Pair Trading Report

A quantitative strategy on the S&P500

Lancelot Watelle

January 2026

Contents

1 Introduction

1.1 What is pair trading?

Pair trading, although an old quantitative strategy, is still widely used in the financial industry today.

It can be defined as a **market-neutral strategy** that aims to profit from most market movements (upward, downward, etc.). This approach consists in monitoring the performance of two historically correlated securities[1]. When this correlation temporarily weakens (that is, when one security rises while the other falls) the strategy is to **short the outperforming security** and **buy the underperforming security**, betting that the spread will eventually converge again.

As this strategy is historically used and lies at the core of statistical arbitrage, the goal was to test a quantitative implementation on the globally recognized index: the S&P500.

1.2 Why the S&P500?

The U.S. stock market is the largest and most dynamic market in the world. It accounts for more than 50% of global market capitalization and nearly 40% of worldwide financial transactions.

Its main index, the S&P500, includes the 500 largest listed companies in the United States. With its high sector and industry diversity, it provides a way to identify companies with strong statistical relationships while grouping them by sector.

The S&P500 is therefore an ideal universe for identifying stock pairs that are likely to exhibit significant cointegration.

1.3 Project objective

The objective of this project is to study the cointegration of S&P500 assets and to build a fictitious portfolio enabling the implementation of an optimized pair trading strategy.

[1] Cointegration is a statistical property of time series that makes it possible to detect a long-term relationship between two or more series.

2 Methodology

2.1 Data

Throughout this project, daily market close data (business days) are used. The type of analysis required by such a strategy justifies this choice.

All data come from Python's free *yfinance* API. Its accessibility and ease of use make it an ideal tool for an academic prototype.

The analysed period depends on the project stage. For global tests, a 2–5 year horizon was chosen. However, when validating the strategy, this period is varied to assess its robustness.

To retrieve information on S&P500 companies, web scraping is used (in full compliance with regulations) on Wikipedia, which provides the companies, their tickers, their sectors, and their sub-industries.

2.2 What is pair trading mathematically?

Pair trading aims to bet on the equilibrium of the spread between two assets. If this phenomenon is described mathematically, the fundamental equation linking the two assets is:

$$\text{asset}_1 = \beta \cdot \text{asset}_2 + \text{spread} \tag{1}$$

The stationarity[2] of the *spread* is analysed while waiting for a disturbance in order to take a position. The central idea is that when the *spread* deviates from its equilibrium value, a profitable position can be built by anticipating a return to the mean.

[2] A stationary series is a time series whose statistical properties (mean, variance) do not depend on time.

3 Strategy construction

3.1 Walk-Forward Analysis

To assess the efficiency of a pair over a given period, the **Walk-Forward Analysis** approach was adopted. This method splits a period into two distinct phases:

- **Phase 1 (historical)**: study of past data to compute indicators and parameters
- **Phase 2 (forward)**: live trading based on the previously estimated parameters

This approach makes it possible to optimize performance by choosing a short trading period while analysing a large amount of historical data. A **1-year analysis window** was therefore chosen for a **1-month trading period**.

3.2 Analysis parameters

3.2.1 Weight difference (Risk-Adjusted Weighting)

In the classical pair trading method, weighting depends solely on the regression beta coefficient. However, to reduce risk while improving performance, past volatility of each asset is used to weight the equation.

The equation becomes:

$$\frac{1}{\text{vol}_{\text{asset}_1}} \cdot \text{asset}_1 = \frac{\beta}{\text{vol}_{\text{asset}_2}} \cdot \text{asset}_2 + \text{spread} \tag{2}$$

where volatility is the standard deviation of absolute daily returns.

The final weights are computed as follows:

$$\text{weight}_{\text{asset}_1} = \frac{\frac{1}{\text{vol}_{\text{asset}_1}}}{\frac{1}{\text{vol}_{\text{asset}_1}} + \frac{\beta}{\text{vol}_{\text{asset}_2}}} \tag{3}$$

$$\text{weight}_{\text{asset}_2} = 1 - \text{weight}_{\text{asset}_1} \tag{4}$$

This approach yields a more balanced allocation that takes into account each asset's risk profile.

3.2.2 Volume and liquidity

An asset's volume is a key indicator of its liquidity. In a pair trading strategy, liquidity is essential to enter and exit positions without significant market impact.

The **Average Daily Dollar Volume** (ADDV) is defined as:

$$\text{ADDV} = \frac{\sum_{i=1}^{n}(\text{Price}_i \times \text{Volume}_i)}{n} \tag{5}$$

This parameter is used to filter out pairs with insufficient liquidity. Each month, only the top 20% of pairs ordered by decreasing ADDV are kept.

3.2.3 Cointegration test

A time series is a sequence of numerical values representing the evolution of a quantity over time. Financial price modelling relies heavily on this field, and a key concept is **cointegration**.

Cointegration is a statistical property that detects long-term relationships between several time series. It is important to distinguish:

- **Correlation**: detects similarities in variations (if A increases, B increases as well)
- **Cointegration**: studies whether the difference between the series remains stable in the long run

To assess whether two series are cointegrated, the **Augmented Dickey–Fuller** test is applied to the regression residual.

