Capstone Project: A Causality based trading strategy in European Energy Markets

Julius Gruber

November 2, 2025

Abstract

This thesis explores the integration of causality-based algorithms and quantitative modeling for forecasting returns in commodity markets, with a specific focus on energy and gas sectors. Commodity markets, distinct in their dependency on supply-demand dynamics and external geopolitical factors, present unique challenges for modeling. Using the PCMCI algorithm, alongside traditional econometric and machine learning methods, this work examines predictive relationships in high-dimensional time-series data. A comparative analysis of power and gas markets in the US and Europe highlights the impact of regional policies and infrastructure on pricing mechanisms. Empirical results from backtests demonstrate that causal methods exhibit predictive capabilities for energy markets. The proposed market-neutral trading strategy shows consistent profitability, achieving a Sharpe ratio of 1.7 and a maximum drawdown of 17.5%. These findings underscore the potential of causal methods for forecasting in Power markets.

Contents

Introduction

Commodity markets play a critical role in the global economy, serving as the backbone for industries ranging from energy to agriculture. Unlike traditional financial assets such as equities or bonds, commodities are uniquely influenced by real-world constraints, including supply-demand dynamics, geopolitical tensions, and weather conditions. The intricacies of these markets create both challenges and opportunities for financial modeling and forecasting. Within this context, energy and gas markets are particularly prominent, given their pivotal roles in powering modern economies and their sensitivity to global economic and political shifts.

The quantitative finance community has increasingly turned its attention to commodities, leveraging advanced statistical methods and machine learning algorithms to capture the complexities of these markets. In particular, causal discovery and inference methods have gained traction as tools to disentangle spurious correlations from meaningful relationships. This is especially relevant in power and gas markets, where pricing is affected by regional policies, infrastructure constraints, and interconnected energy sources. For example, the Dutch Title Transfer Facility (TTF) serves as a benchmark for European gas prices, reflecting both local and global influences.

This thesis investigates the application of causal discovery methods to commodity forecasting, focusing on energy and gas markets in the United States and Europe. The primary algorithm explored is the PCMCI (Peter and Clark Momentary Conditional Independence) method, which extends traditional causal discovery techniques to time-series data. The analysis includes a comparison of predictive relationships identified by PCMCI and traditional methods such as Granger causality. Additionally, this work examines the implementation of a market-neutral trading strategy to assess the practical utility of the forecasts.

The study is structured as follows. First, an overview of commodity markets and their relevance to quantitative finance is presented, with a focus on power and gas markets in the US and Europe. Next, the theoretical foundations of causality, including causal discovery and inference methods, are discussed, emphasizing their application to temporal data. Finally, empirical results from backtests on small and large datasets are analyzed to evaluate the performance of causal methods in forecasting commodity returns. The findings highlight both the potential and limitations of causal approaches in this domain, providing insights for future research and practical implementation.

1 Causality

Causality is a foundational concept across multiple disciplines, addressing the relationship between cause and effect. The modern formalism of causality, pioneered by Judea Pearl and collaborators, introduced graphical models, do-calculus, and rigorous criteria for causal reasoning. For a comprehensive introduction, refer to the book by Schölkopf and Peters.
The field of causality can be broadly divided into two major areas:

- *Causal Discovery*: The task of inferring causal structures and relationships from observational data.
- *Causal Inference*: The process of estimating the effects of interventions and analyzing counterfactuals.

This thesis focuses on time-series-based causal discovery algorithms, which aim to integrate causal discovery techniques with temporal data. This is an active area of research. A detailed introduction to this domain can be found in the work by Runge et al. Runge 2022. The primary algorithm investigated in this thesis is the PCMCI algorithm, an extension of the widely-known PC algorithm. Additionally, the Granger causality test and a sequential baseline algorithm are examined.

