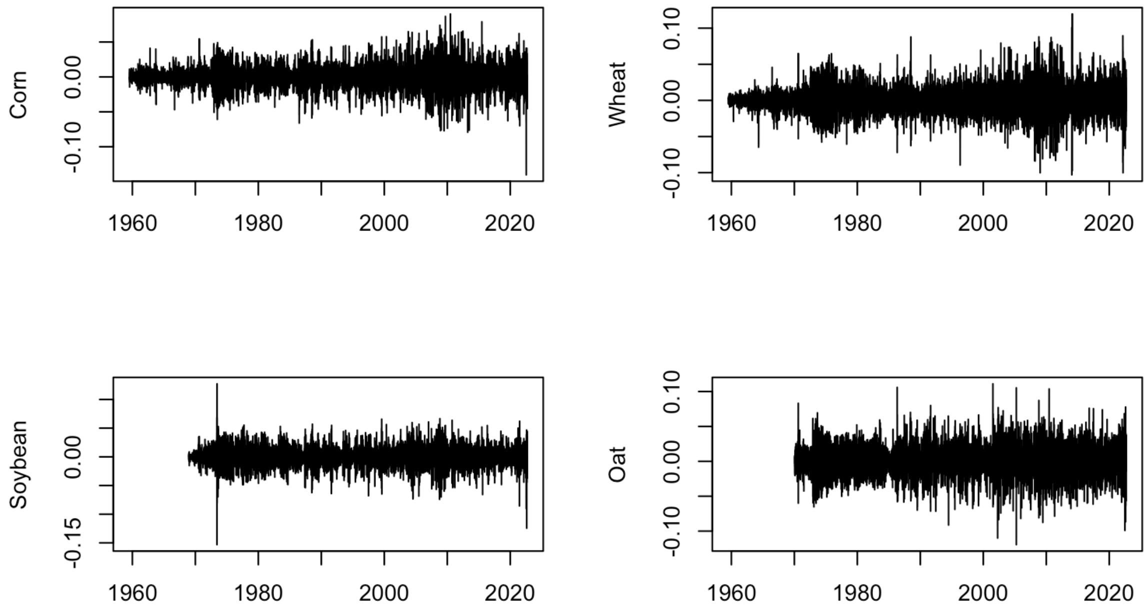

Figure 4. Price return series for the selected grains.

Even under this transformation, it is possible to note the presence of heteroscedasticity in the series.

1.2.2. ARIMA Analysis

The Autoregressive Integrated Moving Average or ARIMA models are a family of statistical models for time series that incorporate linear correlations in the series data, as well as nonstationary effects of integer order. The main goal in the implementation of the ARIMA models, is to describe the conditional mean of the process x_t

$$E[x_t \,|F_{t-1}] = \mu_t \quad (3)$$

where μ_t is the mean function of the process at time t , and $F_{t-1}$is the complete history of the process available at time $t-1$ [13]. These methods were popularized in the work of Box and Jenkins[14]. An ARIMA model is completely characterized by the orders and corresponding coefficients of each of the modeling components, denoted as $ARIMA(p,d,q)$, where p is the order of the autoregressive or $AR(p)$ component

$$x_t = w_t + \sum_{i=1}^{p} \phi_i \, x_{t-i} \quad (4)$$

where x_t is the observed series (it can also represent transformed data such as r_t), w_t is white noise (commonly assumed as Gaussian, but also heavy-tailed noise can be used) and ϕ_i are real coefficients[12]. The index d is the order of the integrated component and is estimated by the ADF test discussed in Sec. 1.2.1. The order d of integration indicates the order of differencing $\Delta^d x_t$ that we need to perform on the series to make it stationary. Finally, the index q represents the order of the moving average or $MA(q)$ component

$$x_t = w_t + \sum_{i=1}^{q} \theta_i \, w_{t-i} \quad (5)$$

which represents a simple description of the series purely in terms of errors w_t that become correlated by the summation and weighting procedure. Put together, for a given differencing order, the $AR(p)$ and $MA(q)$ components conform the $ARIMA(p, d, q)$ models

$$x_t = \mu_t + \sum_{i=1}^{p} \phi_i \, x_{t-i} + \sum_{i=1}^{q} \theta_i \, w_{t-i} + w_t \quad (6)$$

We choose the R language as the computational tool for doing the ARIMA analysis, due to its practical use and reliable, well-maintained libraries containing tested and efficient estimation methods. In particular, the auto.arima() function in R[15] was used to find the best fit for the series according to the ARIMA model. This function performs the following steps to select the best fit based on the AIC (Akaike Information Criterion):

1. It determines if the series have unit root (Dickey Fuller test). If they do, then it differentiates it, so it becomes a stationary time series.
2. Then calculates the Auto Correlation Function (ACF) for the series, to find the MA component (q)
3. It calculates the Partial Auto Correlation Function (PACF) for the series, to find the AR component (p)
4. Afterwards tries different combination of parameters that may adjust the model.
5. Finally calculates some estimators to compare the different possible fits and chooses de best option. The estimator considered for this analysis was the AIC (Akaike Information Criterion); the lower the value, the better is the model. This estimator is likelihood based. The greater the likelihood, the better.

The auto.arima() outcomes for the original series are presented in Table 2

Table 2. Results of auto.arima() fit for all grain original series

Price Series	Corn	Wheat	Soybean	Oat
ARIMA best fit	(0,1,3)	(4,1,0)	(0,1,1)	(0,1,2)
AIC	-48381.52	-33234.3	-16891.91	-44502.87
AR coefficients	-	AR1 0.0260 AR2 -0.0171 AR3 -0.0358 AR4 0.0257	-	-
Integrate order	1	1	1	1
MA coefficients	MA1 0.0699 MA2 -0.0155 MA3 0.0455	-	MA1 0.0756	MA1 0.1335 MA2 -0.0404

1.2.2.1. Analysis of ARIMA residuals

Residuals show all the information not recovered by the proposed model, reflected on the behavior of their ACF. Since the objective for all models is to be able to recuperate most of the behavior of a series, the expected performance for the residuals for a good fit must be like white noise.

The next figure shows the ACF graph for the residuals obtained after fitting the original series to the ARIMA models.

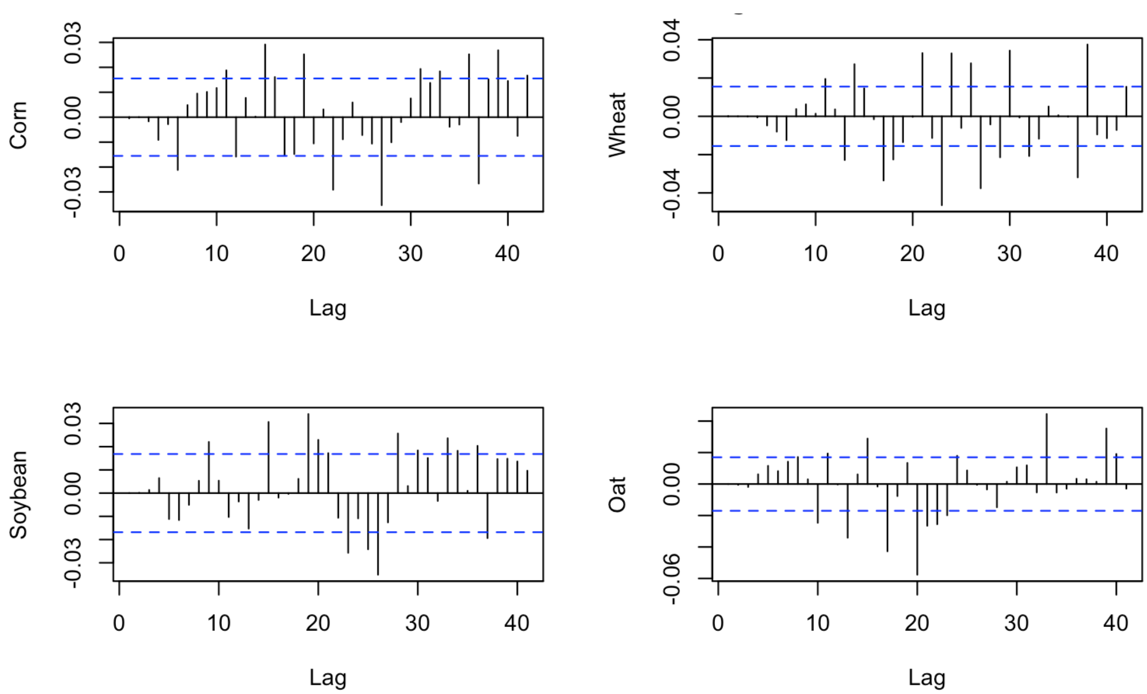

Figure 5. ACF for the ARIMA residuals for the original series

Despite the significant peaks around lag 15th for most plots, we can conclude that the behavior of the residuals is close to a white noise. These significant lags can be explained from spurious correlation in the original data introduced by the finite precision in the price values.

