Statistical arbitrage in multi-pair trading strategy based on graph clustering algorithms in US equities market

University of Warsaw, Faculty of Economic Sciences, Quantitative Finance Research Group

Adam Korniejczuk, Robert Ślepaczuk

October 7, 2024

Working Paper

Korniejczuk A., Ślepaczuk R., 2024, *Statistical arbitrage in multi-pair trading strategy based on graph clustering algorithms in US equities market*, available at arXiv:2406.10695, , https://doi.org/10.48550/arXiv.2406.10695 and available at SSRN: http://ssrn.com/abstract=4849095 or https://.doi.org/10.2139/ssrn.4849095

arXiv

SSRN

Aim

- Develop an effective algorithmic trading strategy.
- Strategy built upon a novel framework involving graph clustering algorithms for statistical arbitrage.
- Study contributes to research on graph theory applications in algorithmic trading.
- Perform detailed performance analysis to assess the strengths and weaknesses of the framework.
- Utilize machine learning classifiers to enhance average trade profitability.
- Test numerous optimizations related to trade execution and risk management

Research questions

- **RQ1: Does the usage of signal quality classifiers improve the quality of the graph-clustering-based strategy?**
- **RQ2: To what extent, does the implementation of the transaction and risk management measures influence the performance of the strategy?**
- **RQ3: What is the sensitivity of the strategy to changes in transaction costs?**
- **RQ4: To what extent does the change of weights in the classifiers ensemble influence the strategy performance?**

Literate Review

- **Historical Context**: Since Markowitz's Modern Portfolio Theory (1952), strategies leveraging relationships between stocks have evolved. Traditional variants of statistical arbitrage and pair trading, have become less effective due to their widespread use.
- **Graph-Based Approaches**: Graph theory has emerged as a promising tool in algorithmic trading. Studies demonstrate that using graphs to model stock relationships (e.g., Zhan et al. 2015, Li et al. 2022) can outperform traditional methods
- **Recent Innovations**: A 2023 study by Cartea et al. introduced a graph clustering approach for statistical arbitrage. Their approach outperformed traditional strategies, achieving higher returns and Sharpe ratios.
- **Machine Learning Integration**: Classification algorithms have proven effective in stock price prediction. However, the use of classifiers to filter trading signals remains relatively unexplored.

Important information about graphs

- Graphs represent relationships between objects, consisting of nodes (vertices) and edges (connections).
- Two types of graphs:
  - Directed Graphs: Ordered pairs (V, E) where edges have a direction.
  - Undirected Graphs: Edges are unordered pairs of vertices.
- Graphs can be:
  - Weighted: Edges have weights indicating the strength of a relationship.
  - Signed: Edges can be positive or negative.
- Vertex degree: Sum of edges connected to a vertex.
- Graphs are often represented with an adjacency matrix, similar to a correlation matrix.
- Graph theory techniques are increasingly applied in investing and algorithmic trading.

2023 study by Cartea et al.

- Cartea, Cucuringu, and Jin (2023) applied graph clustering to statistical arbitrage portfolios.
- US stocks were represented as vertices in an undirected, signed, weighted graph; correlations of residual returns formed edge weights.
- Several clustering algorithms were tested, including:
  - Spectral clustering
  - Signed Laplacian Clustering
  - SPONGE$_{sym}$ (best performer)
- SPONGE$_{sym}$ decomposed the adjacency matrix into positive/negative components and applied generalized eigenvector clustering.
- The SPONGE$_{sym}$ strategy achieved:
  - 12.2% annualized returns
  - Sharpe ratio of 1.1, Sortino ratio of 2.01
- **Strategy used simple mean-return logic without optimization; no transaction costs were considered.**

Data

- Data sourced from Yahoo Finance.
- Historical S&P 500 components used to ensure realistic backtesting.
- Historical components data sourced from GitHub (Farrell 2024).
- Stock universe occasionally below 500 due to data availability.
- Adjusted closing prices used for position opening/closing.

Feature engineering

- Local vertex degree: $\frac{\sum_{n=1}^{S} e_{i,n} - 1}{S-1}$
- Global vertex degree: $\frac{\sum_{n=1}^{G} e_{i,n} - 1}{G-1}$
- Graph density: $\frac{\sum_{i=1}^{S}(\sum_{n=1}^{S} e_{i,n} - 1)}{(S-1)S}$
- Cluster size: $\frac{|v_j|}{G}$
- Number of clusters divided by graph size
- Cumulative return deviation from the cluster (5 days)
- Sign of deviation (long/short)
- Mean cluster returns (10 days)
- Mean stock returns (10 days)

Notes:

- All graph based features are normalized by either cluster size or graph size to ensure comparability.
- This normalization accommodates changes in stock universe size.

Generating training and validation datasets

- Dataset created by backtesting the original strategy with reduced rebalance frequency.
- Only the first 1500 trading days were used (in-sample period).
- In-sample period not used for strategy testing in chapters 4 and 5.
- For each signal, a record is stored with the following:
  - Features described on the previous slide.
  - Binary variables indicating profitability.
- Two conditions for a signal to be profitable:
  - Condition 1: Cumulative returns at the end of trading day since last rebalance $>$ threshold T.
  - Condition 2: Cumulative returns at time of rebalance $>$ transaction costs (no loss incurred).

Classifiers used

Created dataset has been randomly split into two parts - 80% has been used as the training dataset and the remaining 20% as the validation one. 5 classifiers have been trained on the training datasets through grid search cross-validation

- Multi-Layer Perceptron
- Stochastic Gradient Descent Classifier
- Logistic Regression
- Hist Gradient Boosting Classifier
- Ada Boost Clasifier

Classifiers Ensemble

- Performance of the classifiers has been compared on the validation dataset
- Brier score and Precision have been selected as evaluation metrics
- Based on the results of the comparison, soft voting, weighted ensemble has been constructed

Table: Classifiers Performance

Classifier	Brier Score	Precision
MLP	0.243	0.568
ADA boost	0.247	0.544
HistGradientBoosting	**0.218**	**0.653**
SGD	0.247	0.547
Logistic Regression	0.249	0.580

*Source: Own Elaboration.

Comparison of the performance of individual classifiers on validation dataset, created as remaining 20% of the generated dataset.

Kelly criterion and stop functions

- Dynamic Take profit threshold: $Threshold_{tp} = THR \cdot \frac{10 - TD}{10}$
- THR set to 8% to align with classifier training.
- Time-variant stop loss: $Threshold_{sl} = 0.05 \cdot \frac{10 - TD}{10}$
- Reduces risk of extreme losses during market downturns.
- Kelly criterion optimizes position size: $f = \frac{2P - 1}{\lambda}$
- Fractions scaled so long/short sums equal 1, maximizing return under risk.
- Stop loss/take profit thresholds adjusted by signal probability:

$$Threshold_{RW} = Threshold \cdot P \tag{1}$$

- Lower thresholds for risky signals to avoid weak mean reversion losses.