1.1 Causal Discovery

Causal discovery is the study of determining causal relationships from data under specified assumptions. For an in-depth discussion, readers may consult the book by Peters and Schölkopf TODO: cite, which provides a rigorous overview of the subject.
Causal discovery relies on several key assumptions to ensure the validity of inferred causal relationships. **Causal Markov Condition** asserts that every variable is independent of its non-effects given its direct causes, forming the foundation for defining dependencies within a causal graph. **Faithfulness (or Stability)** ensures that the statistical independencies observed in the data align with the underlying causal structure, avoiding contradictions between data and inferred relationships. The assumption of **No Hidden Confounders** requires that all common causes of the variables in the system are accounted for in the analysis, eliminating biases from unobserved variables. Lastly, for time series data, **Stationarity** assumes that the causal relationships remain invariant over time, allowing consistent modeling across temporal datasets. These assumptions collectively provide the framework for robust and interpretable causal inference.

1.2 Temporal Causal Discovery Algorithms

Extending causal discovery to the time-series setting introduces significant challenges due to the inherent temporal dependencies in the data. Temporal causal discovery methods must account for both contemporaneous and lagged causal relationships. These methods aim to disentangle the direct effects from spurious associations introduced by time.

1.3 Multivariate Granger Causality

Granger causality is a statistical hypothesis test used to determine whether one time series is predictive of another, given past information. The method is grounded in the principle that causation implies predictive capability.
Let X_t and Y_t, where $t \in \mathbb{Z}$ be two stationary time series. Define the information sets $\mathcal{I}t^X$ and $\mathcal{I}t^Y$ as the sets of past values of X and Y, respectively. Granger causality tests the following null hypothesis:

$$H_0 : \mathbb{P}(Y_t | \mathcal{I}t - 1^X, \mathcal{I}t - 1^Y) = \mathbb{P}(Y_t | \mathcal{I}_{t-1}^Y), \tag{1}$$

where $\mathbb{P}$ denotes the conditional probability distribution.
Under H_0, the past values of X do not improve the prediction of Y_t beyond what is already achieved by Y's own history. Rejecting H_0 implies that X Granger-causes Y.
Mathematically, this hypothesis is tested by estimating vector autoregressive (VAR) models:

$$Y_t = \sum_{i=1}^{p} a_i Y_{t-i} + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \sigma^2), \tag{2}$$

$$Y_t = \sum_{i=1}^{p} a_i Y_{t-i} + \sum_{j=1}^{q} b_j X_{t-j} + \varepsilon'_t, \quad \varepsilon'_t \sim \mathcal{N}(0, \sigma'^2). \tag{3}$$

Here, p and q are the lag lengths for Y and X, respectively, and a_i, b_j are model coefficients. The null hypothesis is tested using the F-statistic:

$$F = \frac{(RSS_0 - RSS_1)/q}{RSS_1/(n-p-q-1)}, \tag{4}$$

where RSS_0 and RSS_1 are the residual sum of squares for the restricted (no X lags) and unrestricted models, respectively, and n is the sample size.
If the p-value associated with the F-statistic is below a chosen significance level (e.g., $\alpha = 0.05$), we reject H_0 and conclude that X Granger-causes Y. Granger causality provides a useful framework for temporal causal inference, although it does not capture instantaneous causality or distinguish between direct and indirect effects.

1.4 PCMCI

The PCMCI algorithm (Peter and Clark Momentary Conditional Independence) is a state-of-the-art method for causal discovery in time series data. It combines the strengths of constraint-based causal discovery with rigorous statistical testing to identify causal relationships across temporal lags while controlling for confounders.
PCMCI builds on two key components:

- The PC algorithm, which is a graph-based causal discovery method.
- The Momentary Conditional Independence (MCI) test, which allows for robust statistical testing of temporal dependencies.

1.4.1 The PC Algorithm

The PC algorithm is a systematic approach to causal discovery that begins with a fully connected graph of variables and simplifies it iteratively using conditional independence tests. The key steps are as follows Molak and Jaokar 2023:

1. **Fully Connected Graph:** Start with a graph where all variables are connected.
2. **Unconditional Independence:** Remove edges between variables that are unconditionally independent.
3. **Conditional Independence Tests:** For every pair of variables (A, B), iteratively test conditional independence given a conditioning set C. If $A \perp\!\!\!\perp B \mid C$, the edge between them is removed.
4. **Collider Identification:** Identify colliders in structures like $A \rightarrow B \leftarrow C$ where conditional independence does not hold.