1.2.3. ARCH/GARCH Analysis

Analogous to the ARIMA models, the goal of the ARCH/GARCH processes is to model the observed conditional covariance of the time series, encoded in the empirical autocovariance function σ_t

$$Var[r_t \mid F_{t-1}] = E[(r_t - \mu_t)^2 \mid F_{t-1}] = E[\eta_t^2 \mid F_{t-1}] = \sigma_t^2 \quad (7)$$

where $\eta_t = r_t - \mu_t$ are the *excess returns* or *return residuals*. This is done by expressing the returns directly in terms of the conditional variance σ_t and then setting a model for its time evolution as a separate equation, thus obtaining the following extended model

$$\eta_t = \sigma_t w_t \quad (8.1)$$

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{P} \alpha_i \, \eta_{t-i}^2 + \sum_{j=1}^{Q} \beta_j \, \sigma_{t-j}^2 \quad (8.2)$$

where α_0, α_i, β_j are real coefficients, and $w_t \sim N(0,1)$. The equations above are known as GARCH(P,Q) equations.

In Figure 6 below, we show the ACF plot for the squared residuals from the ARIMA fit. This plot gives information to determine if an ARCH-GARCH approach is necessary to as a better model for the series. In this case, the persistent correlation implies that it is necessary to consider an ARCH/GARCH model for the conditional variance. [13]

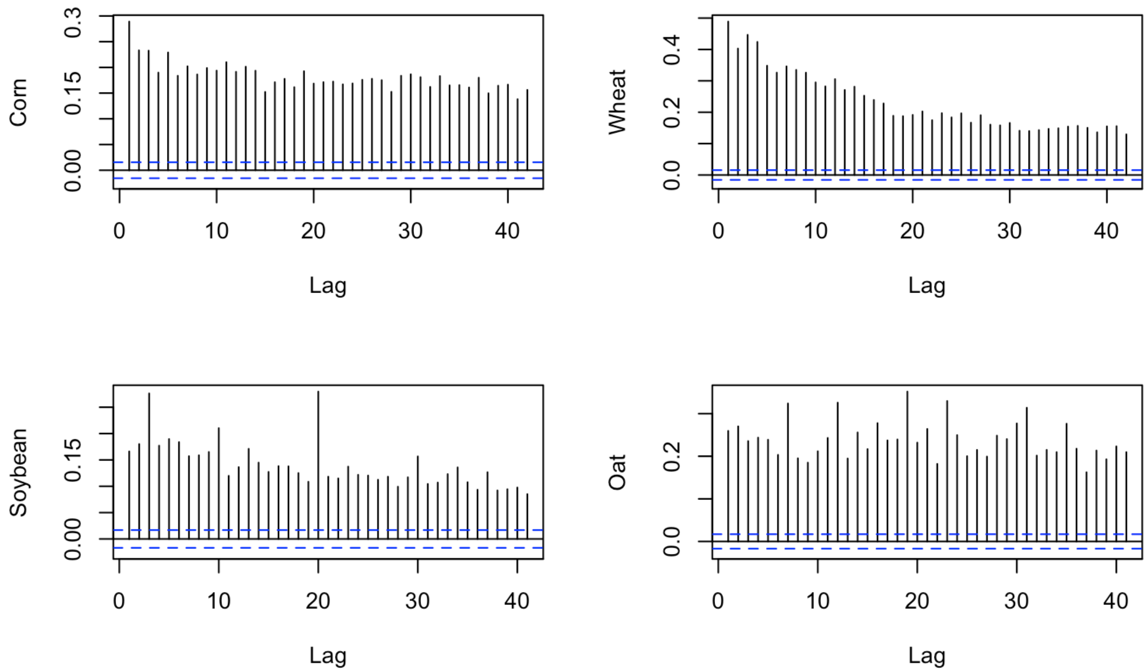

Figure 6. ACF for the ARIMA squared return residuals

Several different GARCH models were evaluated for the differenced log-series for corn. The next chart shows the AIC obtained for each fit.

Table 3. AIC obtained for different GARCH models for the corn series

GARCH Model	AIC
(1,0)	-5.811610
(2,0)	-5.882341
(1,1)	-6.078745
(2,2)	-6.079809
(1,2)	-6.079934
(2,1)	-6.078534
(10,10)	-6.079830

In Table 3 AIC stands for Akaike Information Criterion. It is used as a model selection tool, defined as

$$AIC \; = \; log(\widehat{\sigma_k}) + \frac{n+2k}{n} \quad (9.1)$$

$$\widehat{\sigma_k^2} = \frac{SSE(k)}{n} \quad (9.2)$$

where $SSE(k)$ is the residual sum of squares under the given model with k regression coefficients. AIC is then a balance between the estimation error given by the model, $\widehat{\sigma_k^2}$, and the number of parameters of the model. Then the criterion is to choose the model with the smaller AIC between the proposals, meaning that this would be the model that balances estimation error and complexity.

The smallest AIC obtained corresponds to the GARCH(1,2) model. However, given the small difference between its AIC and the corresponding to the GARCH(1,1) fit, we select the GARCH(1,1) to reduce the complexity of the analysis without jeopardizing the results.

The GARCH(1,1) model was taken to fit all log-return series. In Table 4 below, we show the coefficients obtained for the GARCH(1,1) fit for each grain.

Table 4. Coefficients obtained for the GARCH (1,1) fit for all logarithm series

Coefficient	Corn	Wheat	Soybean	Oat
μ	9.731e-05	5.7192e-05	1.6760e-04	8.3816e-05
α_0	6.657e-07	3.1707e-07	6.8336e-07	4.8887e-06
α_1	8.607e-02	7.0819e-02	7.3940e-02	7.4681e-02
β_1	9.158e-01	9.3247e-01	9.2711e-01	9.1358e-01
AIC	-6.078745	-5.763724	-5.872564	-5.248908

1.2.4. Models for different time scales

The time-series for all four grains have data for more than 50 years. We already adjust the models for the entire series; in this section we intend to model them with different time scales to see how they behave by considering different amount of historical data.

1.2.4.1. ARIMA

Table 5. auto.arima() coefficients for 5-year data

Price Series	Corn	Wheat	Soybean	Oat
ARIMA best fit	(0,1,0)	(1,1,2)	(0,1,0)	(0,1,2)
AIC	-2660.2	-1295.98	-819.19	-2633.54
AR coefficients	-	AR1 -0.6342	-	-
Integrate order	1	1	1	1
MA coefficients	-	MA1 0.6717 MA2 0.1052	-	MA1 0.1425 MA2 -0.0825

Table 6. auto.arima() coefficients for 10-year data

Price Series	Corn	Wheat	Soybean	Oat
ARIMA best fit	(1,2,0)	(2,1,3)	(1,1,1)	(2,1,0)
AIC	-4785.4	-3402.58	-1981.71	-6077.86
AR coefficients	AR1 -0.4891	AR1 0.4056 AR2 0.5190	AR1 -0.2241	AR1 0.1438 AR2 -0.1032
Integrate order	2	1	1	1
MA coefficients	-	MA1 -0.3953 MA2 -0.4747 MA3 -0.9785	MA1 0.2518	-

Table 7. auto.arima() coefficients for 20-year data

Price Series	Corn	Wheat	Soybean	Oat
ARIMA best fit	(1,1,1)	(3,1,3)	(0,1,1)	(0,1,2)
AIC	-10799.65	-5979.22	-3711.45	-13443.13
AR coefficients	AR1 -0.7075	AR1 -0.4080 AR2 -0.4134 AR3 -0.9489	-	-
Integrate order	1	1	1	1
MA coefficients	MA1 0.7609	MA1 0.4161 MA2 0.4192 MA3 0.9239	MA1 0.0359	MA1 0.1360 MA2 -0.0505

