GROUP WORK PROJECT # 1 MScFE 660: RISK MANAGEMENT
Group Number: 8183

FULL LEGAL NAME	LOCATION (COUNTRY)	EMAIL ADDRESS	MARK X FOR ANY NON-CONTRIBUTING MEMBER
Ebenezer Yeboah	Ghana	ebenezeryeboah46@gmail.com	
Christson Hartono	Indonesia	hartono.christson@gmail.com	
Eric Walter Pefura-Yone	Cameroon	pefurayone@gmail.com	

Statement of integrity: By typing the names of all group members in the text boxes below, you confirm that the assignment submitted is original work produced by the group (excluding any non-contributing members identified with an "X" above).

Team member 1	**Ebenezer Yeboah**
Team member 2	**Christson Hartono**
Team member 3	**Eric Walter Pefura-Yone**

Use the box below to explain any attempts to reach out to a non-contributing member. Type (N/A) if all members contributed.
Note: You may be required to provide proof of your outreach to non-contributing members upon request.

Step 2: Overview

a.Problem identification(Alvi)

The thesis attempts to solve several important issues related to the prediction of crude oil prices in the real world. The key problem being tackled here is the complex and unpredictable nature of the oil markets which are clearly influenced by macroeconomics, microeconomic, and geopolitical factors. The volatile nature of the prices of oil has made it necessary to find means to accurately predict its movement for relevant stakeholders such as traders, managers, and even shareholders to make decisions.

It is clearly highlighted in the thesis that it is difficult to learn the structure of the oil market. The traditional method, which has always relied on expert knowledge and related it to various available datasets, is both time-consuming and prone to errors. A new approach is proposed by the thesis which is using the Bayesian Probabilistic Graphical Models (PGMs) that can autonomously learn the relationship that exists between different market factors without necessarily needing expert intervention. The evolving nature of the oil market has made this a significant thing necessitating a flexible model that can adapt to new information and changing market conditions.

Also, this research talks about the challenges in data exploration and exploitation. As clearly seen, the oil market is characterized by a vast amount of data which includes economic indicators, levels of production, and geopolitical events. The thesis also explains the importance of effectively analyzing data to extract or derive insights that can help predict prices. Through the use of computational finance and machine learning methodologies, the research aims to improve the predictive accuracy of the oil price forecast which concludes as a valuable tool for traders and investors.

Lastly, there is also the validation of the constructed models in the thesis. For this model to be relevant (useful in financial markets) it must be tested under various economic scenarios. For that matter, there is stress testing in the models to stimulate economic distress situations which assess their performance so that they can withstand real-world challenges.

b. Bayesian networks in forecasting crude oil prices(Alvi)

Bayesian networks are well suited to forecasting commodity prices, including crude oil prices. These networks are capable of modeling volatile energy markets due to their ability to model complex causal relationships, integrate heterogeneous data and handle uncertainties and missing data.

Modeling Complex Relationships: The Bayesian networks help in analyzing complex relationships under macroeconomic, microeconomics, and geopolitical areas that ultimately affect crude oil prices. As it is well known that the oil market is not influenced by a single variable but rather a web of interconnected factors which includes demand and supply dynamics, geopolitical events, economic indicators, and market sentiments. Using the Bayesian networks, one can graphically model these relationships making it easier to understand how these various factors interact and influence one another. For instance, a

Bayesian network can help illustrate the changes in OPEC production levels which can affect global oil supply and in the long run affect prices.

Handling Uncertainty: With the high level of uncertainty in the oil market as a result of the volatility of prices and the unpredictability of external factors which includes natural disasters and geopolitical tensions, the Bayesian network excels in environments where there is a high level of uncertainty. The Bayesian network allows for the incorporation of uncertainty in the form of probability distributions which enables the model to quantify the likelihood of various outcomes based on the available data. For instance, a Bayesian network can provide a range of possible prices along with their associated probabilities rather than a deterministic forecast.

Learning from Data: This thesis also emphasized the importance of the learning structure of the oil market through data analysis. Through the Bayesian framework, the automatic learning of relationships among or between various variables from the data is facilitated. The ability of the Bayesian network to learn the data means it can evolve and improve over time to make it more accurate as it is exposed to new information.

Incorporating Prior Knowledge: The Bayesian network has the ability to incorporate prior knowledge and opinions from experts in the modeling process. Narrowing it down to crude oil, this means that historical data, insights from experts as well as established economic theories can be integrated into the model to improve or enhance its predictive power. This is important because knowledge from experts and historical data or trends can help improve results from the prediction.

Validation and Robustness: The Bayesian network can be subject to testing through some techniques like cross-validation and stress testing. Its performance can be assessed well under different conditions which helps in ensuring a robust system enough to be deployed in the financial markets.

c. Advantages of using Bayesian network(Alvi)

Several advantages come with the usage of the Bayesian network in forecasting crude oil prices. This advantage defines it as a better methodology as compared to the traditional way of forecasting oil prices.

- Effective Management of Uncertainty: It has been earlier established that the oil market is characterized by a high level of uncertainty that stems from natural disasters, economic fluctuations, and geopolitical tensions. Systematically, the Bayesian network inherently uses probability to manage uncertainty. It generates a range of possible outcomes of which each is associated with a probability. For example, the Bayesian network can provide a forecast of future oil prices along with the likelihood of the various price levels occurring which is important for risk management.
- Incorporation of Prior Knowledge: The Bayesian network has the ability to incorporate prior knowledge and expert opinion which is being used in the traditional way of predicting the oil prices on the market. Nonetheless, historical data or trends can also be incorporated into the Bayesian network which is equally relevant for understanding the current dynamics of the

market. The Bayesian network makes use of the Bayes theorem to give updates on the various probabilities of the prices whenever there is new data. This adaptability can be said to be very crucial in the fast-paced world of trading oil where conditions can change rapidly and significantly therefore, timely information is essential in making sound decisions.

- Facilitating Decision-Making: Ultimately, the importance of the Bayesian network for forecasting crude oil prices culminates in its ability to facilitate decision-making. Through a comprehensive understanding of the various factors that influence oil prices, incorporating prior knowledge as earlier discussed, quantifying uncertainties, and the use of new data, the Bayesian network empowers traders, investors and policymakers to make better decisions in the complex and volatile oil market.

Step 3: Identify, import, structure, and graph data

Given that several types of data may be available from the same data provider, we have written 3 main functions to import and structure macroeconomic data (Student A), geopolitical data (Student A), microeconomic data (Student B) and financial data (Student C) for the period from January 1995 to December 2024. These functions are described and developed in this notebook according to the APIs used. We have used the APIs of the Energy Information Administration(EIA), Federal Reserve Economic Data(FRED), World Bank, Economic Policy Uncertainty(EPU), European Central Bank and Yahoo Finance. The definitions and descriptions of the variables are given in the data dictionary (Table 1) and the table of variables (Table 2). All the data were then combined into a single data table. In the combined data table, the USD/EUR exchange rate was not available for the period 1995 to 1998 (the European currency only came into effect in 1999).

The full retrieved dataset has 360 rows and 27 variables.

a. Graphing macroeconomic data

Figure 1: Time series of original macroeconomic data

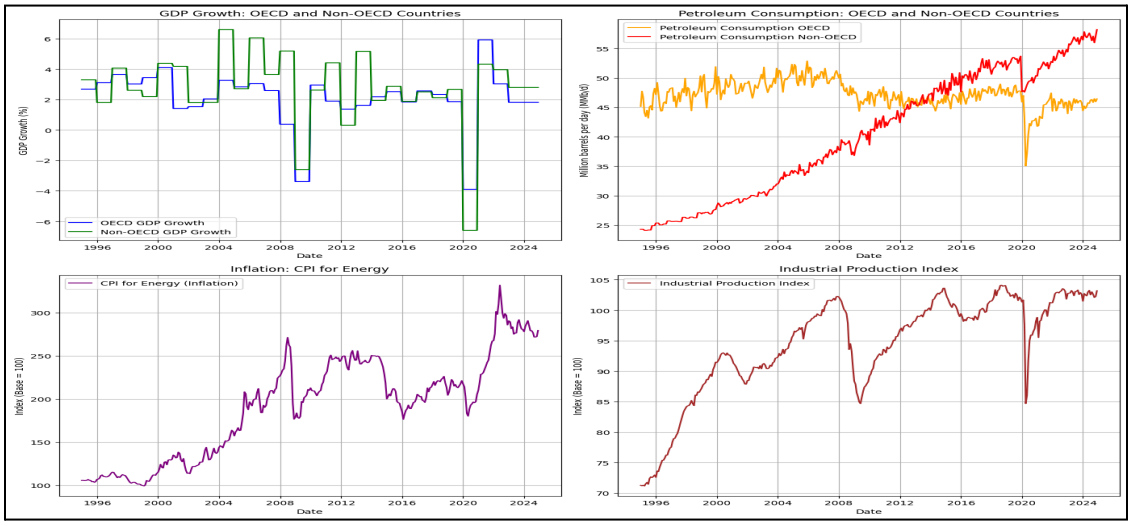


Figure 1 presents four macroeconomic time series from 1995 to 2024. It compares GDP growth in OECD and non-OECD countries, oil consumption in these two groups of countries, inflation via the energy CPI (consumer price index), and the industrial production index. The graphs illustrate the trends, volatilities and interactions between these key indicators over 30 years.