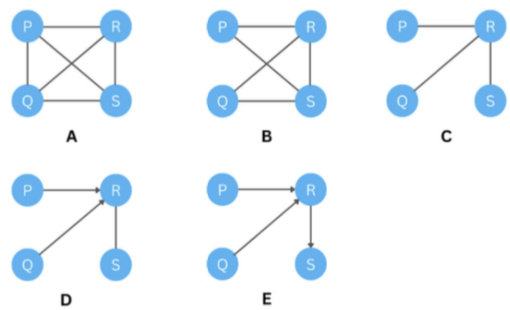


Figure 1: Example of the PC Algorithm, picture taken from Molak and Jaokar 2023

5. **Orientation Rules:** Apply propagation rules to orient remaining edges and ensure a directed acyclic graph (DAG) structure.

Each step systematically refines the causal graph, removing spurious connections and introducing directionality where justified. The result is a partially oriented graph that reflects causal relationships among the variables.

1.4.2 The MCI Algorithm

The MCI (Momentary Conditional Independence) algorithm extends the PC framework to the temporal setting. It incorporates time-lagged variables and adapts the conditional independence tests to account for temporal dependencies. The key features of the MCI algorithm are:

- **Lag-Specific Testing:** Incorporates lagged variables explicitly, testing conditional independence for momentary relationships.
- **False Discovery Rate Control:** Ensures robust inference by controlling for multiple hypothesis testing.

PCMCI integrates these components to efficiently handle large-scale time series datasets. It iteratively refines the causal graph by combining the skeleton identification from the PC algorithm with the temporal specificity of the MCI tests. This makes PCMCI particularly well-suited for high-dimensional time series data with complex causal structures.

PC-MCI Algorithm in Detail

The PC-MCI algorithm consists of two major parts: PC (causal discovery) and MCI (conditional independence testing).

Unconditional Independence

Unconditional independence tests whether $f(x, y) = f(x) \cdot f(y)$. For discrete data, this involves:

1. Calculating the Pearson correlation coefficient:
$$r = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}$$
2. Computing the t-statistic:
$$t = r\sqrt{\frac{n-2}{1-r^2}}$$
3. Determining the p-value using a Student's t-distribution.

Conditional Independence

Conditional independence is defined as $f(x, y|z) = f(x|z) \cdot f(y|z)$. The process involves:

1. Computing Pearson correlations:
   - r_{XY}: Correlation between X and Y
   - r_{XZ}: Correlation between X and Z
   - r_{YZ}: Correlation between Y and Z
2. Applying the partial correlation formula:
$$r_{XY \cdot Z} = \frac{r_{XY} - r_{XZ} r_{YZ}}{\sqrt{(1 - r_{XZ}^2)(1 - r_{YZ}^2)}}$$
3. Computing the t-statistic:
$$t = r_{XY \cdot Z}\sqrt{\frac{n-k-2}{1-r_{XY \cdot Z}^2}}$$
4. Using the t-statistic to compute the p-value.

Types of Conditional Independence Tests

- **Linear:** Straightforward as described above.
- **Kernel:** Maps data to higher dimensions to capture nonlinear causalities. For example, CausalAI from Salesforce implements linear and exponential kernels. More kernels can be added as needed.

The test outputs the test statistic and the corresponding p-value. For more details, refer to: `https://opensource.salesforce.com/causalai/latest/models.common.CI_tests.kci.html`

1.5 Forecasting Models

Forecasting models aim to estimate future outcomes based on a set of predictors. Given predictors $\mathbf{S}_t \in \mathbb{R}^r$, where $r < d$, the target variable Y_{t+1} can be forecasted using a general function $g(\cdot)$ as:
$$Y_{t+1} = g(\mathbf{S}_t) + \epsilon_{t+1},$$
where ϵ_{t+1} represents the residuals. The function $g(\cdot)$ can take various forms depending on the chosen forecasting model.
In this study, the focus is on linear models to evaluate the performance of causal feature selection. Specifically, the target variable is modeled as:
$$Y_{t+1} = \mathbf{S}_t \boldsymbol{\beta} + \epsilon_{t+1},$$
where $\boldsymbol{\beta}$ represents the regression coefficients. These coefficients are estimated using the ordinary least squares (OLS) method. This approach ensures simplicity and interpretability, making it suitable for assessing the contribution of selected predictors.