Results

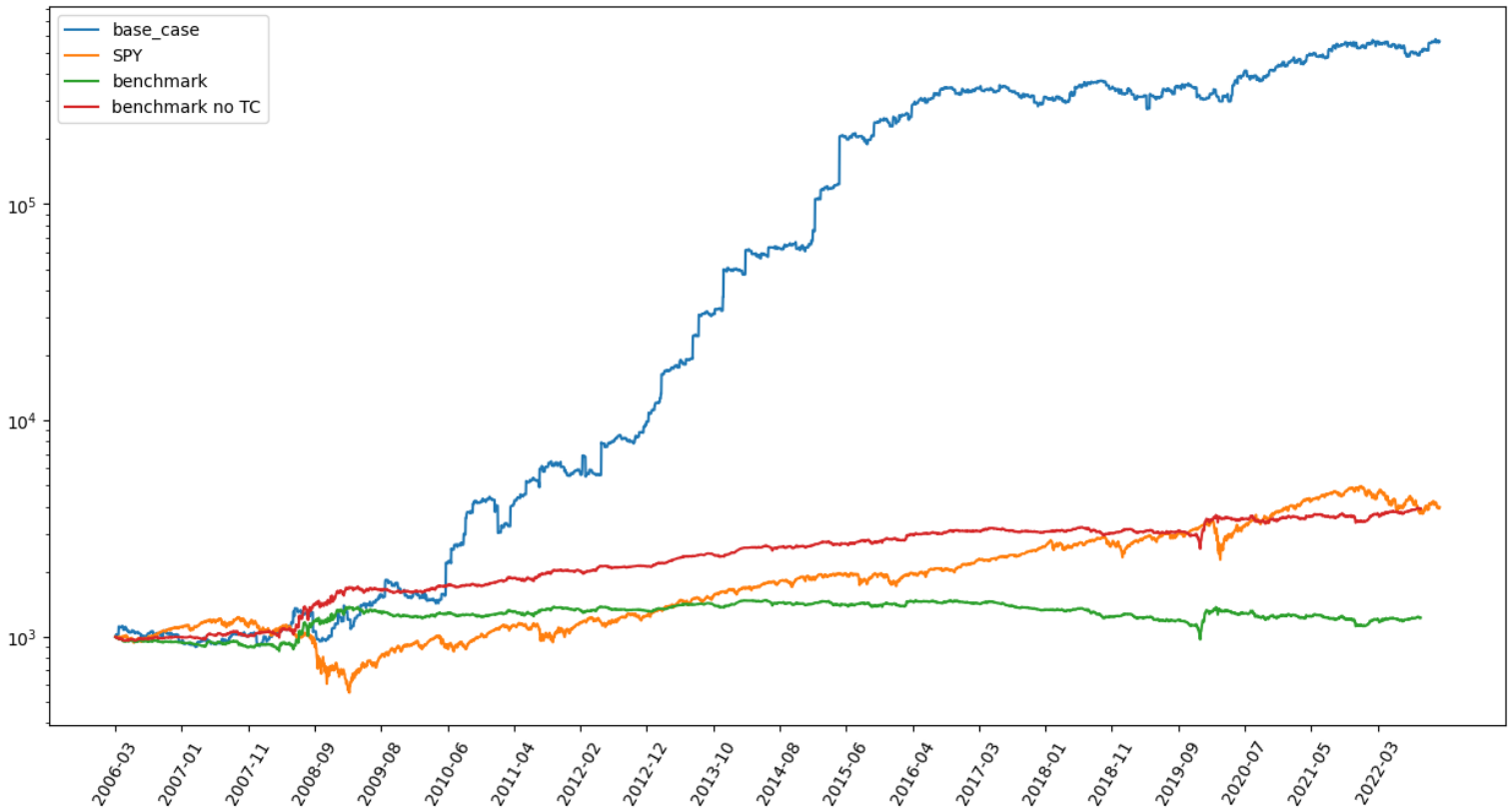
Table: Performance metrics of the proposed strategy and the relevant benchmarks

Metric	Strategy	Benchmark	Benchmark W/o tc	SPY
ARC	**49.33%**	1.13%	9.01%	9.12%
ASD	38.01%	9.16%	**9.14%**	20.11%
IR*	**1.30**	0.14	0.96	0.45
SORTINO	**3.38**	0.29	1.77	0.70
MDD	31.98%	34.30%	**20.68%**	55.19%
MLD	2.10 years	2.59 years	**1.54 years**	4.85 years
CR	**1.54**	0.04	0.44	0.17
IR**	**2.00**	0.01	0.42	0.08

*Source: Own Elaboration. **Strategy**: proposed approach, with weighted ensemble classifier, stop loss and take profit function modifications and Kelly criterion. **Benchmark**: Own implementation of the strategy proposed by Cartea et al. (2023), with 0.05% transaction costs. **Benchmark W/o tc**: Own implementation of the strategy proposed by Cartea et al. (2023), with no transaction costs. **SPY**: SPY ETF. Bolded values indicate the best metric of all presented strategies.

Results

Figure: equity curves of the proposed strategy and the relevant benchmarks



*Source: Own Elaboration. Own backtesting implementation, out-of-sample performance between 03.2006 and 12.2022

Sensitivity analysis

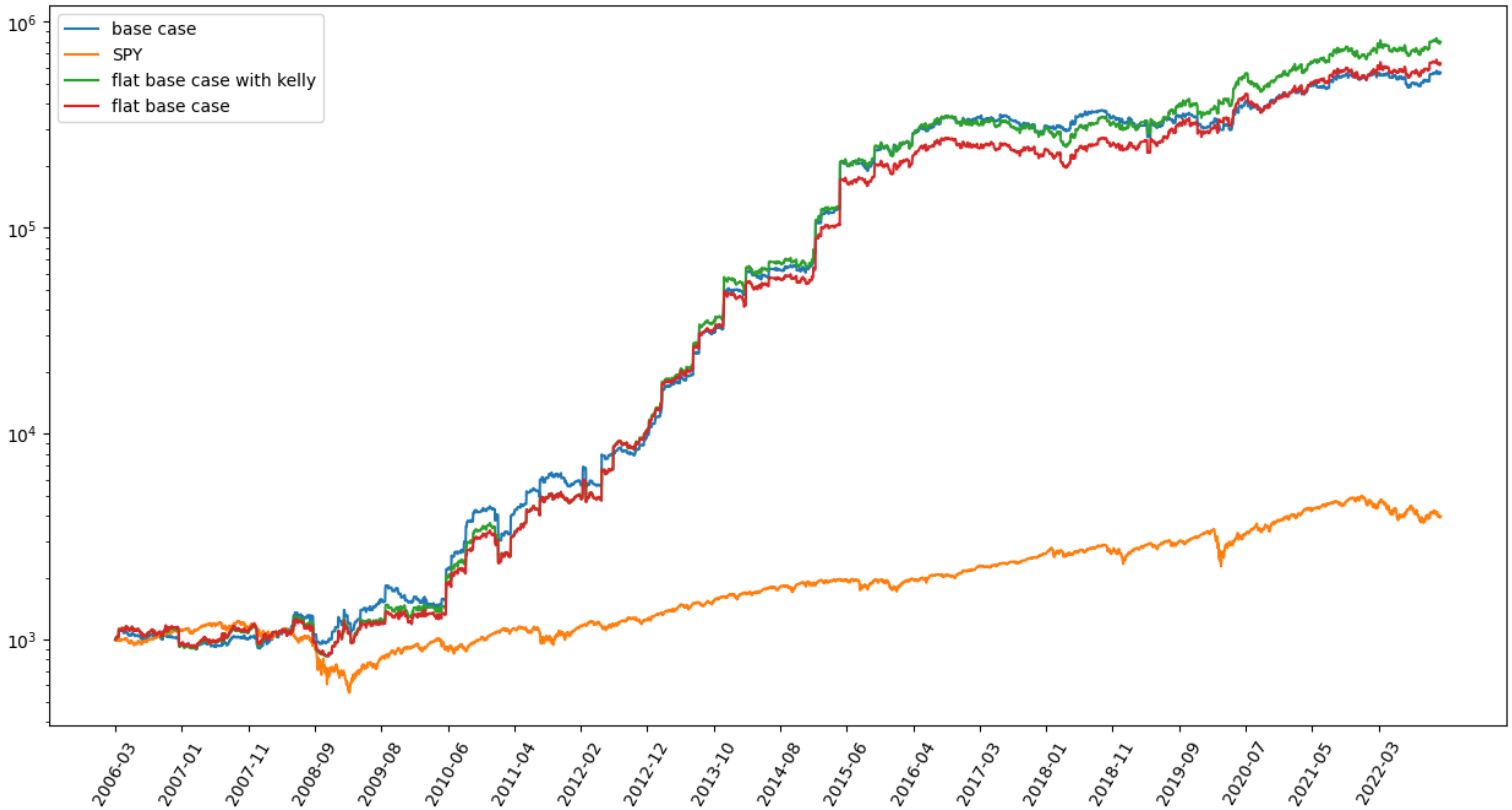
Table: Performance metrics for sensitivity analysis of the stop loss and take profit functions modifications and the Kelly criterion

