Center for Food
and Land Use Research
at Kyiv School of Economics

THE EFFECT OF THE WAR IN UKRAINE ON THE EUROPEAN RAPESEED OIL MARKET

September, 2023

Photo: Europeanseed

KSE Agrocenter information products are available at https://agrocenter.kse.ua and at telegram channel t.me/kseagrocenter

KSE Agrocenter

Center for Food
and Land Use Research
at Kyiv School of Economics

AUTHOR INFORMATION

Pavlo Martyshev

Pavlo is a senior researcher at the Center for Food and Land Use Research of Kyiv School of Economics (KSE Agrocenter). He is specializing in applied research of food markets and agricultural policy, econometric modelling of agricultural prices. Pavlo received a Master's degree in agricultural economics from the Kyiv National Economic University (2015) and a PhD degree at the Institute of Economics and Forecasting of the National Academy of Sciences of Ukraine (2020). Before joining KSE, Pavlo worked as a grain market analyst at Offre & Demande Agricole company. In addition, he participated in consulting projects organized by the World Bank and the German-Ukrainian Agricultural Policy Dialogue.

Hryhorii Stolnikovych

Hryhorii is a junior researcher at the Center for Food and Land Use Research. He graduated Kyiv School of Economics in 2023 and received a Master's degree in the Economic Analysis program. Hryhorii is specialized on applying econometric and machine learning models on food markets research.

TABLE OF CONTENT

INTRODUCTION

Rapeseed is an essential commodity used for production of edible oil, biodiesel, lubricant, and feed. Over the last years, the EU rapeseed oil production was relatively stable at about 9-10 mln. tons annually. Meanwhile, the role of Ukraine as a supplier of rapeseed to the EU countries increased more than twice[1]. Therefore, the full-scale Russian invasion to Ukraine affected essentially European rapeseed and rapeseed oil markets.

Current research is aimed to estimate the effect of the war in Ukraine on the European rapeseed market. Our working hypothesis is that the war affected unusually strong price growth of rapeseed oil prices which was beneficial for the EU processing industry. To reach our research goal we conducted descriptive and fundamental market analysis. Besides, econometric and machine learning methods were used to detect the price anomalies related to the war in Ukraine.

[1] https://apps.fas.usda.gov/psdonline/app/index.html#/app/advQuery

SECTION 1. FUNDAMENTAL ANALYSIS OF THE GLOBAL VEGETABLE OILS MARKET DURING THE WAR IN UKRAINE

Since all mail vegetable oils (palm oil, soybean oil, rapeseed oil, sunflower oil) are substitutes, the fundamental analysis of global vegetable oils market allows to understand indirectly the main factors of rapeseed oil price formation in the EU. To track the main fundamentals for the period from January 2022 to July 2023 we use the variety of sources, in particular: United States Department of Agriculture (USDA), AgRural, APK-Inform, UkrAgroConsult, ABARES, SovEcon, IFPRI, Buenos Aires Stock Exchange, Ukrainian Grain Association, FranceAgriMer, Thomson Reuters, AgWeb, and other sources. Through the synthesis of this information, we formulated the key price drivers on vegetable oil market on a monthly basis (Table 1). Drives related to the war in Ukraine are market by red color.

Table 1. Key drivers of vegetable oils price formation

Month	Fundamental factors
January 2022	- Tight soybeans market balance amidst the climate risks in South America (flooding in the North Brazil; drought in South Brazil and Argentina); - Low stocks of palm oil in Malaysia; - High crude oil prices.
February 2022	- Escalation between Russia and Western countries push oil prices to 7-years highs; - Lack of workforce on palm plantations in Malaysia cuts the local production; - Soybean crop in Brazil is revised downwards.
March 2022	- Market expects the reduction of sunflower planting areas to around 5 mln. ha (from 7.1 mln. ha) with expected 2022 crop at around 10 mln. tons; - Further downward revision of soybeans crop in South America; - Argentinian government increased export taxes for oilseeds; - Growth of global sunflower oil prices due to the lack of Black Sea origin; - Massive selling of soybeans from China state stocks; - Slowdown of Chinese demand amidst the COVID-19 lockdown;

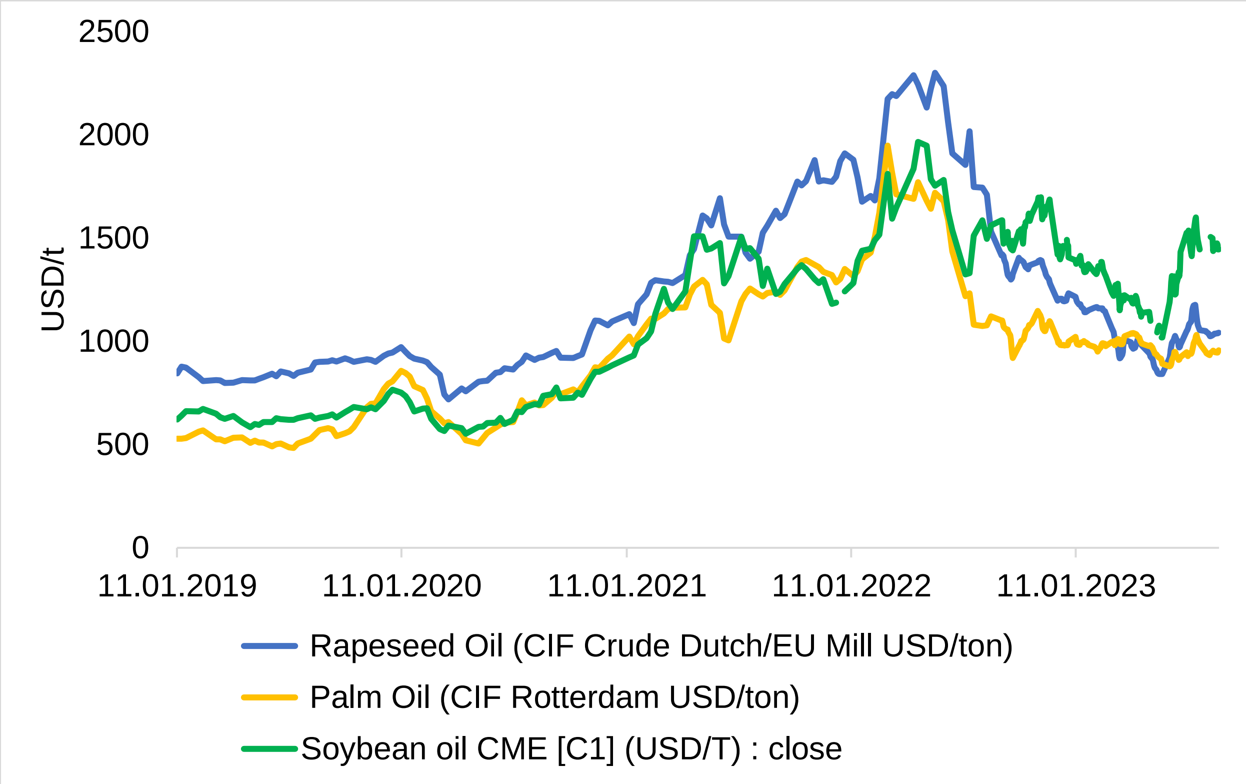
	- First peace negotiations between Ukraine and Russia partially restricts price growth.
April 2022	- Due to the Black Sea ports blockade, the unrealized potential of sunflower oil export in Ukraine is estimated at 3.4 mln. tons; - Expected growth of the US crude oil production restricts the growth of global vegetable oil market; - Estimates of Argentina crop are cut due to the drought.
May 2022	- Rumors about the Russian oil embargo push up energy and vegetable oils markets; - Revealing of the Bucha tragedy in Ukraine collapses peace negotiations in Ukraine; - Malaysia and Indonesia reacts to the price growth by export restrictions for palm oil; this increases the tightness of global vegetable oils market; - Mykolaiv port missile attack increased the panic on the world commodity markets; - USDA expects record soybean crop in 2022.
June 2022	- EU's embargo on Russian crude oil has limited effect on global vegetable oil market; - India reacts on high vegetable oil prices by import intensification; - Lockdown in China limits global soybean demand; - Odessa port missile attack increases the panic on the world commodity markets.
July 2022	- Vegetable oil market declines due to worsened macroeconomic situation in the world and slow China demand; - EU intensifies the purchase of Ukrainian oilseeds to crush them domestically; - The recovery of palm oil production and export in Indonesia and Malaysia eases global market.
August 2022	- Launch of the Grain Deal eased the global market; - Export restrictions for palm oil in Malaysia; - Drought in the USA; - Increasing expectations regarding the soybean crop in Brazil; - Eased relations between the USA and Iran decreases crude oil prices; - EU rapeseed imports have much higher pace than last year.
September 2022	- Crude oil and vegetable oils markets are under pressure from the recessionary processes in the world economy;

	- Euro/USD parity reaches extremely low levels (0,96) due to the termination of North Stream 1; - Downward revision of Canadian canola export to the EU.
October 2022	- Recovery of palm oil production in Malaysia; - Strong progress of the US soybean harvesting; - Slow Chinese demand.
November 2022	- Flooding in Indonesia restrict palm harvesting; - EU rapeseed prices grow due to the disruption of exports from Ukraine and active canola crushing in Canada; - Recovery of China soybean demand; - Sunflower crushing margin is extremely high in Ukraine. Farmers keep sunflower seeds expecting to price increase; - Strong EU demand on Ukrainian rapeseed; - Optimistic prospects of the EU rapeseed crop in 2023.
December 2022	- Easing on the crude oil market; - Growth of the EU rapeseed stocks; - China accelerates soybean and rapeseed import;
January 2023	- Flooding in Malaysia restrict palm harvesting; - Strong China soybean demand; - Slow inspections on Bosporus restrict the supply of Ukrainian sunflower oil to Asian markets; - Low soybean harvest in Argentina due to the drought.
February 2023	- Slow China soybean demand due to the local New Year; - Record estimates of Brazil crop; - High stocks of palm oil in Malaysia.
March 2023	- Proceeding of the Grain Deal to 17 July brings the optimism on the market; - Flooding in Malaysia and Indonesia restrict palm harvesting; - Slow soybean harvesting in Brazil.

Source: authors' analysis

Figure 1 shows the evolution of the global vegetables oil market. Price rally started with the easing of monetary policy in 2020 provoked by the COVID-19 lockdown. The full-scale Russian invasion to Ukraine provided additional boost for prices. Rapeseed oil price increased the most in spring 2022 with price premium to soybean and palm oil reaching unusually high levels. Such sensitivity of rapeseed oil price is explained by two main factors: a) strong correlation between rapeseed oil and crude oil production via biodiesel market; b) expected shortage of Ukrainian rapeseed origin in the EU.

Figure 1. World prices of vegetable oils

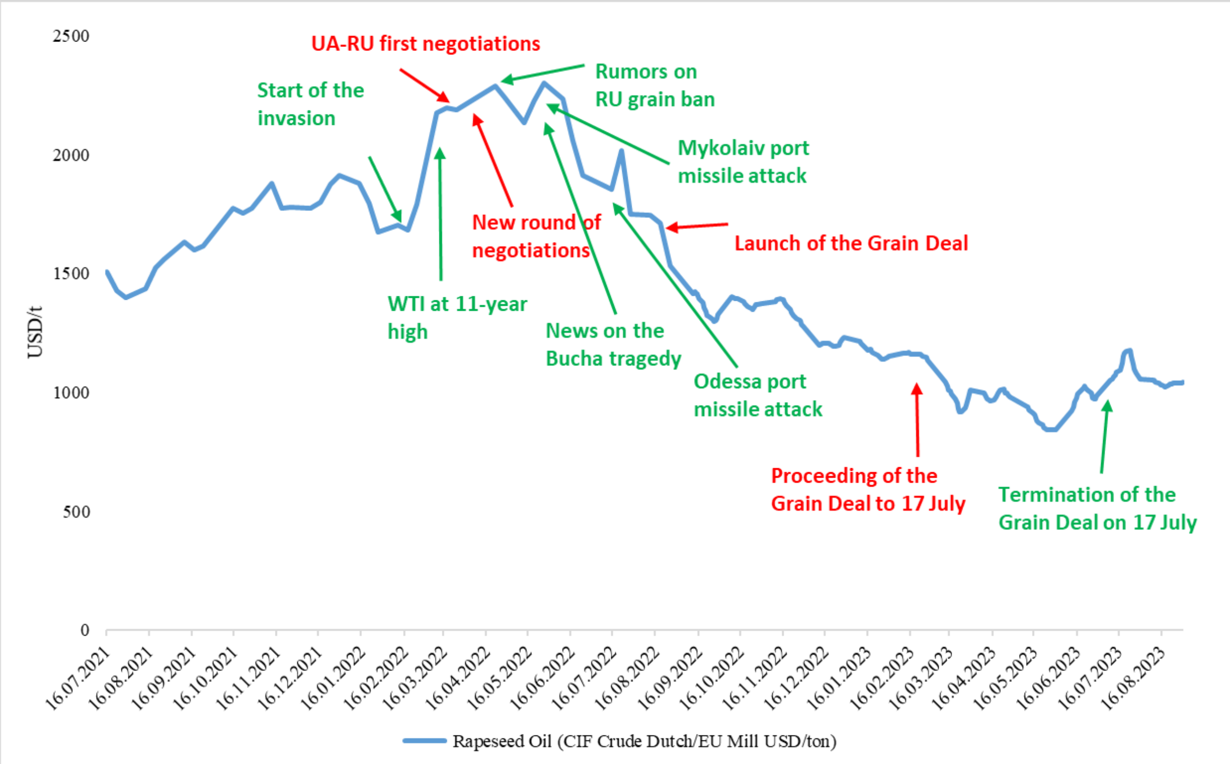


Source: CME, UkrAgroConsult

SECTION 2. DESCRIPTIVE ANALYSIS OF THE WAR IMPACT ON THE EU RAPESEED OIL MARKET

The war-related price drivers in relation to the rapeseed oil price dynamics are presented on Figure 2. Bullish factors (causing price increase) are marked by green; bearish factors (causing price decrease) are marked by red. We can see strong reaction of the price on the news related to Ukraine.

Figure 2. War-related factors of the EU rapeseed oil price formation



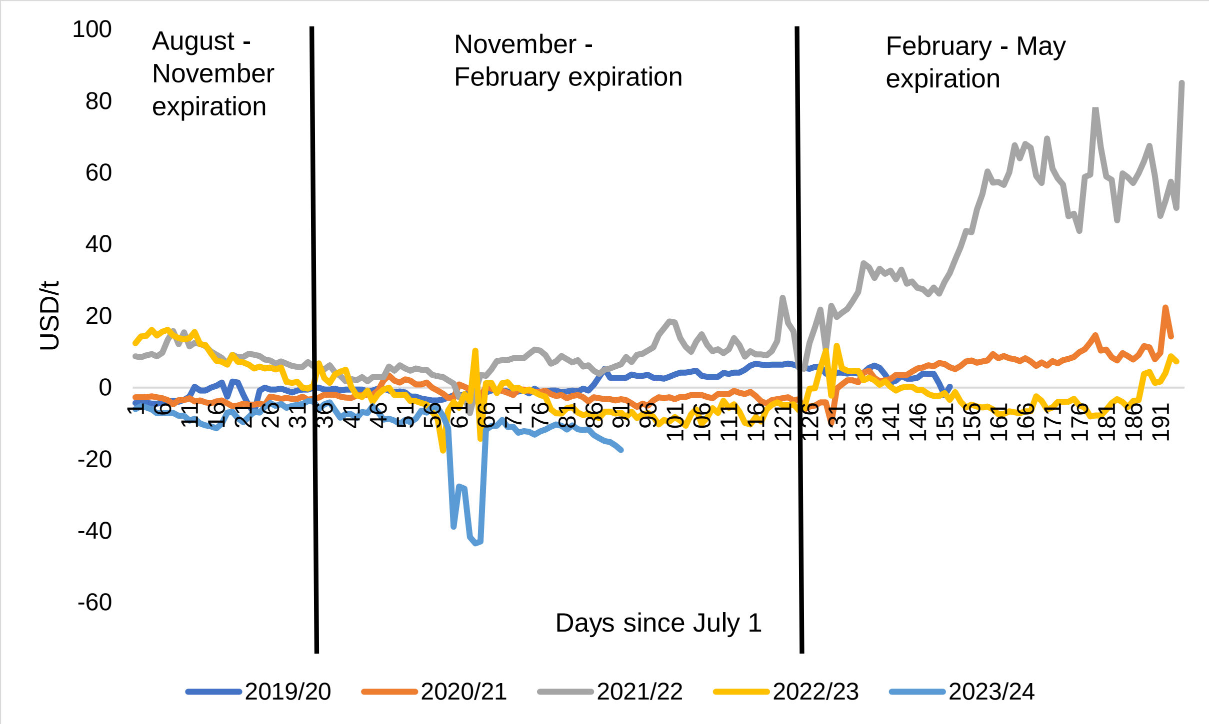
Source: created by the authors

War-related news led to the non-typical dynamics of the EU rapeseed and rapeseed oil markets. This can be confirmed by the analysis of **calendar spreads** of Euronext rapeseed futures market. Along the marketing year for the rapeseed started 1 July and ended 30 June, four futures contracts are listed with expiration in August, November, February, and May. A **normal market** shows futures prices increasing within the one season as the time to maturity increases (so-called contango). This means that price spread between earlier contracts and following expirations tends to be negative. The reason for that are storage costs and financing costs. By contrast, an **inverted market** occurs when near-term maturity futures contracts are priced higher than future maturity contracts[2]. These unusual dynamics signalize that current deficit of products cannot be covered by the available stocks.

[2] https://www.investopedia.com/terms/i/invertedmarket.asp

Figure 3 depicts the price spread for rapeseed futures on Euronext along the last seasons. It shows that inverted market was rare in the pre-war period. But after the full-scale invasion in February, price spread reached record highs (grey line), signalizing about the expected shortage of rapeseed in the EU.

Figure 3. Price spreads for Euronext rapeseed contracts



Source: authors' calculations based on Euronext data

Despite the panic expectations on futures market, the supply-demand balances show that EU markets of rapeseeds and rapeseed oil did not face shortage in two wartime seasons: 2021/22 and 2022/23. Import and stock volumes of both products were close to the pre-war levels (Table 2, Table 3).