2 Commodity Markets

Commodities form a crucial segment of financial markets, encompassing raw materials and products. Within the finance space, commodities are unique due to their dependency on physical supply-demand dynamics, geopolitical factors, and weather conditions. Unlike equities or fixed-income securities, commodities are heavily influenced by real-world constraints such as storage, transportation, and production cycles.

2.1 Commodities in the Quantitative Finance Space

Quantitative models applied to commodities often account for mean reversion tendencies, seasonality, and spillover effects across related markets. For example, oil price shocks frequently influence natural gas markets, while agricultural commodities like corn and soybeans display intricate interdependencies. The quantification of risk in commodities markets also involves complex modeling of tail risks due to the occurrence of extreme events such as droughts, geopolitical conflicts, or regulatory changes.

2.2 Power and Gas Markets: US vs. Europe

Power and gas markets are critical sub-sectors within commodities, each exhibiting unique structures and behaviors shaped by regional policies, energy mix, and infrastructure.

2.2.1 The United States

The US power market is characterized by a diverse energy mix, including natural gas, coal, nuclear, renewables (wind, solar, and hydro), and oil. The country's power grid is managed through regional transmission organizations (RTOs) and independent system operators (ISOs), which ensure competitive wholesale electricity markets. Natural gas has been a dominant force, driven by the shale gas revolution, which has led to significant price decreases and its increased adoption for power generation.

The US gas market benefits from extensive pipeline infrastructure and a well-integrated market, allowing for efficient price discovery and arbitrage. The Henry Hub serves as a benchmark for natural gas pricing, with futures and options traded on exchanges such as NYMEX.

2.2.2 Europe

In contrast, the European power market is heavily influenced by regulatory frameworks emphasizing decarbonization and renewable energy integration. The European Union's Emission Trading Scheme (ETS) adds a layer of complexity, as carbon pricing directly impacts power generation costs. Gas markets in Europe are less homogenous, with reliance on imports from regions such as Russia, North Africa, and Norway, leading to vulnerability to geopolitical tensions.

Europe's interconnected grid promotes cross-border trading but also exposes it to issues such as capacity constraints and regulatory disparities among member states. Unlike the US, Europe lacks a single benchmark like Henry Hub, relying instead on regional hubs such as the Dutch TTF for natural gas pricing.

2.3 Geographical Developments in Power Markets

Geographical and policy-driven developments significantly shape the evolution of power markets. One of the most impactful examples is Germany's energy transition (*Energiewende*). The country's decision to phase out nuclear energy by 2022 has increased reliance on renewable sources such as wind and solar while simultaneously heightening dependence on natural gas imports.

This transition has created a challenging balancing act between ensuring energy security, maintaining grid stability, and meeting decarbonization targets.
Other regions exhibit contrasting dynamics. For instance, Nordic countries leverage their abundant hydroelectric resources to maintain low-cost and stable power markets, while countries such as France retain a strong nuclear presence to achieve energy independence and low-carbon goals. In the US, the growth of distributed energy resources (DERs), such as rooftop solar and battery storage, is transforming traditional utility models, enabling consumers to play an active role in electricity markets. Regional disparities in renewable adoption reflect state-level policy differences, highlighting the decentralized nature of the American energy landscape.