b. Graphing geopolitical data

Figure 2: Time series of original geopolitical data

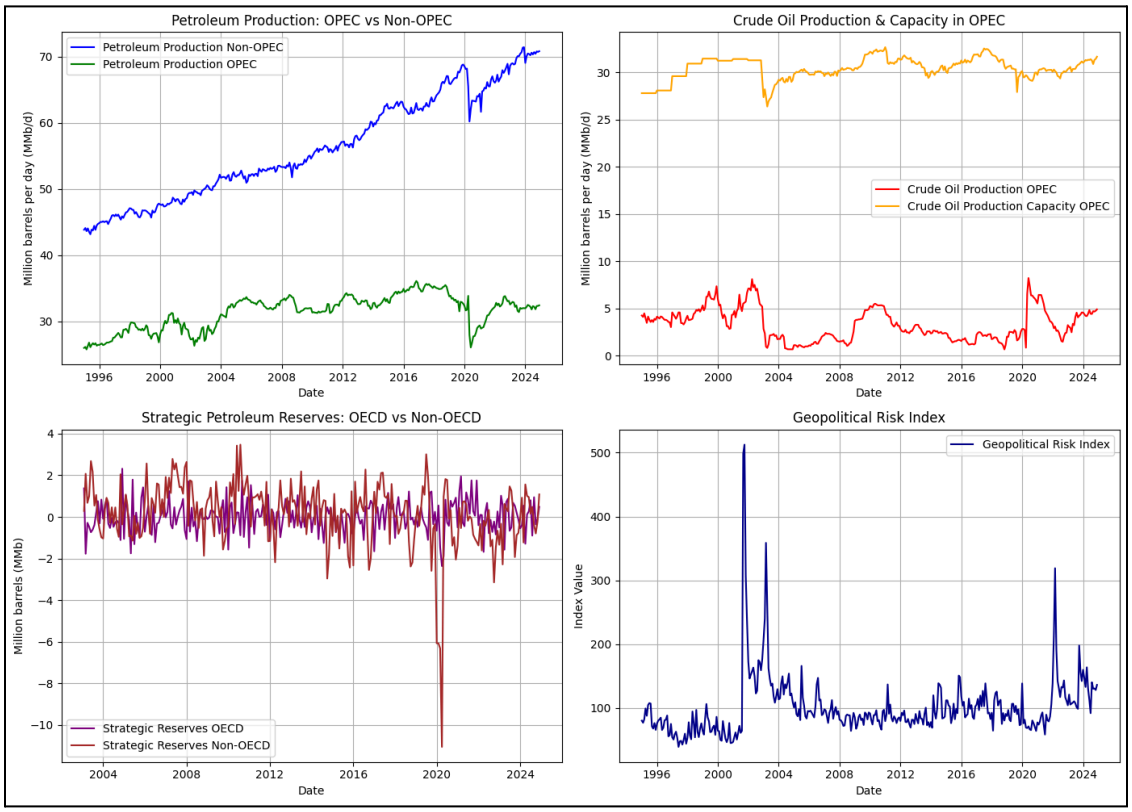


Figure 2 shows the geopolitical dynamics related to oil from 1995 to 2024. OPEC countries' production fluctuates with geopolitical tensions while production capacity largely exceeds actual production. In addition, strategic reserves of non-OECD countries vary greatly as they are impacted by global financial and health crises. The Geopolitical Risk Index (GPR) peaks during major events such as the Russo-Ukrainian war.

c. Graphing microeconomic data

Figure 3: Time series of original microeconomic data

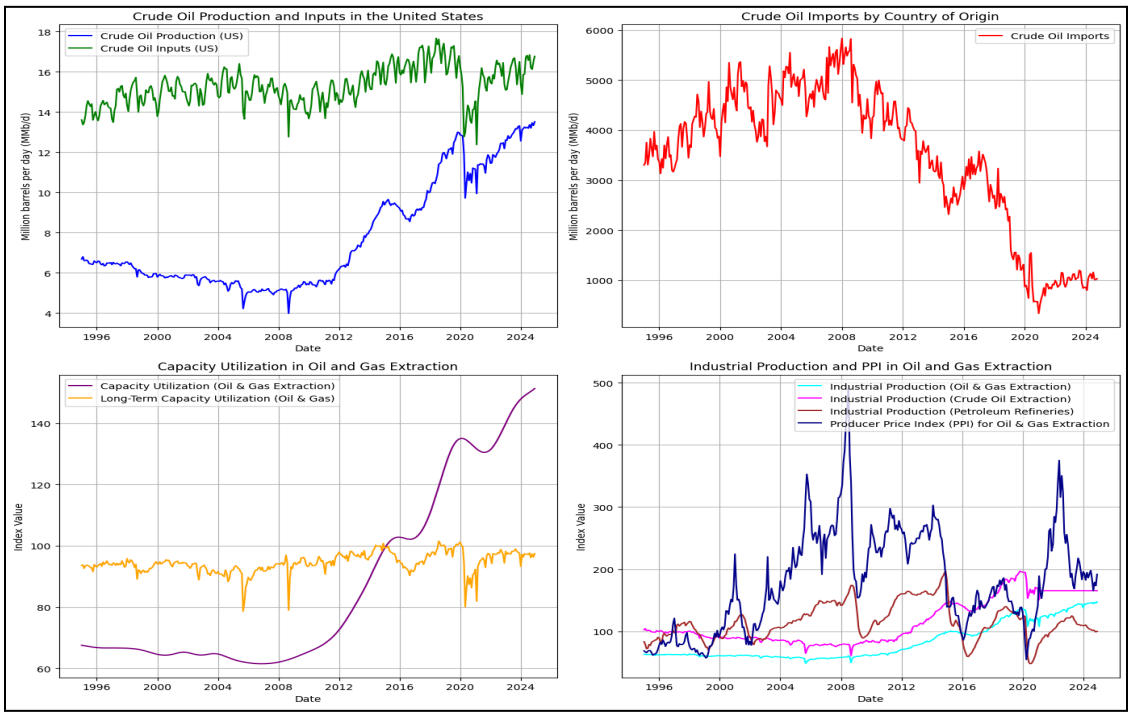
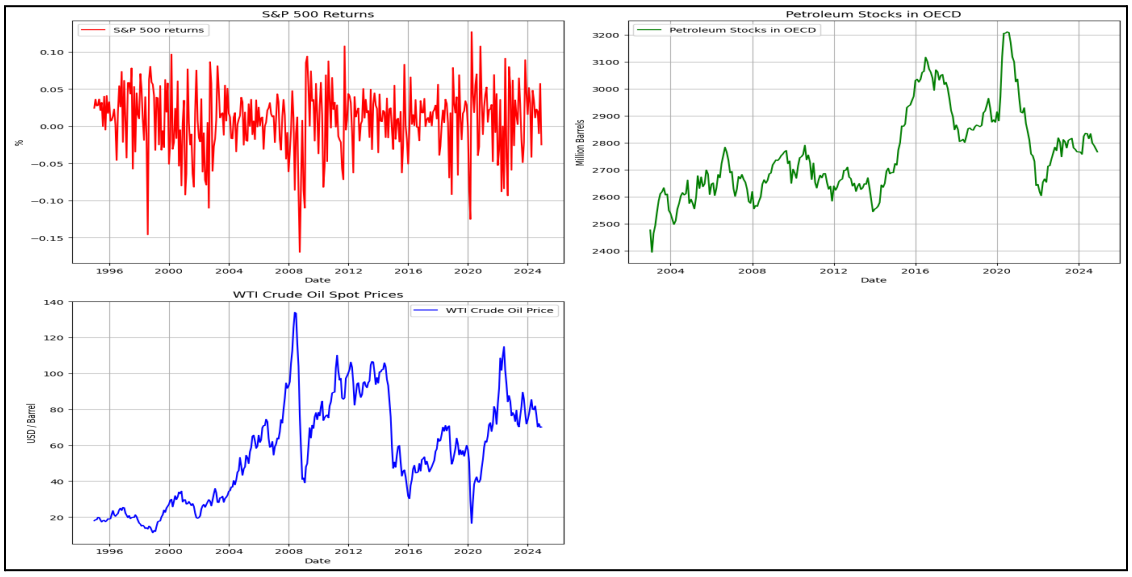


Figure 3 shows microeconomic oil data covering the period 1995-2024. We can note that US production and imports vary according to economic cycles. Capacity utilization (oil extraction) shows a decline after the 2008 crisis and a recovery after COVID. Industrial production indices reflect supply shocks (e.g. 2020 crisis) and increasing demand from refineries.

d. Graphing financial data

Figure 4: Time series of original financial data



Financial data (1995-2024) are shown in Figure 4. S&P500 returns show high volatility during the 2008 and COVID19 crises. The variation in the price of WTI crude oil indicates the presence of peaks and crashes. S&P 500 returns show increased volatility during crises (2008, COVID). WTI oil price experiences peaks (2008, 2022) and crashes (2020). OECD oil stocks track supply shocks.

Step 4: Data dictionary and Data table

a. Data Dictionary

The data dictionary is presented in Table 1. The macroeconomic, geopolitical, microeconomic and financial variables are presented in this table.