Table 2. Supply-demand balance for the EU rapeseed market

Attribute	2019/2020	2020/2021	2021/2022	2022/2023	2023/2024
Beginning Stocks	1,86	1,18	0,72	0,83	1,8
Production	15,25	16,73	17,39	19,59	20,2
Imports	6,08	5,79	5,57	6,85	5,1
Total Supply	23,19	23,72	23,68	27,27	27,1
Exports	0,33	0,17	0,44	0,57	0,45
Domestic Consumption	21,67	22,82	22,4	24,9	25,2
Ending Stocks	1,19	0,72	0,83	1,8	1,45
Total Distribution	23,19	23,71	23,67	27,27	27,1

Source: USDA

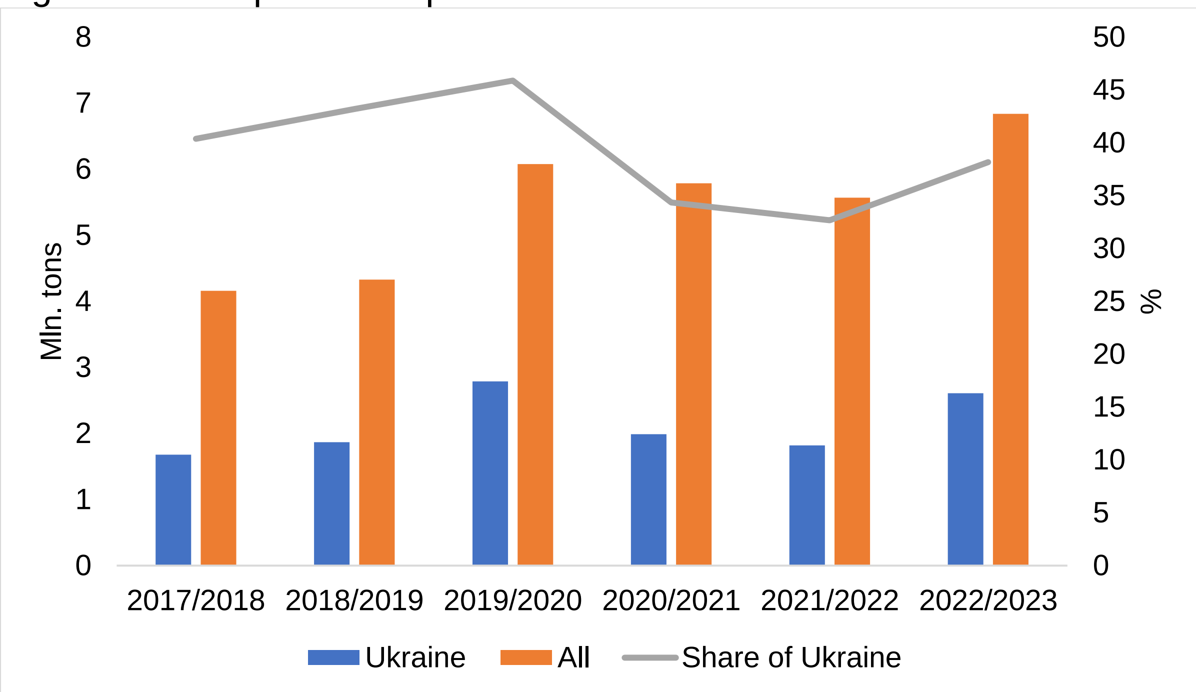
Table 3. Supply-demand balance for the EU rapeseed oil market

Attribute	2019/2020	2020/2021	2021/2022	2022/2023	2023/2024
Crush	21,1	22,3	21,8	24,1	24,4
Beginning Stocks	318	378	211	398	400
Production	8,862	9,366	9,156	10,122	10,248
Imports	467	314	593	400	375
Total Supply	9,647	10,058	9,96	10,92	11,023
Exports	369	722	337	700	750
Food Use Dom. Cons.	2,25	2,4	2,575	2,82	2,775
Feed Waste Dom. Cons.	50	50	50	50	50
Domestic Consumption	8,9	9,125	9,225	9,82	9,875
Ending Stocks	378	211	398	400	398
Total Distribution	9,647	10,058	9,96	10,92	11,023

Source: USDA

Ukraine’s exports to the EU during the wartime was moderate comparing to the last years (Figure 4). Usually, the cycle of rapeseed exports in Ukraine is relatively short; it starts in July and ends in December. Therefore, only 2022/23 marketing year can be considered as wartime season in Ukraine.

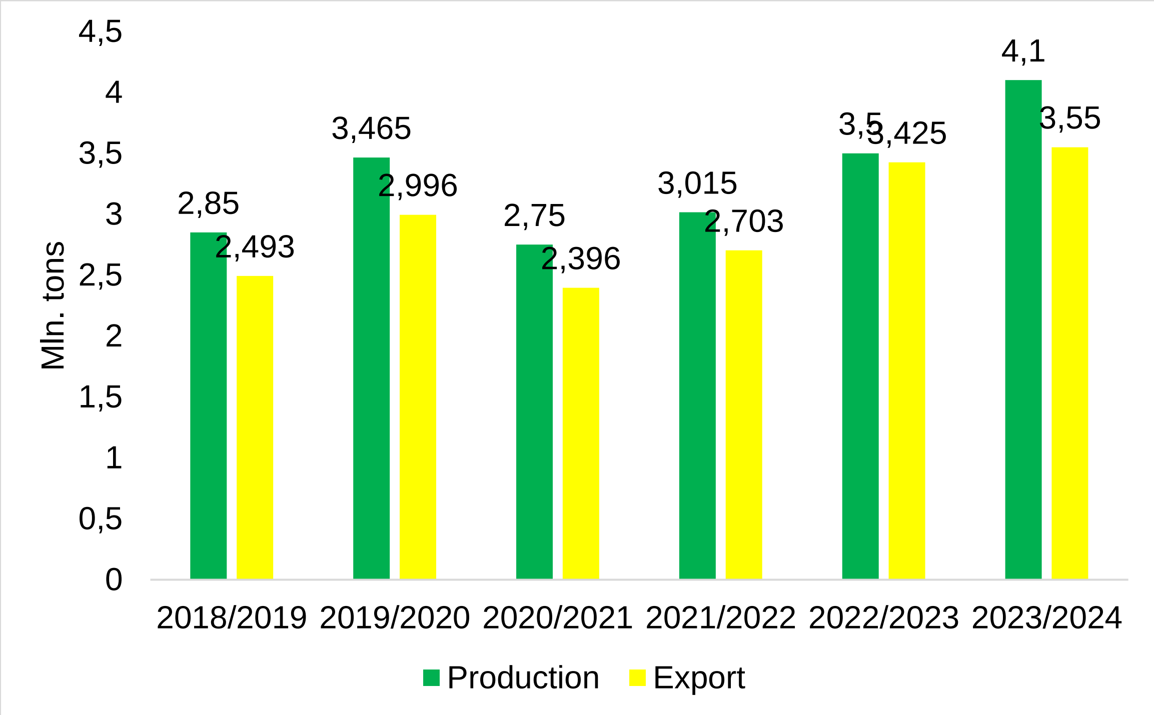
Figure 4. EU rapeseed imports



Source: European Commission

As for the rapeseed production in Ukraine, it was quite high in 2022 (3.5 mln. tons); almost 100% of this production were exported (Figure 5). Due to the massive switching of Ukrainian farmers from grain to oilseeds[3] and good climate conditions[4], the 2023 rapeseed harvest is expected at more than 4 mln. tons.

Figure 5. Production and export of rapeseed in Ukraine



Source: USDA

The overview of rapeseed market balances and trade volumes shows that the EU had no strong shortage of rapeseed during the war in Ukraine. Meanwhile, the raise of rapeseed and rapeseed oil price was caused by **global rather than local factors**. The main of them are growth of crude oil and biodiesel prices as well as price growth for substituting products – other vegetable oils and corn (the main input for bioethanol production). This effect will be modelled with econometric and machine learning methods in the next section.

3 https://ukragroconsult.com/en/news/ukrainian-farmers-plan-to-allocate-record-areas-for-winter-rape/

4 https://uga.ua/en/news/ukraine-warm-winter-to-increase-the-harvest-of-winter-wheat-by-20-30-hydrometeorological-center/

SECTION 3. ECONOMETRIC ANALYSIS OF THE EU RAPESEED OIL PRICES

I. RAPESEED OIL PRICES ANOMALIES DETECTION (UNIVARIATE APPROACH)

Anomaly detection is a fundamental concept in time series analysis that plays a pivotal role in identifying unusual patterns or observations within a sequence of data points recorded over time. In the context of price time series data, an anomaly refers to a data point or a series of data points that deviate significantly from the expected or normal behavior of the price trajectory. These anomalies can manifest as sudden price spikes, sharp declines, or persistent fluctuations that cannot be explained by regular market forces or historical trends. Anomalies can be both indicative of underlying issues, such as supply disruptions or speculative trading, and opportunities, such as arbitrage possibilities. In this report, we apply several methodological approaches to detect the anomalies in rapeseed oil price data.

1. Anomalies definition

A time series anomaly, also known as an outlier, refers to a data point or a sequence of data points within a time-ordered dataset that exhibit behaviors significantly different from what is considered normal or expected within the context of the given data. Anomalies within time series data are characterized by their departure from the prevailing patterns, trends, or statistical properties exhibited by the majority of data points. These deviations can take various forms, including sudden spikes or drops, recurring irregularities, seasonality shifts, or long-term drifts, depending on the specific application and domain.

Accurate identification of time series anomalies involves statistical and machine learning models, as well as domain-specific knowledge, to distinguish between genuine anomalies and natural fluctuations or noise in the data. Typology of time series anomalies includes three main classes: point, contextual, and collective anomalies.

Point anomalies are the simplest type in which each data point can be analyzed by the anomaly detector without considering any other data points in the input dataset. Point anomaly is an observation x or y that deviates remarkably from X according to some predefined criteria, where $x \in X$ and $y \notin X$ (Teng et. al, 2017). Limits can be placed on to automatically detect these point anomalies when the specified threshold is violated. A **contextual anomaly** is a point, or sequence of points, that might not be considered a point anomaly, but is an outlier in the context of the data where this point or sequence occurs. For instance, a rapid fluctuation in a price time series. A more significant challenge is contextual anomalies that occur within the normal operating range but which are not conforming to the expected temporal pattern. (Fahim and Sillitti, 2019) **Collective anomalies** are sets of data points that themselves can be normal, but together they represent an anomaly (Shaukat et al., 2011).

<u>The occurrence of anomalies in prices of the given product on the market, point, contextual or collective, make the market highly unpredicted and creates accidental winners and looser.</u>

2. Methodological approach to anomaly detection

Under the umbrella of anomaly detection, the methodologies are typically categorized into three groups: 1) statistical (econometric) models, 2) supervised, and 3) unsupervised machine learning algorithms. Anomaly detection with the supervised or semi-supervised machine learning algorithms is not suitable for anomalies detection in this case with rapeseed oil prices. These algorithms require having a label for each data point in the training dataset, whether it is an anomaly or not, to learn the differences. Such labeling is not possible due to the small size of the available dataset and no prior knowledge of the anomalies over the period covered by the data.

Unsupervised machine learning algorithms are more suitable in this case, as they do not require data labeling. An algorithm has to analyze the dataset to infer the real concept of abnormality or make an assumption of the concept. A concrete example of this type is the set of clustering-based anomaly detection methods which presume the data that rest inside small clusters are prone to be anomalous. Another example is the isolation forest, an algorithm, which uses an ensemble of decision trees to isolate anomalous points in the data. The issue with unsupervised ML algorithms is the fact that they cannot make a concrete decision whether a point is an anomaly, or give a specific probability of the point occurrence. In other words, these algorithms produce an ordinal data, or ranking, of points based on some criteria that acts as a measure of their anomality. To classify specific points as anomalies, a *contamination* parameter should be set by a researcher, based on educated guess about the proportion of outliers in the dataset.

Clustering approaches work well with the univariate data. In case of multivariate data, a chosen number of variables are examined, often a maximum of two or three. However, the approaches that use decision trees or neural network, require little modeling effort and are known to be relatively easy to build and implement (Kacprzyk et. al, 2016). On the other hand, approaches that use decision trees or neural networks, perform much better with the multivariate data. The anomaly detectors based on residuals has the advantage that the algorithm easily can be trained to capture seasonal and cyclic variations, where the relationship between the target variable and input features for different seasons is captured, assuming that the data is available.

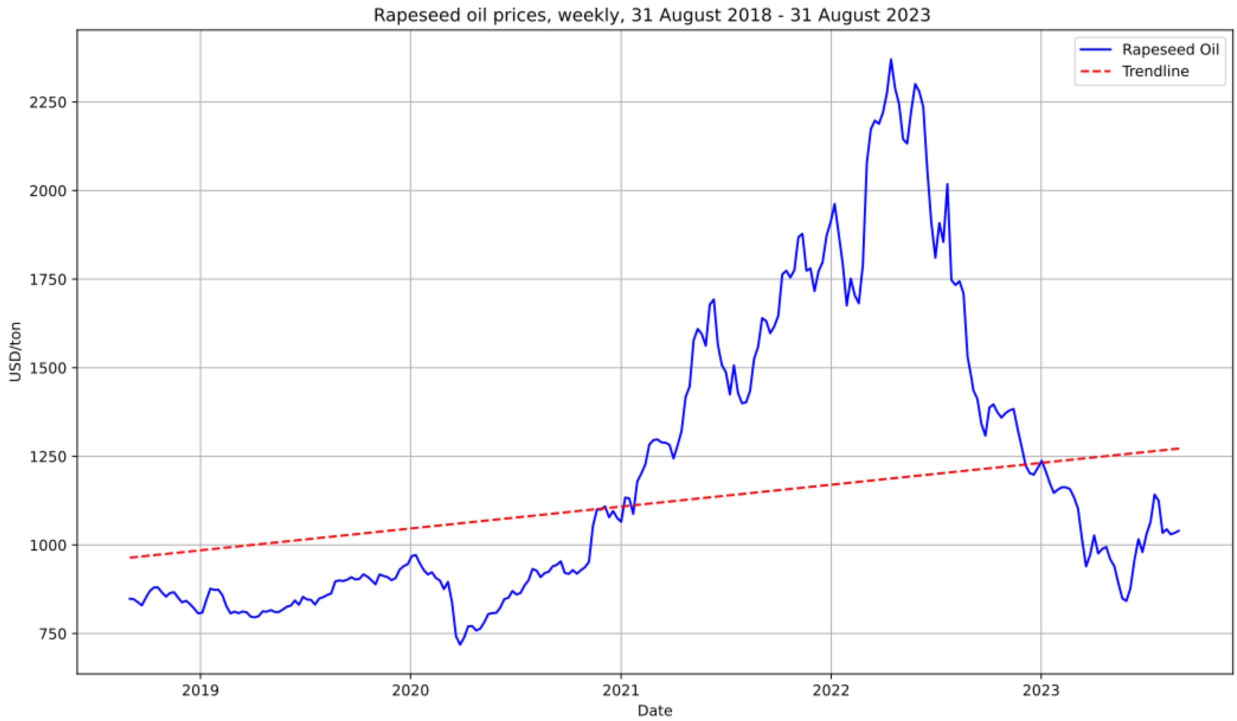
Statistical (econometric) models are mostly based on the comparison of some statistical properties to test whether outliers exist. During the training phase, the distribution parameter is optimized under specific evaluation criteria. The resultant distribution will define the boundary with a probabilistic threshold. Furthermore, to test the generalization of unseen data, the learned model is assumed to have potential outliers to lie in the low-probability density region (Garcia-Font et. al, 2016). As can be seen, for training, a suitable probabilistic model should be defined.

Based on the abovementioned information, a set of models for this analysis was defined. From the econometric models group, ARIMA model was chosen. Isolation forest, local outlier factor (LOF) and autoencoder models were chosen from the unsupervised ML algorithms. In the further sections, more detailed review of each model is provided, as well as their limitations, fitting, and results.

3. Data description and exploratory analysis

Data used for this analysis is a single variable weekly time series of rapeseed oil prices.[5] Data covers a period from 31 Aug 2018 to 31 August 2023 with a total of 282 observations. From the Figure 6 it is clearly seen that in 2021-2022 a significant increase in prices occurred. It peaked in March-July 2022 and rapidly dropped to late-2020 level.

Figure 6. Rapeseed oil prices, weekly, 2018-2023



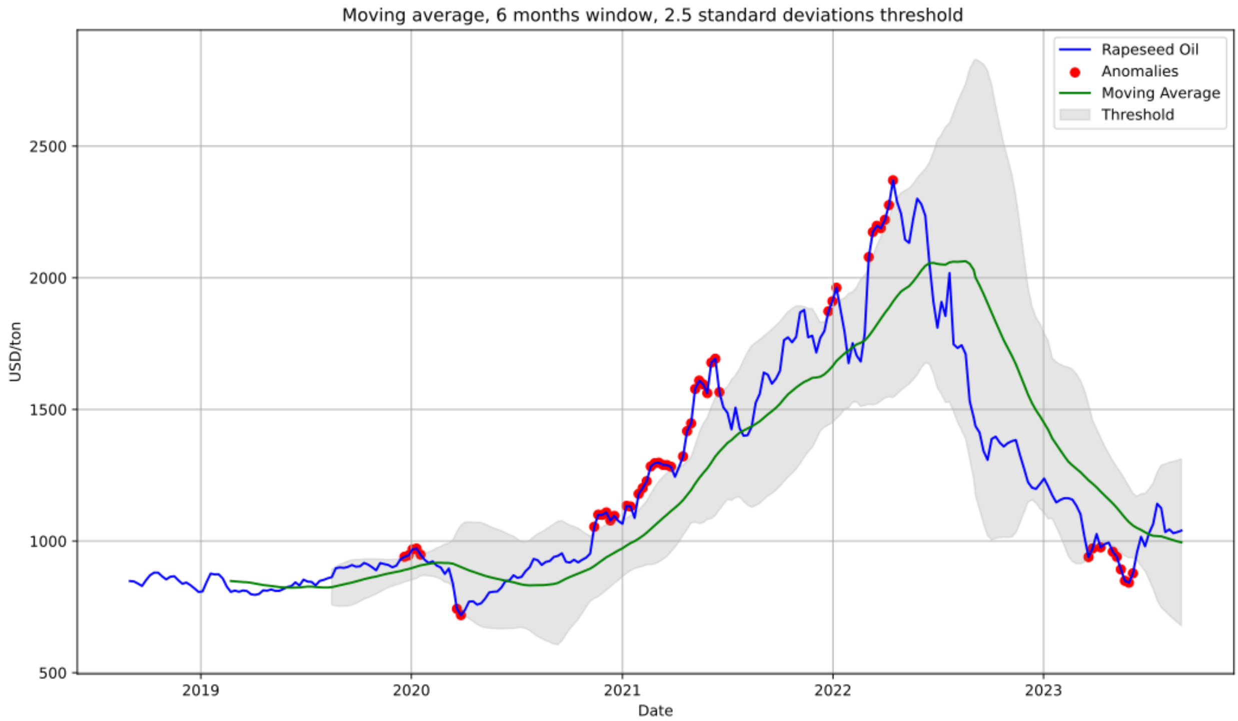
Source: own visualization based on Thomson Reuters data.

[5] https://www.neste.com/investors/market-data/palm-and-rapeseed-oil-prices#f52a96cc

To perform a simple check for outlying observations we calculate the moving average of the data. For a moving window of 6 months (26 weeks/observations) the mean value and its standard deviation are calculated. Those observations, which deviate from the moving average further than the set threshold could be considered anomalies. In this case the threshold is set at the level of 2.5 standard deviations, which corresponds to p-value equal to 0.01. This approach is best suited to detect the short-term point anomalies, which significantly deviate from the regular pattern. It struggles to detect the contextual and collective (sequential anomalies). It could be seen on the Figure 2, that in the periods of high volatility the confidence interval of the moving average expands significantly, which makes it inefficient in capturing outliers. To capture the long-term anomalies, the window of moving average should be increased, which reduces the precision of detection and is simply not possible due to limited amount of data.

As it is seen from the Figure 7, plenty of outlying observations have been found. They are located in the period of rapid growth in early 2021, at the peak in spring of 2022, and in spring of 2023, when prices had rapidly dropped.

Figure 7. Preliminary anomalies check with moving average



Source: own calculations.

4. Econometric approach. ARIMA model

To improve the precision of the Moving Average approach, we apply the ARIMA model, which adds the autoregressive (AR) component, which models the relationship between the observation and past values. It accounts for the idea that future values of a time series are influenced by its own previous values. At the same time the MA component models the relationship between the current observation and past forecast errors, emphasizing the short-term effects of past disturbances.

ARIMA models are adept at capturing the underlying structure of time series data, including trends, seasonality, and cyclic patterns. By fitting an ARIMA model to training sample and extrapolating it into the test sample, we can make forecast, which is compared with the actual data to identify deviations from expected values, which are indicative of anomalies.

The model takes form of:

(1) $ARIMA(p,d,q)\colon \Delta^d X_t = c + \sum_{i=1}^{p} a_i \Delta^d X_{t-i} + \sum_{j=1}^{q} b_j \varepsilon_{t-j} + \varepsilon_t,$

Where c, a_i, b_j are the model parameters, Δ^d is differencing operator, ε is an error term, and X is the time series observation. In other words, prediction equals the constant term, plus the linear combination of lags of X, plus the linear combination of lagged forecast errors.

Parameters *p*, *d*, and *q* reference the quantity of AR, I (differencing), and MA components of the model and should be determined before its fitting.