Metric	Base case	Flat Base Case Kelly	Flat Base Case	SPY
ARC	49.33%	**52.63%**	50.33%	9.12%
ASD	38.01$	40.29%	38.95%	**20.11%**
IR*	1.30	**1.31**	1.29	0.45
SORTINO	**3.38**	3.25	3.26	0.70
MDD	**31.98%**	36.05%	33.94%	55.19%
MLD	2.10 years	2.82 years	**1.99 years**	4.85 years
CR	**1.54**	1.46	1.48	0.17
IR**	**2.00**	1.91	1.90	0.08

*Source: Own Elaboration. **Base Case**: proposed approach, with weighted ensemble classifier, stop loss and take profit function modifications and Kelly criterion. **Flat Base case Kelly**: Base case strategy without time variant stop function modifications. **Flat Base case**: Base Case strategy without time variant takes profit and stop loss functions and Kelly criterion. **SPY**: SPY ETF. Bolded values indicate the best metric of all presented strategies.

Sensitivity analysis

Figure: equity curves for sensitivity analysis of the stop loss and take profit functions modifications and the Kelly criterion



*Source: Own Elaboration. Own backtesting implementation, out-of-sample performance between 03.2006 and 12.2022

Sensitivity analysis

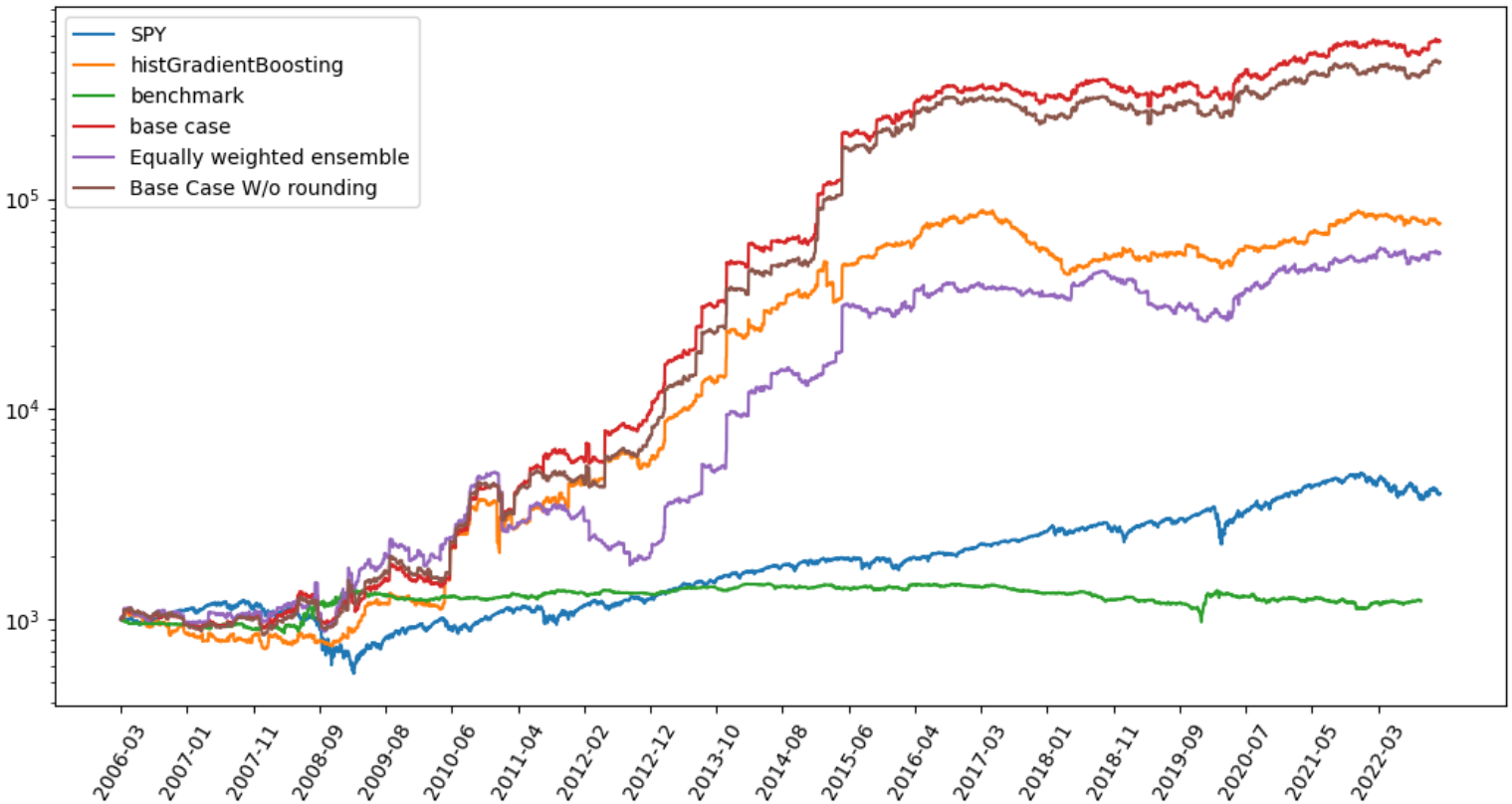
Table: Performance metrics for sensitivity analysis of the ensemble construction

Metric	Base Case	HistGradientBoosting	equally weighted	W/t rounding	SPY
ARC	**49.33%**	31.58%	28.91%	47.15%	9.12%
ASD	38.01$	34.82%	38.22%	38.99%	**20.11%**
IR*	**1.30**	0.91	0.76	1.21	0.45
SORTINO	**3.38**	1.89	1.54	3.11	0.70
MDD	**31.98%**	50.7%	63.95%	38.83%	55.19%
MLD	**2.10 years**	3.01 years	2.63 years	3.22 years	4.85 years
CR	**1.54**	0.62	0.45	1.38	0.17
IR**	**2.00**	0.56	0.34	1.67	0.08

*Source: Own Elaboration. **Base case**: proposed approach, with weighted ensemble classifier, stop loss and take profit function modifications and Kelly criterion. **HistGradientBoosting**: Base case but using only one model rather than an ensemble. **equally weighted**: The base case with equally weighted ensemble. **W/t rounding**: Base case but without rounding the threshold used in determining whether a signal is profitable. **SPY**: SPY ETF. Bolded values indicate the best metric of all presented strategies.