Table 1: Data dictionary

Type	Sub-Type	Dataset ID	Description
Macroeconomic	GDP Growth	OECD.GDP.GROWTH	Real GDP for OECD countries (quarterly, monthly data).
	GDP Growth	Non-OECD.GDP.GROWTH	Real GDP for non-OECD countries (quarterly, monthly data)
	Petroleum Consumption	STEO.PATC_OECD.M	Petroleum consumption in OECD countries
	Petroleum Consumption	STEO.PATC_NON_OECD.M	Petroleum consumption in non-OECD countries
	Forex Rates	exchange_rate	Global foreign exchange rates (USD/EUR), reflecting macroeconomic stability
	Inflation	CPIENGSL	Consumer Price Index (CPI) for energy, reflecting inflation in energy prices
	Industrial Production	INDPRO	Industrial Production Index, measuring output in the industrial sector
Geopolitical	Petroleum Production	STEO.PAPR_NONOPEC.M	Petroleum production in non-OPEC countries
	Petroleum Production	STEO.PAPR_OPEC.M	Petroleum production in OPEC countries
	Crude Oil Production	STEO.COPS_OPEC.M	Crude oil production in OPEC countries

Type	Sub-Type	Dataset ID	Description
	Crude Oil Production Capacity	STEO.COPC_OPEC.M	Crude oil production capacity in OPEC countries
	Strategic Reserves	STEO.T3_STCHANGE_OOECD.M	Strategic petroleum reserve changes in OECD countries
	Strategic Reserves	STEO.T3_STCHANGE_NOECD.M	Strategic petroleum reserve changes in non-OECD countries
	Geopolitical Risk	GPR	Geopolitical Risk Index.
Microeconomic	Crude Oil Production	STEO.COPRPUS.M	Crude oil production in the United States
	Crude Oil Inputs	STEO.CORIPUS.M	Crude oil inputs to refineries in the United States
	Crude Oil Imports	PET.MCRIMXX2.M	Crude oil imports by country of origin
	Capacity Utilization	CAPG211S	Capacity utilization in the oil and gas extraction industry
	Capacity Utilization	CAPUTLG211S	Capacity utilization in the oil and gas extraction industry (long-term)
	Industrial Production	IPG211S	Industrial production index for oil and gas extraction
	Industrial Production	IPG211111CN	Industrial production index for crude oil extraction
	Industrial Production	IPN213111N	Industrial production index for petroleum refineries
	Producer Price Index	PCU211211	Producer Price Index (PPI) for oil and gas extraction
Financial	Crude Oil Prices	WTISPLC	West Texas Intermediate (WTI) crude oil spot price, a key financial benchmark
	Petroleum Stocks	STEO.PASC_OECD_T3.M	Petroleum stocks in OECD countries, influencing oil prices
	Market Indices	SP500 (^GSPC)	S&P 500 Index, reflecting overall market performance

Type	Sub-Type	Dataset ID	Description
	Forecasted Prices	forecast_oil_price	Forecasted oil price, used for financial planning and analysis

b. Data table

Table 2 shows the data structure with types, their frequency and their sources.

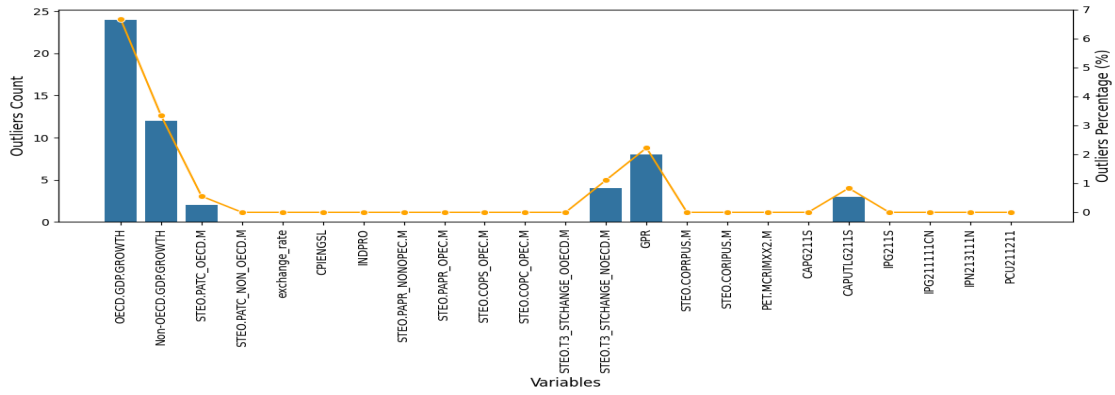
Table 2: Data table

Type	Sub-Type	Dataset ID	Description	Frequency	Source	Start Date	End Date	Unit of Measurement	Usual Range
Macroeconomic	GDP Growth	OECD.GDP.GROWTH	Real GDP for OECD countries	Quarterly	World Bank	01/01/1995	31/12/2024	Percentage (%)	-2% to 5%
	GDP Growth	Non-OECD.GDP.GROWTH	Real GDP for non-OECD countries	Quarterly	World Bank	01/01/1995	31/12/2024	Percentage (%)	-3% to 7%
	Petroleum Consumption	STEO.PATC_OECD.M	Petroleum consumption in OECD countries	Monthly	EIA	01/01/1995	31/12/2024	Million barrels per day (MMb/d)	40-50 MMb/d
	Petroleum Consumption	STEO.PATC_NON_OECD.M	Petroleum consumption in non-OECD countries	Monthly	EIA	01/01/1995	31/12/2024	Million barrels per day (MMb/d)	50-60 MMb/d
	Forex Rates	exchange_rate	Global foreign exchange rates (USD/EUR), reflecting macroeconomic stability	Monthly	ECB	01/01/1999	31/12/2024	USD/EUR	0.8-1.2 USD/EUR
	Inflation	CPIENGSL	Consumer Price Index (CPI) for energy, reflecting inflation in energy prices	Monthly	FRED	01/01/1995	31/12/2024	Index (Base = 100)	80-120
	Industrial Production	INDPRO	Industrial Production Index, measuring output in the industrial sector	Monthly	FRED	01/01/1995	31/12/2024	Index (Base = 100)	80-120
Geopolitical	Petroleum Production	STEO.PAPR_NONOPEC.M	Petroleum production in non-OPEC countries	Monthly	EIA	01/01/1995	31/12/2024	Million barrels per day (MMb/d)	50-60 MMb/d
	Petroleum Production	STEO.PAPR_OPEC.M	Petroleum production in OPEC countries	Monthly	EIA	01/01/1995	31/12/2024	Million barrels per day (MMb/d)	30-40 MMb/d
	Crude Oil Production	STEO.COPS_OPEC.M	Crude oil production in OPEC countries	Monthly	EIA	01/01/1995	31/12/2024	Million barrels per day (MMb/d)	25-35 MMb/d
	Crude Oil Production Capacity	STEO.COPC_OPEC.M	Crude oil production capacity in OPEC countries	Monthly	EIA	01/01/1995	31/12/2024	Million barrels per day (MMb/d)	30-40 MMb/d
	Strategic Reserves	STEO.T3_STCHANGE_OOECD.M	Strategic petroleum reserve changes in OECD countries	Monthly	EIA	01/01/1995	31/12/2024	Million barrels	±10 million
	Strategic Reserves	STEO.T3_STCHANGE_NOECD.M	Strategic petroleum reserve changes in non-OECD countries	Monthly	EIA	01/01/1995	31/12/2024	Million barrels	±5 million
	Geopolitical Risk	GPR	Geopolitical Risk Index	Monthly	EPU	01/01/1995	31/12/2024	Index	50-150
Microeconomic	Crude Oil Production	STEO.COPRPUS.M	Crude oil production in the United States	Monthly	EIA	01/01/1995	31/12/2024	Million barrels per day (MMb/d)	10-15 MMb/d
	Crude Oil Inputs	STEO.CORIPUS.M	Crude oil inputs to refineries in the United States	Monthly	EIA	01/01/1995	31/12/2024	Million barrels per day (MMb/d)	15-20 MMb/d
	Crude Oil Imports	PET.MCRIMXX2.M	Crude oil imports by country of origin	Monthly	EIA	01/01/1995	31/12/2024	Thousand barrels	200-300 thousand
	Capacity Utilization	CAPG211S	Capacity utilization in the oil and gas extraction industry	Monthly	FRED	01/01/1995	31/12/2024	Percentage (%)	70-90%
	Capacity Utilization	CAPUTLG211S	Capacity utilization in the oil and gas extraction industry (long-term)	Monthly	FRED	01/01/1995	31/12/2024	Percentage (%)	70-90%
	Industrial Production	IPG211S	Industrial production index for oil and gas extraction	Monthly	FRED	01/01/1995	31/12/2024	Index (Base = 100)	80-120
	Industrial Production	IPG211111CN	Industrial production index for crude oil extraction	Monthly	FRED	01/01/1995	31/12/2024	Index (Base = 100)	80-120
	Industrial Production	IPN213111N	Industrial production index for petroleum refineries	Monthly	FRED	01/01/1995	31/12/2024	Index (Base = 100)	80-120
	Producer Price Index	PCU211211	Producer Price Index (PPI) for oil and gas extraction	Monthly	FRED	01/01/1995	31/12/2024	Index (Base = 100)	80-120
Financial	Crude Oil Prices	WTISPLC	West Texas Intermediate (WTI) crude oil spot price, a key financial benchmark	Monthly	EIA	01/01/1995	31/12/2024	USD per barrel	20-120 USD
	Petroleum Stocks	STEO.PASC_OECD_T3.M	Petroleum stocks in OECD countries, influencing oil prices	Monthly	EIA	01/01/1995	31/12/2024	Million barrels	2500-3000 million
	Market Indices	SP500 returns	S&P 500 Index, reflecting overall market performance	Monthly	Yahoo	01/01/1995	31/12/2024	Percentage (%)	-10% to 10%
	Forecasted Prices	forecast_oil_price	Forecasted oil price, used for financial planning and analysis	Monthly	EIA	01/01/1995	31/12/2024	USD per barrel	20-120 USD

Step 5: Cleaning data

a. Extreme Outliers

Figure 5: Number and Percentage of extreme outliers for each variable



Extreme outliers were identified via the interquartile range (IQR) method with a strict threshold (boundaries: Q1 - 3×IQR and Q3 + 3×IQR), Figure 5. The approach revealed extreme outliers mostly explain by specific type of crisis:

- Geopolitical Risk Index (GPR): peak in 2001 and 2022 reflecting geopolitical tensions;
- Non-OECD Gross Domestic Product (Non-OECD.GDP.GROWTH); troughs in 2008 and 2020 marking recessions;
- Strategic Petroleum Reserve Changes (STEO.T3_STCHANGE_NOECD.M): COVID-19 related variations;
- Capacity utilization in the oil and gas extraction industry (long-term) (CAPUTLG211S) with post-crisis falls of 2008; These outliers, uncorrected, capture real events (crises, wars), considered essential for the analysis of oil shocks.