ARIMA model has a requirement of data stationarity. From the visual inspection the dataset seems to be non-stationary. The Augmented Dickey-Fuller test confirms this conclusion:

$$ADF\ Statistic\colon -1.882032$$

$$p-value\colon 0.340560$$

P-value > 0.05 does not allow to reject the null hypothesis that the data is non-stationary. To achieve the stationarity, the dataset is differenced. The Augmented Dickey-Fuller test results:

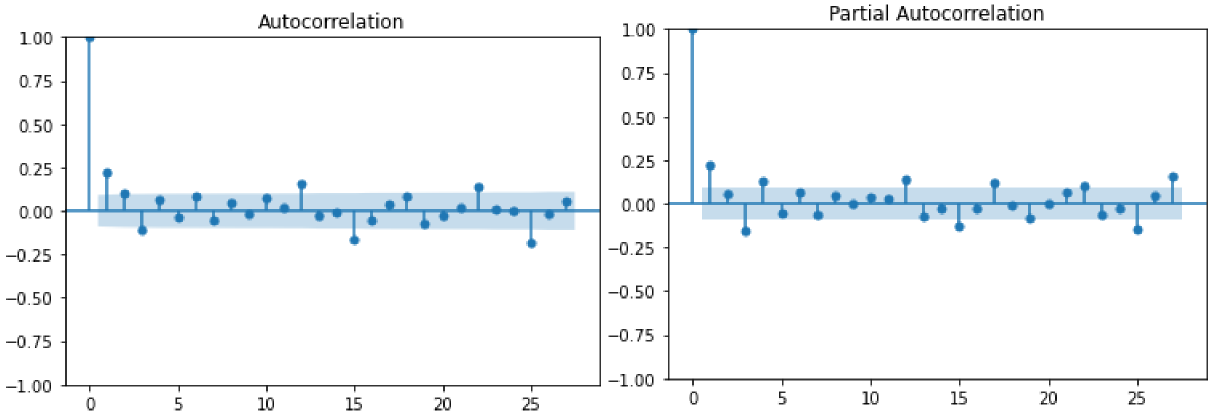
$$ADF\ Statistic\colon -4.699316$$

$$p-value\colon 0.000084$$

Given the p-value < 0.01 we can reject the null hypothesis and conclude that the differenced data is stationary. Thus, the “d” parameter of ARIMA model is 1.

To determine the *p* and *q* we inspect the autocorrelation and partial autocorrelation functions (Figure 8).

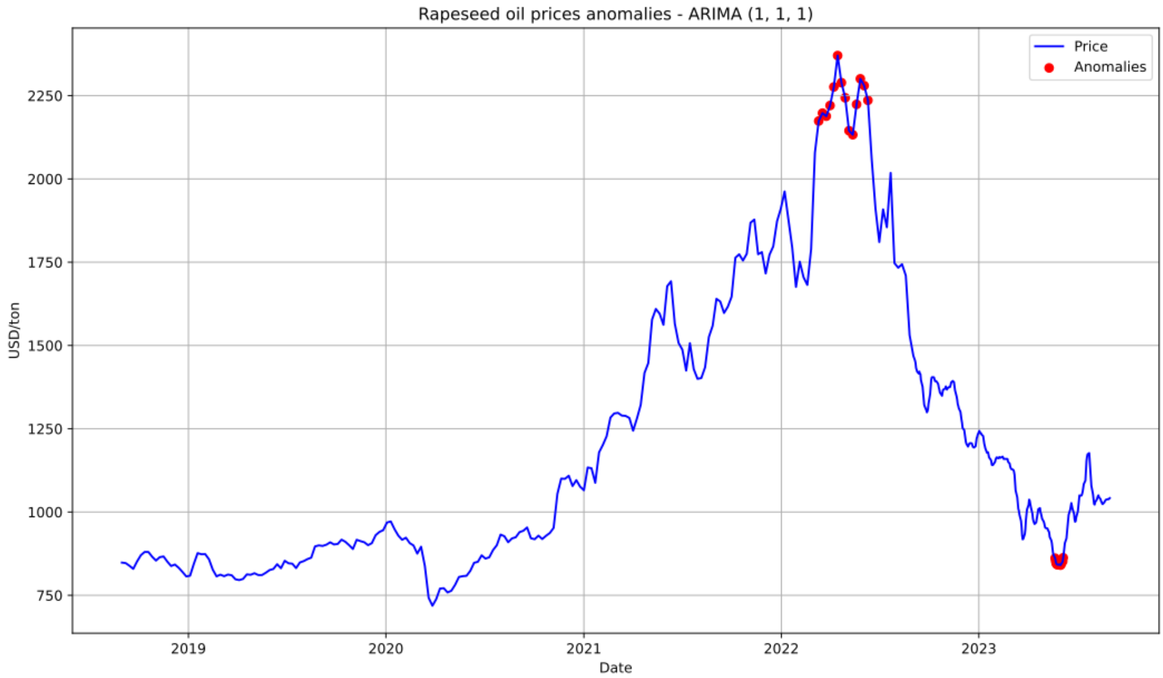
Figure 8. Autocorrelation and partial autocorrelation functions for EU rapeseed oil prices



From the PACF we can see that the first lag is significantly out of the limit and the second one close to the threshold, so we can select the value of *p* (AR) to be 1. From the ACF we can see that only 1 of the lags is out of the significance limit so we can say that the optimal value of our q (MA) is 1.

To identify the anomalies, we fit the model with the training sample, make a projection, and compare it to the test sample. Observation which are outside the confidence interval of the projection are considered anomalies (outliers). Training sample includes data covering 31 August 2018 – 31 August 2022. Test sample covers the period of testing interest: 25 February 2021- 31 August 2023. Two samples overlap intentionally due to small amount of data and increased precision for the purpose of the anomalies detection. Additionally, it allows to assess the model fitting on the part of the training data as well. The chosen threshold for anomaly is 2.5 standard deviations (0.99 confidence interval). Results are presented on the Figure 9.

Figure 9. ARIMA (1, 1, 1) anomalies detection results.



Source: own estimation.

The results suggest that the periods of 3 March 2022 – 10 June 2022, and 22 May 2023 – 06 June 2023 are the anomalies. These periods correspond to peak and rapid drop in prices in spring 2022 and spring 2023, respectively.

5. Approach with machine learning algorithms

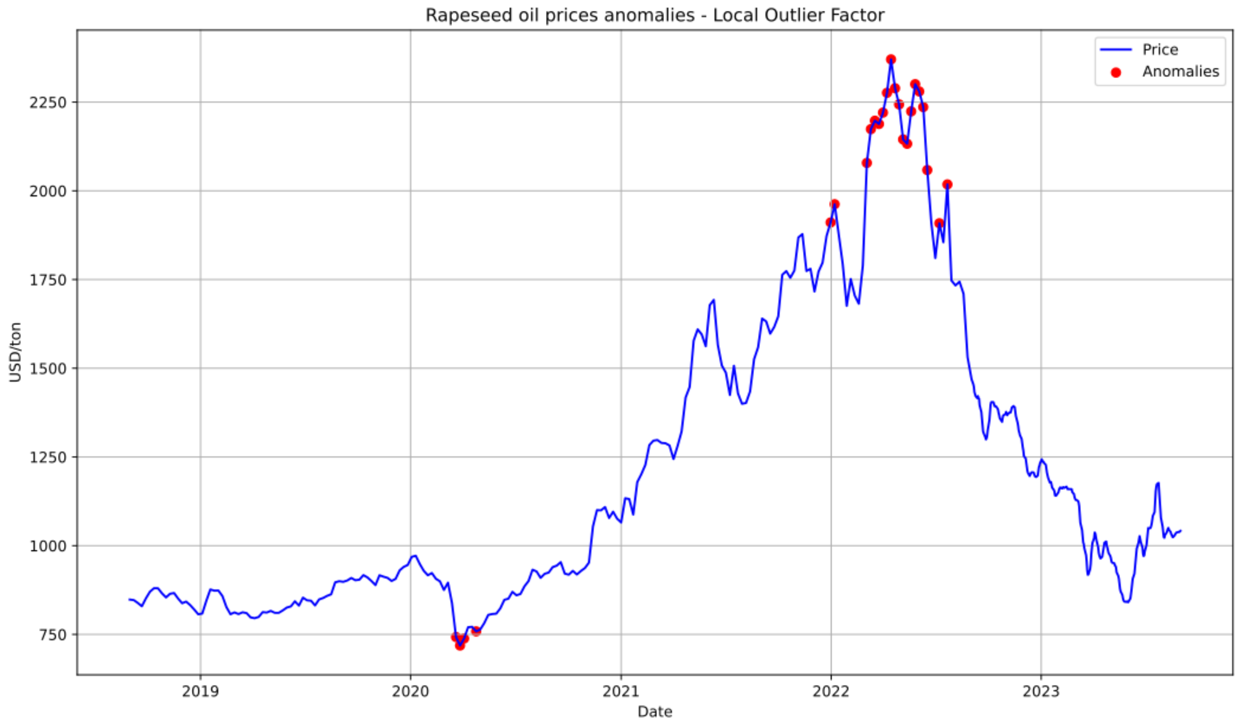
5.1. Isolation Forest

Isolation forest uses an ensemble of decision trees to isolate anomalous points in the data. Developed by Liu et al. in 2008, the Isolation Forest algorithm stands out for its ability to effectively isolate and identify anomalies, making it particularly well-suited for applications where anomalies are rare and distinct from normal observations.

The algorithm starts by randomly selecting a feature and a split value within the range of observed values for that feature. This random partitioning effectively isolates a subset of the data. The partitioning process is repeated recursively, where at each step, a new feature and split value are chosen randomly to further subdivide the data. This process continues until the isolation tree isolates individual data points or reaches a predefined depth. To detect anomalies, the Isolation Forest algorithm assigns a score to each data point based on its path length in the isolation tree. Data points with shorter path lengths are considered more likely to be anomalies, while those with longer path lengths are considered normal.

The main drawback of this algorithm is the fact that is does not clearly classify points as anomalies or normal ones. Instead, it only gives the information whether a point is more anomalous than others or vice versa. The key parameter that is used to label specific points as anomalies is *contamination* value – it represents the expected share of anomalies in the dataset. The model labels this share of observations, which have the shortest paths, as anomalies. In our case we contamination value of 0.05, in order to label those points, which stand out the most. Results are presented on the Figure 10.

Figure 10. Anomalies identified with the isolation forest algorithm.



Source: own estimation.

Results are somewhat similar to ARIMA: a period of March-June 2022 is labeled as a clear anomaly. In addition to it, price drop in spring 2020 is found to be an anomaly, as well as price peaks in January and August 2022.

5.2. Local Outlier Factor (LOF)

Local Outlier Factor (LOF) is an unsupervised machine learning algorithm for anomaly detection based on K-Nearest Neighbors clustering algorithm. It works by measuring the local density of each data point and comparing it to the densities of its neighbors. It is based on the idea that anomalous data points are often located in low-density regions of the feature space.

LOF calculates a local density for each data point by measuring how densely its neighbors are located within a specified radius. Data points with higher local densities are considered more "normal," while those with lower local densities are potential outliers. For each data point, LOF compares its local density to that of its neighbors. If a point has a considerably lower density compared to its neighbors, it is likely an outlier. LOF quantifies this by computing the LOF score, which is a measure of how much the local density of the point deviates from the local densities of its neighbors.

The local reachability density of a point A is defined by:

(2) $lrd_k(A) := 1/(\Sigma_{B \in Nk(A)} reachability - distance_k(A, B)/|N_k(A)|)$, where $Nk(A)$ denotes the set of k nearest neighbors of A

Then, a Local Outlier Factor score of a point A is calculated as

$LOFk(A) := (\Sigma B \in \frac{Nk(A) lrdk(B)}{\boldsymbol{lrdk}(A)})/|\boldsymbol{Nk}(\boldsymbol{A})| = (\Sigma \in Nk(A) lrdk(B))/(|\boldsymbol{Nk}(\boldsymbol{A})| \cdot \boldsymbol{lrdk}(\boldsymbol{A}))$**,** where:

LOF(k) ~ 1 means Similar density as neighbors,

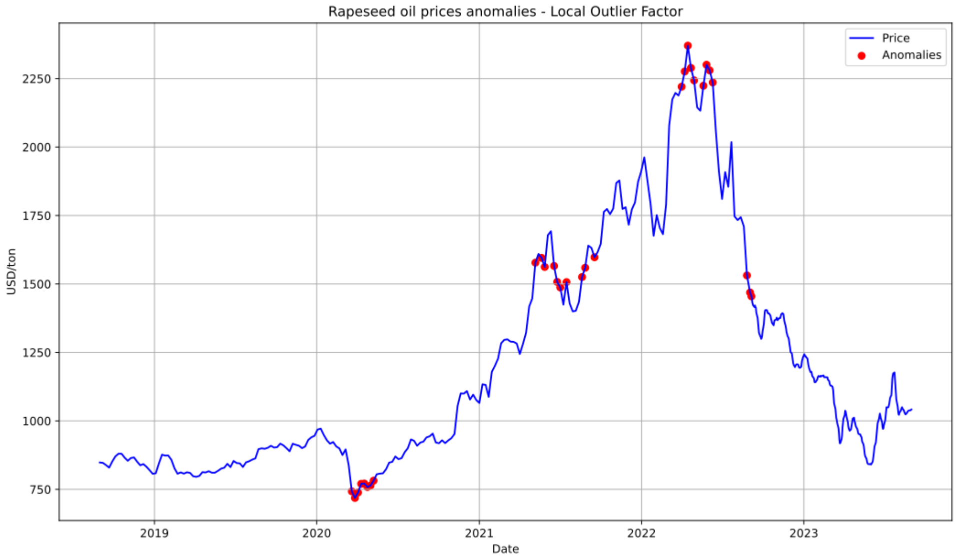
LOF(k) < 1 means Higher density than neighbors (Inlier),

LOF(k) > 1 means Lower density than neighbors (Outlier).

As in the Isolation Forest, LOF uses the *contamination* parameter to label points as anomalies. In this case, value of the parameter indicates share of points with LOF(k) > 1, which are labeled as anomalies. One more important parameter is ***k***, which indicated the number of neighbors considered when calculating the LOF score.

For estimation values of parameters were chosen as *k=40, contamination=0.1.* Results are presented on the Figure 11.

Figure 11. Anomalies detected with the LOF algorithm.



Source: own estimation.

As in the two previous models, period of March-June 2022 is clearly labeled as an anomaly. In addition, LOF labeled as anomalies the following periods:

- Price drop in March-April 2020;
- June-August 2021;
- A few points in October 2023.

5.3. Autoencoder

Autoencoder belongs to unsupervised machine learning algorithms based on neural network models. An autoencoder has two parts: encoder and decoder. The encoder takes an input and maps it to a lower-dimensional representation. The decoder takes this representation and tries to reconstruct the original input. The drawback of the algorithm is the fact that it acts as a complete "black-box", i. e. there is no way to explain why a specific point was labeled as an anomaly.

Results of the fitted model are presented on the Figure 12. Model was fitted for 3 different values of *window size* parameter: 4, 13, and 26, which correspond to approx. 1, 3, and 6 months. Higher values of window size allow to capture the long-term anomalies, and vice versa. The results show short-term anomalies in 2021 and the mix of short-term and long-term anomalies in 2022. Therefore, anomalies in 2022 were stronger and did not dissipate quickly.

Figure 12. Anomalies detected by the Autoencoder.

Source: own estimation

II. ANOMALIES IN PRICES OF RELATED COMMODITIES

For a better understanding of the market context, we apply the same methodology to price time series for related commodities.

Data description and exploratory analysis

Commodities of interest and description of their prices time series datasets are presented in the Table 4. There are missing values in the dataset, but their number is relatively low. High number of missing values negatively affects the precision of anomaly detection, so these variables are dropped from the estimation.

Table 4. List of commodities and description of the data

Commodity	Period covered by the data	Frequency	Number of observations	Number of NA values
Palm Oil (Malaysian Rotterdam)	1 January 2019 – 1 September 2023	Weekly	262	0
Rapeseed Euronext	1 January 2019 – 1 September 2023	Daily, except for weekends	1196	27
Soybean CME	1 January 2019 – 1 September 2023	Daily, except for weekends	1170	53
Soybean oil CME	1 January 2019 – 1 September 2023	Daily, except for weekends	1160	63
Crude oil Brent ICE	1 January 2019 – 1 September 2023	Daily, except for weekends	1197	26
Crude palm oil Kuala Lumpur	1 February 2019 – 1 September 2023	Daily, except for weekends	1105	118

Sources: Thomson Reuters, Euronext, UkrAgroConsult

Anomaly detection

To detect the anomalies in prices time series for these 5 commodities, we apply the same methodology as it was for rapeseed oil in the section (I) of the report. Models used are ARIMA, Isolation forest, Local Outlier Factor, and Autoencoder. Models are fitted for each commodity separately.

1. Palm oil (Malaysian, Rotterdam)

Results of anomaly detection for Palm oil (Malaysian, Rotterdam) are presented in the Figures 13-16.

Figure 13. Palm oil (Malaysian, Rotterdam) price anomaly detection with ARIMA

Figure 14. Palm oil (Malaysian, Rotterdam) price anomaly detection with Isolation forest