1.2.4.2. GARCH (1,1)

Table 8. GARCH (1,1) coefficients for 5-year data

Coefficient	Corn	Wheat	Soybean	Oat
μ	5.761e-04	5.443e-04	4.177e-04	6.360e-04
α_0	5.121e-06	1.257e-05	9.346e-07	3.684e-10
α_1	8.422e-02	8.951e-02	6.186e-02	1.719e-02
β_1	8.977e-01	8.747e-01	9.377e-01	9.842e-01
AIC	-5.672156	-5.225858	-6.003328	-5.149235

Table 9. GARCH (1,1) coefficients for 10-year data

Coefficient	Corn	Wheat	Soybean	Oat
μ	9.330e-05	-1.417e-04	9.680e-05	6.913e-05
α_0	7.362e-06	9.533e-06	2.060e-06	5.918e-06
α_1	8.299e-02	7.435e-02	6.177e-02	3.152e-02
β_1	8.861e-01	8.966e-01	9.285e-01	9.522e-01
AIC	-5.670364	-5.306098	-5.970762	-5.146528

Table 10. GARCH (1,1) coefficients for 20-year data

Coefficient	Corn	Wheat	Soybean	Oat
μ	2.172e-04	6.169e-05	3.296e-04	2.319e-04
α_0	4.097e-06	6.341e-06	2.083e-06	1.907e-05
α_1	6.186e-02	5.914e-02	5.8370e-02	5.574e-02
β_1	9.258e-01	9.260e-01	9.339e-01	8.961e-01
AIC	-5.404566	-5.07244	-5.713074	-5.040362

1.2.5. Model validation (5, 10 and 20 years)

We divide the data into an initial training segment and a validation final segment. We use the first 98% of the data to train the models, then make predictions and compare them with the last 2% of the data that we set aside for testing. This 2% corresponds to 30 days for predictions.

1.2.5.1. Short-term forecasting with ARIMA – 10 years

The following graphs show the training segment of the time series in black, and the predictions in color blue with the confidence interval highlighted in gray. It is easy to visually compare the forecast obtained by comparing these blue lines versus the test segment which is shown in red[3].

Figure 7. Forecast Corn Arima (1,2,0)

[3] See appendix A for graphs with zoom in the predict segment.

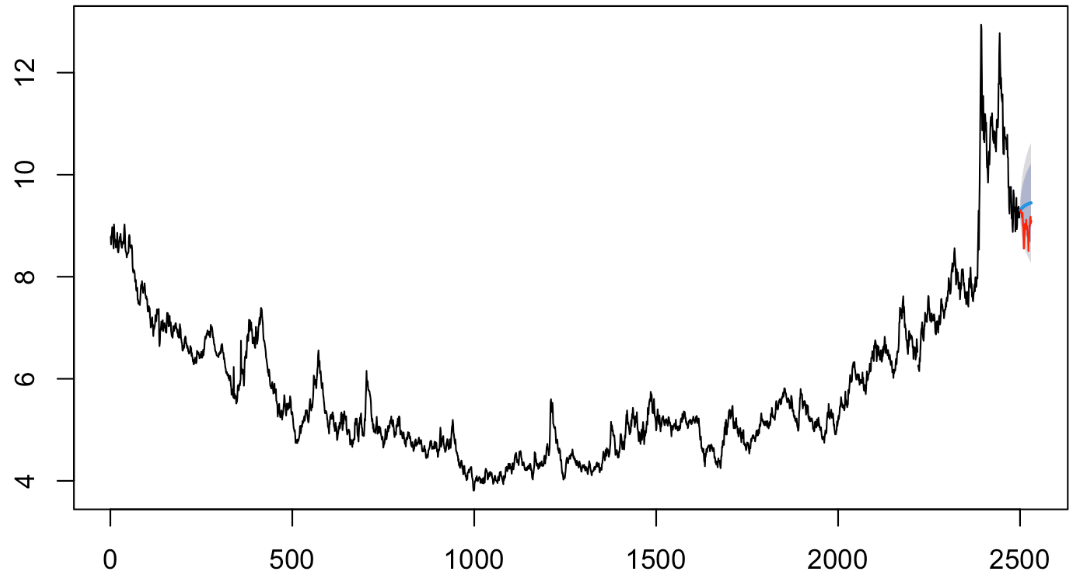

Figure 8. Forecast Wheat Arima (2,1,3)

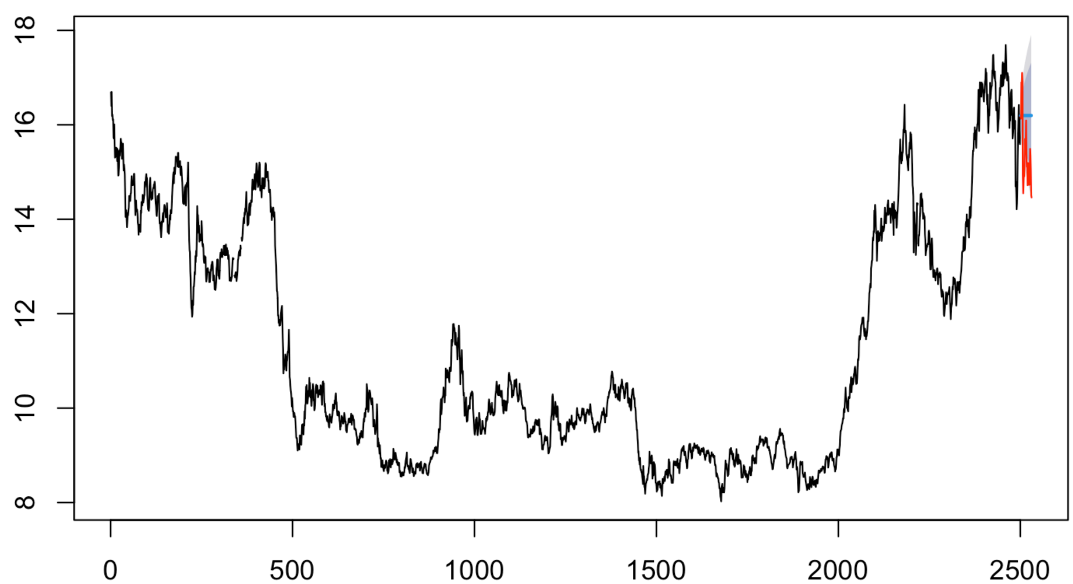

Figure 9. Forecast Soybean Arima (1,1,0)

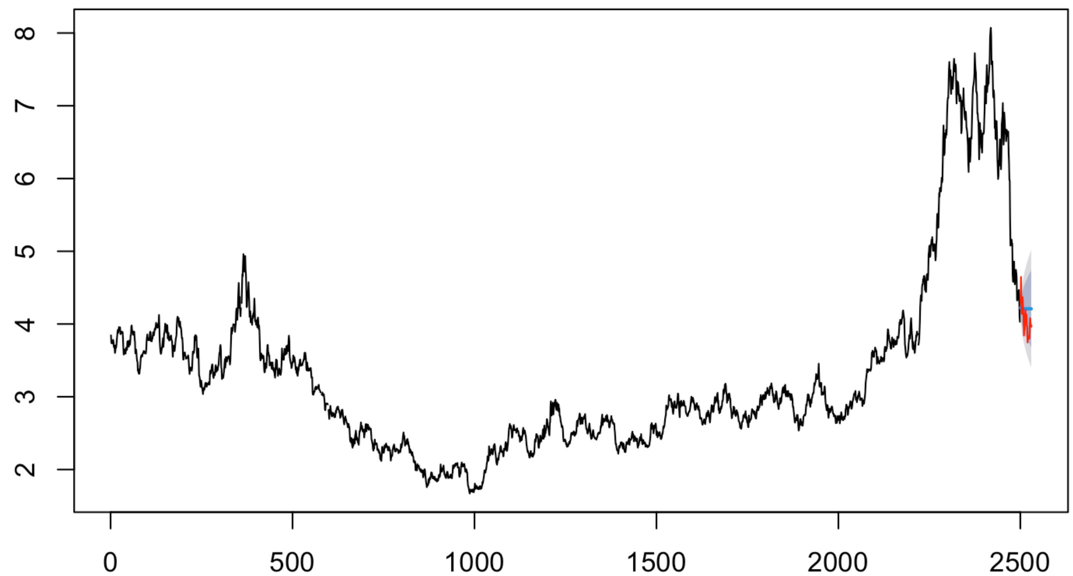

Figure 10. Forecast Oat Arima (2,1,0)