The numerical summary of the data with the main measures of central tendency, dispersion and range is presented in Table3.

Table 3: Summary of data

	count	mean	std	min	25%	50%	75%	max
OECD.GDP.GROWTH	360.0	2.119563	1.837880	-3.897883	1.820987	2.427798	3.032759	5.921184
Non-OECD.GDP.GROWTH	360.0	2.718742	2.443459	-6.594182	1.942653	2.754989	4.182768	6.601726
STEO.PATC_OECD.M	360.0	47.247891	2.215432	35.059801	45.935597	47.244754	48.490546	52.875347
STEO.PATC_NON_OECD.M	360.0	40.443306	10.496034	24.116678	29.934819	40.910417	50.362398	58.213118
exchange_rate	312.0	1.184252	0.155141	0.853167	1.083932	1.174693	1.303195	1.576970
CPIENGSL	360.0	190.944819	58.228942	99.200000	130.025000	200.828000	236.188000	331.738000
INDPRO	360.0	94.451548	8.208665	71.153800	90.399700	97.107250	101.217325	104.103800
STEO.PAPR_NONOPEC.M	360.0	56.245322	7.901313	43.167173	49.527025	55.009761	62.919111	71.432773
STEO.PAPR_OPEC.M	360.0	31.389250	2.550202	25.763118	29.158556	32.030459	33.143192	36.095361
STEO.COPS_OPEC.M	360.0	3.282241	1.683560	0.660000	1.982522	2.933800	4.427500	8.220000
STEO.COPC_OPEC.M	360.0	30.421498	1.176779	26.383232	29.795229	30.475500	31.278895	32.666521
STEO.T3_STCHANGE_OOECD.M	263.0	0.009907	0.757680	-2.365900	-0.481332	0.038710	0.485772	2.317742
STEO.T3_STCHANGE_NOECD.M	263.0	0.154556	1.551727	-11.062917	-0.534475	0.355280	1.004795	3.472694
GPR	360.0	100.600402	49.320375	39.045624	76.338326	89.767509	111.934097	512.529724
STEO.COPRPUS.M	360.0	7.713102	2.704763	3.973586	5.599746	6.426183	9.521985	13.510967
STEO.CORIPUS.M	360.0	15.287523	0.936090	12.370929	14.614784	15.267032	15.864813	17.665667
PET.MCRIMXX2.M	358.0	3479.382682	1437.693574	336.000000	2674.000000	3825.500000	4579.250000	5836.000000
CAPG211S	360.0	86.958046	29.438242	61.482200	64.567075	66.952650	104.748675	151.332400
CAPUTLG211S	360.0	94.468506	3.285230	78.636000	92.929725	94.327650	96.906775	101.544100
IPG211S	360.0	82.563805	29.419131	48.814100	60.424300	63.057300	99.940425	147.480400
IPG211111CN	360.0	115.115433	35.962131	60.981300	85.867900	98.715950	146.165325	196.833100
IPN213111N	360.0	115.105345	28.809123	47.994700	94.232700	111.056100	132.792600	196.388800
PCU211211	360.0	179.260775	80.592148	54.600000	112.050000	171.450000	242.650000	490.400000

b. Bad data

b.1. Wrong and questionable data

The outliers have not been corrected as analysis shows that these data are associated with major crises and reflect systemic risks that can impact the price of crude oil (Table 4).

Table 4: Wrong and questionable data

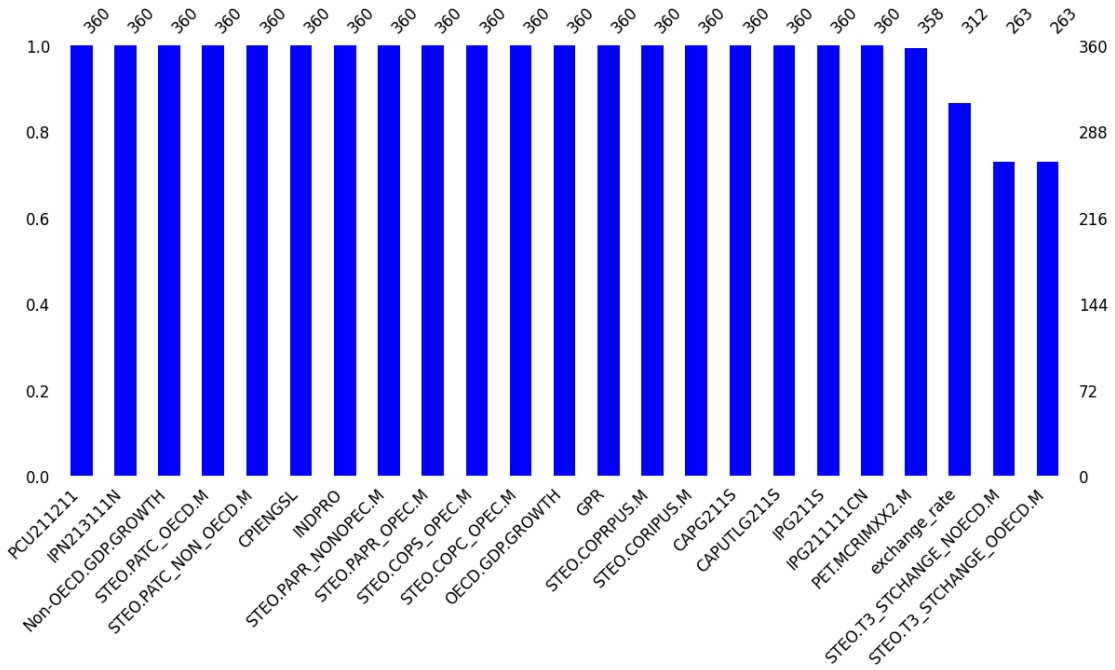
Type	Dataset ID	Description	Potential Wrong Data Issues
Microeconomic	CAPUTLG211S	Capacity utilization in the oil and gas extraction industry (long-term)	**Extreme outliers** Significant drops in capacity utilization due to economic downturns
Geopolitical	GPR	Geopolitical Risk Index	**Extreme outliers** Spikes in risk index due to events like the 2001 US terrorism attack or the 2022 Russo-Ukrainian war
	STEO.T3_STCHANGE_NOECD.M	Strategic petroleum reserve changes in non-OECD countries	**Extreme outliers** Major fluctuations during early COVID-19 pandemic
Macroeconomic	Non-OECD.GDP.GROWTH	Real GDP for non-OECD countries (quarterly, monthly data)	**Extreme outliers** Economic contractions in non-OECD countries due to global downturns.
	OECD.GDP.GROWTH	Real GDP for OECD countries (quarterly, monthly data)	**Extreme outliers** Significant GDP fluctuations due to economic shocks or booms.

b.2. Duplicated data

We found no duplicate data.

b.3. Missing values

Figure 6: Count and proportion of Missing data



Missing data are visualized in Figure 6. Thus:

- Missing USD/EUR exchange rate from 1995-1998.
- Missing OECD/non-OECD strategic reserves (1995-2003).
- Incomplete oil imports (PET.MCRIMXX2.M) at end-2024. The Little test ($p< 0.00001$) confirms non-random missingness, impacting the analysis.

Step 6: Sterilized data

We applied KNN imputation to the covariates. The extreme outliers have not been corrected since these values seem to correspond to periods of crises or major phenomena. We also add crude oil return (oil_price_return) in the dataset. The missing values are now filled.