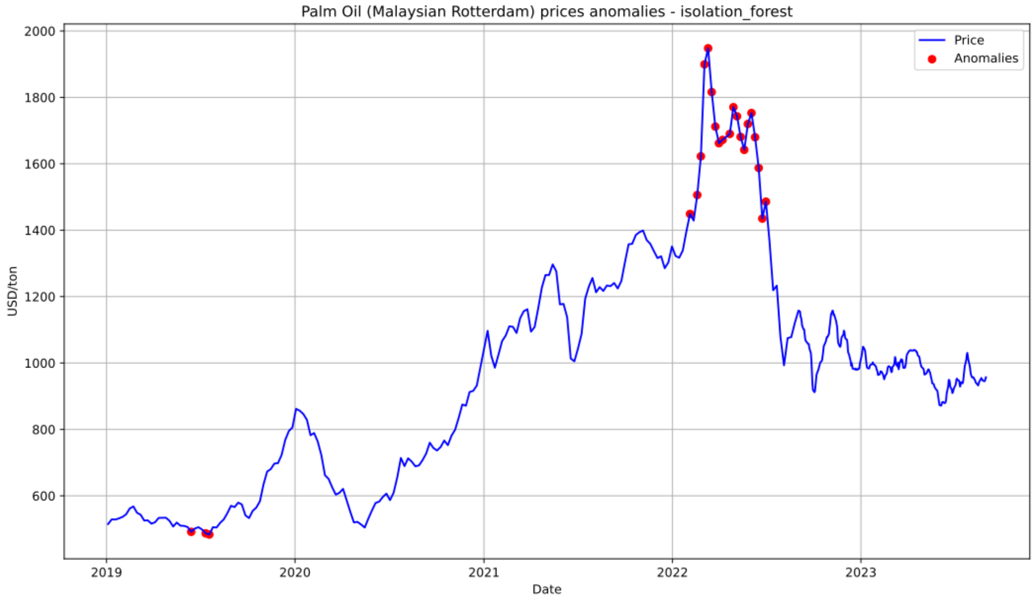


Figure 15. Palm oil (Malaysian, Rotterdam) price anomaly detection with LOF

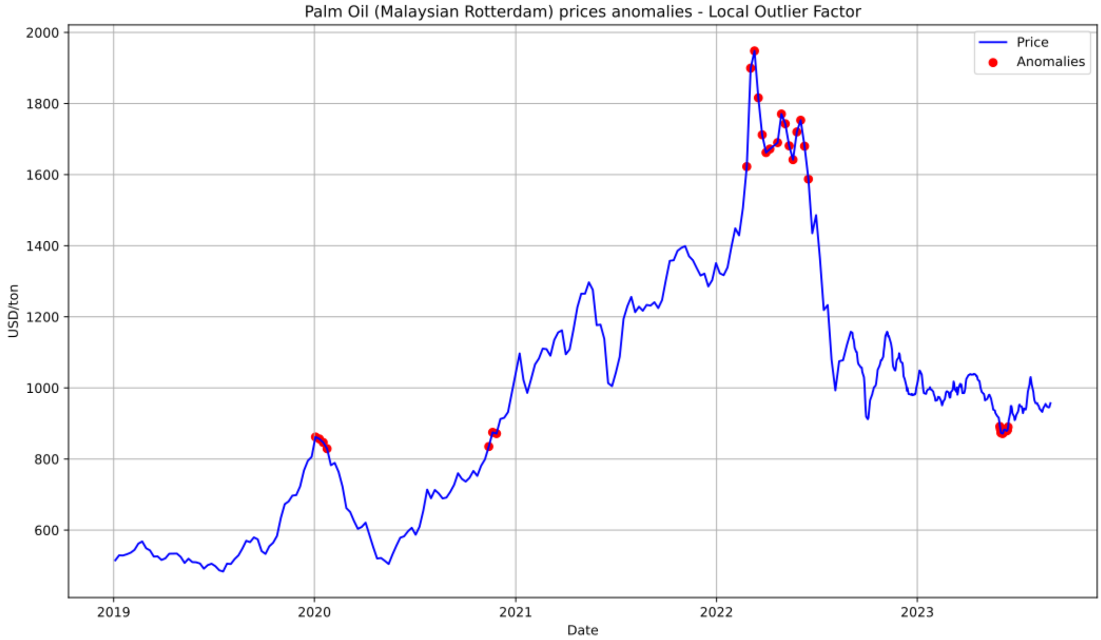
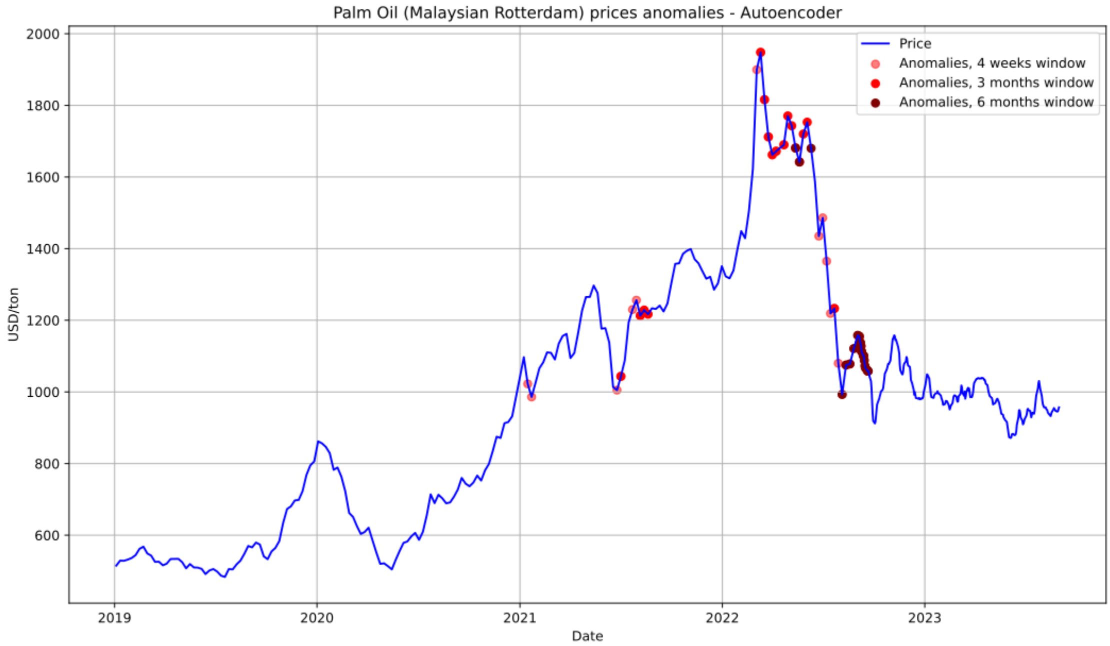


Figure 16. Palm oil (Malaysian, Rotterdam) price anomaly detection with Autoencoder



More than 80% of detected anomalies are located in the war-time period. We can explain this by strong price volatility connected to the instability of crude oil and biofuel markets.

2. Rapeseed (Euronext)

Results of anomaly detection for Rapeseed (Euronext) are presented in the Figures 17-20.

Figure 17. Rapeseed (Euronext) price anomaly detection with ARIMA

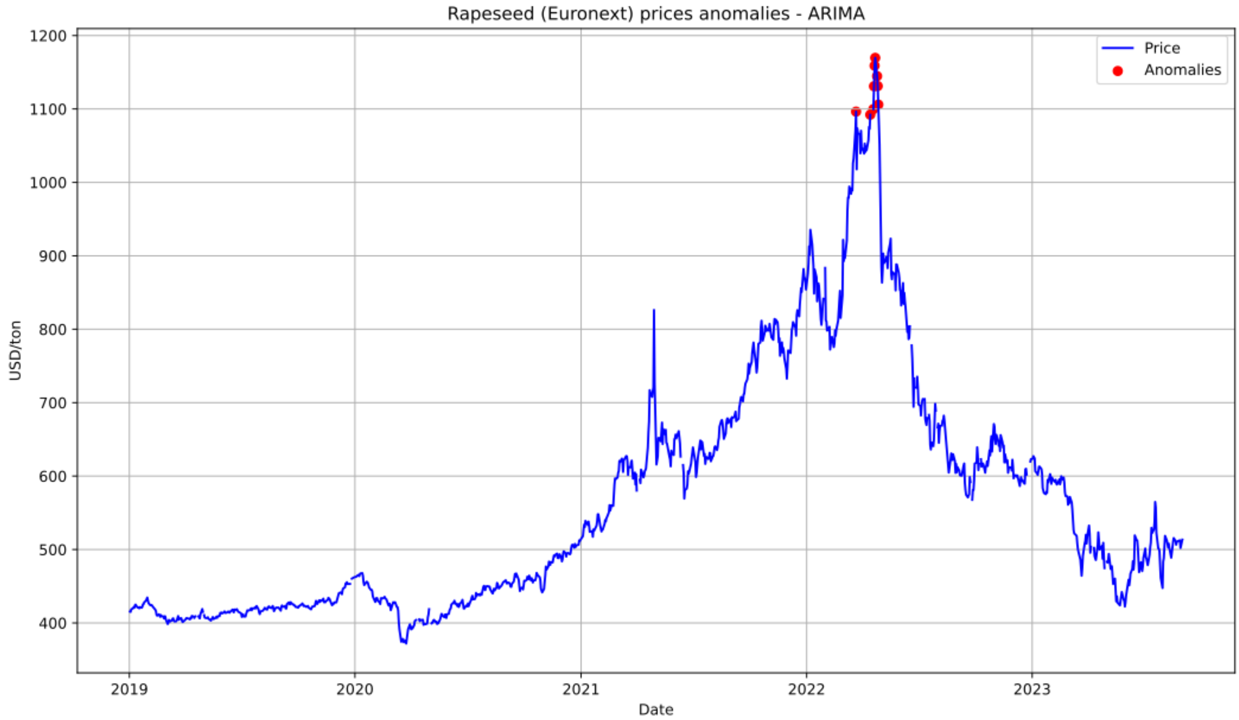


Figure 18. Rapeseed (Euronext) price anomaly detection with Isolation forest

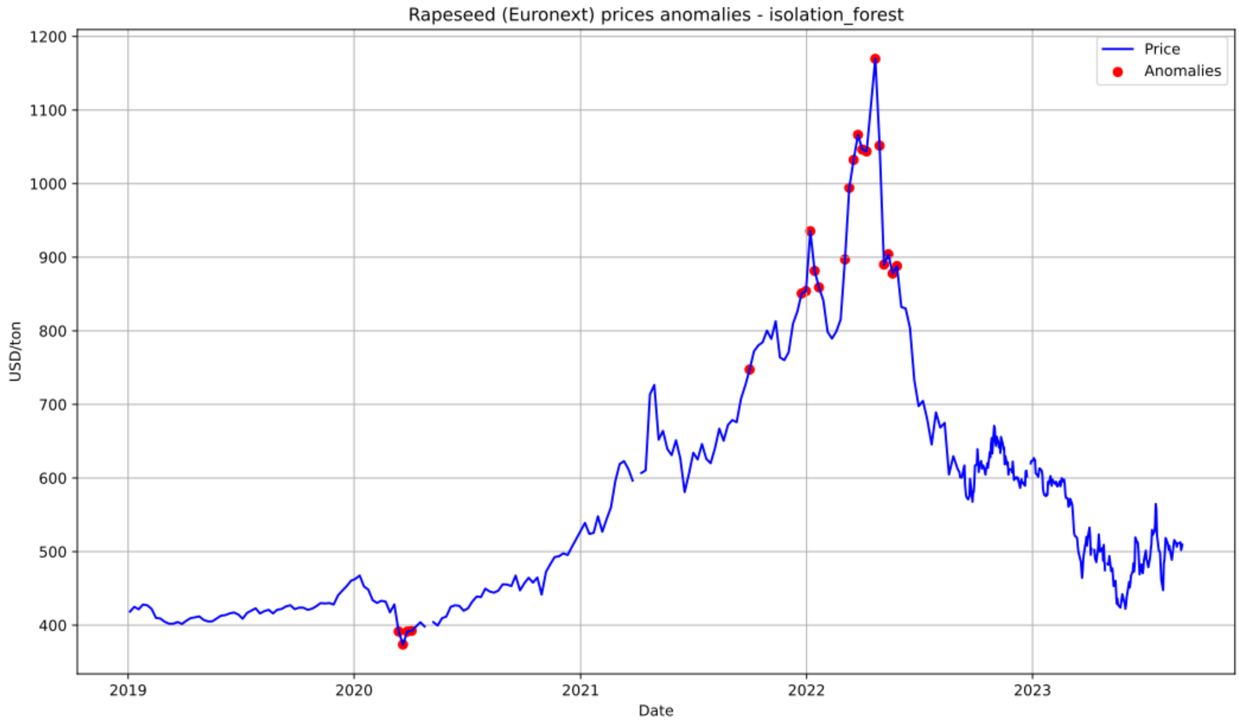


Figure 19. Rapeseed (Euronext) price anomaly detection with LOF

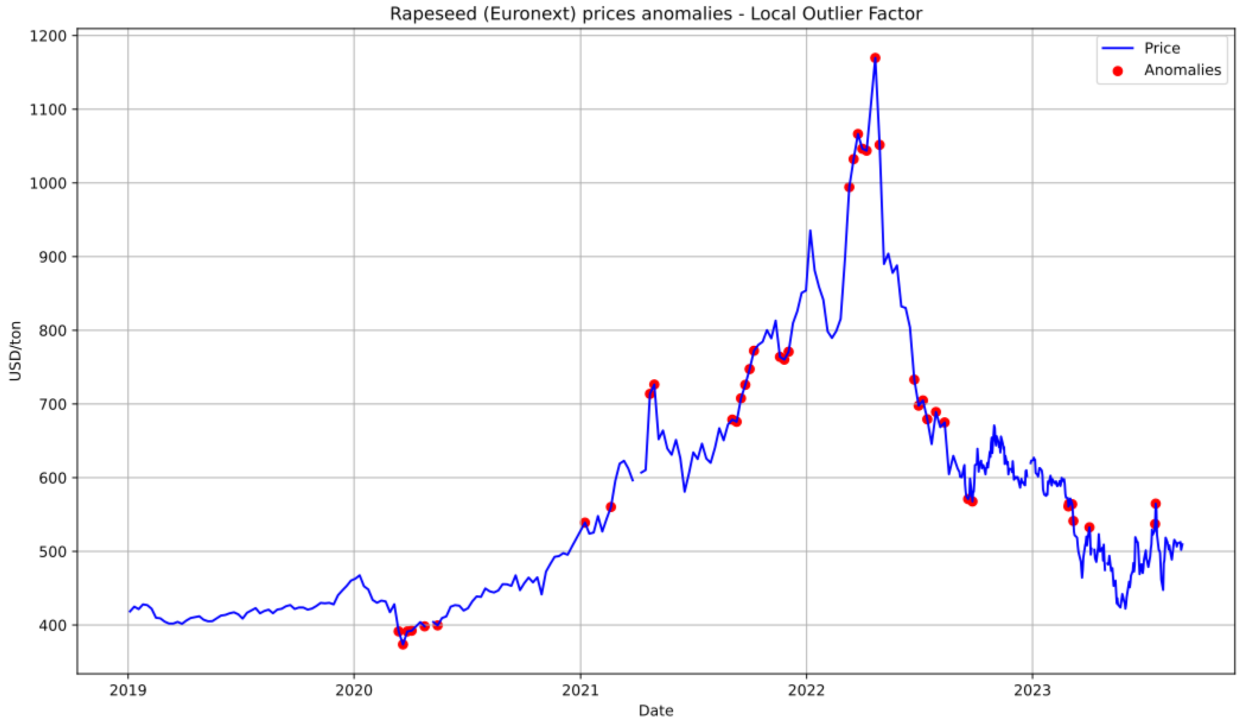
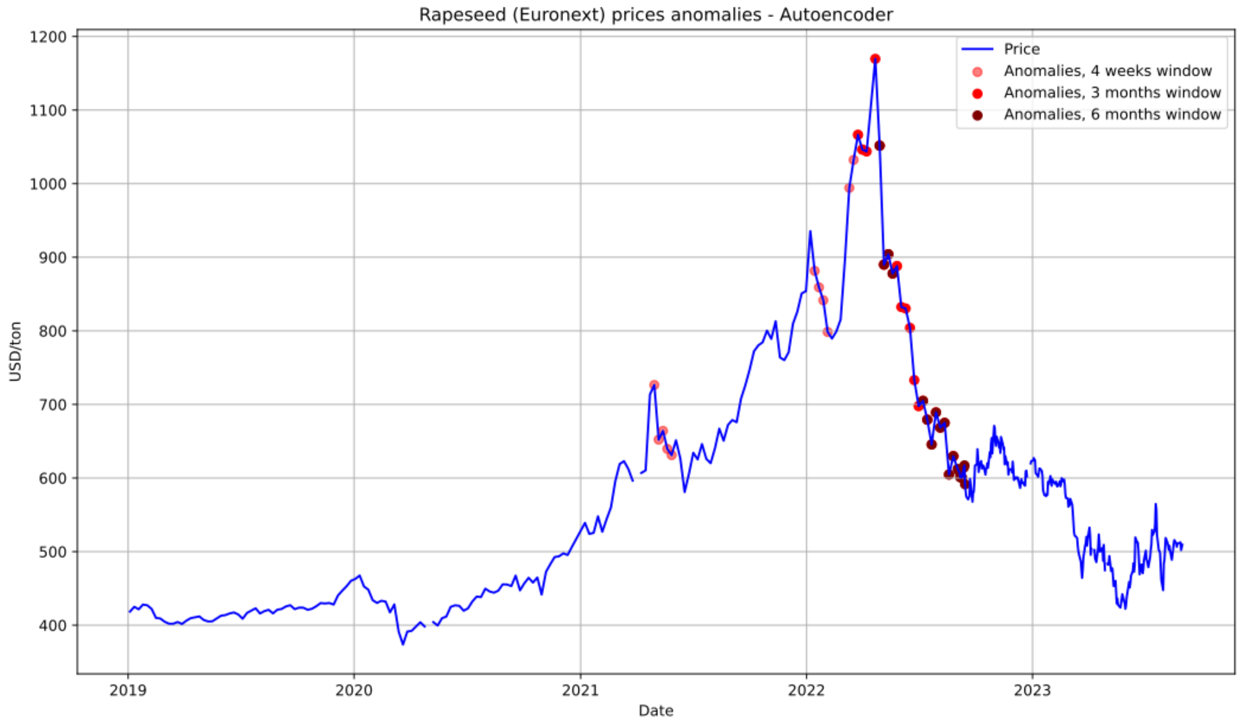


Figure 20. Rapeseed (Euronext) price anomaly detection with Autoencoder



<u>Generally, rapeseed prices on Euronext exchange show many long-term anomalies since the start of Russian invasion in Ukraine.</u>

3. Soybean (CME)

Results of anomaly detection for Soybean (CME) are presented in the Figures 21-24.