Sensitivity analysis

Figure: equity curves for sensitivity analysis of the ensemble Weights



*Source: Own Elaboration. Own backtesting implementation, out-of-sample performance between 03.2006 and 12.2022.

Sensitivity analysis

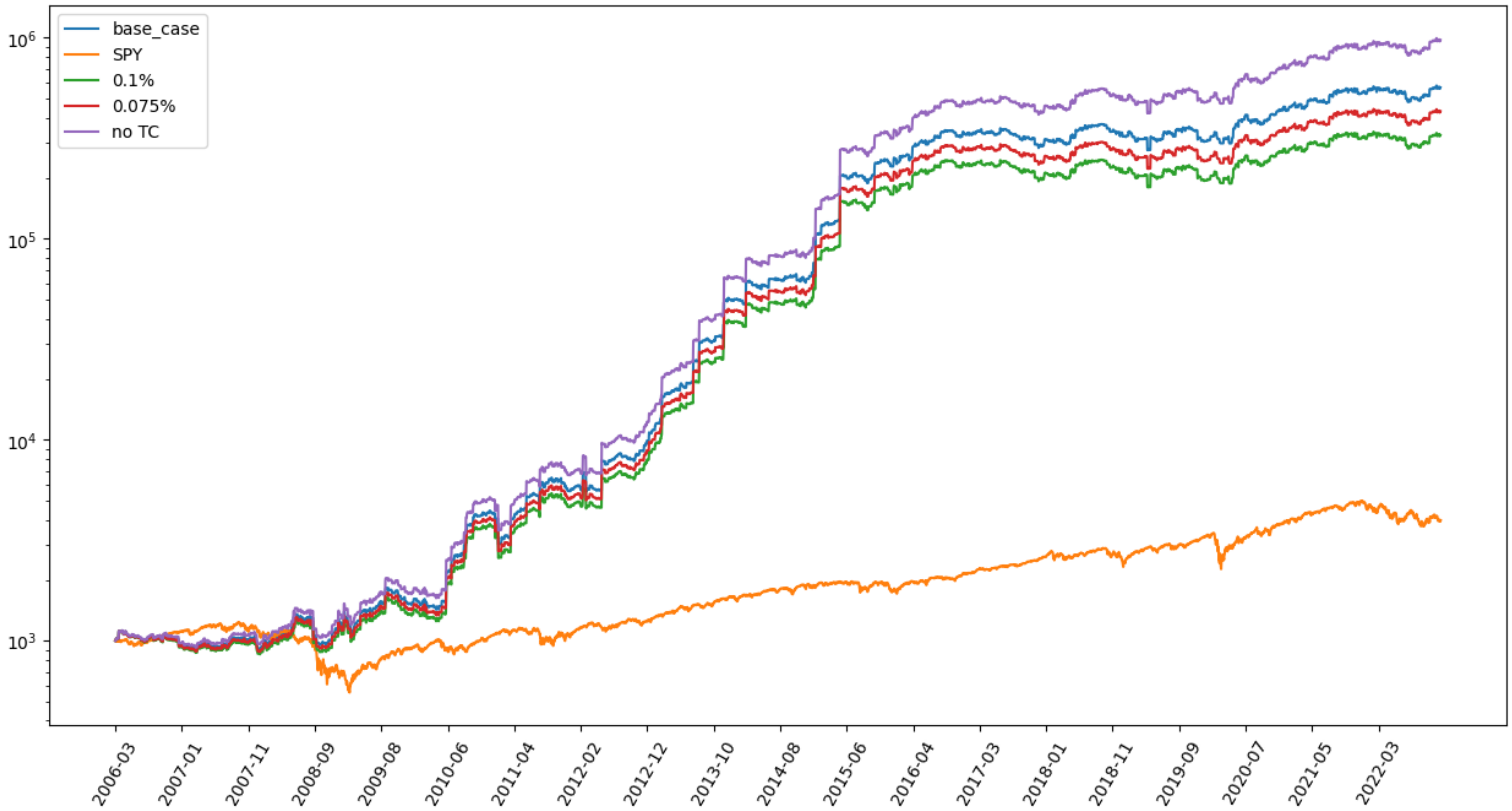
Table: Performance metrics for sensitivity analysis of the transaction cost rate

Metric	0%	Base Case	0.075%	0.1%	SPY
ARC	**54.55%**	49.33%	46.78%	44.27%	9.12%
ASD	38.02%	**38.01%**	38.02%	38.02%	20.11%
IR*	**1.43**	1.30	1.23	1.16	0.45
SORTINO	**3.74**	3.38	3.20	3.027	0.70
MDD	**31.72%**	31.98%	32.11%	32.24%	55.19%
MLD	**1.98 years**	2.10 years	2.14 years	2.17 years	4.85 years
CR	**1.72**	1.54	1.46	1.37	0.17
IR**	**2.46**	2.00	1.80	1.59	0.08

*Source: Own Elaboration. **0%**: Base Case with no transaction costs. **Base case**: proposed approach, with weighted ensemble classifier, stop loss and take profit function modifications and Kelly criterion. **0.075%**: Base Case with 0.075% transaction costs. **0.1%**: Base Case with 0.1% transaction costs.**SPY**: SPY ETF. Bolded values indicate the best metric of all presented strategies.