The comparison between the complexities of the models for the 5, 10, and 20-year periods would give us a reasonable parsimony argument for picking the 10-year period models, since they have the best compromise of parameter orders among the three groups. In other words, the 10-year models are not as simple as the 5-year models and not as complicated as the 20-year models. However, we resort in an additional argument to help us in the selection. We consider the MAD (Mean Absolute Deviation) between two series X and Y of length n, defined as

$$MAD(X,Y) = \frac{1}{n}\sum_{i=1}^{n}|X_i - Y_i| \quad (10)$$

as a simple criterion to evaluate the performance of these models. The results are as follows:

Table 11. MAD for all grains

Grain	5 years	10 years	20 years
Corn	0.4766	0.6121	0.4759
Wheat	0.3293	0.4062	0.3136
Soybean	0.9783	0.9981	0.9777
Oat	0.2759	0.2213	0.2770

We evaluate the MAD since it makes the results between different time periods comparable. The choice of mean squared errors (MSE) would not be the best choice since the variance itself is a dynamical quantity for the present series, as shown by the GARCH(1,1) fit. In contrast, the MAD is a more robust estimator[16].

Despite not having the best MAD (see table 11), we decide to pick the models for the 10-years period for benchmarks in Section 2.4 with the MCMC estimation, for two main reasons: 1. We need a longer period of data to have better statistics and 2. The models for the 10-year period are the more parsimonious, as discussed previously.

1.2.5.2. GARCH(1,1)

As the GARCH analysis intend to predict volatility, the following graphs show the volatility of the training segment of the time series in black, and the predictions in color red, with the confidence interval marked with blue.