Step 7: Exploratory Data Analysis

a. Distributional plots

Figure 7: Histograms of variables with Kurtosis and Skewness

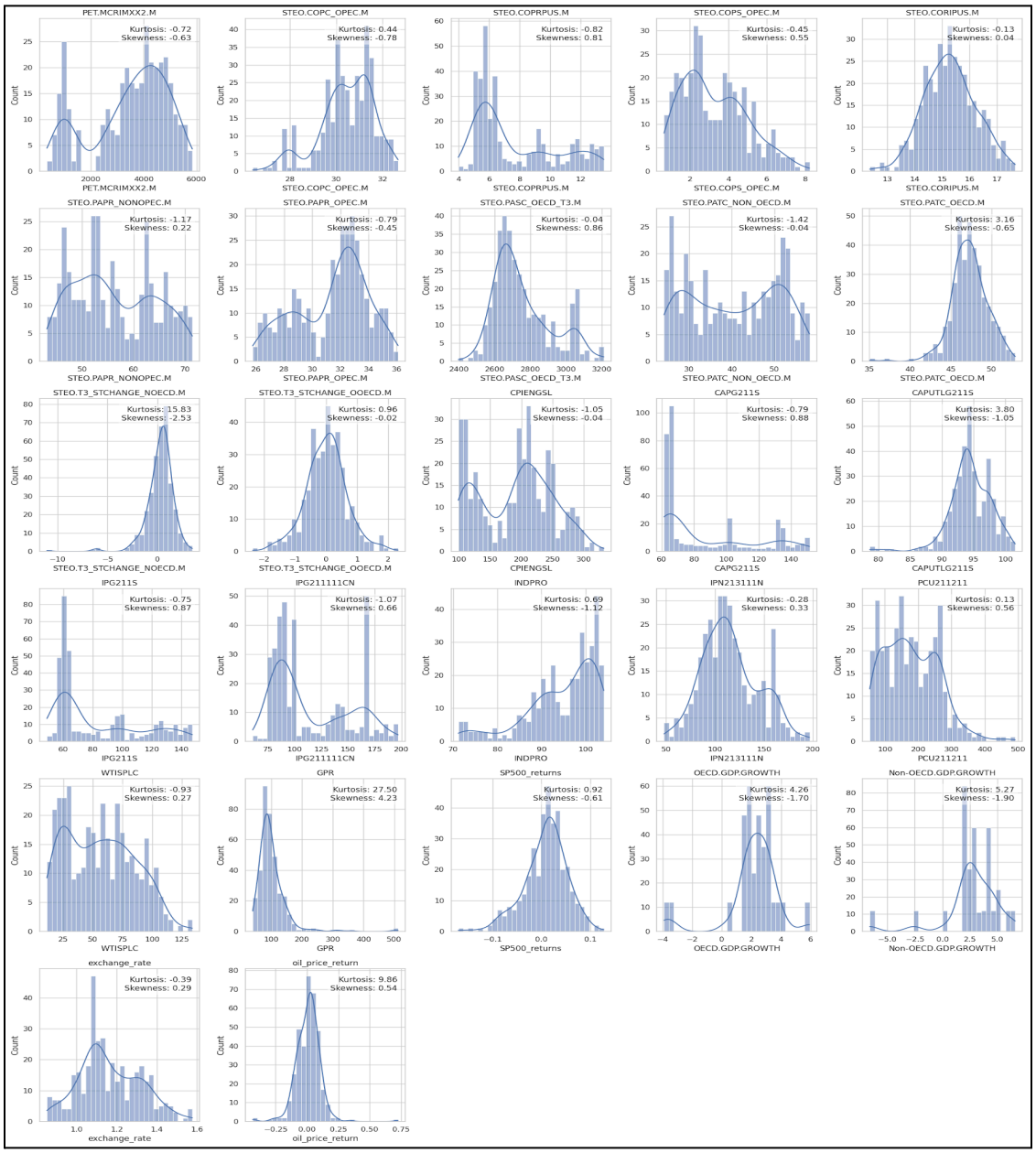


Figure 7 explores the distribution of variables via histograms, enriched with measures of kurtosis and skewness, revealing key insights on oil and economic dynamics. Most variables are skewed and some

have a highly leptokurtic distribution (GRP, GDP). The crude oil price return is leptokurtic with a kurtosis of 9.86.

b. Time series plots

b1. Plotting macroeconomic time series data

Figure 8: Time series of macroeconomic sterilized data

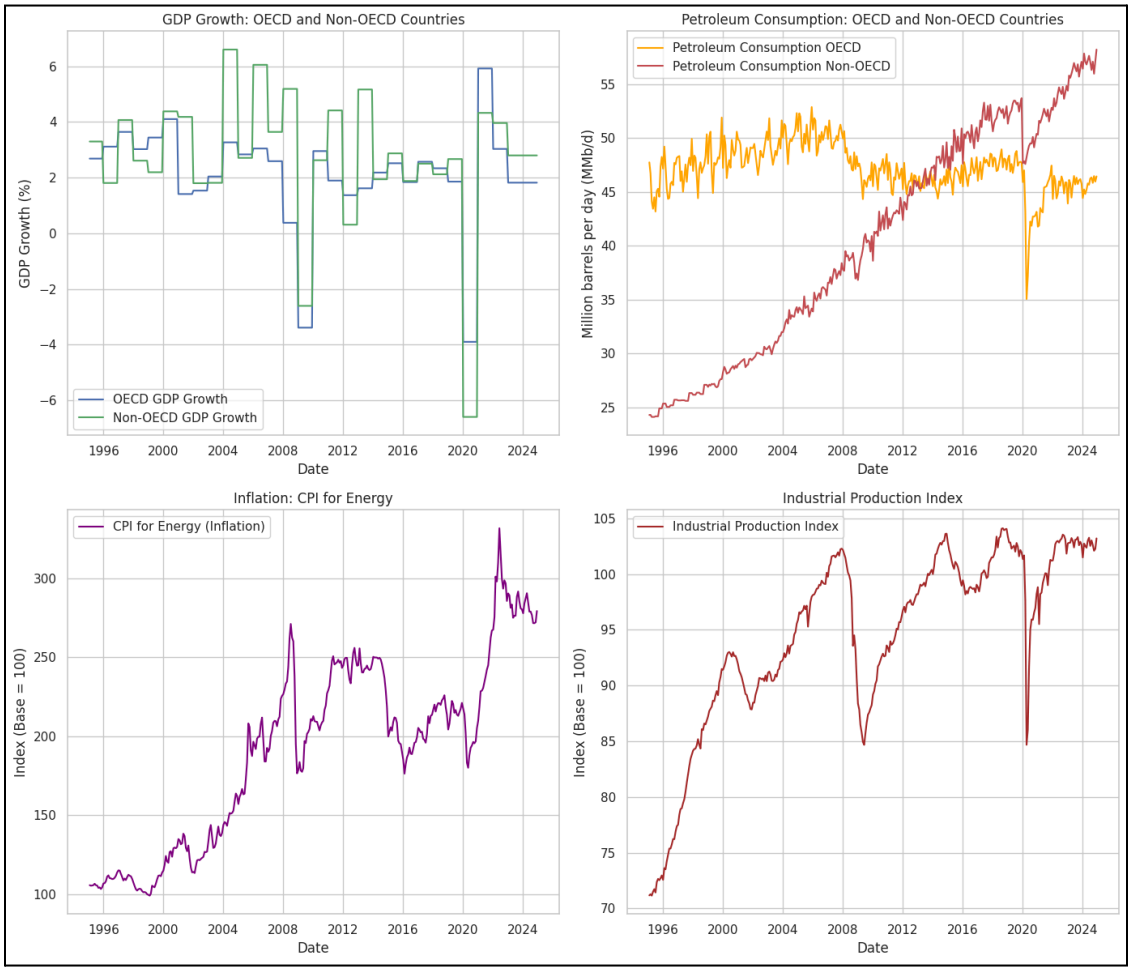


Figure 8 shows the macroeconomic time series with cleaned data from 1995 to 2024. The OECD GDP series has a stable trend despite two troughs in 2008 and 2020 related to global financial and health crises. Volatility is significantly higher for GDP in non-OECD countries. For oil consumption, it increases steadily in non-OECD countries and stagnates in OECD countries. There is also a surge in energy inflation in 2008 and 2022.