Sensitivity analysis

Figure: equity curves for sensitivity analysis of the transaction cost rate



*Source: Own Elaboration. Own backtesting implementation, out-of-sample performance between 03.2006 and 12.2022

Sensitivity analysis

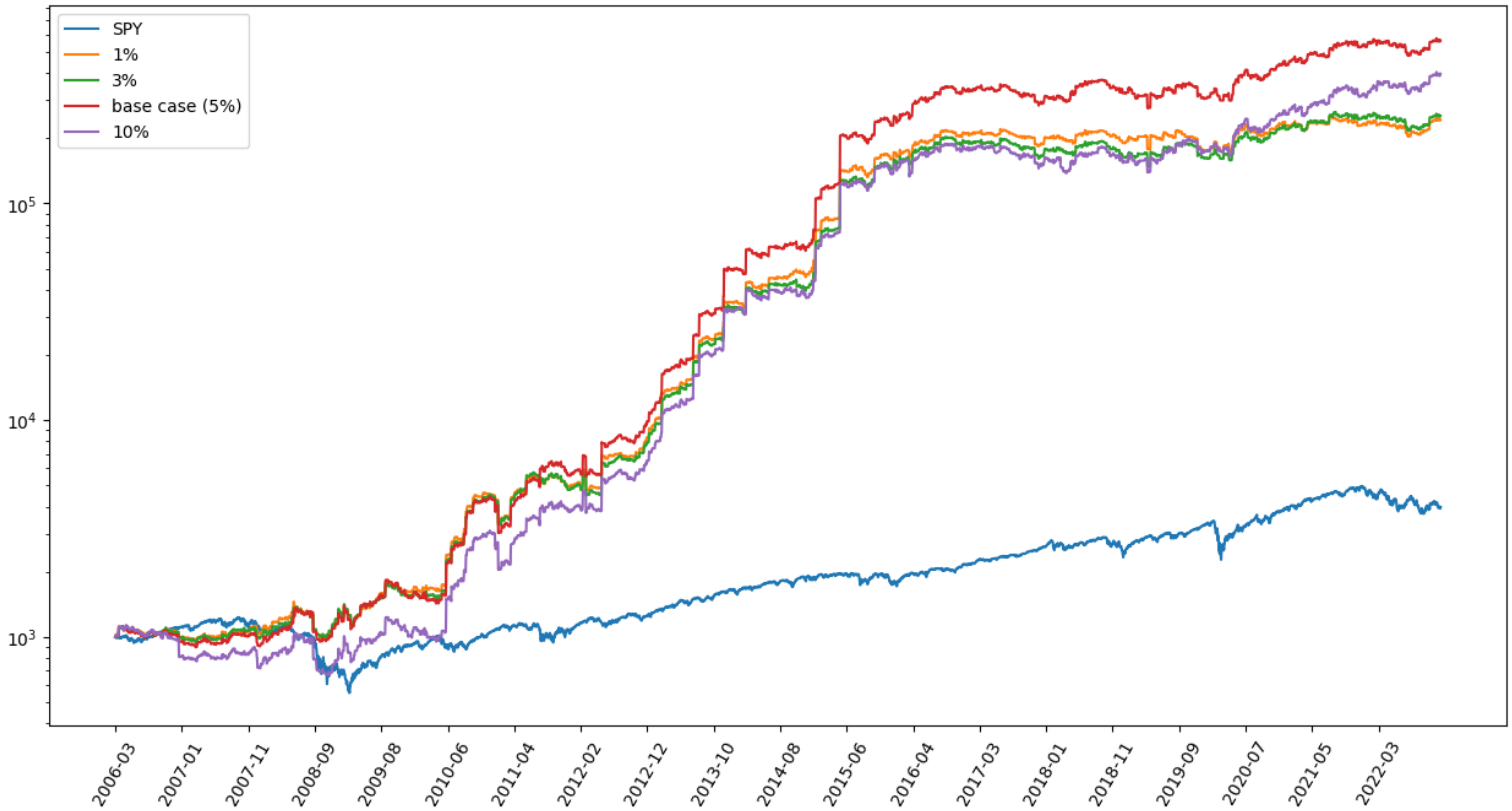
Table: Performance metrics for sensitivity analysis of stop-loss threshold value

Metric	1%	3%	base case	10%	SPY
ARC	41.54%	41.93%	**49.33%**	46.00%	9.12%
ASD	**35.29%**	35.75%	38.01%	39.12%	20.11%
IR*	1.18	1.17	**1.30**	1.18	0.45
SORTINO	3.21	3.14	**3.38**	2.84	0.70
MDD	32.08%	**28.58%**	31.98%	41.05%	55.19%
MLD	3.01 years	3.75 years	**2.10 years**	3.28 years	4.85 years
CR	1.30	1.52	**1.54**	1.12	0.17
IR**	1.53	1.78	**2.00**	1.32	0.08

*Source: Own Elaboration. **1%**: Base Case with stop loss threshold equal to 1%. **3%**: Base Case with stop loss threshold equal to 3%. **Strategy**: proposed approach, with weighted ensemble classifier, stop loss and take profit function modifications and Kelly criterion. **10%**: Base Case with stop loss threshold equal to 10%. **SPY**: SPY ETF. Bolded values indicate the best metric of all presented strategies.

Sensitivity analysis

Figure: equity curves for sensitivity analysis of stop loss threshold value



*Source: Own Elaboration. Own backtesting implementation, out-of-sample performance between 03.2006 and 12.2022

Sensitivity analysis

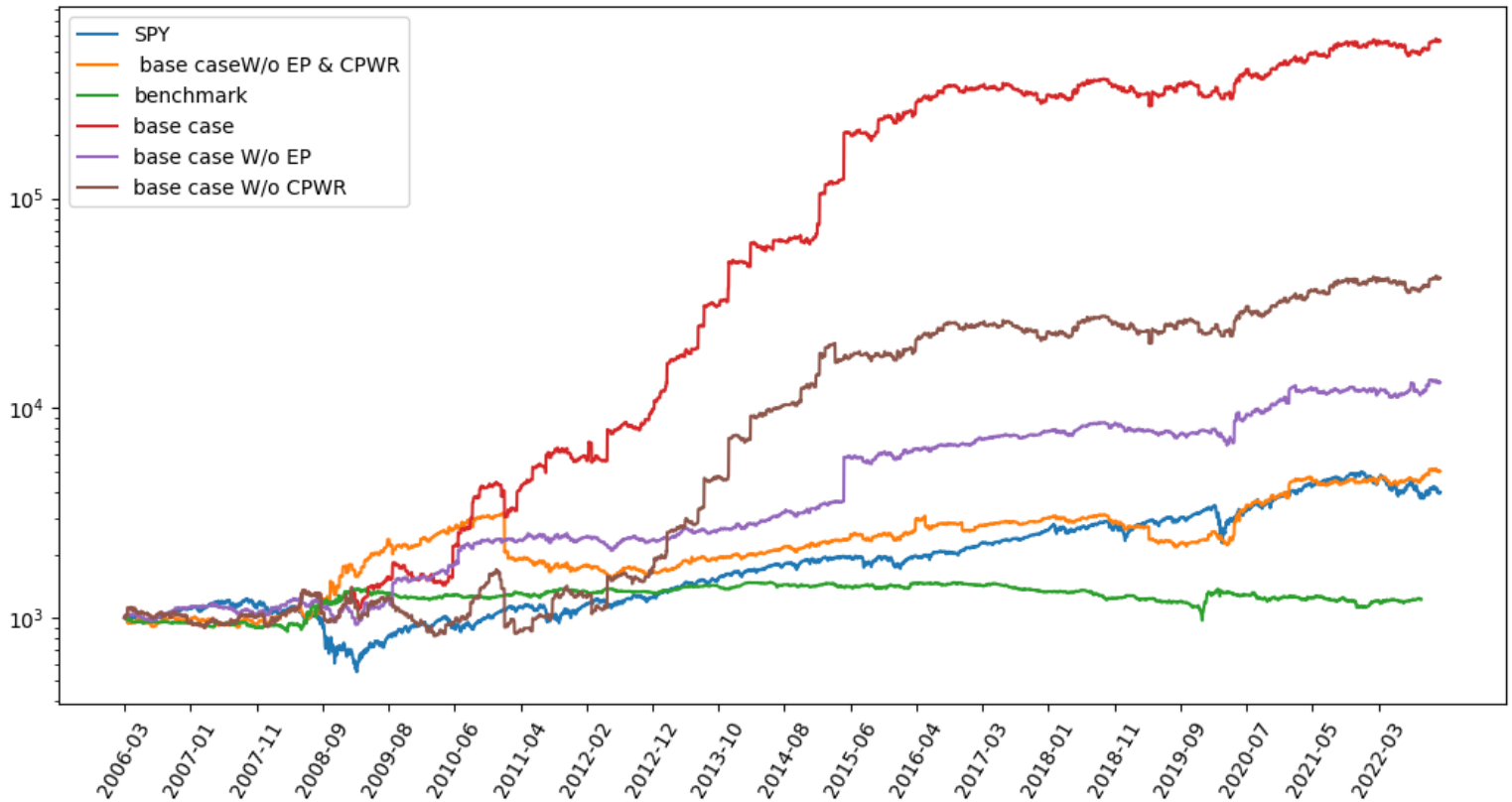
Table: Performance metrics for sensitivity analysis of over-performing stocks exclusion