2.4 The Importance of LNG for European Energy Markets

Liquefied Natural Gas (LNG) has become a cornerstone of European energy security, particularly in the context of reducing dependence on pipeline gas from geopolitically sensitive regions such as Russia. LNG provides a flexible and scalable alternative, allowing European countries to diversify their energy supply sources. The strategic importance of LNG has been amplified in recent years by disruptions in traditional supply chains and increasing geopolitical tensions.
The United States, and in particular the Texas region, plays a pivotal role in the global LNG market. Texas is home to some of the largest LNG export facilities, such as those located in the Gulf Coast region. The proximity of these facilities to major shale gas basins ensures a steady and cost-effective supply of natural gas, which is processed and liquefied for export. The LNG terminals in Texas, such as Sabine Pass and Corpus Christi, have emerged as critical infrastructure in bridging the gap between US production and European demand.
For Europe, LNG imports from Texas and other US regions have been instrumental in mitigating supply disruptions and balancing the energy mix. The availability of LNG allows for rapid adjustments in supply routes, as tankers can be rerouted to meet demand spikes. European countries have invested heavily in LNG import terminals to accommodate the influx of shipments, with hubs like Spain, the Netherlands, and Germany acting as key entry points.
The geopolitical aspect of LNG cannot be understated. By strengthening ties with the US as a reliable supplier, Europe gains leverage in negotiations with traditional pipeline providers and enhances its resilience against supply shocks. Moreover, LNG facilitates the transition to a cleaner energy system by offering a lower-carbon alternative to coal and oil, particularly during periods of renewable energy intermittency.
In summary, LNG from the Texas region not only supports European energy security but also underpins efforts to achieve greater flexibility and sustainability in the energy system. Its importance is expected to grow as Europe continues its transition toward a low-carbon future.

3 Data Description

The data used was queried from Bloomberg, which consists of daily on-the-run futures. The time series is automatically back-adjusted by Bloomberg to ensure consistency across contract rolls. Additionally, data was sourced from the European Energy Exchange (EEX) in Leipzig. Unfortunately, EEX only provided real price data for costs, and as a result, these data were excluded from backtesting purposes. The data from EEX is generated with an auction process and is not liquidly traded in an orderbook.

3.1 Data Pre-Processing

Since the futures data was already back-adjusted, no further adjustments were necessary. To ensure stationarity, the returns of the time series were calculated.

3.2 Training Procedure

The dataset was divided into training and backtesting subsets. The first three years of data were used for model training, ensuring the model captures long-term dependencies and patterns. The final year was reserved for backtesting to evaluate out-of-sample performance and robustness.

3.3 Performance Evaluation

Model performance was evaluated using three key metrics. Directional accuracy was used to evaluate the model's ability to predict the correct direction of returns, highlighting its predictive capacity. The Sharpe ratio, a classical metric for comparing backtesting performance, was employed to assess risk-adjusted returns, ensuring the model's practical utility for decision-making. Additionally, maximum drawdown was analyzed to measure the largest peak-to-trough decline in the portfolio's value during the backtesting period, providing insight into the strategy's downside risk and resilience. These metrics collectively offer a comprehensive evaluation of the model's effectiveness and robustness.

4 Empirical Results

This section introduces the results gathered throughout the thesis. We follow the steps outlined in the paper by Oliveira et al. Oliveira et al. 2024 from a broader methodological standpoint.

4.1 Structure of the Backtest

The backtest is structured in the following way. The algorithm processes each contract iteratively:

1. Run the causal algorithm on the training set to identify the causing elements.
2. Construct a new dataset with features corresponding to the selected elements from the previous step and the target variable as the contract chosen in the current iteration.
3. Run a linear regression with the causing elements on the in-sample data to train the model weights. Input data is clipped to reduce sensitivity to outliers.
4. Run the linear regression on the out-of-sample data using the weights obtained in the previous step.
5. Compare and contrast the performance of the out-of-sample prediction against the realized values.