b2. Plotting geopolitical time series data

Figure 9: Time series of geopolitical sterilized data

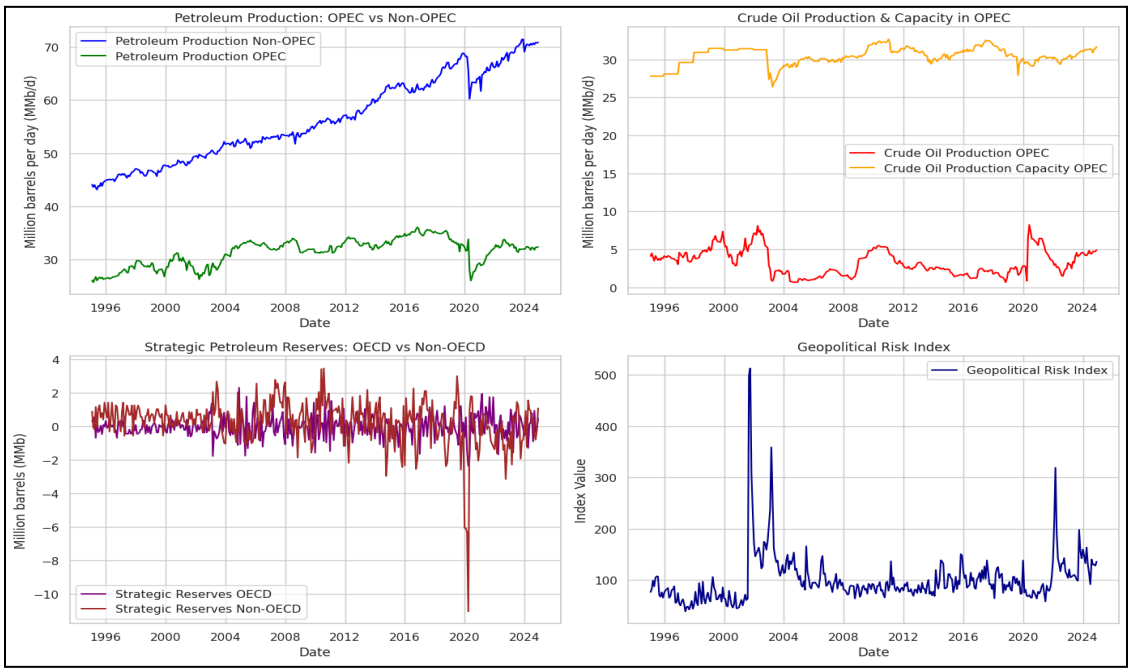
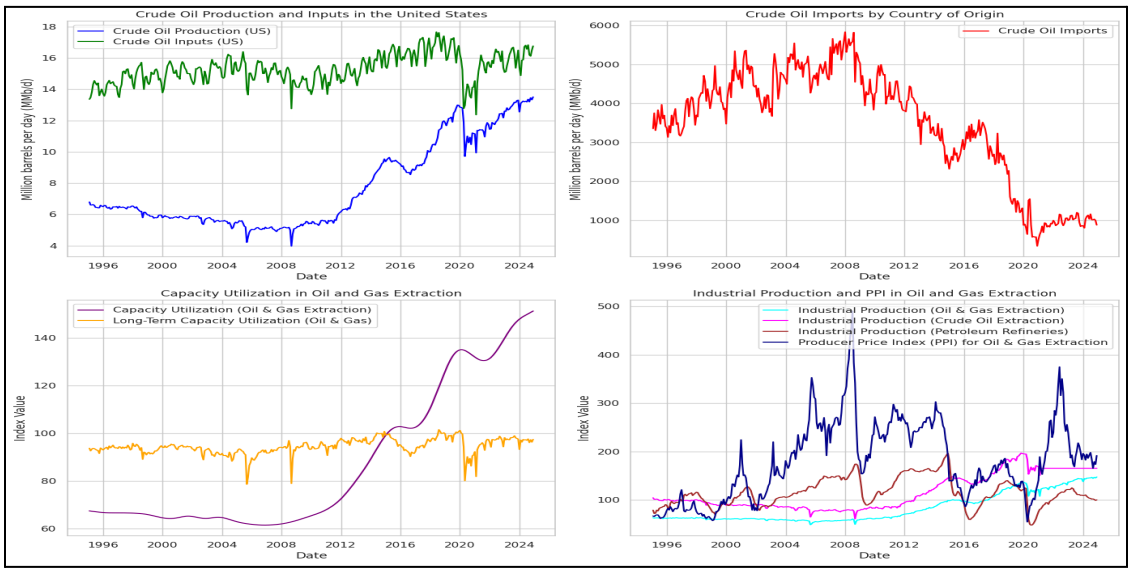


Figure 9 shows the geopolitical dynamics related to oil from 1995 to 2024. OPEC countries' production fluctuates with geopolitical tensions while production capacity largely exceeds actual production. In addition, strategic reserves of non-OECD countries vary greatly as they are impacted by global financial and health crises. The Geopolitical Risk Index (GPR) peaks during major events such as the Russo-Ukrainian war.

b3. Plotting microeconomic time series data

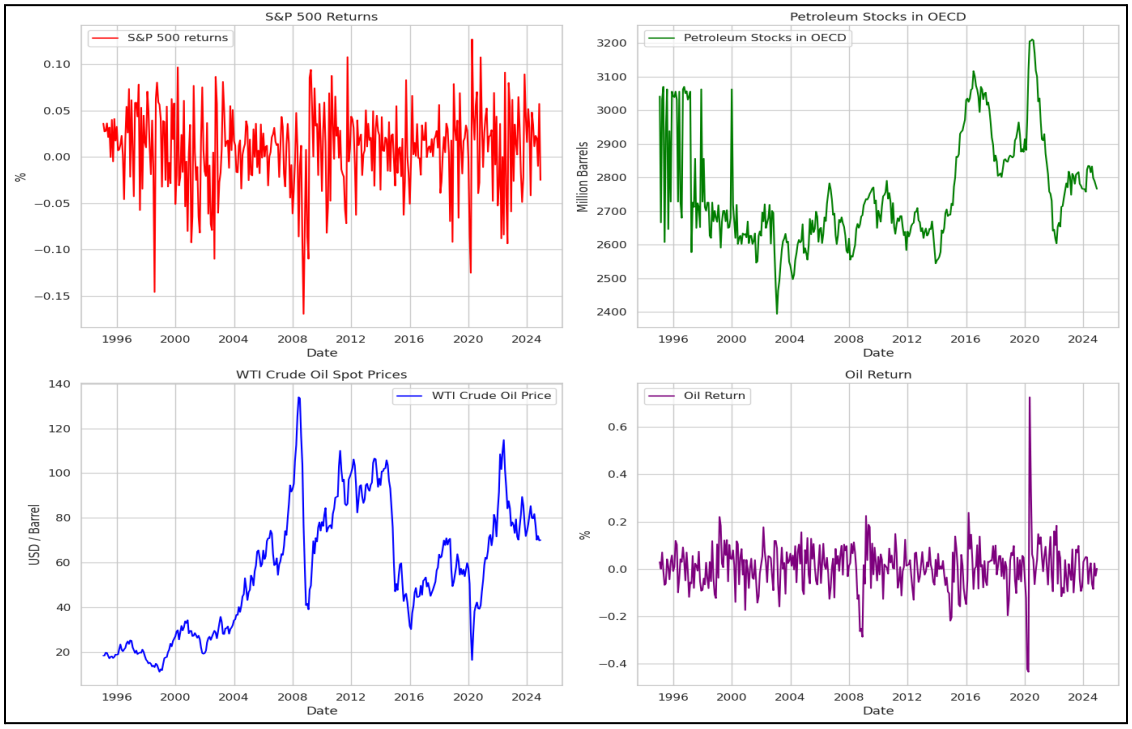
Figure 10: Time series of microeconomic sterilized data



Microeconomic oil data covering the period 1995-2024 are shown in Figure 10. We can note that US production and imports vary according to economic cycles. Capacity utilization (oil extraction) shows a decline after the 2008 crisis and a recovery after COVID. Industrial production indices reflect supply shocks (e.g. 2020 crisis) and increasing demand from refineries.

b4. Plotting financial time series data

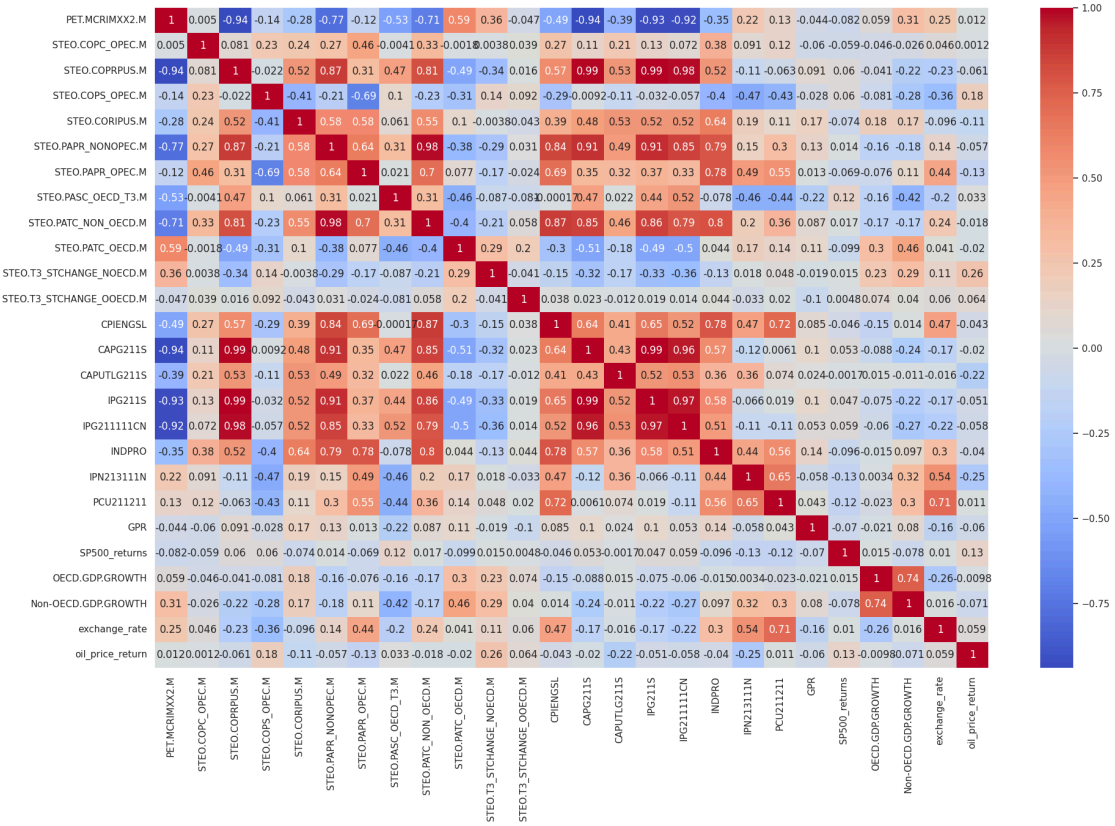
Figure 11: Time series of financial data



Financial data (1995-2024) are shown in Figure 4. S&P500 returns show high volatility during the 2008 and COVID19 crises. The variation in the price of WTI crude oil indicates the presence of peaks and crashes. S&P 500 returns show increased volatility during crises (2008, COVID). WTI oil price experiences peaks (2008, 2022) and crashes (2020). OECD oil stocks track supply shocks. Oil return shows relatively marked volatility with major peaks in 2008 and 2020.

c. Multivariate plots (Heatmap of Correlations)

Figure 12: Correlation Heatmap on sterilized data



The correlation heatmap shows a complex correlation between the variables. Some variables are strongly positively or negatively correlated while others are very weakly or not at all correlated. The crude oil price performance is very weakly correlated or not correlated with almost all the variables. The strongest correlation of this variable is with the change in strategic reserves (r = 0.26). These complex relationships between the variables demonstrate the usefulness of models that can capture this complexity for modeling oil prices.