Figure 21. Soybean (CME) price anomaly detection with ARIMA

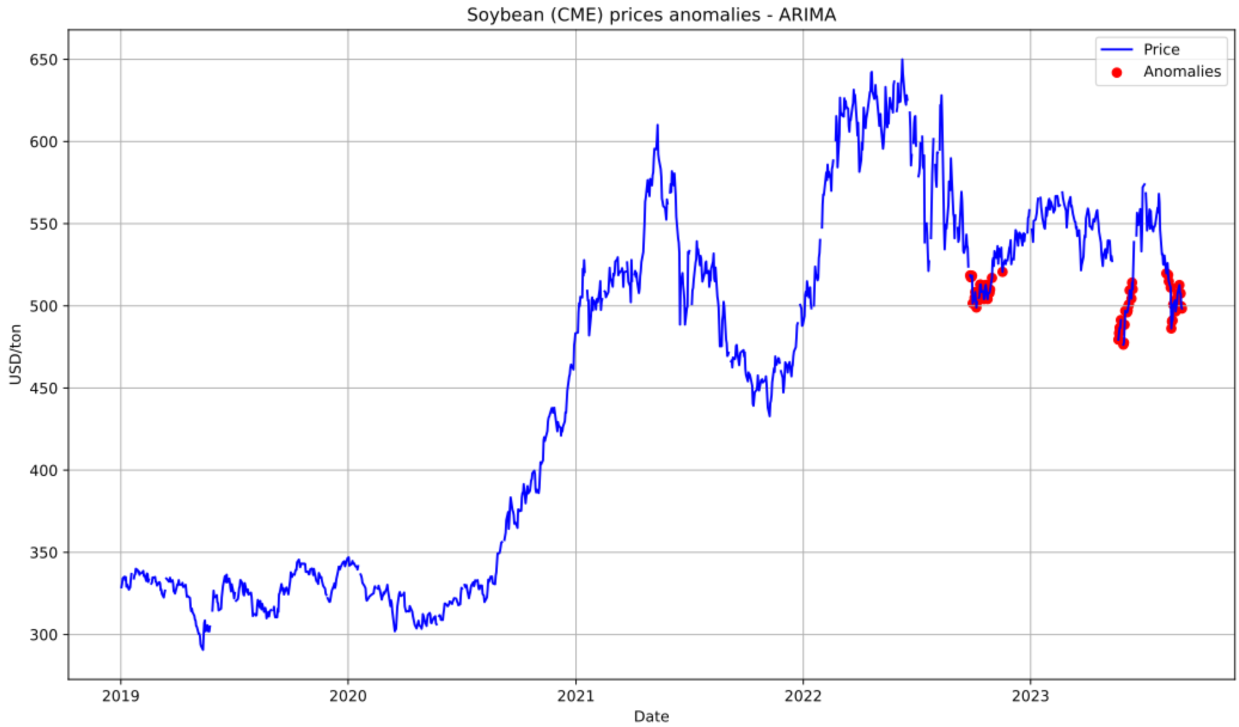


Figure 22. Soybean (CME) price anomaly detection with Isolation forest

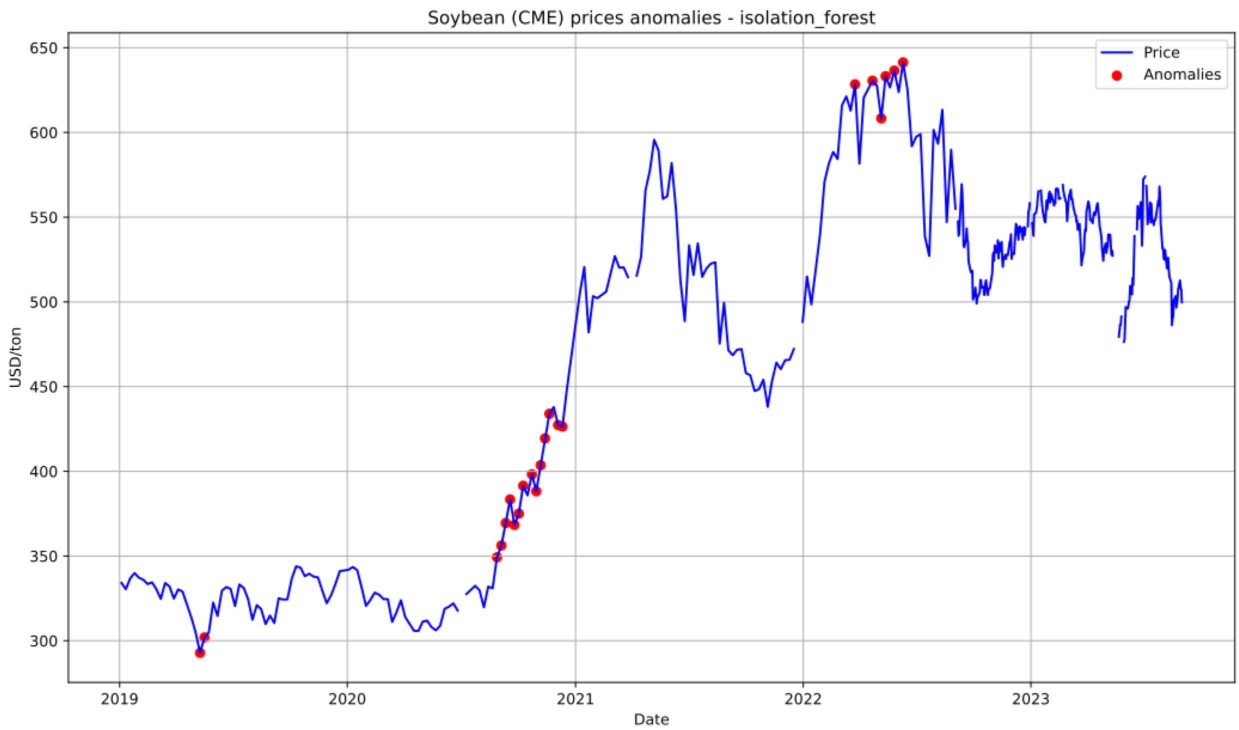


Figure 23. Soybean (CME) price anomaly detection with LOF

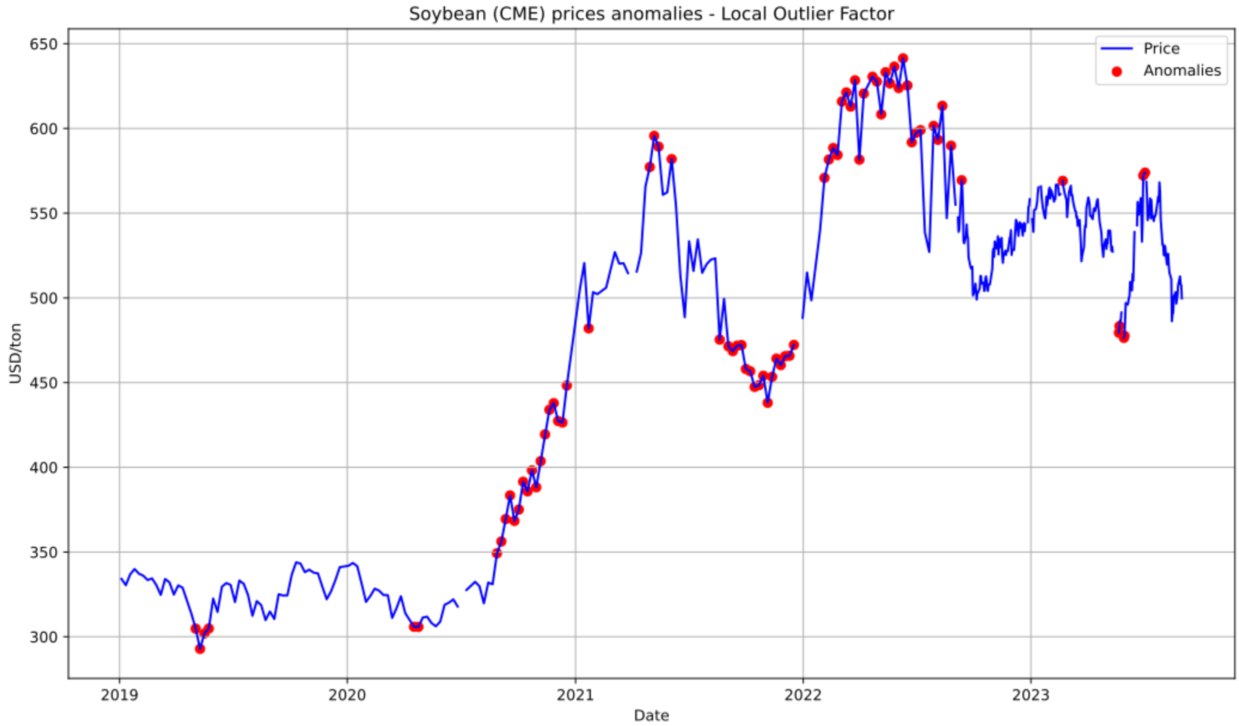
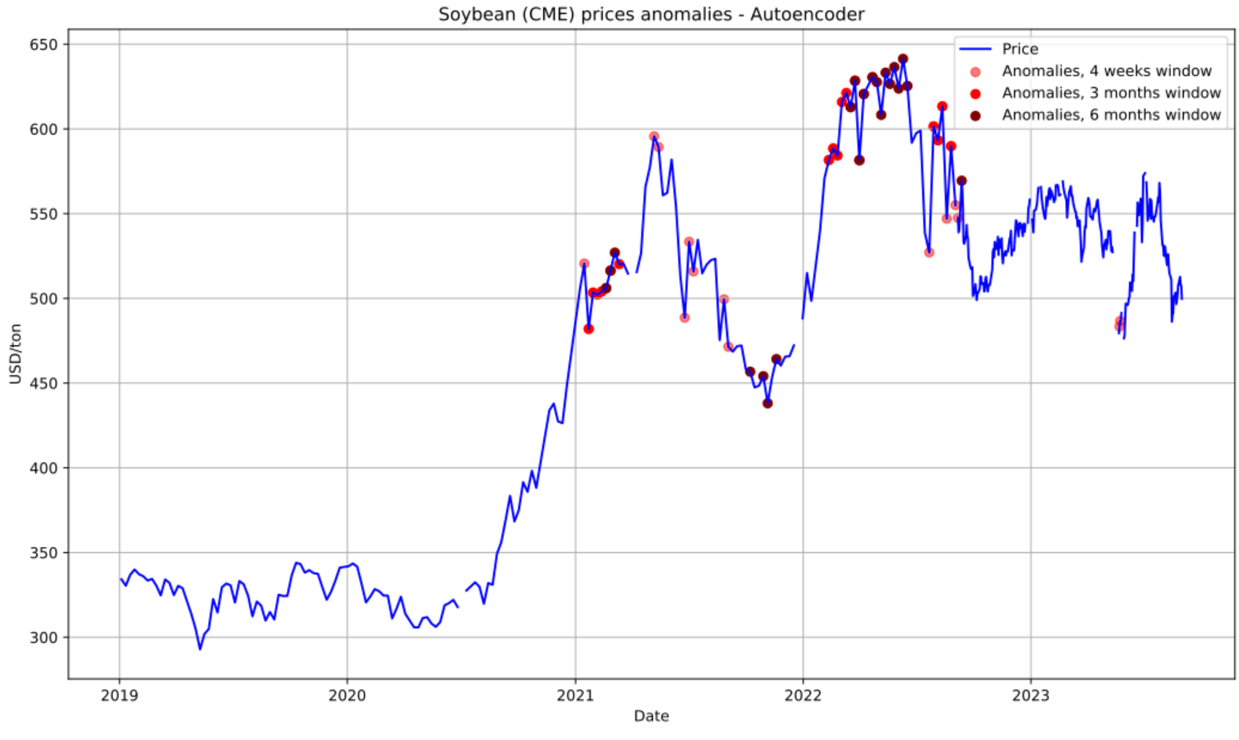


Figure 24. Soybean (CME) price anomaly detection with Autoencoder



Soybean prices were to some extent less affected by the war in Ukraine than other oilseeds and vegetable oils. This is because Black Sea region is not the key supplier of this commodity. Besides, soya beans prices are less related to energy markets than rapeseed oil and palm oil sectors due to their feed use.

4. Soybean oil (CME)

Results of anomaly detection for Soybean oil (CME) are presented in the Figures 25-28.

Figure 25. Soybean oil (CME) price anomaly detection with ARIMA

Figure 28. Soybean oil (CME) price anomaly detection with Isolation forest

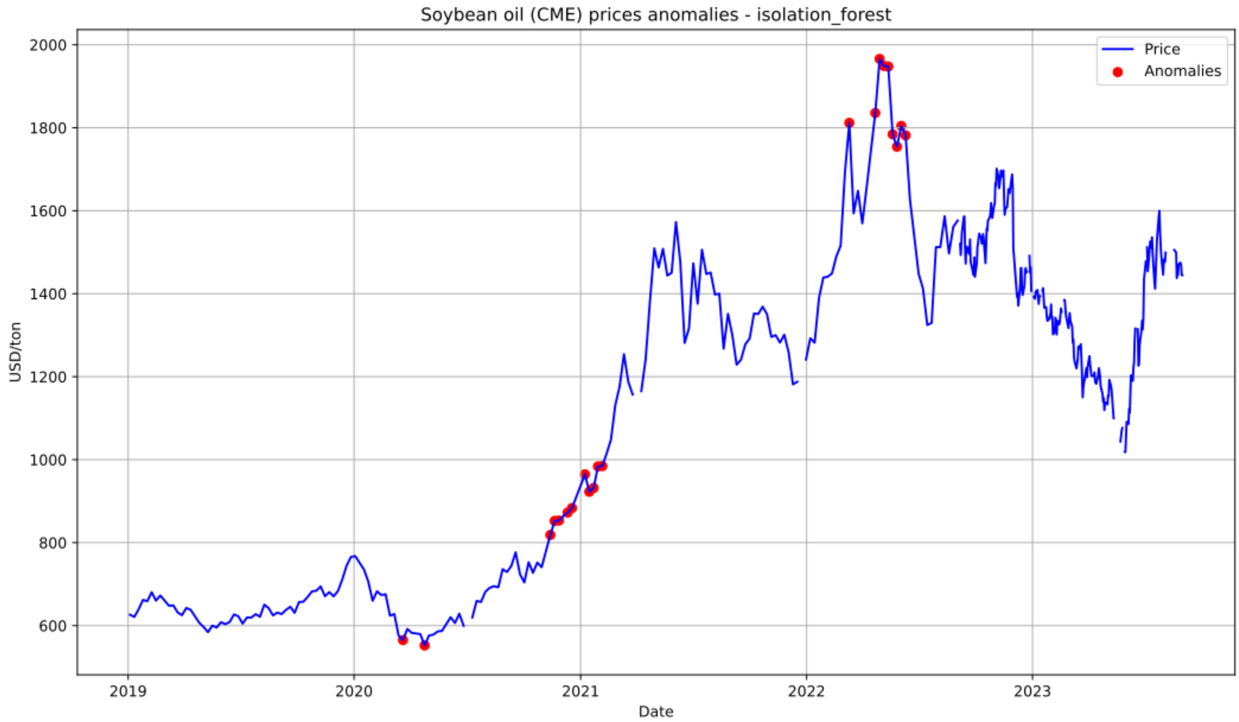


Figure 29. Soybean oil (CME) price anomaly detection with LOF

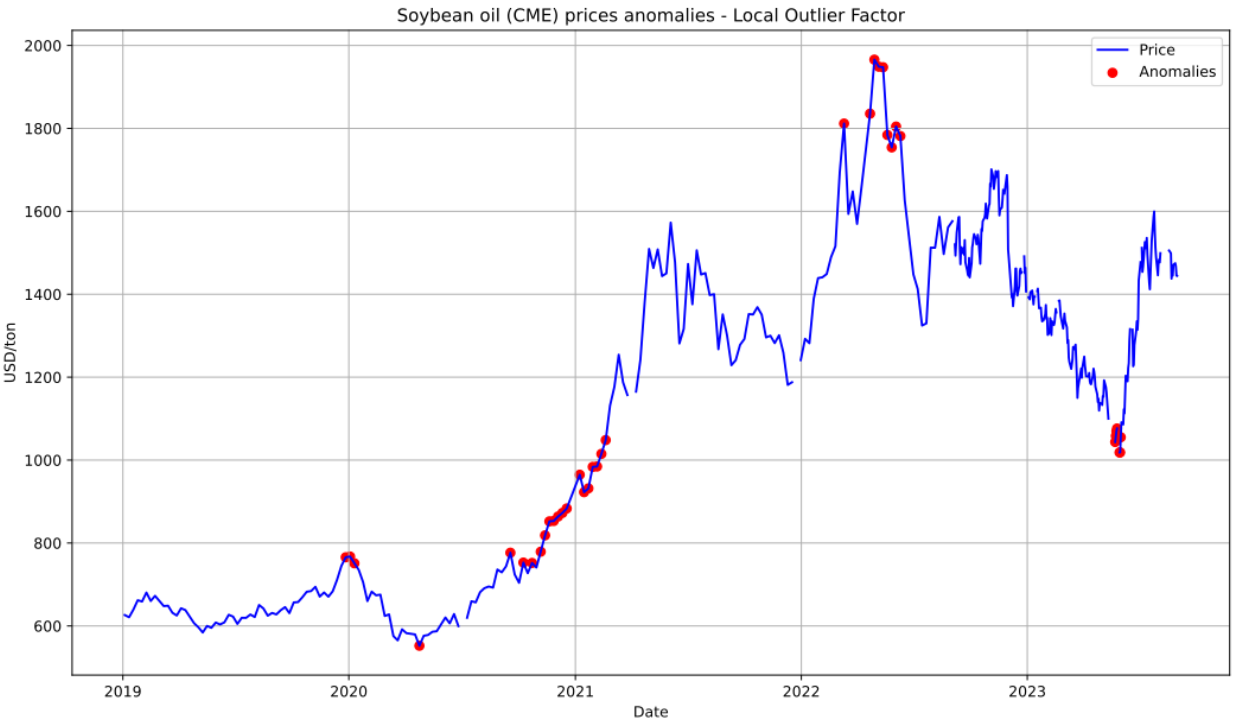
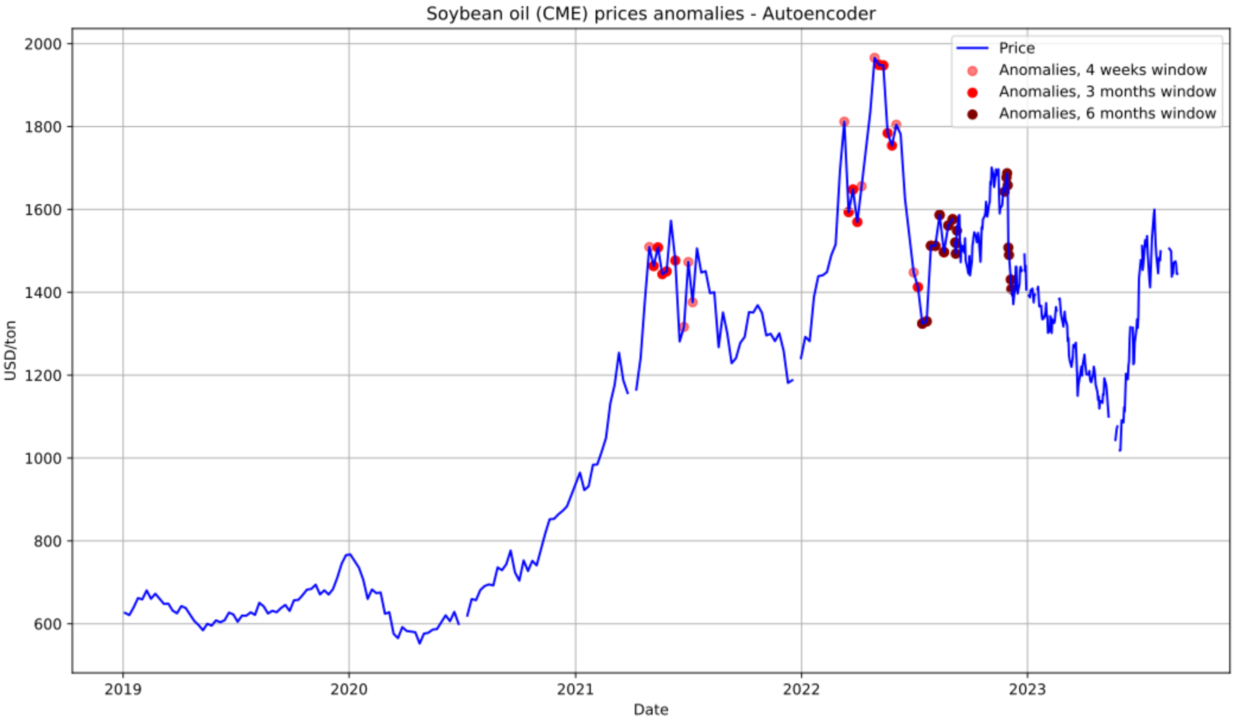


Figure 30. Soybean oil (CME) price anomaly detection with Autoencoder



<u>Soybean oil anomalies were strong and long-lasting in 2022.</u>

5. Crude oil Brent (ICE)

Results of anomaly detection for Crude oil Brent (ICE) are presented in the Figures 28-31.