Metric	Base case	Benchmark	W/o EP & CPWR	W/o EP	W/o CPWR	SPY
ARC	**49.33%**	1.13%	10.73%	17.78%	26.65%	9.12%
ASD	38.01$	**9.16%**	19.71%	22.72%	33.38%	20.11%
IR*	**1.30**	0.14	0.54	0.78	0.80	0.45
SORTINO	**3.38**	0.29	0.83	2.09	1.57	0.70
MDD	31.98%	34.30%	51.44%	**24.95%**	50.51%	55.19%
MLD	2.10 years	2.59 years	9.38 years	**1.67 years**	2.10 years	4.85 years
CR	**1.54**	0.04	0.21	0.71	0.53	0.17
IR**	**2.00**	0.01	0.11	0.55	0.42	0.08

*Source: Own Elaboration. **Strategy**: proposed approach, with weighted ensemble classifier, stop loss and take profit function modifications and Kelly criterion. **Benchmark**: Own implementation of the strategy proposed by Cartea et al. (2023), with 0.05% transaction costs. **W/o EP & CPWR**: Base Case with EP and CPWR stocks excluded from the dataset. **W/o EP** : Base Case with EP stock excluded from the dataset. **W/o CPWR**: Base Case with CPWR stock excluded from the dataset. **SPY**: SPY ETF. Bolded values indicate the best metric of all presented strategies.

Sensitivity analysis

Figure: equity curves for sensitivity analysis of over-performing stocks exclusion



*Source: Own Elaboration. Own backtesting implementation, out-of-sample performance between 03.2006 and 12.2022

Sensitivity analysis

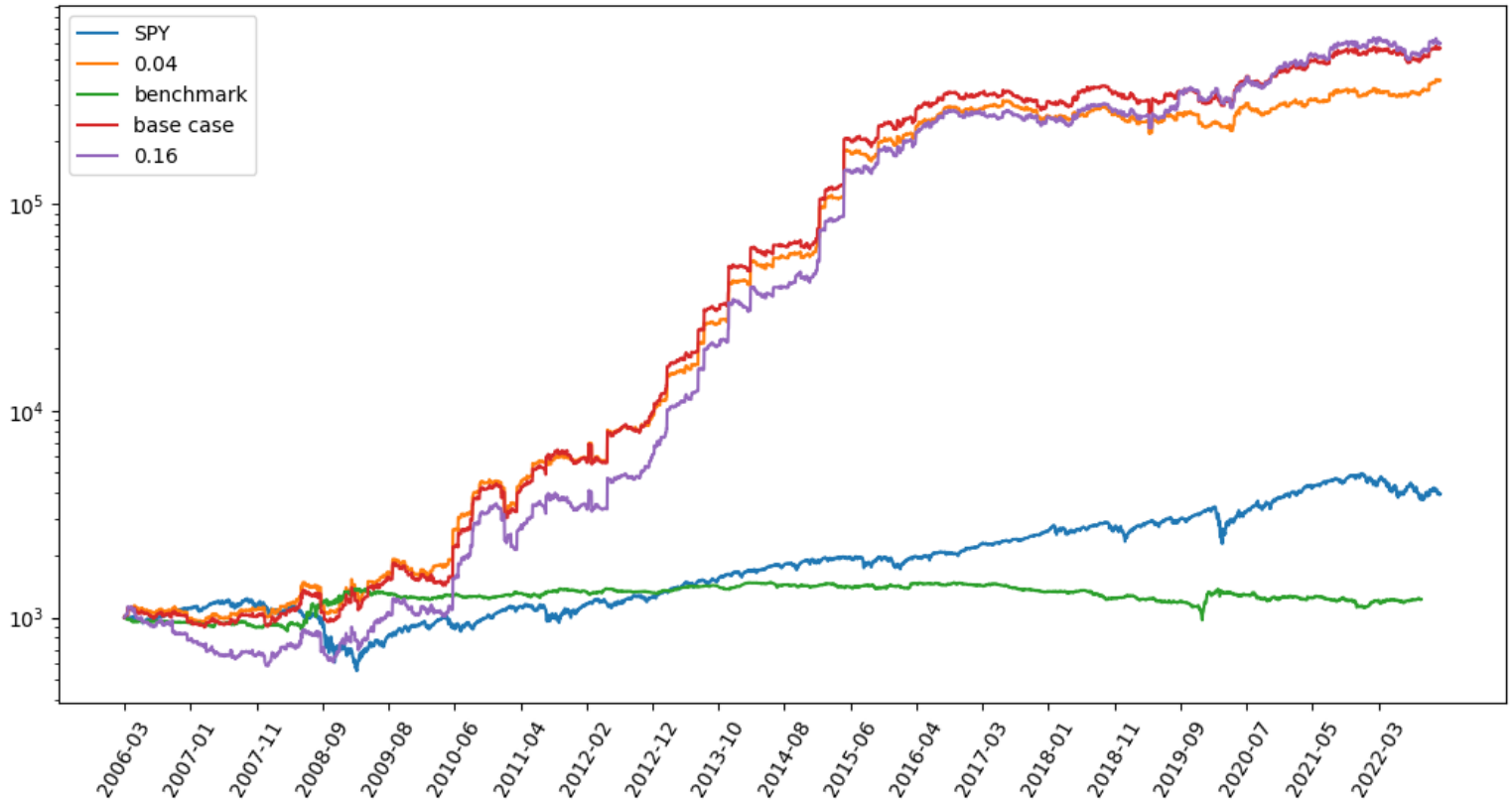
Table: Performance metrics for sensitivity analysis of different take profit scaling factors

Metric	Base Case	Benchmark	0.04	0.16	SPY
ARC	49.33%	1.13%	45.99%	**49.86%**	9.12%
ASD	38.01$	**9.16%**	36.91%	39.46%	20.11%
IR*	**1.30**	0.14	1.24	1.26	0.45
SORTINO	**3.38**	0.29	3.32	3.10	0.70
MDD	31.98%	34.30%	**31.75%**	48.43%	55.19%
MLD	**2.10 years**	2.59 years	3.72 years	3.37 years	4.85 years
CR	**1.54**	0.04	1.45	1.02	0.17
IR**	**2.00**	0.01	1.80	1.23	0.08

*Source: Own Elaboration. **Base Case**: proposed approach, with weighted ensemble classifier, stop loss and take profit function modifications and Kelly criterion. **Benchmark**: Own implementation of the strategy proposed by Cartea et al. (2023), with 0.05% transaction costs. **0.04**: Base Case with take profit scaling factor equal to 0.04. **0.16**: Base Case with take profit scaling factor equal to 0.16.**SPY**: SPY ETF. Bolded values indicate the best metric of all presented strategies.