Step 8: Facts on oil returns

a. Why Oil Prices Stand Out: Spikes, Volatility Clusters, and Seasonality

Oil prices experience spikes (e.g. 2008, 2022), clustered volatility and potentially seasonality not evident in oil price curves and returns. Oil prices are heavily influenced by:

- geopolitical shocks (wars, decisions by oil exporting countries);
- supply-demand imbalances (COVID-19 demand collapse, significant increase in shale production);
- storage constraints (negative prices in 2020);
- The series of oil returns is stationary according to Augmented Dickey-Fuller (ADF) Test; this contrasts with other financial series which are often non-stationary.

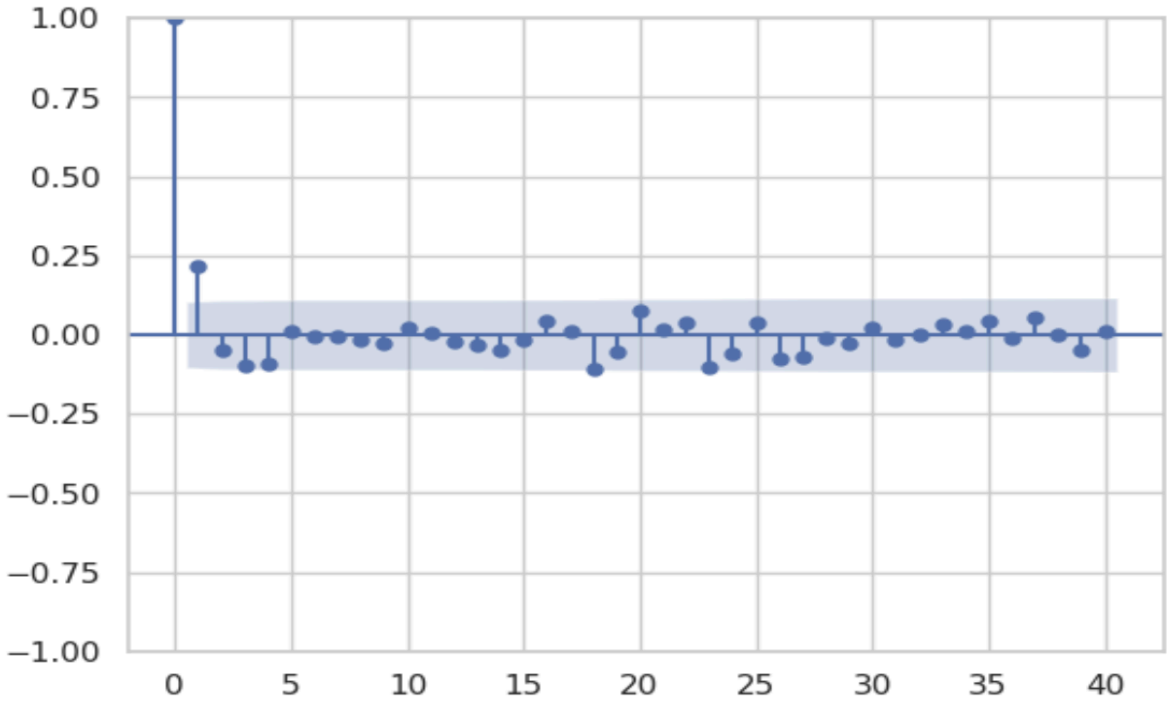
These factors create abrupt and asymmetric reactions absent from most financial assets.

b. Distributions of oil returns

Oil returns follow a non-normal, leptokurtic distribution (kurtosis >3) with slight positive skewness (0.54).

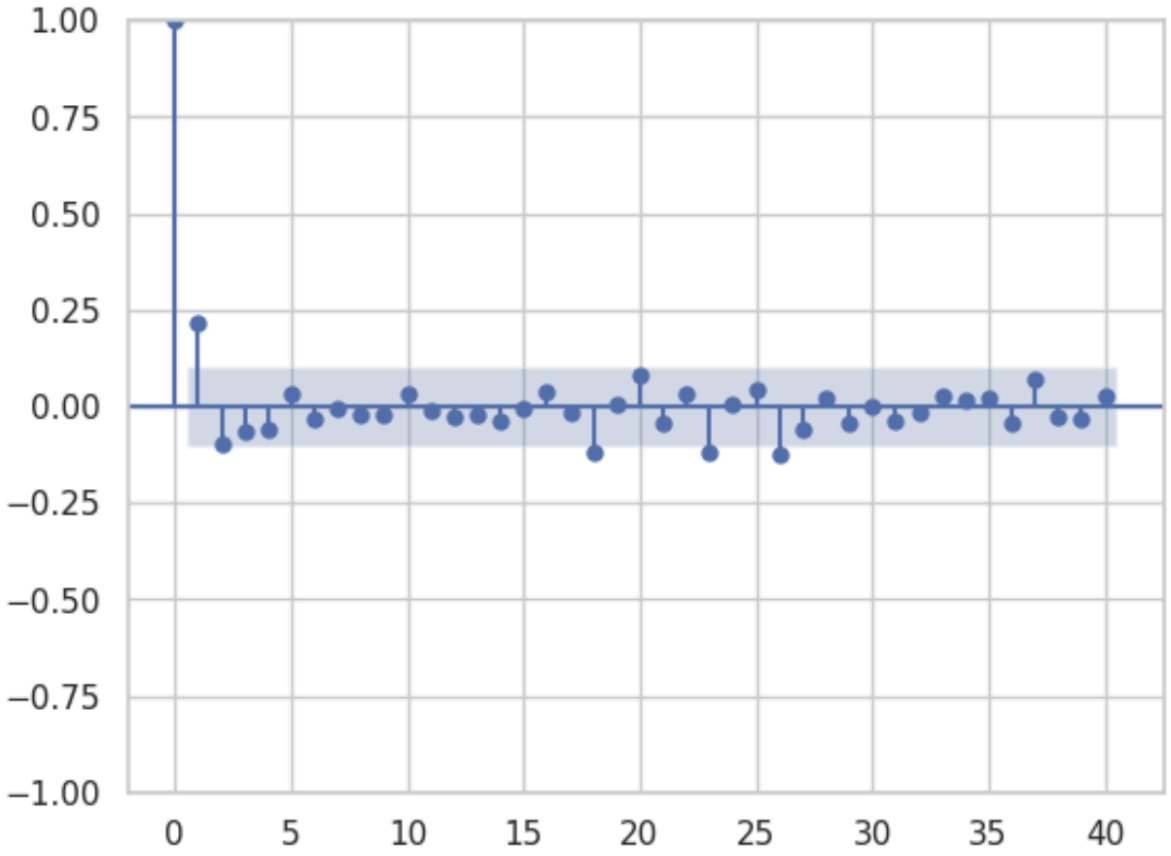
c. Autocorrelation in oil returns

Figure 13: Autocorrelation Function (ACF) of Oil Returns



The lag at lag 1 shows a positive and significant autocorrelation. Beyond lag 1, the autocorrelation values are within the confidence limits. These results indicate some short-run dependence and limited long-run dependence of oil yields. Thus, oil yields appear unpredictable beyond short-run movements. This underlines the stochastic nature of oil market dynamics (Figure 13).

Figure 14: Partial Autocorrelation Function (PACF) of Oil Returns



The partial autocorrelation plot also indicates short-term dependence (lag 1) and no long-term dependence (partial autocorrelation within confidence limits beyond lag 1, figure 14.

The Ljung-Box Test indicates autocorrelation until lag 20.

d. Other stylized facts about oil prices

- Volatility increases more after a price drop;
- Macro-geopolitical sensitivity;
- Jumps: discontinuous price movements.

Step 9: Model

a. Student A: Probabilistic Graphical Models (PGMs)(Larrañaga et al.; Pernkopf et al.)

Probabilistic Graphical Models(PGMs) are a framework that comes about as a combination of the probability theory and the graph theory to help in modeling systems that have multiple variables and interdependencies. These models are created to provide compact and intuitive representations of joint probability distributions which enables efficient reasoning and decision making which is effective under

periods of uncertainty. Probabilistic Graphical Models become useful in addressing problems where the relationship between the variables is complex and the data is also incomplete, noisy, or needs to be cleaned.

Probabilistic Graphical Models play an integral role in the field of machine learning, artificial intelligence, and data science. They have the ability to capture dependencies and interdependencies between or among variables which makes them essential tools for tasks like predictive modeling (forecasting demand or diagnosing diseases), decision-making under uncertainty (autonomous systems), and learning relationships from data.

Because of the dual strength of the graphs and the probability, Probabilistic Graphical Models can compactly represent systems while retaining computational efficiency.