Figure 31. Crude oil Brent (ICE) price anomaly detection with ARIMA

Figure 32. Crude oil Brent (ICE) price anomaly detection with Isolation forest

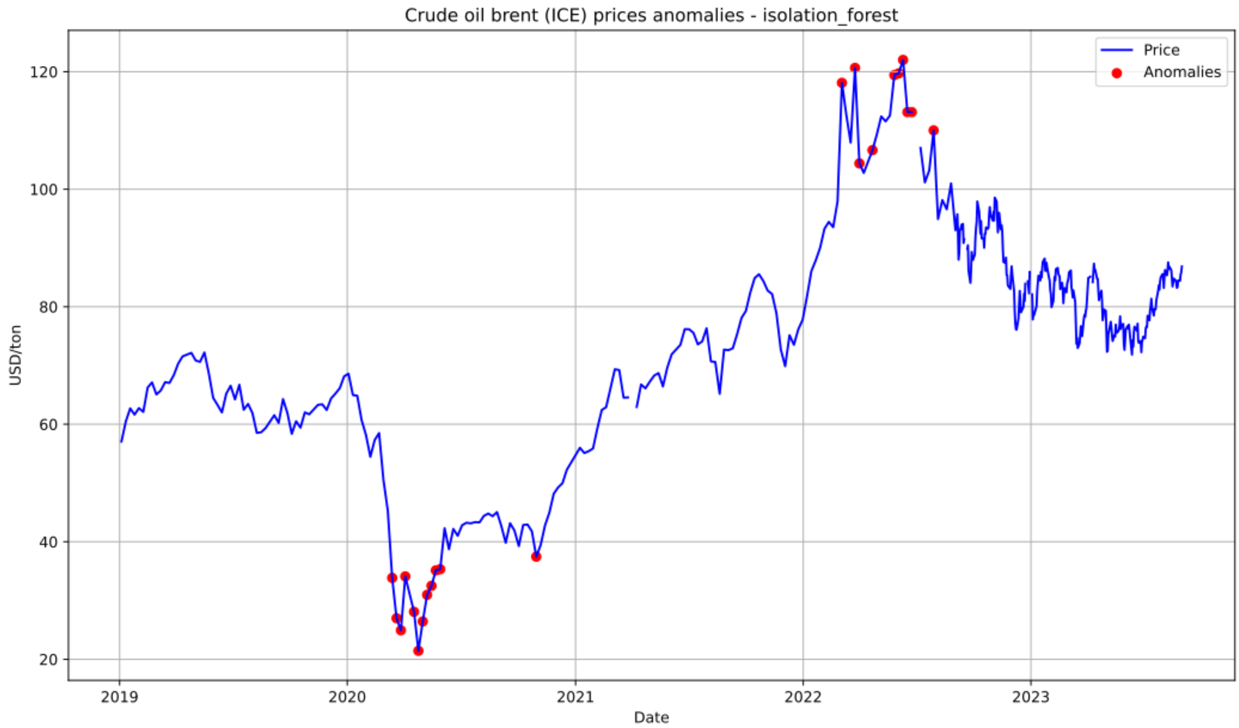


Figure 33. Crude oil Brent (ICE) price anomaly detection with LOF

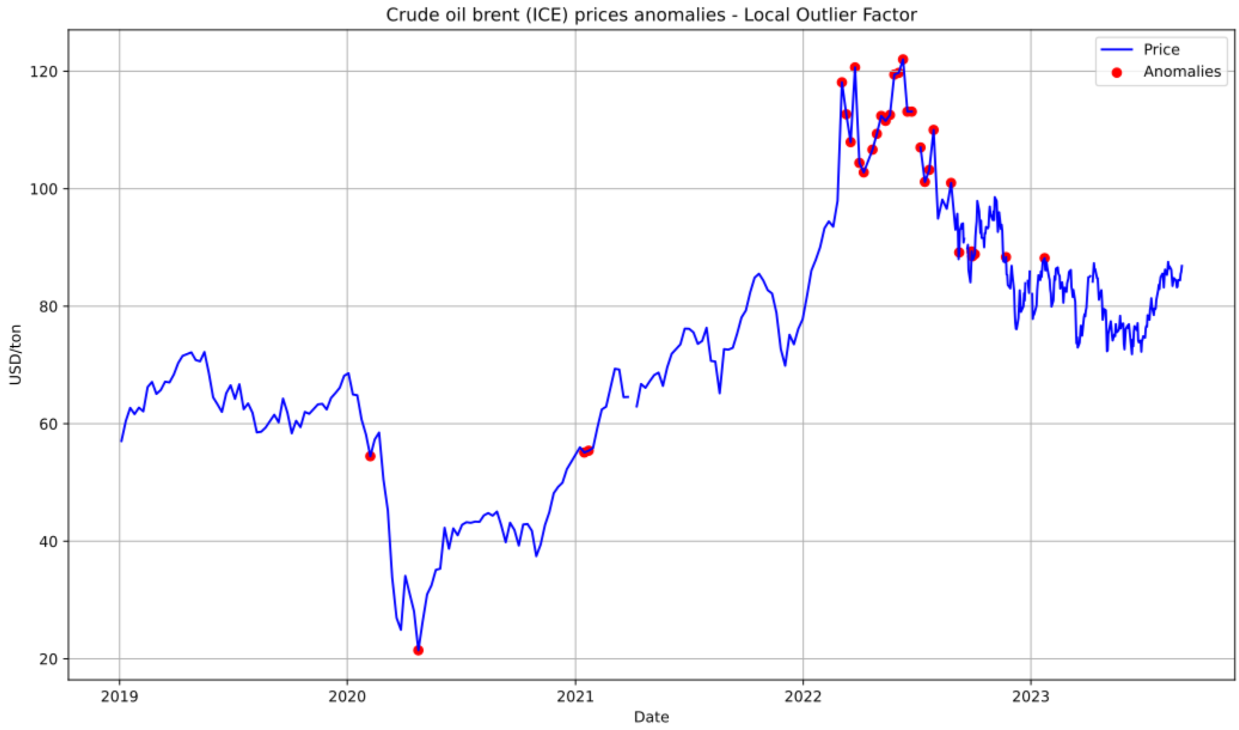
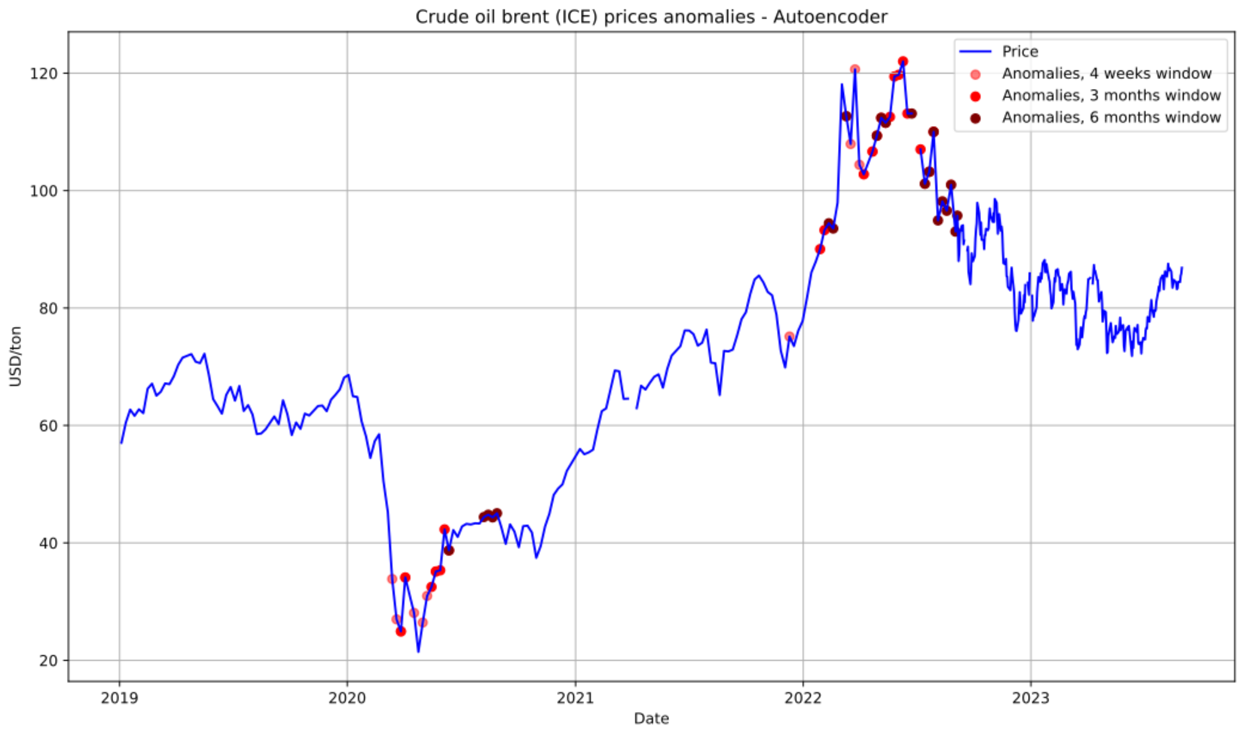


Figure 34. Crude oil Brent (ICE) price anomaly detection with Autoencoder



<u>Crude oil prices show strong and persistent anomalies in 2022.</u>

III. ANOMALIES IN PRICE SPREADS

The identification of anomalies within the price spreads between rapeseed oil and related commodities offers another perspective to examine market dynamics. Unlike anomalies within individual price time series, anomalies in spreads provide insights into the relationship between the two related commodities. These deviations from expected spread values serve as indicators of relative pricing relationships and shifts in market equilibrium. By focusing on these anomalies, we gain a deeper understanding of how unusual the rapeseed oil price dynamics is in a context of related markets.

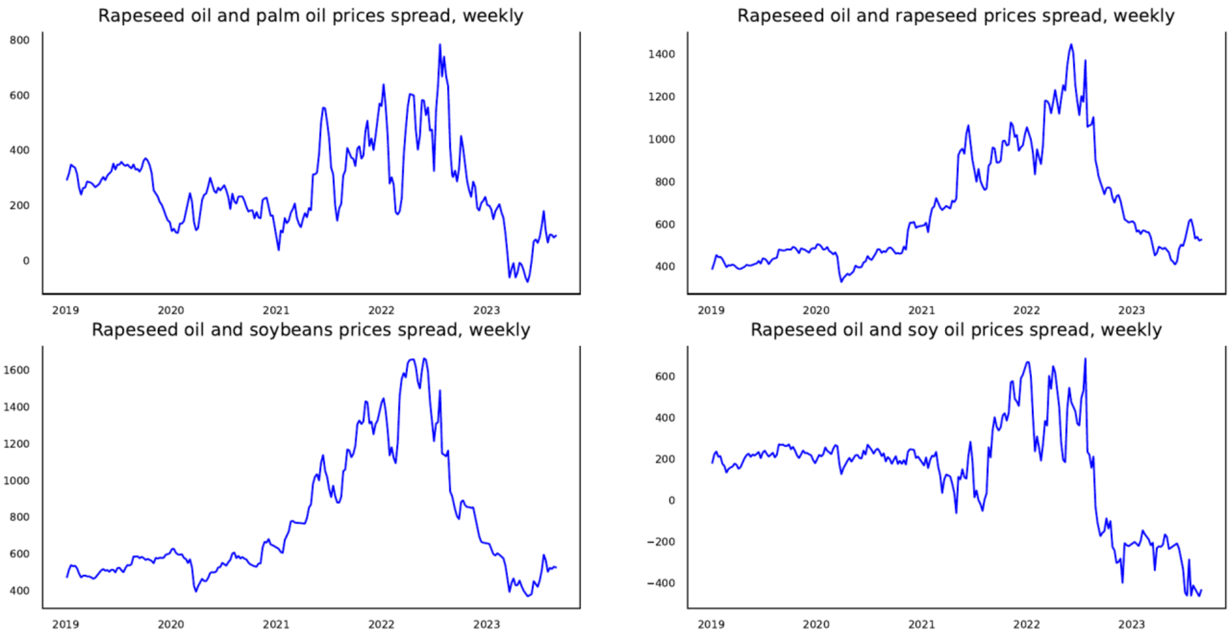
Data description and exploratory analysis

To calculate the prices spreads we use the data described in the previous section. Four new variables are obtained, which correspond to price spreads between the rapeseed oil and the following commodities:

- Palm oil;
- Rapeseed;
- Soybeans;
- Soy oil.

Each variable has 221 observations, representing a weekly spread of prices. Visualization of the data is presented on the Figure 35.

Figure 35. Spreads of prices of rapeseed oil and related commodities.



Despite the time series of these commodities' prices show the similar pattern with a peak at the beginning of 2022, from the spreads it is seen that the increase in price of rapeseed oil exceeds the increase in prices of related commodities. This confirms our findings in Section 1. As mentioned above, such increase was based on two main war-related factors: a) strong dependence of rapeseed oil prices on crude oil market, which skyrocketed during the war due to expectations of ban for Russian oil; b) relatively low share of Ukrainian rapeseed in EU import due to production and logistical problems in this country.

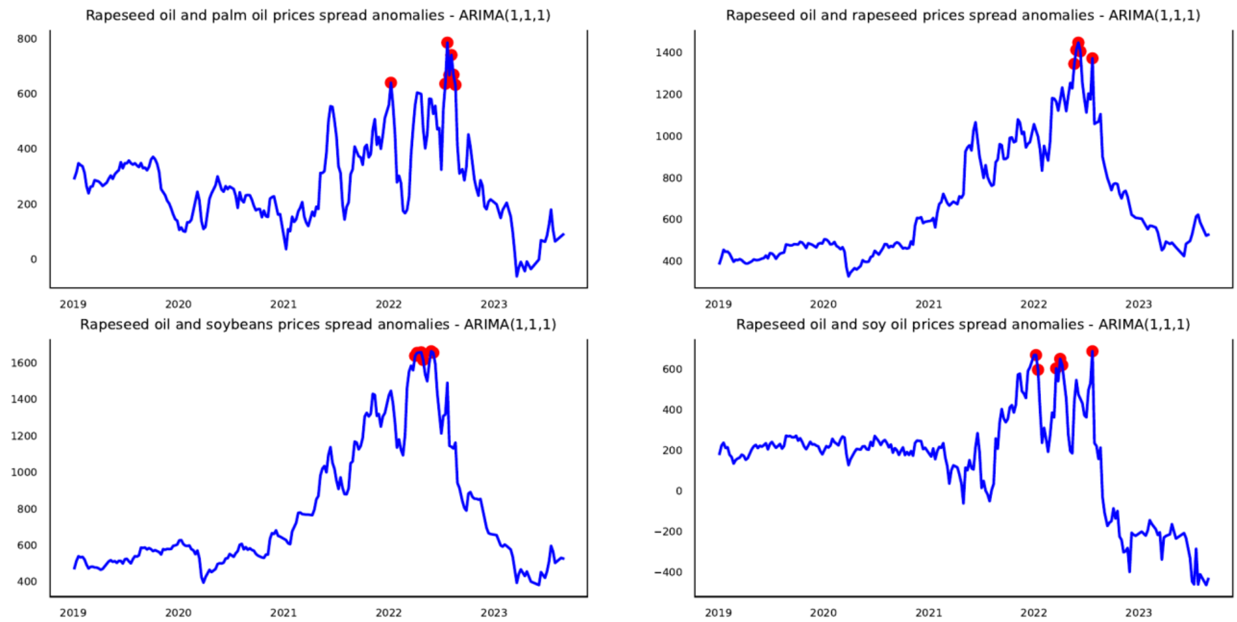
Anomaly detection

1. ARIMA

To estimate the ARIMA model, we apply the similar approach, which have been used in the section 1. For each of the price spreads, stationarity check was performed, and ACF and PACF functions were studied. Due to similar nature of the all four variables, chosen values of model parameters turned out to be the same: q=1, p=1, d=1.

To identify the anomalies, we fit the model with the training sample, make a projection, and compare it to the test sample. Observation which are outside the confidence interval of the projection are considered anomalies (outliers). Training sample includes data covering 1 January 2019 – 1 January 2022. Test sample covers the period of testing interest: 1 January 2022- 1 September 2023. The chosen threshold for anomaly is 1.9 standard deviations (which corresponds to 0.95 confidence interval). Results are presented on the Figure 36. Detected anomalies are located solely in mid-2022.

Figure 36. ARIMA (1, 1, 1) price spreads anomalies detection results.

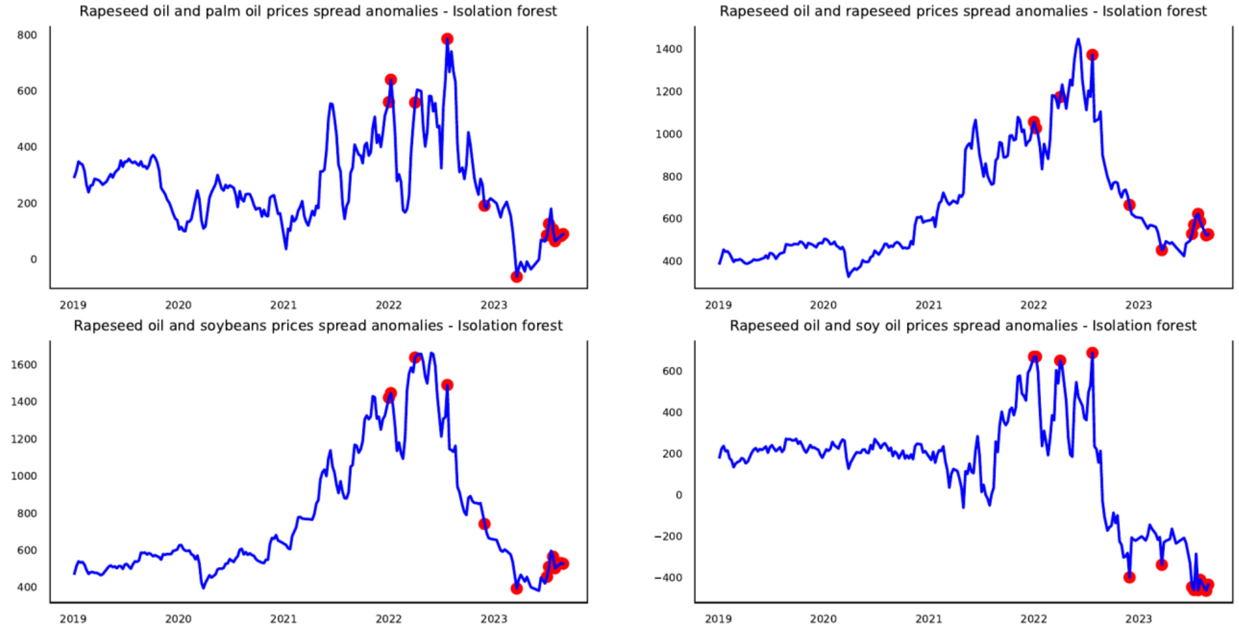


2. Isolation Forest

The same methodology as in the section (I) is applied to detect anomalies in prices spreads using Isolation Forest. In this case the contamination value of

0.05 was chosen, in order to label those points, which stand out the most. Results are presented on the Figure 37. Observed anomalies mostly are located in the period of drop in price spreads in early-mid 2023.

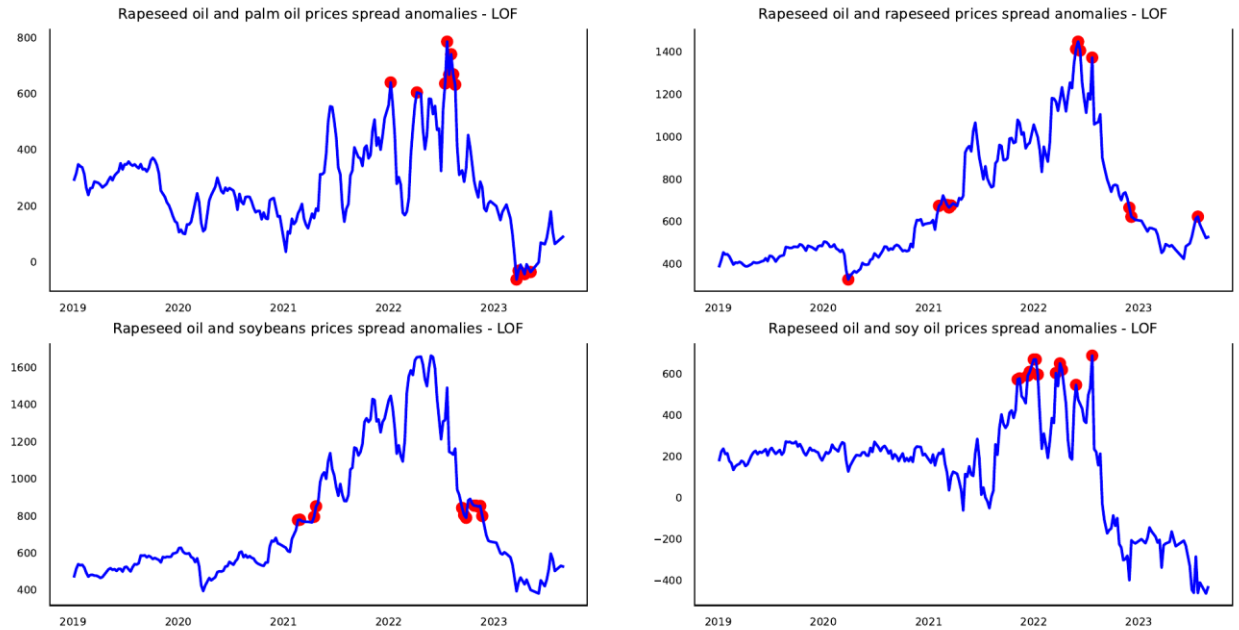
Figure 37. Anomalies identified with the isolation forest algorithm.



3. Local Outlier Factor

The same methodology as in the section (I) is applied to detect anomalies in prices spreads using Local Outlier Factor model. For estimation values of parameters were chosen as *k=100, contamination=0.05.* Results are presented on the Figure 38.