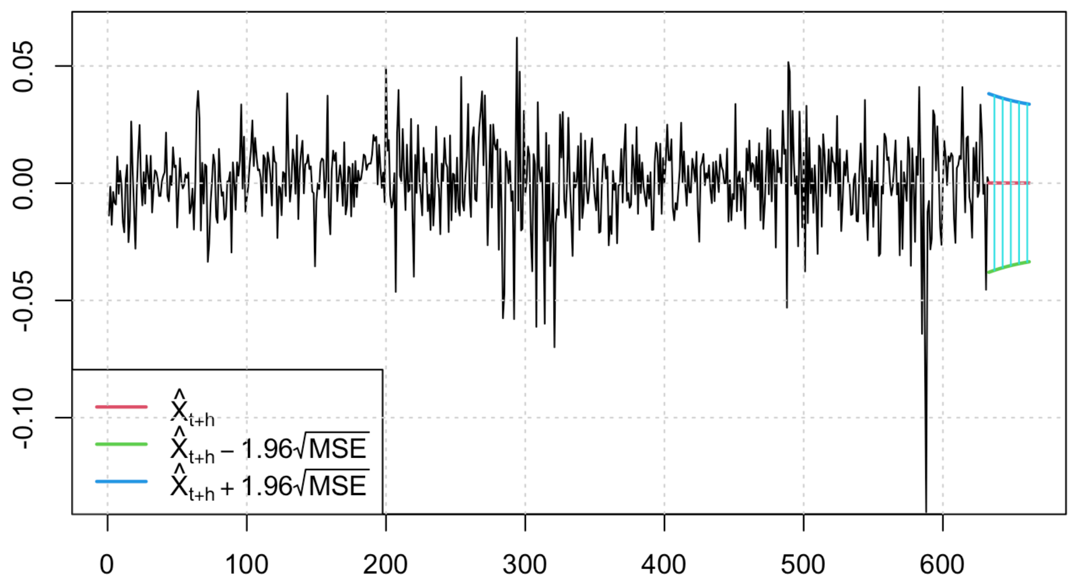

Figure 11. Volatility prediction GARCH (1,1) Corn

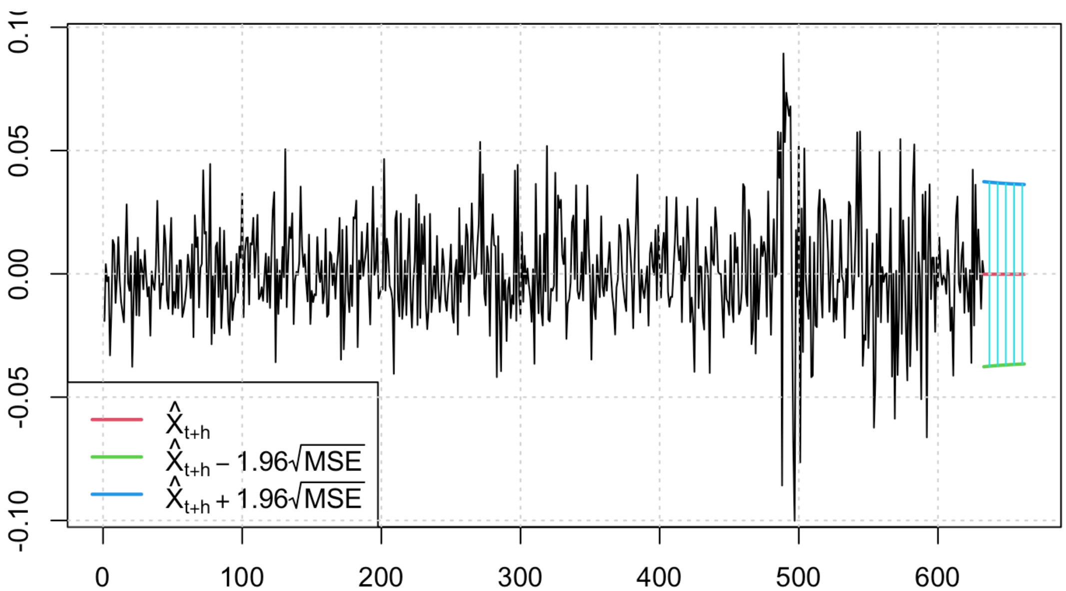

Figure 12. Volatility prediction GARCH (1,1) Wheat

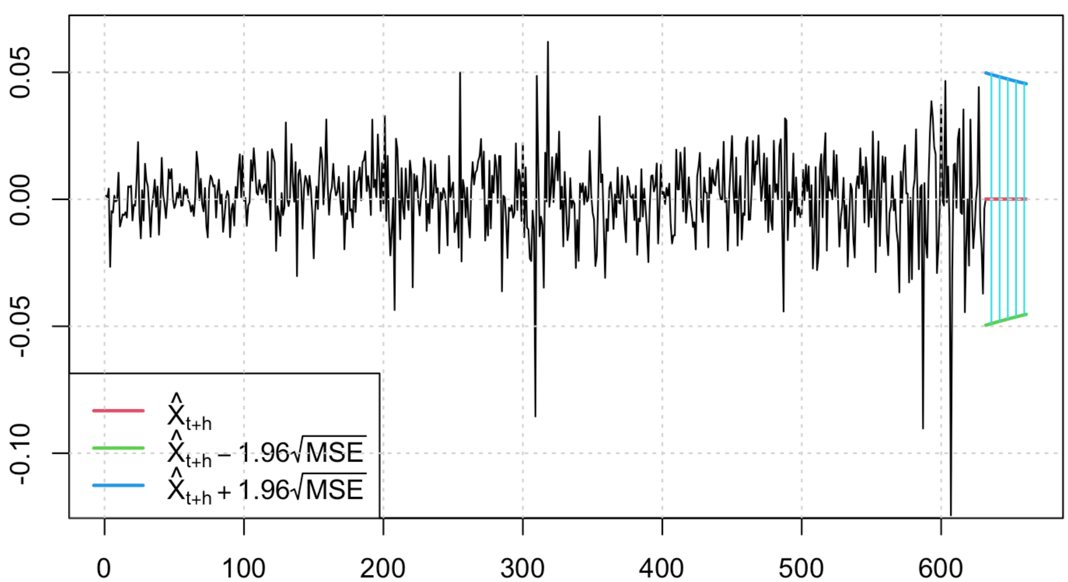

Figure 13. Volatility prediction GARCH (1,1) Soybean

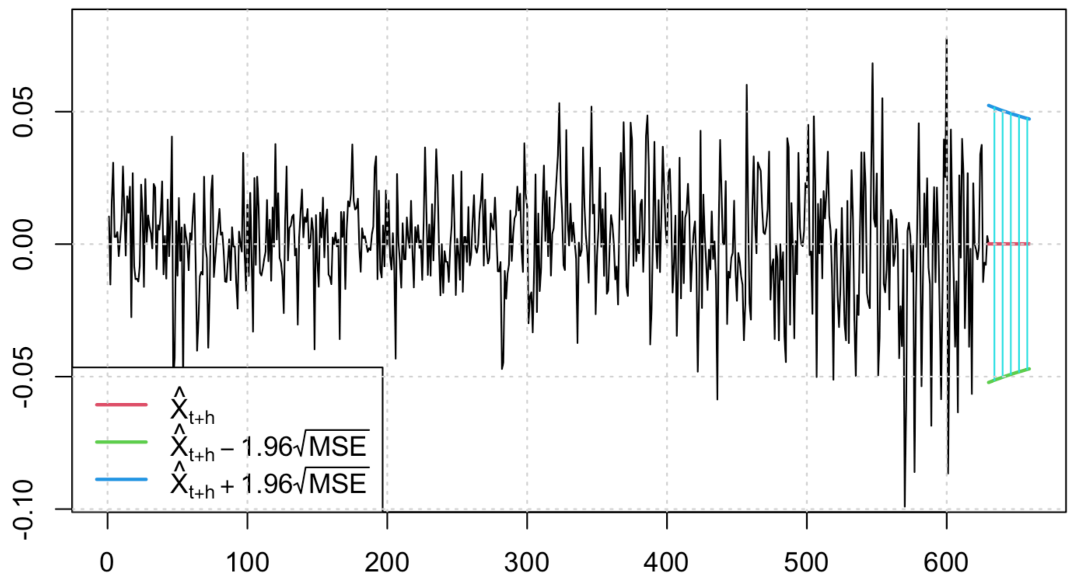

Figure 14. Volatility prediction GARCH (1,1) Oat

The above results show that there is room for another model approach that might capture better the complex dynamics of the volatility. The next part of this study presents an alternative based on a Bayesian approach to estimate the parameters of a Stochastic Volatility (SV) model via Markov Chain Monte Carlo (MCMC).

PART 2. BAYESIAN APPROACH: MARKOV CHAIN MONTE CARLO

2.1 Bayesian statistics and sampling

The central idea in Bayesian statistics is to update our knowledge on some conditional aspects of a model, such as the conditional probability of observing a set of parameters Θ and a collection of state variables S given a set of observations X, this is $p(\Theta,\ S\,|X)$. We then consider that it is possible to have a pre-existent *a priori* knowledge or belief about the distribution of Θ, encapsulated in the so-called *prior distribution* $p(\Theta)$*.* The former means that, in the Bayesian view, we treat parameters and other components of the model as random variables themselves.

The goal is now to take this prior knowledge to estimate the distribution of the unknown parameters of the model conditional on the observations X and take samples from it and performing the computation of the likelihood function $p(X \mid \Theta,\ S)$. The conditional probability distribution $p(\Theta,\ S \mid X)$ corresponding to the updated target information is called *posterior probability distribution* and represents the updated knowledge of the model parameters and state variables*.* The previous statements are summarized in the formula

$$p(\Theta,\ S\,|X) \ = \ \frac{p(X \mid \Theta,\ S)p(S \mid \Theta)p(\Theta)}{p(X)} \qquad (11)$$

known as Bayes' theorem or Bayes' formula[17]. We can build powerful models for analysis and forecasting by merging the Bayesian approach with dynamical models from the theory of stochastic processes and combining with time series analysis techniques.

2.2 Dynamic asset pricing models

To take advantage of the updating scheme of the Bayesian approach, we need to incorporate a specific pricing model for finding its parameters. The best candidate lies within the family of dynamic asset pricing models. In dynamic asset price modelling, the objective is to propose dynamical models to explain the general asset price behavior in time: The parameters Θ and state variables S are considered as given, and the dynamic model describes the evolution of the observations X . The converse problem emerges when analyzing data: Given a set of observations X , a model must be proposed. Hence, the goal is to obtain information about the corresponding parameters Θ and other variables of interest (state variables S), such that they explain best the observations X[18]. This means that the empirical analysis is an inverse statistical inference problem. A realistic but minimal *(i.e., parsimonious)* stochastic volatility model is given by

$$dX_t = X_t(r_t + p_t)dt + X_t\sqrt{V_t}\, dW_t^x + d\left(\sum_{j=1}^{N_t} X_{\tau_j} - (e^{Z_j} - 1)\right) \quad (12.1)$$

$$dV_t = \kappa_v(\theta_v + V_t)dt + \sigma_t\sqrt{V_t}\, dW_t^v \quad (12.2)$$

where W_t^x and W_t^v are independent Brownian motions for the asset price and the volatility, respectively, N_t counts the number of jump times τ_j, prior to time t, Z_t are the jump sizes, p_t is the equity risk premium, and r_t is the spot interest rate[18].

Despite their solid theoretical basis, continuous time models like Eq. (12.1) are tricky to implement in real world situations due to the discrete nature of the actual financial time series. Although methods for making a proper connection between data and models can be utilized[18], the interpretations of the results can be obscured by details and the parameter estimation is many times not robust even with MCMC methods due to the very large times necessary for the chains to convergence[13].

Therefore, instead of Eq. (12.1) we resort in the equivalent discrete-time version of SV models, resulting in something like a model of the GARCH family. We could stick to the GARCH(1,1) model studied in Section 1.2.3, but the poor predictive power and wide confidence intervals provided by the GARCH(1,1) model (see Figs 11 to 14) conform sufficient justification for the comparison with a more complete approach.

This is not a surprise, since the modeling and forecasting of financial volatility is inherently difficult due to the fact that the volatility is a latent quantity and thus not possible to observe directly. This fact is one of the main reasons (and necessity) for the incorporation of MCMC into our analysis: A latent variable can be easily incorporated into a Markov model, more precisely a Hidden Markov Model (HMM)[19] with the important advantage that discrete-time SV models incorporate two independent innovations[20] [13], and model both the process value (mean equation) and its volatility (variance equation) through the following pair of expressions common in the literature (with variations) [13][21]

$$y_t = \epsilon_t \exp\left(\frac{h_t}{2}\right) \quad (13)$$

$$h_{t+1} = \mu + \phi(h_t - \mu) + \sigma\delta_t \quad (14)$$

$$h_1 \sim N\left(\mu, \frac{\sigma}{\sqrt{1-\phi^2}}\right) \quad (15)$$

$$\epsilon_t \sim N(0,1)$$

$$\delta_t \sim N(0,1)$$

where μ is the mean log volatility; ϕ represents the persistence of the volatility term (as in a usual AR model); ϵ_t is a white noise shock on the asset return at time t and δ_t models the corresponding error on the volatility at time t. Finally, the most interesting term and the key to the SV model is h_t, that represents the log-volatility as a latent variable. We introduce this model in the R-language[22] implementation of Stan[23] the multiplatform Software for Bayesian analysis. The results of these simulations are discussed later in Section 2.4.2

2.3 Markov Chains and Monte Carlo methods: MCMC

It can be argued that the origin of MCMC methods lies within physics, with the development of the Monte Carlo method by von Neumann, Ulam and Metropolis for the study of Neutron diffusion during the late 1940's[24] [25] in Los Alamos, culminating in a paper demonstrating the final form of the Metropolis algorithm in 1953[26], and an important later extension due to Hastings[27].

The use of MCMC methods gained strength in the mathematical statistics literature until 1990's due to Gelfand, Smith and others[28].

Nowadays, MCMC are well-established methods of increasing popularity in many applications for its power to sample unknown dynamic variables for complex, high dimensional systems[17].

Our goal is to choose an appropriate dynamic asset pricing model, such as a minimal stochastic volatility model that coupled to MCMC can be solved for its parameters and state variables describing the price series treated in the previous sections. A candidate stochastic volatility model is given in previous section in Eq. (12.2)

We can perform a comparison between the model in Eqs. (12), with the distributions in the denominator in Eq. (11) sampled by MCMC, and the classical approach treated in previous sections. Finally, we will be able to conclude whether there is an improvement in capabilities or if the classical models are sufficient for forecasting and volatility monitoring.