4.2 Results

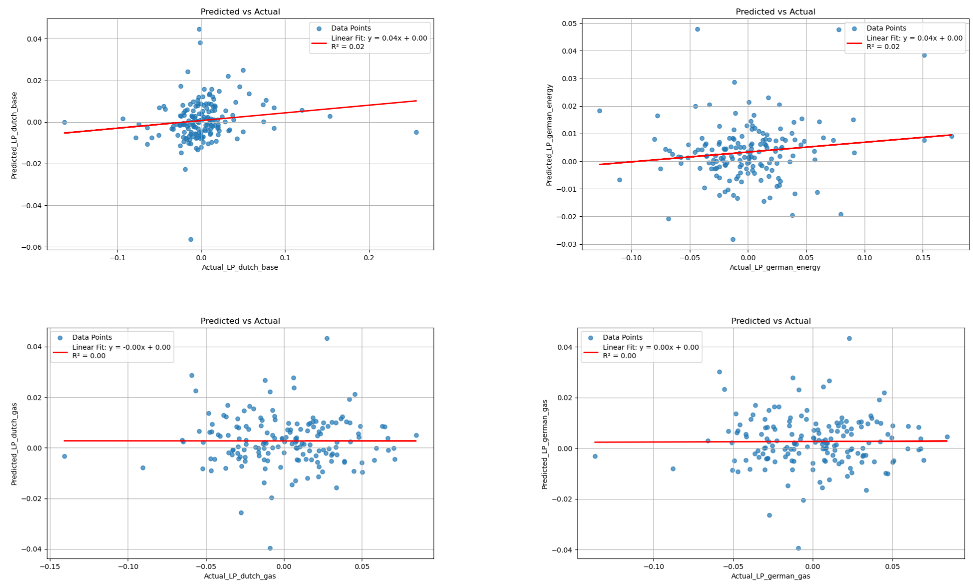
Two backtests were conducted for result gathering: one on a small dataset and one on a larger dataset. The small dataset consists of German Energy, Dutch Energy, Dutch Gas, and German Power futures. These two markets are interconnected, as the Dutch Gas contract (TTF) serves as the baseline index for European gas prices Kristensen 2024. As a next step, we apply the same methodology to a larger dataset to assess whether it improves performance and will look at some backtesting results for it.

4.3 Results on the Small Dataset

After running the described procedure on the data set with German Energy, Dutch Energy, Dutch Gas and German Power futures, the following results were found when comparing predicted and actual return values. Unfortunately, the Granger Causality test did not find any statistically significant pairs.
The following relationship pairs were identified by the PCMCI algorithm:

- **German Energy:** German Energy (-3), Dutch Energy (-5), Dutch Base (-1)
- **Dutch Energy:** German Gas (-2), Dutch Gas (-2)
- **German Gas:** German Energy (-1), German Gas (-3), Dutch Gas (-3), Dutch Gas (-2), Dutch Energy (-5)
- **Dutch Gas:** German Energy (-1), German Gas (-3), Dutch Gas (-2), Dutch Gas (-3), German Gas (-2), Dutch Energy (-5)

In this notation, the emphasized keyword is the variable of interest, while the ones following are the influencer variables with the corresponding lag in parentheses.



The graphs indicate that no predictive performance was found for the gas contracts. This aligns with expectations, as many omitted variables may play a role in the analysis.
Interestingly, a positive R^2 is associated with the energy variables being forecasted. The accuracy rates for these variables were 48%, 59%, 48%, and 47% in the order presented.

4.4 Results on the Bigger Dataset

In this part, I present the results on the larger dataset. The contracts analyzed include German Energy, German Gas, Dutch Energy, Dutch Gas, Polish Energy, Austrian Energy, Czechia Energy, and French Energy.
The PCMCI model identified the following accuracies:

Contract	Accuracy
German Energy	0.5333
German Gas	0.5091
Dutch Energy	0.6061
Dutch Gas	0.5030
Polish Energy	0.4909
Austrian Energy	0.6424
Czechia Energy	0.5394
French Energy	0.5030

Table 1: Aggregated Results for Various Energy and Gas Contracts

The results indicate predictive capabilities for the causal method applied to the energy contracts, although the gas markets remain challenging for the model due to missing variables.

4.4.1 Backtest Methodology

The backtest followed these steps:

1. Obtain predicted and actual values using the steps described earlier.

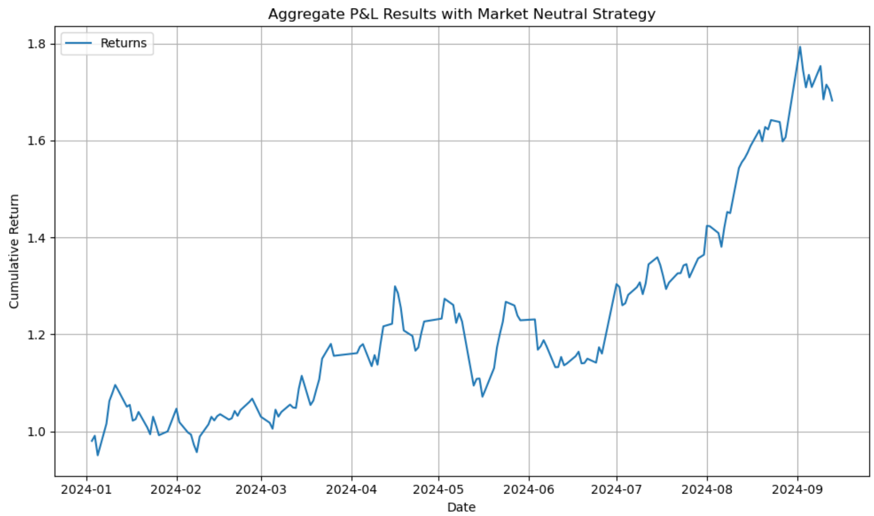


Figure 2: Results for the Market-Neutral Trading Strategy, Including Transaction Costs of 1 Basis Point

2. Enter a long position for the next day if the predicted value is above 0; enter a short position if the predicted value is below 0. In both cases, go long/short a share of the future contract.

3. Normalize weights daily to ensure the portfolio is market-neutral.

The backtest results show that this methodology consistently generates profits. The Sharpe ratio for the portfolio over 165 trading days is 1.7, and the maximum drawdown is 17.5%.

4.5 Final Remarks

The empirical results demonstrate that the causal approach holds promise for certain energy markets, as evidenced by predictive capabilities in German and Dutch Energy contracts. However, the gas markets remain a challenge, likely due to omitted variables and external influences that the current model cannot capture. The market-neutral strategy applied to the backtest further showcases the viability of this methodology, achieving a respectable Sharpe ratio and manageable drawdowns. Future work could focus on enriching the dataset with additional features and external variables to improve predictive performance, particularly for gas markets.

5 Conclusion

This thesis explored the application of causality-based methods for forecasting returns in commodity markets, focusing on energy and gas sectors. Commodity markets are inherently complex, shaped by supply-demand dynamics, geopolitical factors, and infrastructure constraints. A particular emphasis was placed on power and gas markets in the United States and Europe, highlighting key regional differences such as the US's reliance on shale gas and Europe's dependence on LNG and regulatory frameworks like the Emission Trading Scheme.
The study leveraged the PCMCI algorithm to uncover causal relationships in time-series data, comparing its performance against traditional methods such as Granger causality. The empirical analysis included backtests on both small and large datasets of energy and gas futures. The results revealed that the PCMCI algorithm identified predictive relationships in energy contracts, particularly German and Dutch markets, while its performance in gas markets was hindered by missing variables and external factors.
A market-neutral trading strategy was implemented to evaluate the practical utility of the forecasts, achieving a Sharpe ratio of 1.7 and a maximum drawdown of 17.5%. These findings underscore the potential of causal discovery methods for forecasting in commodity markets but also highlight the challenges of modeling gas markets, where external influences are significant. Future research could focus on incorporating additional features and external variables to enhance predictive accuracy, especially in highly volatile and interconnected markets such as gas.

References

Kristensen, Alexander (2024). “TTF Article”. In: *online*. URL: https://www.time2market.dk/blog/why-has-ttf-become-the-european-wholesale-natural-gas-benchmark.

Molak, A. and A. Jaokar (2023). *Causal Inference and Discovery in Python: Unlock the Secrets of Modern Causal Machine Learning with DoWhy, EconML, PyTorch and More*. Packt Publishing, Limited. ISBN: 9781804612989. URL: https://books.google.at/books?id=XNf8zwEACAAJ.

Oliveira, Daniel Cunha et al. (2024). “Causality-Inspired Models for Financial Time Series Forecasting”. In: *arXiv preprint arXiv:2408.09960*.

Runge, Jakob (2022). *Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets*. arXiv: 2003.03685 [stat.ME]. URL: https://arxiv.org/abs/2003.03685.