Sensitivity analysis

Figure: Equity curves for sensitivity analysis of the take profit scaling factor



*Source: Own Elaboration. Own backtesting implementation, out-of-sample performance between 03.2006 and 12.2022

Sensitivity analysis

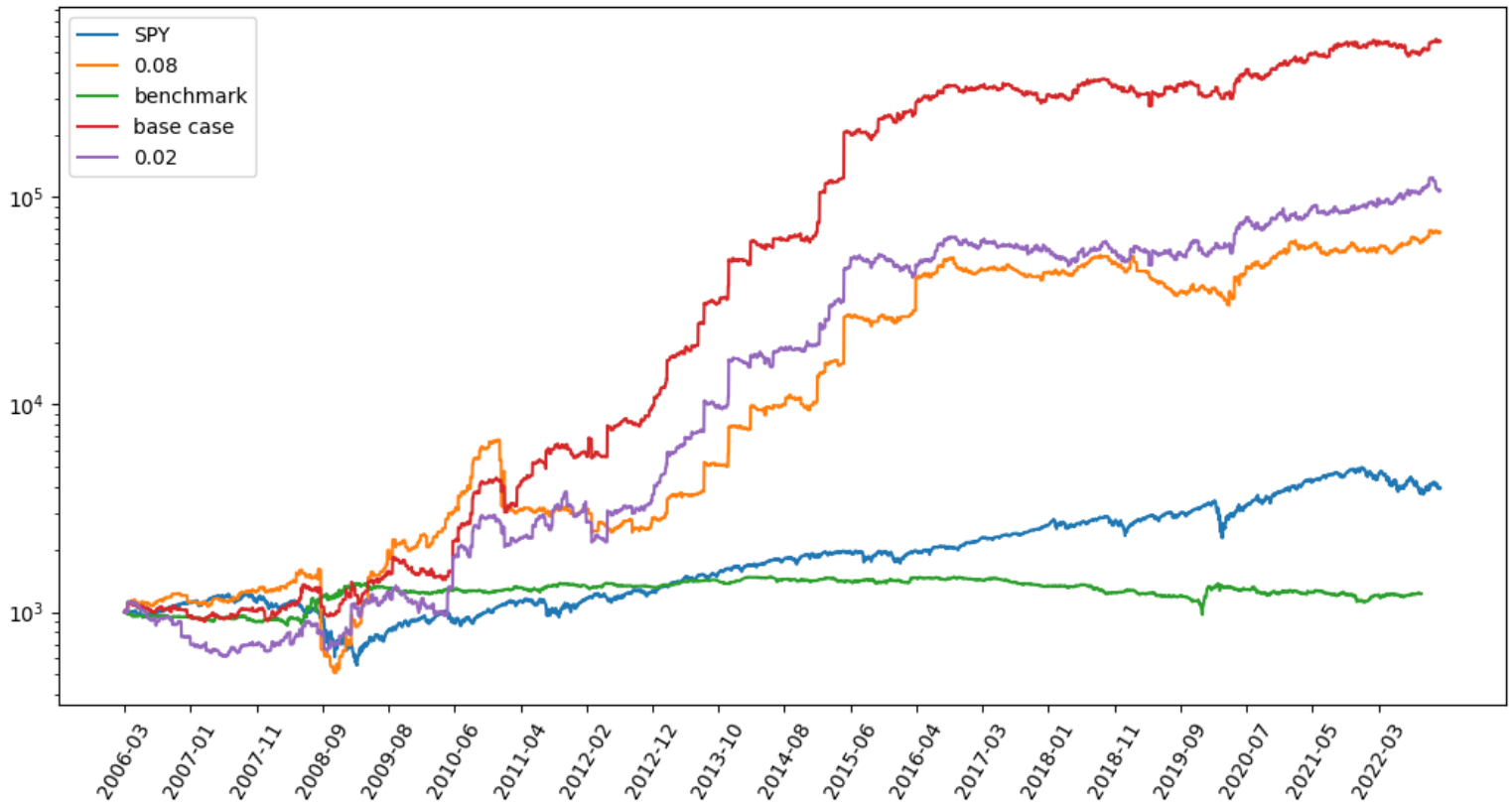
Table: Performance metrics for sensitivity analysis of threshold applied in the construction of in-sample dataset

Metric	Base Case	Benchmark	0.02	0.08	SPY
ARC	**49.33%**	1.13%	34.47%	30.57%	9.12%
ASD	38.01$	**9.16%**	37.88%	43.33%	20.11%
IR*	**1.30**	0.14	0.91	0.71	0.45
SORTINO	**3.38**	0.29	2.05	1.32	0.70
MDD	**31.98%**	34.30%	45.70%	68.40%	55.19%
MLD	**2.10 years**	3.56 years	3.72 years	2.92 years	4.85 years
CR	**1.54**	0.04	0.75	0.45	0.17
IR**	**2.00**	0.01	0.68	0.32	0.08

*Source: Own Elaboration.**Base Case**: proposed approach, with weighted ensemble classifier, stop loss and take profit function modifications and Kelly criterion. **Benchmark**: Own implementation of the strategy proposed by Cartea et al. (2023), with 0.05% transaction costs. **0.02**: Base Case with halved threshold used in the construction of in-sample dataset. **0.08**: Base Case with doubled threshold used in the construction of in-sample dataset. **SPY**: SPY ETF. Bolded values indicate the best metric of all presented strategies.