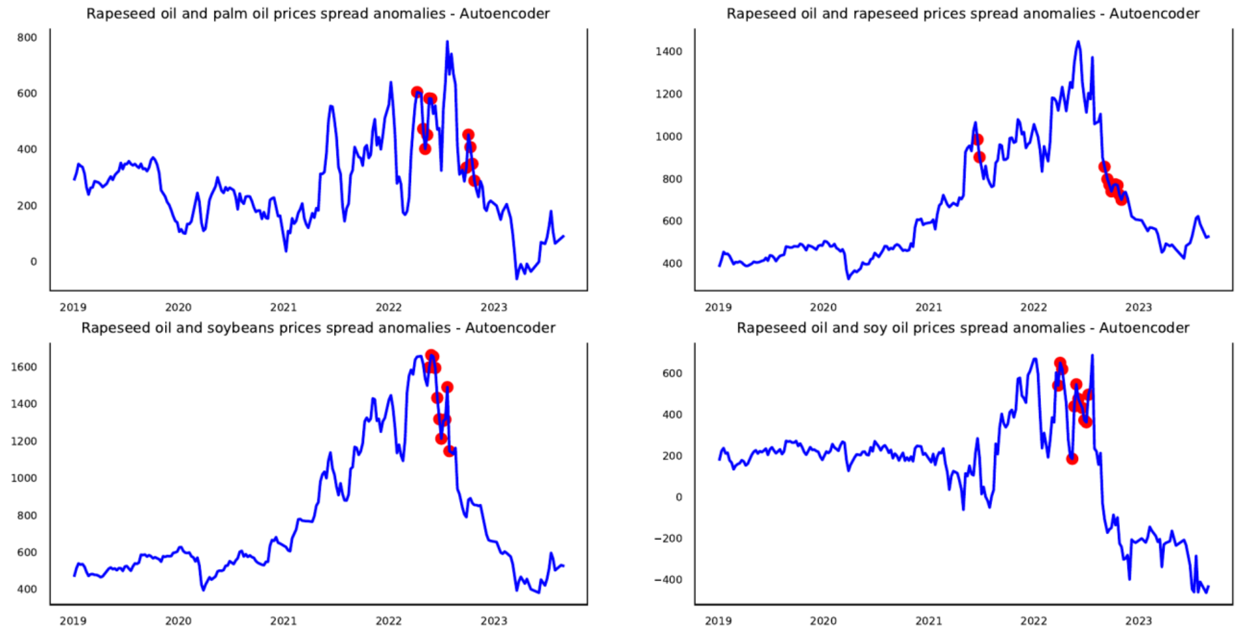
Figure 38. Anomalies detected with the LOF algorithm.



4. Autoencoder

The same methodology as in the section (I) is applied to detect anomalies in prices spreads using Autoencoder model. Results are presented on the Figure 39.

Figure 39. Anomalies detected with the Autoencoder algorithm.



<u>All four methods show that price spread between rapeseed oil prices and price for related commodities show strong anomalies in 2022. Again, this is because rapeseed oil prices are more sensitive to the war-related crude oil market shocks than prices for other oils.</u>

IV. RAPESEED OIL PRICES ANOMALIES DETECTION - MULTIVARIATE APPROACH

To observe the anomalies in prices of rapeseed oil in the context of connected markets and their dynamics, we adapt and apply models used in univariate analysis for multivariate data. In addition to rapeseed oil prices, we introduce to our models time series of soybeans, soybeans oil, palm oil, rapeseed and crude oil (Brent) prices. Instead of ARIMA model, the VAR model is used, as it is designed to work with multivariate time series.

Using additional variables in anomaly detection can enhance the accuracy of anomaly detection. This approach makes sense for several reasons:

Firstly, we assume that there might be a causality relationship between rapeseed oil and mentioned commodities, i. e., there are with various other agricultural and energy markets, which influence rapeseed oil prices. By incorporating additional variables, you gain a holistic view of the market dynamics and context. Anomalies in one market can often be linked to anomalies in others, reflecting systemic factors or global events that affect multiple commodities simultaneously. To observe the anomalies in the context of related markets, we apply the previously used Isolation Forest, LOF, and Autoencoder algorithms. Use of multiple models helps to cross-validate the results produced by different models.

Secondly, multivariate approach leverages the idea that anomalies in one variable can manifest as abnormal patterns in the relationships between variables. In the context of rapeseed oil prices, this means that unusual price movements in rapeseed may be correlated with unexpected changes in the prices of connected commodities. To account for these relationships, we estimate the VAR model, together with the Granger causality and cointegration tests.

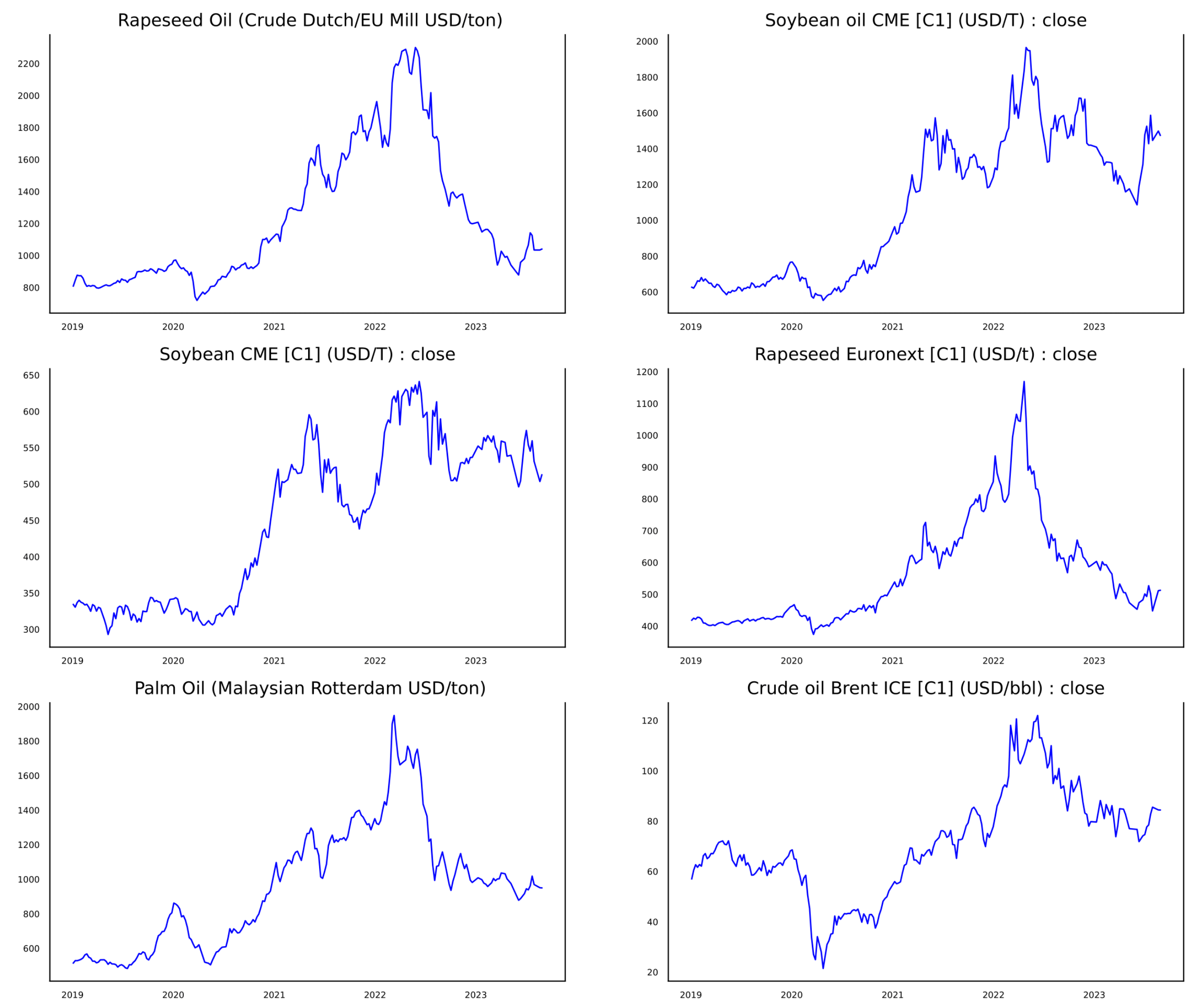
For each of the ML models (Isolation Forest, LOF, Autoencoder), 6 separate estimations were conducted on different datasets: 5 estimation on pairs of rapeseed oil price with each of the related commodities, and 1 fit with a full dataset. Estimation of the VAR model is discussed separately in the corresponding section.

Data overview

Data used in the estimation consists of the 220 observations of weekly prices of rapeseed oil, rapeseed, soybean oil, soybeans, palm oil, and crude Brent oil.

As expected, rapeseed and rapeseed oil prices follow the similar pattern. Trajectory of the palm oil price is somewhat similar, with a slight decrease at the beginning of 2020, growth in mid-2020-early 2022, peak and high volatility in spring-summer of 2022, and a rapid decrease to the 2021 level. Soybeans and soybean oil demonstrate slightly different trajectory with two distinct peaks in early 2021 and 2022. After the peak of 2022 its price decreased at much lower rate, as compared to rapeseed and palm oil. Crude oil prices dropped sharply in the beginning of 2020, peaked in 2022 and have been decreasing gradually throughout late 2022-2023. Visualization of the variables is presented on Figure 40.

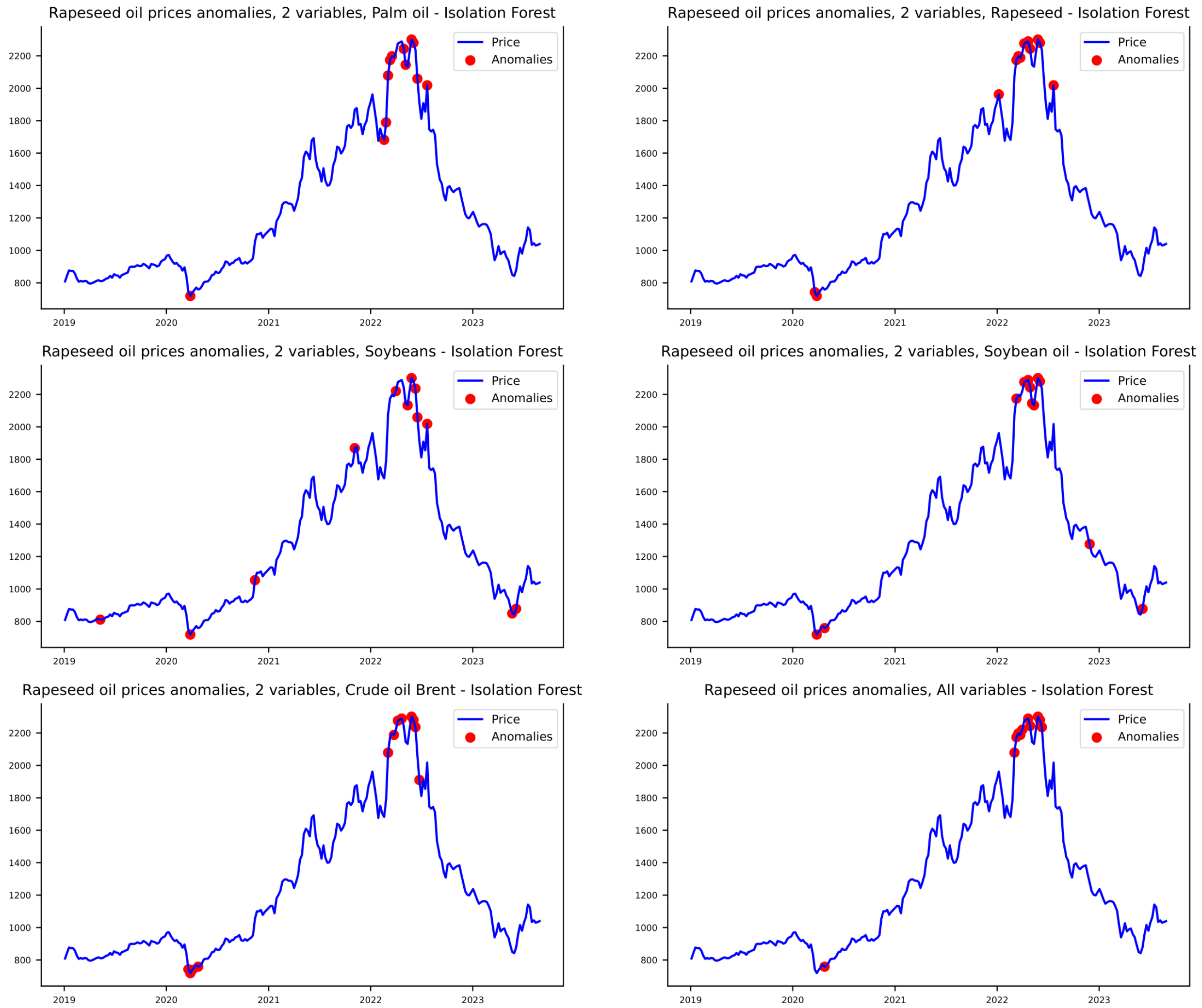
Figure 40. Plots of the variables in the dataset



1. Isolation Forest results

All estimations of Isolation Forest clearly indicate that the prices peak in April-June 2022 is an anomaly (Figure 41). Outside of this period, all models label a price drop in 2020 as an anomaly as well.

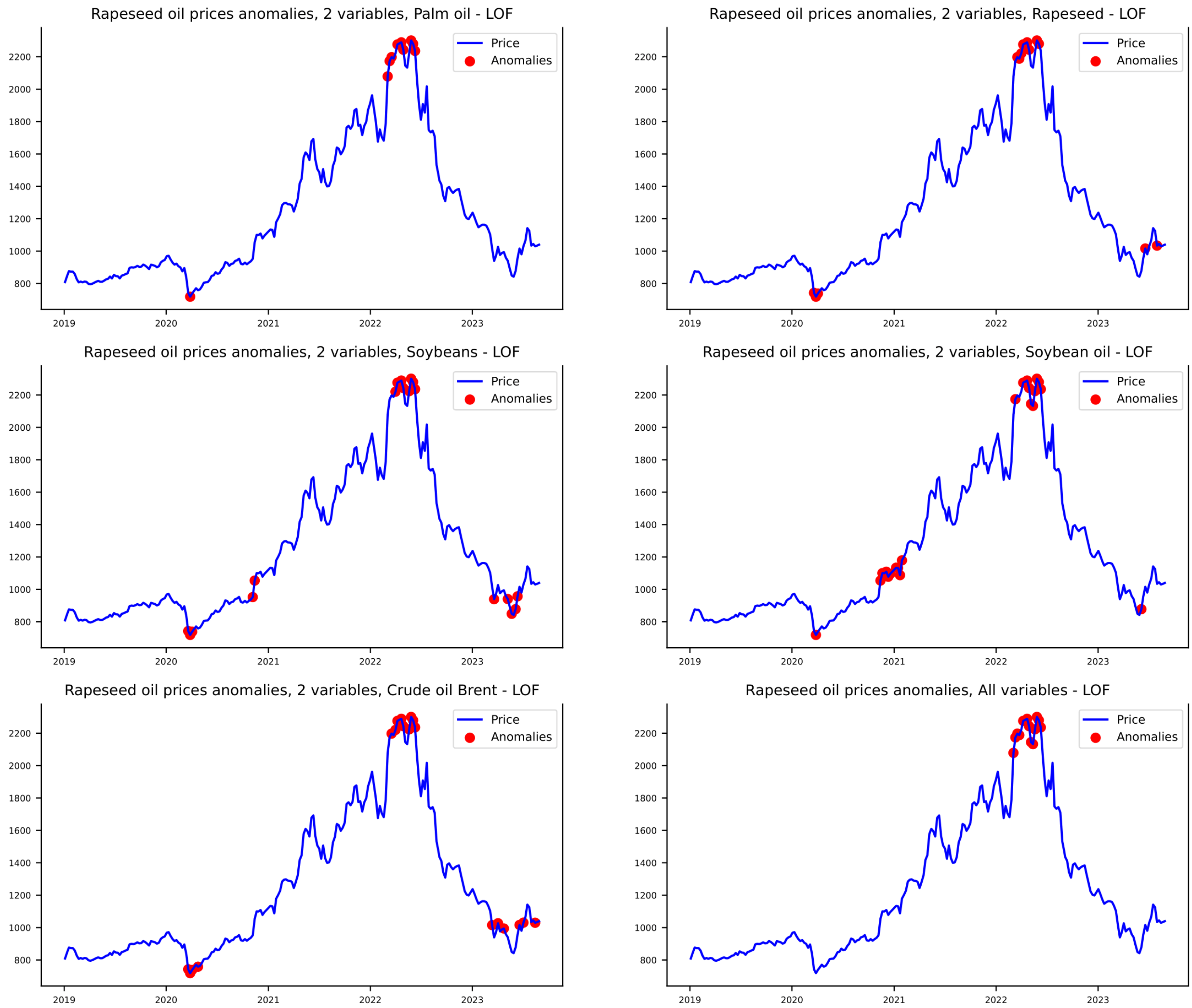
Figure 41. Results of multivariate anomalies detection with Isolation Forest



2. Local Outlier Factor results

Results produced by the Local Outlier Factor model are similar to Isolation Forest (Figure 42). Price peak of April-June 2022 is clearly labeled as an anomaly by all estimated models. Additionally, pairwise estimations with rapeseed, soybeans, soybean oil, and crude Brent oil prices suggest that the price decrease in spring of 2023 was an anomaly as well. However, these results do not show up in the estimation with all 6 variables.

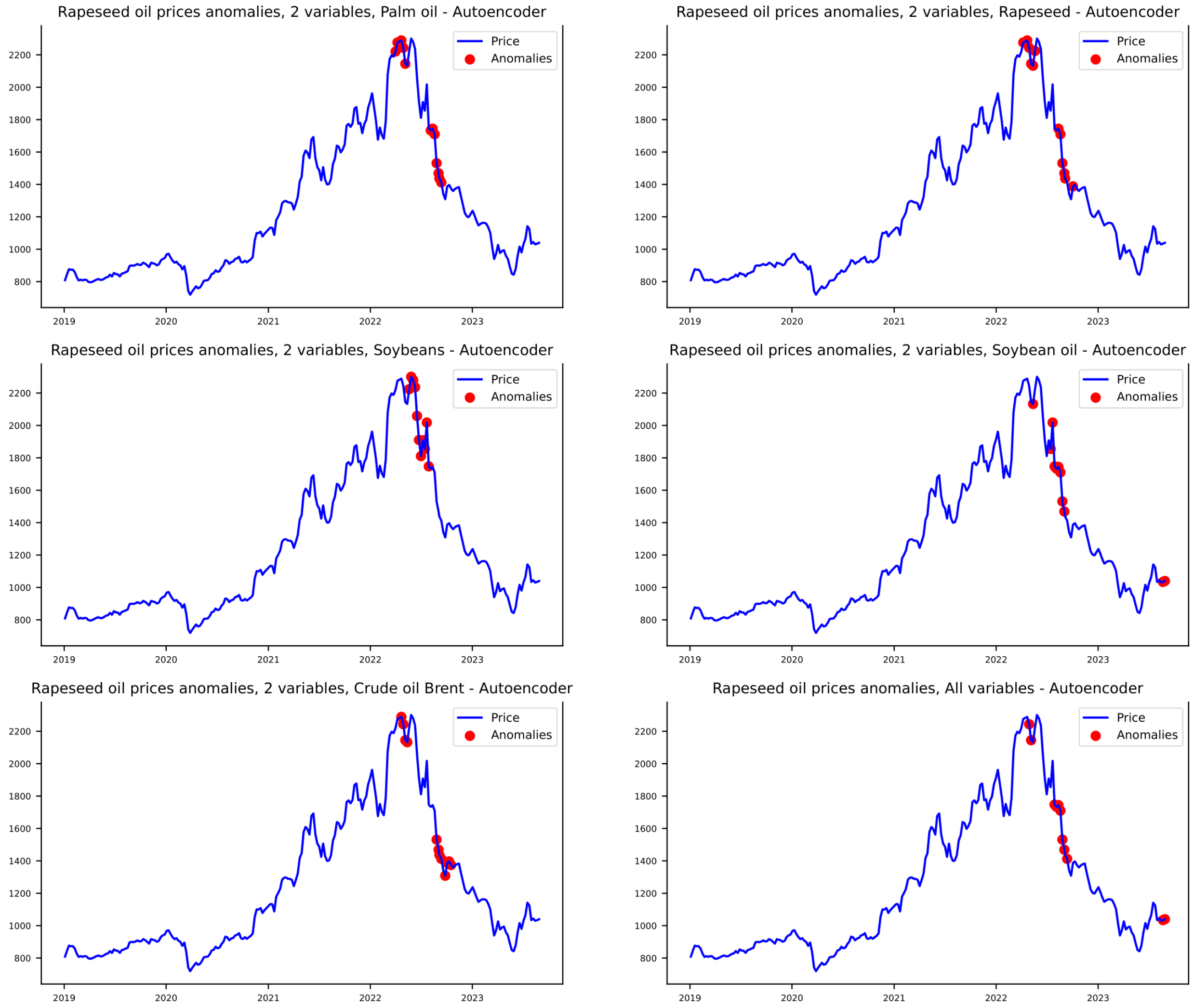
Figure 42. Results of multivariate anomalies detection with LOF



3. Autoencoder results, 3 months window

Autoencoder model similarly labels the prices peak in the spring of 2022 as an anomaly (Figure 43). Besides that, model labels as an anomaly a period of sharp price decrease in the summer of 2022. Results are similar for all 6 estimations.