2.4 RStan

Just like the first part of this study, we choose the R language as the computational tool for running the SV models described in section 2.2. Particularly, we used the RStan library, which is the R interface to Stan.

Stan is a C++ probabilistic library that implements Bayesian statistical inference via Markov Chain Monte Carlo[29].

2.4.1 Benchmark of Libraries

To compare the results between the function auto.arima() from the stats library and the RStan package (the implementation of Stan in R), we first run the selected ARIMA models fitted from the 10-year datasets for all four grains and compare the coefficients obtained versus the ones calculated in Table 6. section 1.2.4.1 with the auto.arima() function.

Below are the results for each grain price series. The tables include the coefficients fitted from both approaches to make the comparison easy. Tables also include the standard error for each parameter calculated with the auto.arima() function. This metric helps understand the level of looseness of the coefficients from the frequentist approach and let us confirm that the Bayesian approach delivers indeed similar results in most cases when considering the standard error of the estimation in the auto.arima() fit.

The results from MCMC through Stan were all computed with 4 chains and 1000 iterations for each chain, giving the corresponding average values and histograms below. We discuss the results afterwards.

Table 12. RStan CORN Arima coefficients vs Auto.arima()

CORN 1,2,0			
Approach	Auto.arima()		RStan
	Coef	Stand error	
Phi	-0.4891	0.0174	-0.4878
Sigma	-	-	0.0938

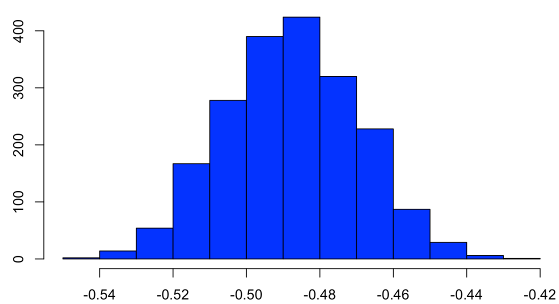

Figure 15. RStan Phi histogram (corn)

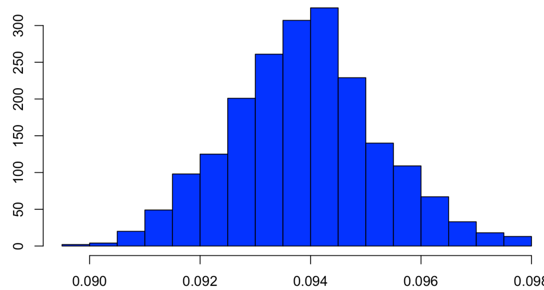

Figure 16. RStan Sigma histogram (corn)

We can clearly appreciate that the result from Stan MCMC estimation provides an AR coefficient ϕ that is approximately equal to the one given by maximum likelihood estimation (MLE) in auto.arima().

Table 13. RStan WHEAT Arima coefficients vs Auto.arima()

WHEAT 2,1,3			
Approach	Auto.arima()		RStan
	Coef	Stand error	
Phi1	0.4056	0.1419	
Phi2	0.5190	0.1392	
Theta1	-0.3953	0.1420	
Theta2	-0.4747	0.1429	
Theta3	-0.0785	0.0209	
Sigma	-	-	

We were not able to run the code for this grain. This might be an issue of convergence of the chains in the MCMC method.

Table 14. RStan SOYBEAN Arima coefficients vs Auto.arima()

SOYBEAN 1,1,1			
Approach	Auto.arima()		RStan
	Coef	Stand error	
Phi	-0.2241	0.6432	-0.1454
Theta	0.2518	0.6317	0.1679
Sigma	-	-	0.1633

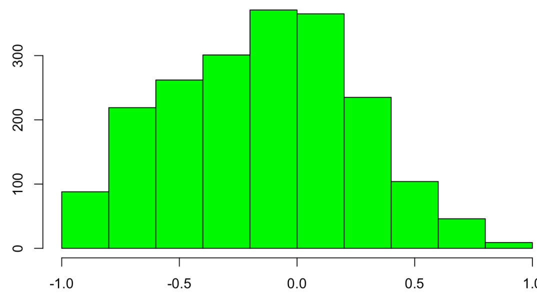

Figure 17. Rstan Phi histogram (soybean)

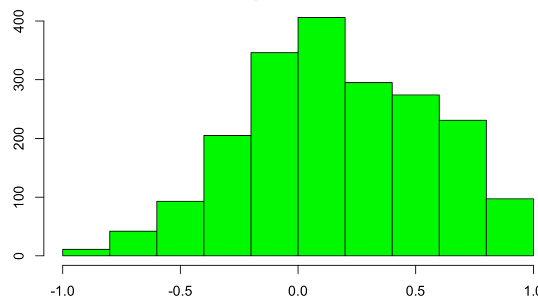

Figure 18. RStan Theta histogram (soybean)

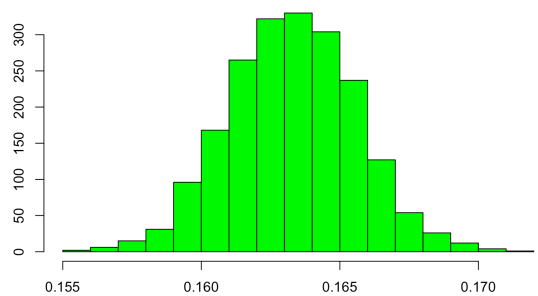

Figure 19. RStan Sigma histogram (soybean)

For this case, the MCMC chains computed with Stan give a result that, even if not close, it is within the margin of error provided by the standard error of MLE from auto.arima(). As it can be appreciated in this example, the MLE estimation gives a very large standard error, contrasting shockingly with the very small value for the first example, but we are talking of the same kind of time series. This just tells us how complicated these time series are and how tricky is to make interpretations based only in one type of models and with one method of parameter estimation. We can observe the skewness and large spread of the histograms, a symptom of bad convergence.

Table 15. RStan OAT Arima coefficients vs Auto.arima()

OAT 2,1,0			
Approach	Auto.arima()		RStan
	Coef	Stand error	
Phi1	0.1438	0.0198	0.1443
Phi2	-0.1032	0.0198	-0.1035
Sigma	-	-	0.0725

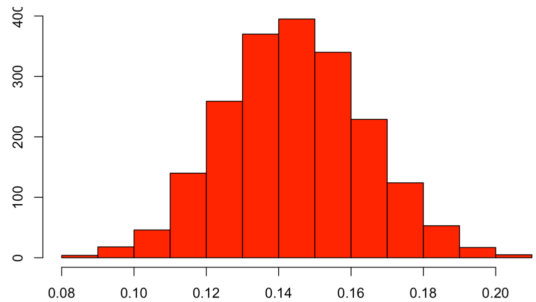

Figure 20. RStan Phi1 histogram (oat)

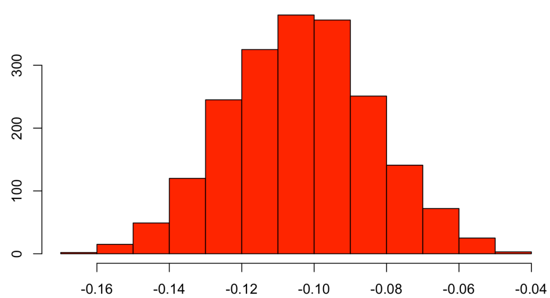

Figure 21. RStan Phi2 histogram (oat)

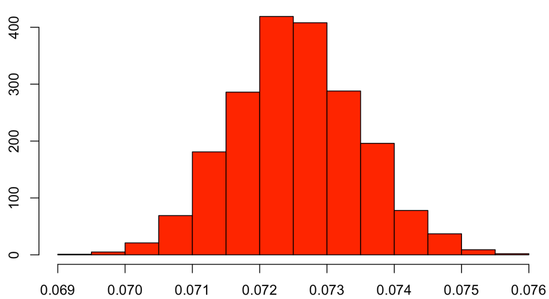

Figure 22. RStan Sigma histogram (oat)

We observe matching coefficients very well contained within the standard error margin. It is worth noting that we observed convergence also when treating the whole time series.