Sensitivity analysis

Figure: Equity curves for sensitivity analysis of threshold applied in the construction of in-sample dataset



*Source: Own Elaboration. Own backtesting implementation, out-of-sample performance between 03.2006 and 12.2022

Sensitivity analysis

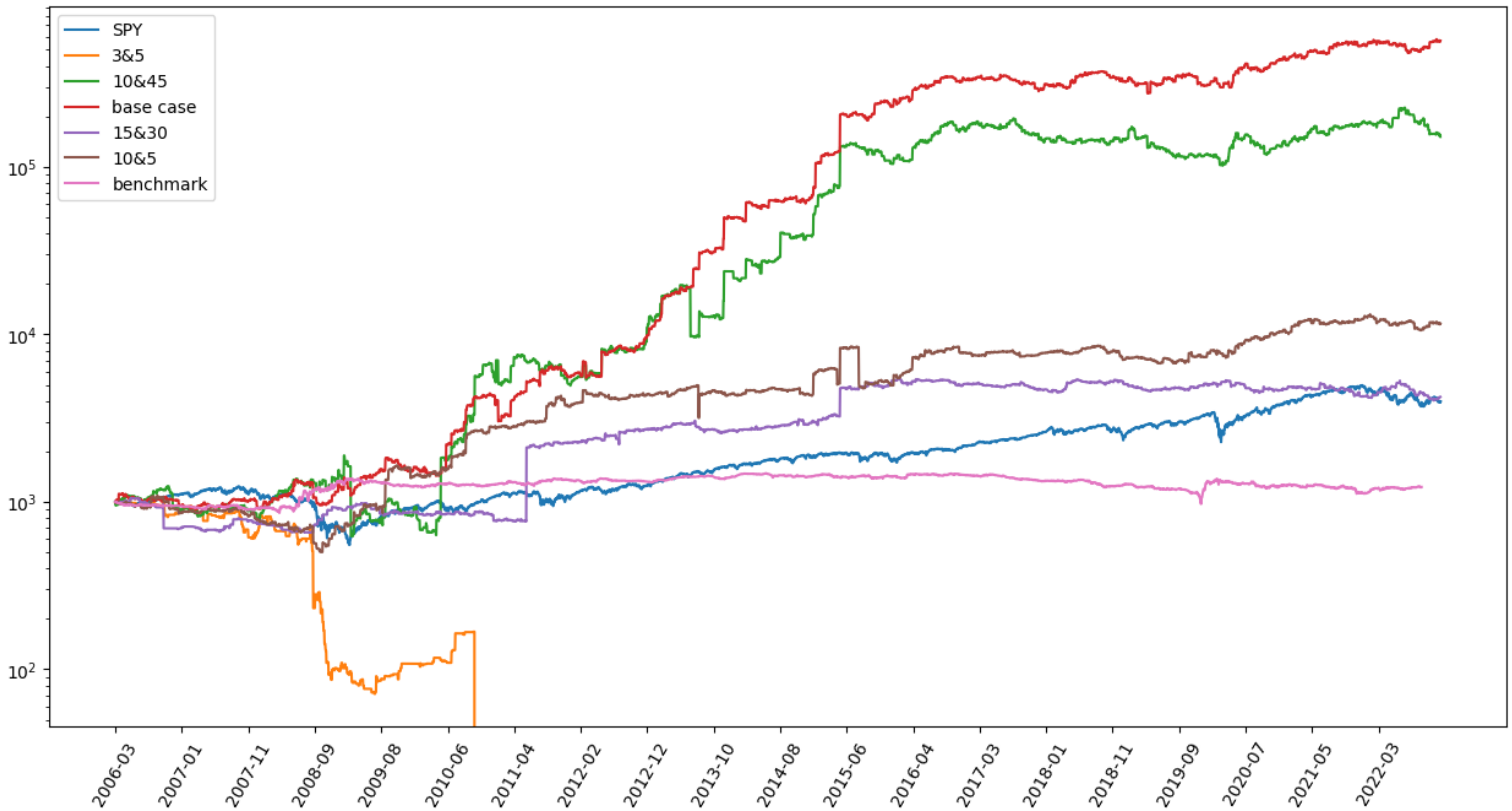
Table: Performance metrics for sensitivity analysis of rebalance frequency and length of clustering lookback window

Metric	Base Case	Benchmark	3&5	10&5	15&30	10&45	SPY
ARC	**49.33%**	1.13%	-100%	16.78%	9.54%	37.34%	9.12%
ASD	38.01$	**9.16%**	35.11%	33.48%	46.83%	59.92%	20.11%
IR*	**1.30**	0.14	-2.85	0.50	0.20	0.62	0.45
SORTINO	**3.38**	0.29	-3.18	0.94	0.95	1.42	0.70
MDD	**31.98%**	34.30%	100%	54.02%	37.67%	67.02%	55.19%
MLD	**2.10 yrs.**	3.56 yrs.	11.17 yrs.	2.92 yrs.	4.99 yrs.	4.79 yrs.	4.85 yrs.
CR	**1.54**	0.04	-1	0.31	0.25	0.58	0.17
IR**	**2.00**	0.01	-2.85	0.16	0.05	0.35	0.08

*Source: Own Elaboration. **Base Case**: proposed approach, with weighted ensemble classifier, stop loss and take profit function modifications and Kelly criterion. The values of a and b are equal to 10&30. **Benchmark**: Own implementation of the strategy proposed by Cartea et al. (2023), with 0.05% transaction costs.**SPY**: SPY ETF. Bolded values indicate the best metric of all presented strategies.

Sensitivity analysis

Figure: Equity curves for sensitivity analysis of rebalance frequency and length of clustering lookback window



*Source: Own Elaboration. Own backtesting implementation, out-of-sample performance between 03.2006 and 12.2022

Summary of sensitivity analysis

Stop loss and take profit modifications		**Take Profit Threshold**	
No time-variant, no Kelly	1	0.04	2
No time Variant, Kelly	2	0.08 (Base Case)	5
Time variant & Kelly (Base Case)	5	0.16	1
Ensemble weights		**Threshold for constructing in-sample dataset**	
Equally Weighted	0	0.02	1
HistGradientBoosting weight doubled (Base Case)	7	0.04 (Base Case)	7
Only HistGradientBoosting	1	0.08	0
Transaction costs		**Stop loss threshold**	
0%	7	1%	1
0.05% (Base Case)	1	3%	1
0.075%	0	5% (Base Case)	6
0.1%	0	10%	0
Rebalance frequency and length of clustering look back window		**Overperforming stocks exclusion**	
3&5	1	With EP and CPWR (Base Case)	5
10&5	0	Without EP and CPWR	1
10&30 (Base Case)	7	Without EP	2
10&45	0	Without CPWR	0
15&30	0		

Souce: Own elaboration. The table shows the number of performance metrics on which a given strategy performed the best within the sensitivity analysis of a given parameter.

Answers to the research questions

- **RQ1**: *Does the usage of signal quality classifiers improve the quality of the graph-clustering-based statistical arbitrage strategy?*
  - Classifiers improve strategy performance, particularly when using a weighted soft voting ensemble, by reducing unprofitable signals.
- **RQ2**: *To what extent does the implementation of transaction and risk management measures influence the performance of the strategy?*
  - Kelly criterion had limited impact, but stop-loss functions based on signal probability reduced downside risk.
- **RQ3**: *What is the sensitivity of the strategy to changes in transaction costs?*
  - Strategy remained robust, with only a 10% drop in the Information Ratio after doubling transaction costs.
- **RQ4**: *To what extent does the change of weights in the classifiers ensemble influence the strategy?*
  - Weighted ensembles with emphasis on top-performing models led to better results than equal weights or single-classifier setups.

Further research ideas

- Explore strategy performance in volatile markets like cryptocurrencies.
- Test multilabel classification (profitable shorts, longs, or non-profitable signals).
- Apply the framework to higher-frequency data, commonly used in statistical arbitrage.

Thank you!

Q&A