Types of Probabilistic Graphical Models

- **Bayesian Networks (Belief Networks):** Here, nodes represent random variables and the edges indicate direct interdependencies which often capture casual relationships. Each node in the Bayesian Network is associated with a conditional probability distribution which specifies how the variable depends on the parent node or the relationship between the variable and the parent node. This Bayesian Network is particularly effective or useful for handling missing data and reasoning of casual relationships. Let's assume we have a medical diagnostic system. In this scenario, the nodes can be represented as the diseases and symptoms, and the edges represent the causal links between them.
- **Markov Networks (Markov Random Fields):** With Markov Networks, nodes represent random variables but the edges represent the symmetrical dependencies between variables. Unlike how the Bayesian network works, the Markov networks do not not impose directionality on relationships which makes them suitable for modeling undirected dependencies such as correlations.

Features of Probabilistic Graphical Models

- Represents Independence: Explicitly, Probabilistic Graphical Models encode independencies among the various variables which helps to reduce the complexity of the probabilistic computations. With the Bayesian Networks, independencies are based on the parent-child relationship as earlier shared. With the Markov Networks, independencies are based on the graph separation.
- Learning Ability: Probabilistic Graphical Models can learn from the data which involves discovering the graph structure that best represents the dependencies in the data or estimating the clique potentials.

Distinction between Belief Network and Markov Networks

- Bayesian Networks use directed graphs, whereas Markov Networks use undirected graphs.
- In contrast to Markov Networks, all Bayesian Networks explicitly model causal relationships between variables.

- Bayesian Networks represent probability explicitly using Conditional Probability Distributions (CPDs), while Markov Networks implicitly represent probability using clique potentials.
- Bayesian Networks use parent-child relationships and d-separation to determine independence. Markov Networks used the important property of graph separation, employing the precise paths between nodes, to rigorously define independence.
- Bayesian Networks excel at modeling cause-and-effect relationships. They are also superb for making decisions under uncertainty. Markov Networks are superb for modeling correlations and they are also well-suited for symmetrical, undirected relationships.

b. Student B: Parameter Learning and Structure Learning(Pernkopf and Bilmes; Wang et al.)

Parameter learning finds the optimal values that govern the relationships within a probabilistic graphical model. In Bayesian Networks, these parameters are fundamentally probabilities; in Markov Networks, they are, instead, meaningful potentials. Observed data will be used to determine appropriate parameter values, thus guaranteeing realistic model behavior.

Steps for learning necessary parameters

- Parameter learning initiates with a large quantity of data, frequently a complete dataset of all observed variables. Sufficient quality and a sufficient quantity of data are both important for reliable parameter estimation.
- Sufficient data availability determines the optimal learning method.

  Complete variable observation is often necessary. Maximum likelihood estimation and Bayesian estimation are frequently used in these cases.

  Severely incomplete data considerably obstructs thorough analysis; therefore, methods such as Expectation-Maximization iteratively estimate missing values and parameters.

- Parameter tuning in the learning process maximizes the likelihood of observed data which ensures that output models reflect the real-world patterns.

An illustration of a Bayesian network

Structure learning would entail figuring out whether the presence of an umbrella is directly influenced by the weather or if both factors have an impact on the wet ground using a dataset that contains variables like "Weather", "Umbrella," and "Wet Ground.".

Key distinctions between Parameter learning and Structure Learning

- Parameter learning adjusts a model's settings, while structure learning determines how its variables relate.
- Parameter learning requires a known structure to estimate probabilities or potentials, while structure learning uses the data to find the model's structure.

- Parameter learning uses Maximum Likelihood Estimation (MLE) and Bayesian Estimation and it also uses the Expectation-Maximization (EM) Algorithm. Structure learning finds all dependencies using constraint-based, score-based, or hybrid methods.
- Detailed parameter learning yields precise probabilities or potentials that definitively define model behavior; thorough structure learning defines the graph's topology providing thorough connections between the variables.
- In parameter learning, all graph structures are predefined while structure learning methods require no prior structural knowledge.
- Parameter learning faces hardships with overfitting and limited data and structure learning is computationally expensive and susceptible to local optima.

C. Student C: Markov Chains and Markov Blankets(Bruineberg et al.; Liu and Liu)

Markov Chains: This is used to describe the sequence of events where the probability of each event or situation will depend solely on the state of the previous one. So for one event to happen, it will solely depend on the previous one.

In the world of finance, the Markov Chains is used to predict stock price movement and even credit ratings as the previous one can help determine the future one.
Also, in weather forecasting, weather transitions can or are predicted based on historical data or previous data. So for instance in a weather forecast model, if today happens to be sunny, there can be a higher chance that the next day will be sunny in the morning and later a chance of rain late in the day. This can be mapped as a transition matrix which can help in prediction about the future states.

Markov Blankets: This represents a set of nodes in a probabilistic graphical model that renders a specific node conditionally independent from other nodes in the same network. There are components that form the Markov Blanket and they are:

- Parents: theses are nodes that have a direct influence on the node
- Children: these are nodes that are directly influenced by the node
- Co-Parents: these are nodes that share the same children with the selected node

Markov blankets let you efficiently analyze large, complex networks by focusing only on the most important nodes. Grasping the Markov blanket of a node in Bayesian Networks is important because it reveals all and only the factors that affect or are affected by that node.

Markov Chains and Markov Blankets

- Markov Chains precisely describe how all sequential processes evolve over time or space.
- Markov Blankets simplify static networks. They identify the smallest set of nodes that affect a given node, isolating the relevant variables.
- Time-series prediction uses Markov chains and feature selection uses Markov blankets.

Step 10: pseudocode version of Algorithm1

Pseudocode

```
Algorithm InferredCausality(Data):

    Input: Data - A dataset with observed variables

        Output: CausalGraph - A graphical representation of causal relationships



    Step 1: Initialize

     Begin by initializing several parameters. This is the first step in the process.

     An empty graph G should be created; its nodes will represent each variable in Data. For all statistical importance thresholds, set each threshold ε to 0.05.



      Step 2: The second phase concludes with execution of all Markov Blanket Learning procedures.

    Considering each node X within graph G:

        1. Find all variables in the Markov blanket of X.

           - Include Parents: Variables that directly influence X

           - Include Children: Variables directly influenced by X

           - Include Co-Parents: Variables that share children with X

         2. Determine subject dependencies and perform strict statistical tests such as conditional independence tests.

        3. Add edges to G between X and every node in X's Markov blanket.



    Step 3: Learning parameters is the third step

     Consider each edge (X, Y) in graph G and this is relevant to the analysis.:

        1. Approximate P(Y | X) and use this approximation.:

           - Maximum likelihood estimation

           - Bayesian inference
```

2. Considerably update all edge weights using the learned parameters.

Step 4: Several Important Learning Structures

A scoring function evaluates possible structures. BIC is an example of such a function.:

1. Candidate graphs are generated by adding, deleting, or reversing edges.

2. The graph structure possesses the considerably highest score; therefore, retain that structure

3. All Bayesian networks require acyclic structures, while all Markov networks should be looped

Step 5: Use Markov Chains for Sequential Dependencies

When data include sequential information:

1. Creating a transition probability matrix carefully is therefore necessarys

2. Observed sequences provide several important data points. These data allow for the accurate estimation of transition probabilities.

3. Sequentially update each of the graph's relationships.

Step 6: Validate and refine the results thoroughly and carefully.

Strict validation of causal relationships is analytically important and including intervention data, when available, considerably improves this validation:

1. Substantially intervene to explore the cause-effect relationship.

2. Thoroughly test this relationship using do-calculus, for example.Refine graph structure inference considerably using the observations.

Step 7: Output Causal Graph

```
    Return G as the inferred CausalGraph


End Algorithm
```

References

Alvi, Danish A. *Application of Probabilistic Graphical Models in Forecasting Crude Oil Price*. arXiv:1804.10869, arXiv, 29 Apr. 2018. *arXiv.org*, https://doi.org/10.48550/arXiv.1804.10869.

Bruineberg, Jelle, et al. "The Emperor's New Markov Blankets." *Behavioral and Brain Sciences*, vol. 45, Jan. 2022, p. e183. *Cambridge University Press*, https://doi.org/10.1017/S0140525X21002351.

Larrañaga, Pedro, et al. "A Review on Probabilistic Graphical Models in Evolutionary Computation." *Journal of Heuristics*, vol. 18, no. 5, Oct. 2012, pp. 795–819. *Springer Link*, https://doi.org/10.1007/s10732-012-9208-4.

Liu, Xu-Qing, and Xin-Sheng Liu. "Markov Blanket and Markov Boundary of Multiple Variables." *Journal of Machine Learning Research*, vol. 19, no. 43, 2018, pp. 1–50.

Pernkopf, Franz, et al. "Chapter 18 - Introduction to Probabilistic Graphical Models." *Academic Press Library in Signal Processing*, edited by Paulo S. R. Diniz et al., vol. 1, Elsevier, 2014, pp. 989–1064. *ScienceDirect*, https://doi.org/10.1016/B978-0-12-396502-8.00018-8.

Pernkopf, Franz, and Jeff Bilmes. "Discriminative versus Generative Parameter and Structure Learning of Bayesian Network Classifiers." *Proceedings of the 22nd International Conference on Machine Learning*, Association for Computing Machinery, 2005, pp. 657–64. *ACM Digital Library*, https://doi.org/10.1145/1102351.1102434.

Wang, William Yang, et al. "Structure Learning via Parameter Learning." *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, Association for Computing Machinery, 2014, pp. 1199–208. *ACM Digital Library*, https://doi.org/10.1145/2661829.2662022.