2.4.2 SV model analysis

As the RStan library can also fit the coefficients for the classic models and the results are reasonable whenever there is an agreement, and, furthermore, since it gives us a more complete and detailed perspective thanks to the ensemble nature of the MCMC approach (histograms), we will use it for the estimation of the parameters of the SV models, for which we don't have a way to estimate the parameters via frequentist (MLE) approach to cross validate. The results are as follows:

Table 16. SV model coefficients

	Corn	Wheat	Soybean	Oat
Phi	0.9957	0.9955	0.9937	0.9960
Sigma	0.1402	0.1439	0.1246	0.1473

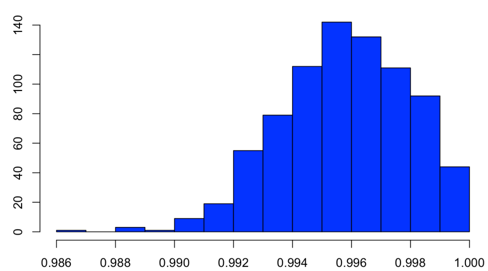

Figure 23. SV model Corn Phi histogram

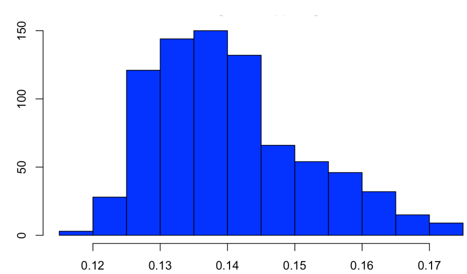

Figure 24. SV model Corn Sigma histogram

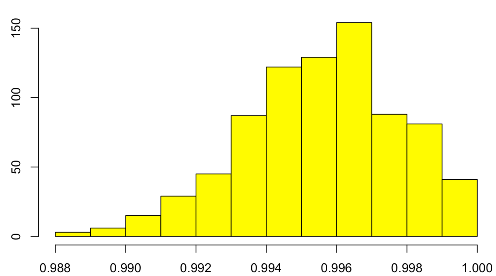

Figure 25. SV model Wheat Phi histogram

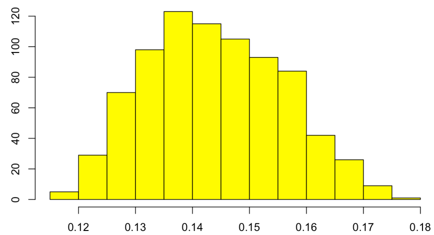

Figure 26. SV model Wheat Sigma histogram

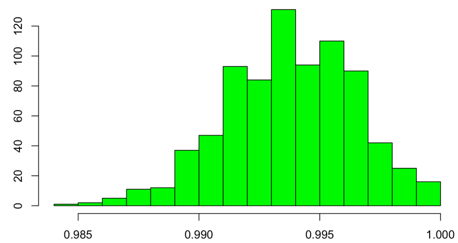

Figure 27. SV model Soybean Phi histogram

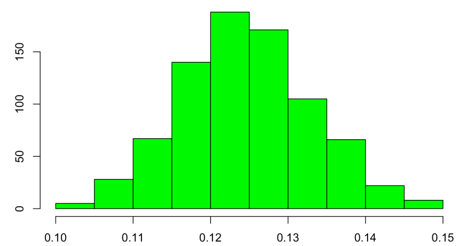

Figure 28. SV model Soybean Sigma histogram

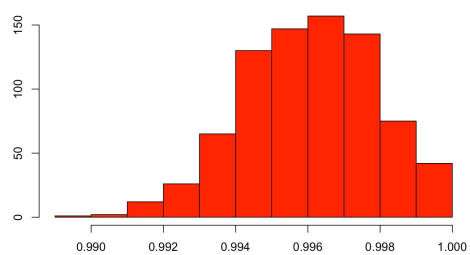

Figure 29. SV model Oat Phi histogram

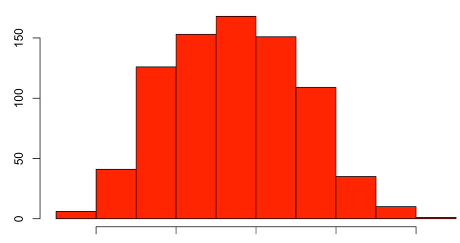

Figure 30. SV model Oat Sigma histogram

The skewness and large spread of the histograms show limited convergence of the iterations of the 4 chains of 1000 iterations used to estimate this SV model parameters. We also know that for the specific algorithm used by Stan called NUTS (stands for no U-turn sampler) [21] [13] there is significant autocorrelation between the MCMC samples, and thus the convergence is compromised.

The parameter phi in the SV used indicates the persistence of the volatility. Here this persistence is indicated as very strong for all the series (significantly close to 1) and that means that when the volatility is high in the price series, it tends to stay high, and when it is low, it tends to stay low, for an unpredictable length of time, until some random event or shock makes it change behavior again.

The parameter sigma is the instantaneous. Sigma has a less intuitive interpretation, being the standard deviation of the latent variable h_1 in the SV model, which corresponds to the log-volatility, that as mentioned previously, it is not directly observable. However, the direct observation we can draw from table 16 is that that σ, the expected amplitude of the logarithm of the volatility is similar for all the price series. This is something that is not necessarily intuitively clear, but it is also not a crazy idea, considering that all of them are the same kind of product and, although they are produced in different regions in the world, they are so fundamental in the economy and in the financial system in general, that they are part of the global market with a similar strength. This intuition makes sense as they are even very correlated, as it can be appreciated from the plot in Fig. 2. Thus, it is not that surprising that they display the same scale of the log- volatility even though the particular dynamics of their volatilities are different.

CONCLUSIONS

Conclusions

We have achieved a (perhaps deceivingly) satisfactory classical analysis of the four series. The classical analysis performed in the first part of the Thesis provides better results statistically speaking because of the larger series considered for the implementation of the MLE estimation of the orders and parameters of the autoregressive models and the volatility analysis through the GARCH models. However, a word of caution is in order, and although the estimations are reasonable and statistically significant, we should not merely accept the first results only because we had large series to work with and easy methods of estimation that give simple outputs, as it is the typical use and "easy" interpretation of the ARIMA or GARCH models.

As the contrasted MCMC results in the second part showed, time series analysis is not a simple task in general, and specifically in finance is a very tricky endeavor. The very reason of this is the extremely complex nature of financial time series, due among many things to the multiple and mixed stochastic processes coupled to each other at several time and even different spatial scales.

Therefore, it is better to resort on more than one model and, in turn, on more than one method of estimation of the parameters of those models, and then make benchmarks, which is basically the approach taken in this Thesis.

Finally, the person or persons performing the modeling will decide based on information that might escape the capabilities of the methods and the mathematical analysis and careful comparison of approaches will serve as a conscious ground for decision making, but not as an absolute truth to follow blindly. As it could be seen in Part II, the Bayesian approach through the MCMC method for estimating the same parameters of the ARIMA models gives us a more detailed picture of the parameter estimation problem, and it helped us to corroborate some of the models and to take a second look to others, or even to disregard them until a more detailed analysis is made.

Although more detailed in principle, we observed that there were potential issues of convergence of the MCMC method, attributed in part to the short chains and low number of them, and the short original data series of 10 years. But it is worth noting that we observed convergence of the MLE approach from auto.arima() when treating the whole time series is not completely appropriate. Using the whole time series of more than 50 years for forecasting the prices and/or volatilities ahead for few days would give nonsensical results, since the full time series contains very mixed information, from very different periods in time, very different historical events and different technology, different economic policies and even a different dynamics in the financial markets and different rules in the financial operations since various kinds of financial instruments have been implemented in the financial markets in the last decades[30].

That is so far for the price series itself. The issue of volatility modeling is even more complex and the field of research on this direction is very active and rather young[13][20], since standard modeling techniques are not capable of capturing the huge complexity of the problem, as could be seen by simple comparison of two simple modeling approaches of volatility.

Nevertheless, these simplified models can give reasonable approximations when properly selected, estimated, and benchmarked, and a monitor for volatility could be built from a combination between GARCH and SV models in the future, once convergence problems are solved and the reasons for the discrepancies, when they occurred, are properly clarified. I consider this a satisfactory first exploration of the fascinating world of mathematical finance and its empirical problems, thanks to which I realize more and more how rich, and complex are the problems in real-world data analysis.