Figure 43. Results of multivariate anomalies detection with Autoencoder



4. VAR estimation and results

The ARIMA model, used in the univariate analysis section, does not allow estimation with the multivariate data. Its multivariate alternative, Vector Autoregressive model, is applied. It is being chosen for several reasons. Firstly, they are particularly suited for modeling multivariate time series data, making them ideal for capturing the interdependencies and dynamic relationships that exist among variables, such as the prices of rapeseed oil and connected commodities like soybeans, palm oil, and crude Brent oil. Secondly, VAR models provide a flexible framework for studying how shocks or anomalies in one variable can affect the entire system over time, allowing for a more comprehensive understanding of market dynamics.

In the VAR model, each variable is modeled as a linear combination of past values of itself and the past values of other variables in the system. Since you have multiple time series that influence each other, it is modeled as a system of equations with one equation per variable (time series). The model equation of a

VAR(p) model of k variables takes form of:

(1) $Yt = c + \Phi 1 Yt - 1 + \Phi 2 Yt - 2 + \cdots + \Phi_p Yt - p + \varepsilon_t$, where p is a number of lags, c is a k-dimensional constant vector, Φ_p are a k x k coefficient matricies for each of the lags p, ε_t is an error term.

The basis behind VAR is the fact that each of the time series in the system influences each other. To check for such relationship before estimating the model, we perform Granger causality test (Table 5).

Table 5. Results of the Granger causality test

	Rapeseed Oil (Crude Dutch/EU Mill USD/ton)_x	**Soybean oil CME [C1] (USD/T) : close_x**	**Soybean CME [C1] (USD/T) : close_x**	**Rapeseed Euronext [C1] (USD/t) : close_x**	**Palm Oil (Malaysian Rotterdam USD/ton)_x**
Rapeseed Oil (Crude Dutch/EU Mill USD/ton)_y	1.0000	0.0112	0.0552	0.0000	0.0000
Soybean oil CME [C1] (USD/T) : close_y	0.0026	1.0000	0.0012	0.0049	0.0000
Soybean CME [C1] (USD/T) : close_y	0.0000	0.0868	1.0000	0.0425	0.0001
Rapeseed Euronext [C1] (USD/t) : close_y	0.0000	0.0000	0.0010	1.0000	0.0000
Palm Oil (Malaysian Rotterdam USD/ton)_y	0.1962	0.2337	0.1420	0.0065	1.0000

Causality (with respect to rapeseed oil price, 1st row in the table) is significant at the 0.05 level for all variables except for soybean CME price. However, the significance barely exceeds the threshold of 0.05 (0.0552), and is significant at the 0.1 level. Thus, all 5 variables will be used in the model.

Pairwise cointegration tests are conducted to establish the presence of a statistically significant connection between the time series. If there exists a linear combination of two or more time series that has an order of integration (d) less than that of the individual series, then the series are said to be cointegrated. The results of the cointegration tests:

$$Palm\ oil\ (test\ statistic\ value, p - value)\text{: } (-2.41, 0.3206)$$

$$Rapeseed\ (test\ statistic\ value, p - value)\text{: } (-4.18, 0.0039)$$

$$Soybeans\ (test\ statistic\ value, p - value)\text{: } (-1.66, 0.6949)$$

$$Soy\ oil\ (test\ statistic\ value, p - value)\text{: } (-1.24, 0.8460)$$

Results suggest that no cointegration is found for palm oil, soybean, and soy oil prices. A statistically significant (at 0.01 level) cointegration relationship is found for rapeseed price. This means that usually rapeseed and rapeseed oil prices tend move synchronically and deviations from this dynamic indicate the presence of extraordinary market shocks. Indeed, the spread between rapeseed oil and rapeseed prices is a proxy of rapeseed crush margin. When the margin is high, crushing plants increase demand for rapeseed and narrow crushing margin, and vice versa. Therefore, spread between rapeseed oil and rapeseed tend to fluctuate around some average values (in the normal market conditions).

While a Vector Error Correction model (VECM) should be used for a pair of rapeseed oil and rapeseed prices, for palm oil, soybeans, and soybean oil VAR model should be used. To identify the anomalies, we fit the model with the training sample, make a projection, and compare it to the test sample and its confidence interval. Observation which are outside the confidence interval of the projection are considered anomalies (outliers). Training sample includes data covering 1 January 2019 – 1 February 2022. Test sample covers the period of testing interest: 1 February 2022- 1 September 2023.

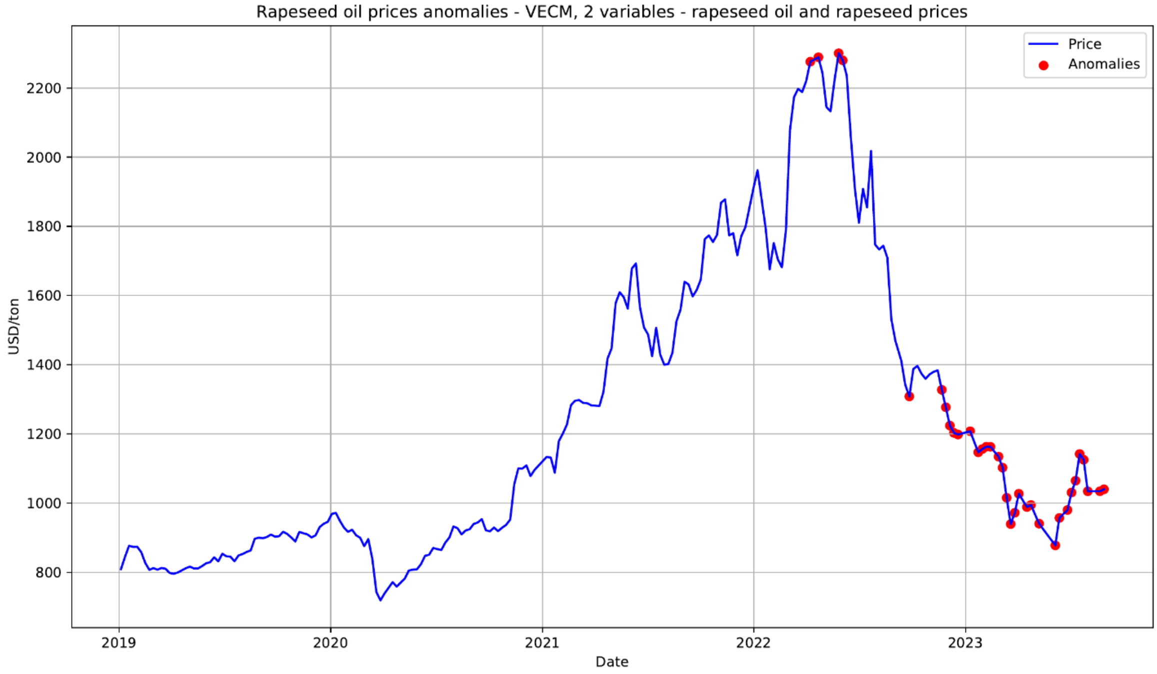
The optimal number of lags for the VECM model (rapeseed oil and rapeseed) is chosen as 1 based on AIC and BIC criteria.

For VAR models (rapeseed oil pairwise with palm oil, soybeans, and soy oil), to achieve stationarity data was differenced once. It was not enough to achieve the stationarity for soy oil, so in the model with soy oil, degree of differencing is equal to 2. Similarly, based on AIC and BIC criteria, number of lags is chosen to be 2 for palm oil and soy oil models. For soybeans, optimal number of lags is 1.

The threshold for anomaly detection was set as 2 standard deviations of residuals.

Results of the anomalies detection with the VECM model are presented on the Figure 44. Anomalies are detected in 2022 and the first half of 2023; they indicate the asynchronous movement of rapeseed oil and rapeseed prices when rapeseed oil prices increased faster or decreased slower than rapeseed prices.

Figure 44. Results of multivariate anomalies detection with VECM(1), rapeseed oil and rapeseed prices.



Results of the anomalies detection with VAR models are presented on the Figures 45-47.

Figure 45. Results of multivariate anomalies detection with VAR(2), rapeseed oil and palm oil.

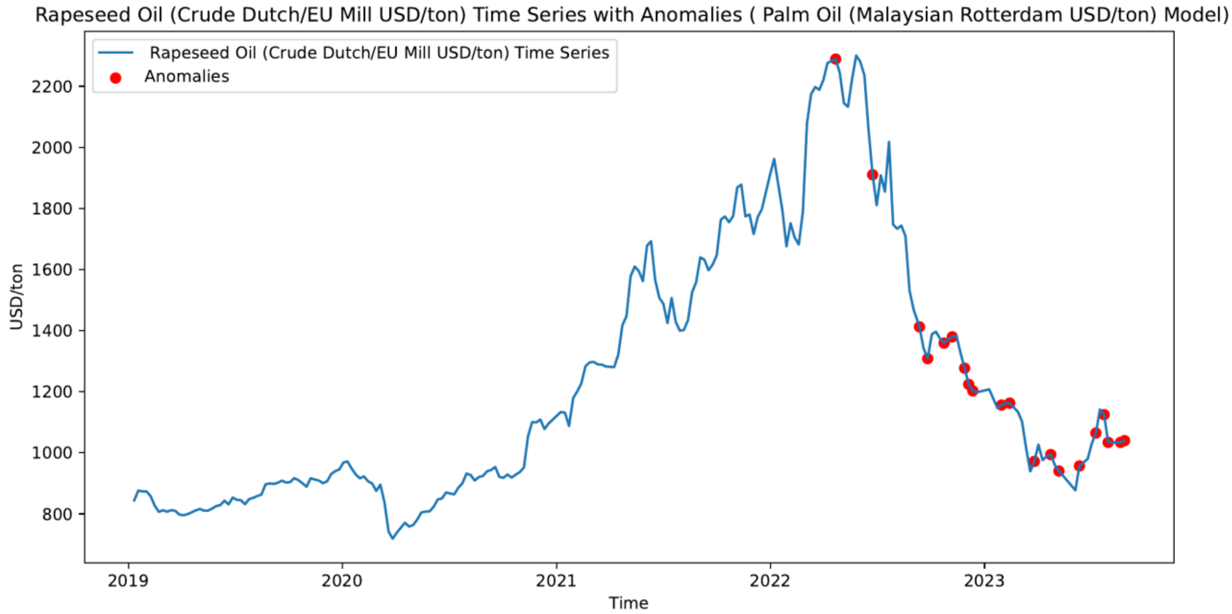


Figure 46. Results of multivariate anomalies detection with VAR(1), rapeseed oil and soybeans.

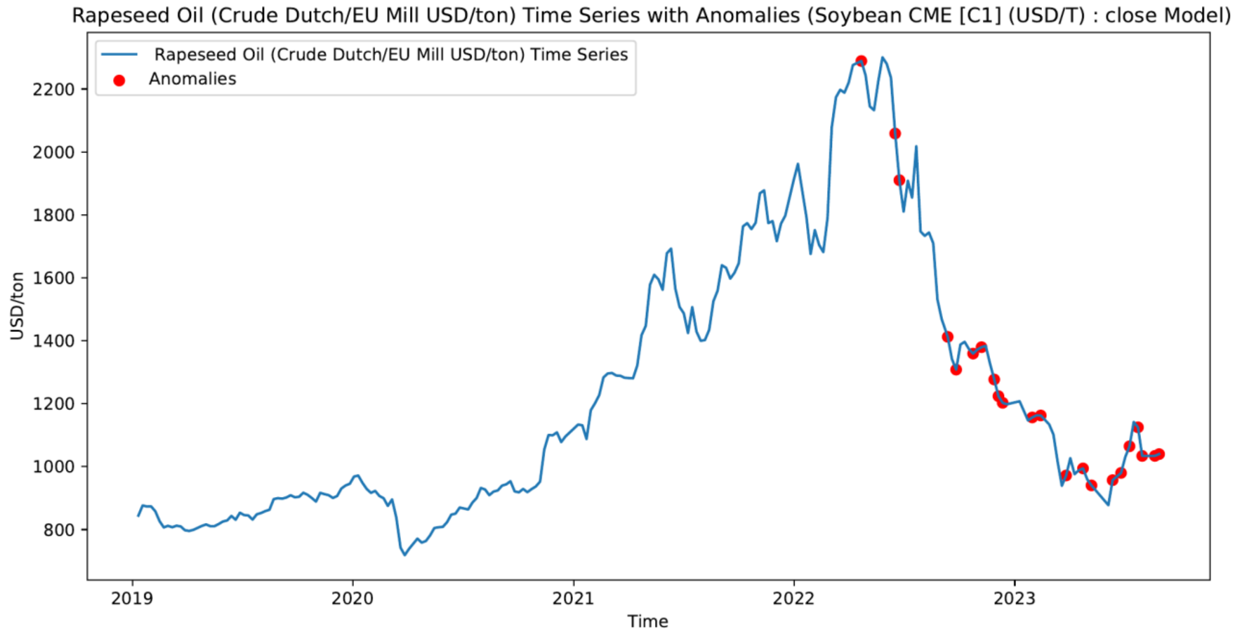
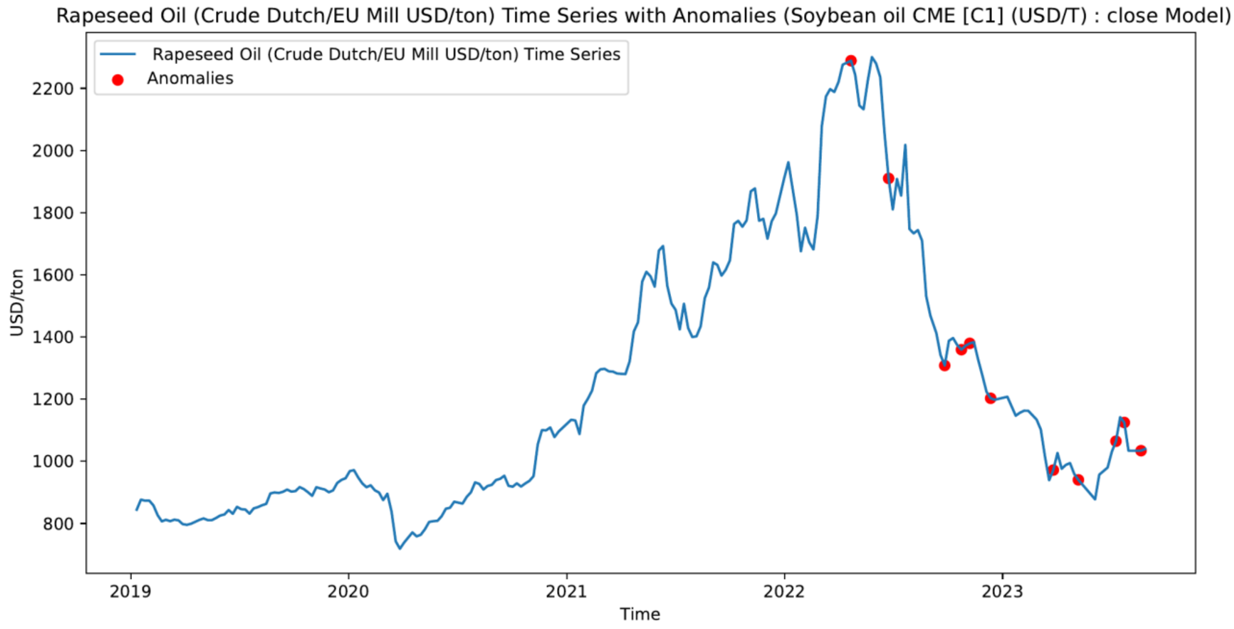


Figure 47. Results of multivariate anomalies detection with VAR(1), rapeseed oil and soy oil



VAR model anomalies could be interpreted as extraordinary strong shocks transmitted from price of exogenous variables (palm oil, soybeans, soy oil) to rapeseed oil prices. This partially confirm our hypothesis that rapeseed oil prices were pushed by the war-related growth of global vegetable oils market.

SUMMARY

The full-scale war in Ukraine has strong effect on the whole range of commodities, in particular, vegetable oils. Our analysis reveals that the war-related growth of the rapeseed oil in the EU was more pronounced compared to other vegetable oils. First of all, this could be explained by the rapid growth of the EU demand for rapeseed oil. Such growth was caused by the extremely high crude oil prices, which make rapeseed oil more attractive for biodiesel production. Second, blockade of Black Sea ports soared global prices for agricultural commodities, many of which are substitutes for rapeseed oil (corn as a main input for bioethanol production, sunflower oil as an important food ingredient).

Third, the war negatively affected exports of Ukrainian rapeseed to the EU countries. However, the effect of this factor was limited compared to the first two. This is confirmed by the review of market balances which shows sufficient amount of shocks. Therefore, the increase of rapeseed and rapeseed oil prices has more exogenous and speculative nature in the wartime. Tracing the price dynamics for rapeseed oil in Rotterdam port shows the high market fluctuations connected to the war dynamics. For example, the main spikes were observed after the start of the invasion, news on the Bucha tragedy, missile attacks on Odessa and Mykolaiv seaports, termination of the Grain Deal. The non-normality of rapeseed oil price dynamics could be partially confirmed by the inverted calendar spreads for rapeseed futures markets on Euronext.

To conduct the anomaly detection in price time series for rapeseed oil and related products, we employed a set of econometric and machine learning techniques. For the single time series, four methods are used: ARIMA model, Isolation forest, Local Outlier Factor (LOF), Autoencoder. They show a large number of anomalies in rapeseed oil price and prices of related commodities (rapeseed, palm oil, soya beans, soybean oil, crude oil). According to the autoencoder algorithm, anomalies in the wartime were more persistent in time than anomalies before the war. Also, the price spreads between rapeseed oil and connected commodities show strong anomalies in the wartime. This means that rapeseed oil prices were not moving synchronically with prices of other commodities. In particular, price spikes of the spread between rapeseed oil and rapeseed (proxy for crush margin) are anomalies. This means that the EU rapeseed oil producers benefit from the market volatility in the wartime.

For the multivariate analysis, we also use Isolation forest, Local Outlier Factor (LOF), Autoencoder methods, but also added Vector Autoregressive model (VAR) and Vector Error Correction model (VECM). Results shows abnormally strong effect of related commodities on rapeseed oil. This unidirectional effect is confirmed by the Granger test. All related commodities except soybeans affected the disturbance in rapeseed oil price dynamics. Moreover, cointegration tests and VECM results confirm the disruption of synchronous movement between rapeseed oil and rapeseed price time series in the wartime. Given the high crush margin in 2022, this confirms the fact that EU rapeseed oil producers received extra-profits on the volatile market during the war.

REFERENCES

1. Shaukat, K. et al. (2021). A Review of Time-Series Anomaly Detection Techniques: A Step to Future Perspectives. In: Arai, K. (eds) Advances in Information and Communication. FICC 2021. Advances in Intelligent Systems and Computing, vol 1363. Springer, Cham.

2. Teng, X., Lin, Y.-R., Wen, X.: Anomaly detection in dynamic networks using multi-view timeseries hypersphere learning. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, pp. 827–836 (2017)

3. Fahim, M., Sillitti, A.: Anomaly detection, analysis and prediction techniques in iot environment: a systematic literature review. IEEE Access 7, 81664–81681 (2019)

4. Kacprzyk, J., Owsinski, J.W., Viattchenin, D.A., Shyrai, S.: A new heuristic algorithm of possibilistic clustering based on intuitionistic fuzzy relations. In: Novel Developments in Uncertainty Representation and Processing. Springer (2016)

5. Garcia-Font, V., Garrigues, C., Rifà-Pous, H.: A comparative study of anomaly detection techniques for smart city wireless sensor networks. Sensors 16, 868 (2016)

6. Liu, F. T., Ting, K. M., Zhou, Z. H. "Isolation Forest," 2008 Eighth IEEE International Conference on Data Mining, Pisa, Italy, 2008, pp. 413-422, doi: 10.1109/ICDM.2008.17.