Future work

1. *Clarify the reasons for the nonconvergence or the errors from the Stan library for some of the model parameter estimates.*
2. *Run the MCMC analysis on Stan but with much more iterations and more chains. For time and computational constraints, we did not do it for this Thesis.*
3. *Use the results to try to build an actual monitor that can be updated automatically with the daily price data. This of course represented a lot of time and effort that we could not provide for the purpose of this thesis, but remains as a nice idea to implement, if possible, in the future.*

BIBLIOGRAPHY

[1] *International Food Policy Research Institute*. (March 30, 2022). IFPRIUkraine: https://www.ifpri.org/blog/russia-ukraine-war-exacerbating-international-food-price-volatility

[2] *International Monetary Fund*. (April, 2022). IMF: https://www.imf.org/external/pubs/ft/ar/2020/eng/spotlight/covid-19/

[3] (04 de 2022). NatGeo: https://www.nationalgeographic.org/encyclopedia/grain/

[4] *Food Security Portal*. (May, 2022). FSP: https://www.foodsecurityportal.org/tools/excessive-food-price-variability-early-warning-system

[5] *IFPRI*. (s.f.). https://www.ifpri.org/

[6] High-Order Conditional Quantile Estimation Based on Nonparametric Models of Regression. (2015). En F. Y. C. Martins-Filho. Taylor & Francis.

[7] *FAOSTAT*. (2022). https://www.fao.org/faostat/es/#home

[8] *FAO*. (2022). Food and Agriculture oranization of the United Nations: https://www.fao.org/in-action/inpho/crop-compendium/cereals-grains/en/?page=1&ipp=7&no_cache=1&tx_dynalist_pi1[par]=YToxOntzOjE6IkwiO3M6MToiMiI7fQ==

[9] *FAO*. (s.f.). Wheat in the world: https://www.fao.org/3/y4011e/y4011e04.htm

[10] E.J. Stevens, K. A. (s.f.). *FAO*. Food and Agriculture Organizatino of the United Nations: https://www.fao.org/3/y5765e/y5765e06.htm

[11] *Statista*. (January, 2022). https://www.statista.com/statistics/1073550/global-leading-oats-producers/

[12] P. Brockwell, R. (2016). *Introduction to Time Series and Forecasting, 3rd ed.* Springer.

[13] Tsay, R. (2002). *Analysis of Financial Time Series, 2nd ed.* John Wiley & Sons.

[14] Box GEP, J. G. (1970). *Time series analysis, forecasting, and control.* Oakland: Holden-Day.

[15] *R Documentation*. (2022). https://www.rdocumentation.org/packages/forecast/versions/8.16/topics/auto.arima

[16] Yongma Moon, T. Y. (2011). A robust mean absolute deviation model for portfolio optimization. Elsevier.

[17] A. Gelman, J. C. (2013). *Bayesian Data Analysis, 3rd ed.* Chapman & Hall.

[18] M. Johannes, N. P. (2006). *MCMC Methods for Continuous-Time Financial Econometrics.* SSRN Electronic Journal.

[19] Bishwal, J. (1st edition). Parameter Estimation in Stochastic Volatility Models. Springer.

[20] Taylor, S. (1994). Modeling Stochastic Volatility: A Review and a Comparative Study.

[21] Guide, S. U. (November, 2022). *Stan Development Team.* https://mc-stan.org/docs/stan-users-guide/index.html

[22] *The R Project for Statistical Computing*. (2022). https://www.r-project.org

[23] *Stan*. (2022). https://mc-stan.org/

[24] Christian Robert, G. C. (2011). *JSTOR*. A Short History of Markov Chain Monte Carlo: Subjective Recollections from Incomplete Data: https://www.jstor.org/stable/23059158

[25] N. Metropolis, S. U. (1949). *The Monte Carlo Method* . Amer. Stat, Assoc.

[26] N. Metropolis, A. R. (1953). *Equations of State Calculations by Fast Computing Machines.* J. Chem. Phys. 21.

[27] Hastings, W. (1970). *Monte Carlo Sampling methods using Markov chains and their applications.* Biometrika.

[28] A. Gelfand, A. S. (1990). *Sampling-Based approaches to calculating Marginal Densities,.* J. Amer, Stat Assoc.

[29] *rstan: R Interface to Stan*. (November 2022). https://cran.r-project.org/web/packages/rstan/index.html

[30] *Federal Reserve of the United States*. (November 2022). https://www.federalreserve.gov/

APENDIX A.

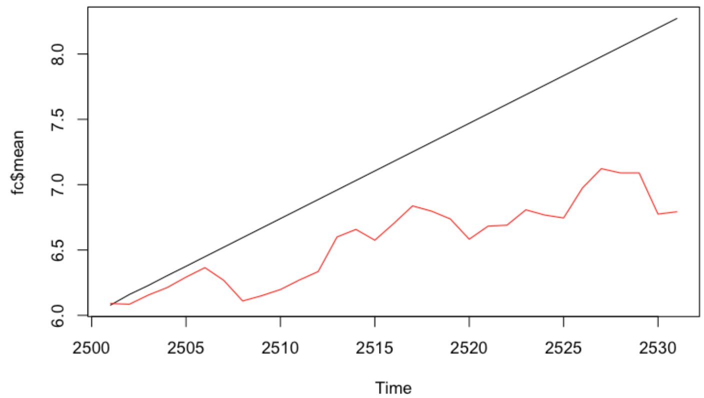

Figure 31. Zoom Forecast Corn ARIMA

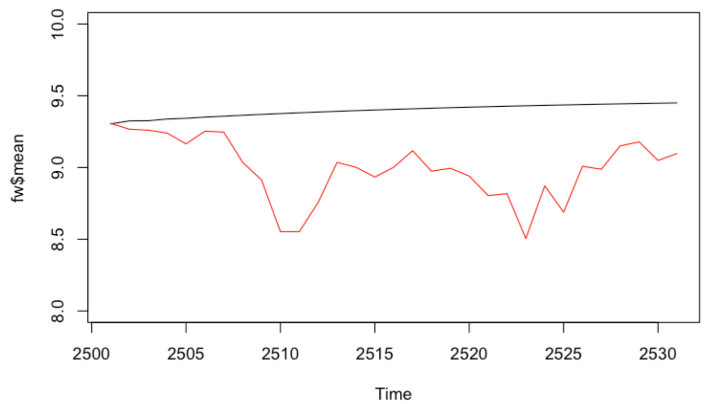

Figure 32. Zoom Forecast Wheat ARIMA

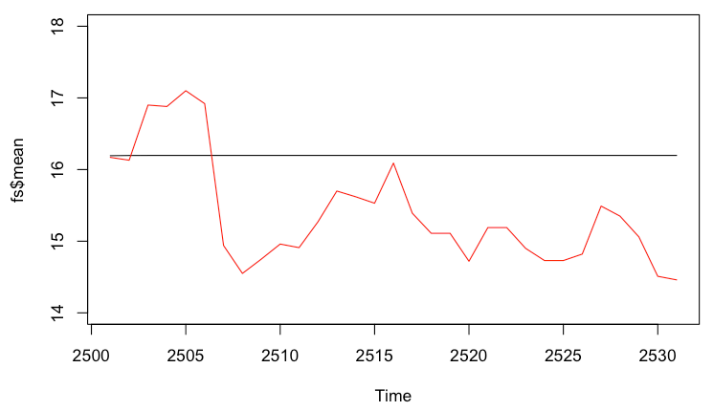

Figure 33. Zoom Forecast Soybean ARIMA

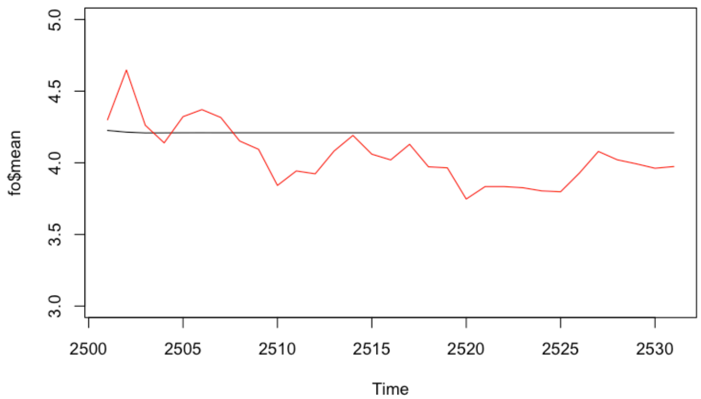

Figure 34. Zoom Forecast Oat ARIMA