Formal definition: Two series y_t and x_t, integrated of order 1 $[I(1)]$, are cointegrated if there exists a vector $\beta = (1, -\beta_1)$ such that:

$$\dots \quad u_t = y_t - \beta_1 x_t \sim I(0)$$

where u_t is stationary. Studying the cointegration of two series is therefore essential for an effective pair trading strategy.

3.2.4 Half-life of the spread (Mean Reversion Speed)

The *half-life of mean reversion* represents the time needed for a disturbed spread to move halfway back toward its equilibrium value. This parameter is crucial to size positions and manage risk.

The **Ornstein–Uhlenbeck** stochastic process is used:

$$ds_t = \lambda(\mu - s_t)\, dt + \sigma\, dW_t \tag{6}$$

where:

- s_t: spread
- λ: speed of mean reversion
- μ: long-term mean

The solution to this process shows that:

$$\text{halfLife} = \frac{\ln(2)}{\lambda} \tag{7}$$

In practice, λ is estimated by a discrete linear regression:

$$\Delta s_t = a \cdot s_{t-1} + b + \epsilon_t \quad \text{with} \quad a = -\lambda \tag{8}$$

3.2.5 Spread volatility

Now that the speed at which the spread returns to its mean is known, its amplitude of variation must be assessed. The **standard deviation of the spread** over the analysis window is simply used.

3.2.6 Pseudo-trading (Backtesting on Historical Data)

To predict a pair's performance in live trading, a fictitious trading simulation is run on the last 6 months of the analysis period. This period is split into six 1-month subperiods to remain as realistic as possible.

The extracted parameters are:

- **Average PnL**: average of PnL samples
- **Prop PnL Pos**: proportion of winning periods (%)
- **Worst PnL**: worst monthly result
- **Avg Max Drawdown**: average monthly drawdown

4 Data analysis

4.1 Initial approach vs final approach

Initially, the idea was to build a linear regression between the parameters and the PnL to predict future performance. However, it quickly appeared that a few extraordinarily performing pairs completely distorted the model, driving the R^2 close to 0 (useless regression).

An alternative approach based on **ranking** was therefore adopted: ranking pairs according to the relative influence of their parameters on the following month's trading.

4.2 Quantile interpretation

Quantile-based segmentation analysis is a statistical method used to understand non-linear relationships between variables.

The method is split into 3 steps:

1. **Ordering**: data are ordered according to the dependent variable (here the monthly PnL)
2. **Splitting**: the sample is divided into quantiles (5 quantiles were chosen)
3. **Analysis**: for each explanatory variable, its average value per quantile is computed and a trend is sought

4.3 Data construction

Over the period **2018-01-01 to 2022-01-01** (48 periods), indicators are computed for each pair within the same sector, as well as the actual PnL of the following month. This creates a dataset enabling the analysis of relationships between indicators and performance.

4.4 Z-scores and normalization

To compare variables coming from different distributions, each variable is standardized:

$$\text{z\_score}_{\text{element}} = \frac{\text{variable}_{\text{element}} - E[\text{variable}]}{\sigma[\text{variable}]} \tag{9}$$

In some cases, extreme values are capped (max cap at 1) to prevent a single variable from dominating the ranking. For variables with a U-shaped distribution (where central values are better), the following is used:

$$z_{\text{final}} = \exp\left(-\frac{z^2_{\text{variable}}}{2}\right) \tag{10}$$

4.5 Information Coefficient (IC Score)

The IC measures the predictive power of a variable for the following month's PnL:

$$\text{IC} = \text{corr}(\text{zscore}_{\text{var}}, \text{pnl\_month\_next}) \tag{11}$$

By computing the IC for each variable, a measure of its relative importance is obtained.

4.6 Final ranking function

Although some variables (p-value, volume) have a low IC, they remain useful as **filters**:

- **Filter 1**: keep only pairs with p-value < 0.05 (cointegrated pairs)
- **Filter 2**: keep only the top 20% of pairs with the highest ADDV

For the remaining variables, a weighted ranking is created:

$$\alpha_i = \frac{|\mathrm{IC}_i|}{\sum |\mathrm{IC}|} \tag{12}$$

$$\mathrm{RANKING}(\mathrm{pair}) = \sum_{i=1}^{n} \alpha_i \cdot z\_\mathrm{score}_{\mathrm{pair},i} \tag{13}$$

To avoid performance-chasing bias, pairs that performed well in the previous month are penalized:

$$\mathrm{RANKING\_FINAL}(\mathrm{pair}) = \frac{\mathrm{RANKING}(\mathrm{pair})}{1 + 0.5 \cdot \mathrm{count}_{m-1} + 0.30 \cdot \mathrm{count}_{m-2}} \tag{14}$$

5 Strategy backtest

5.1 Strategy architecture

The strategy is split into two monthly steps:

Step 1: Data analysis (end of previous month)

1. For each pair within the same sector, retrieve its indicators (based on the last 12 months)
2. Apply filters (p-value, ADDV)
3. Sort the rows using the ranking function
4. Extract the top 20 pairs while respecting: max 2 occurrences of the same stock, max 40% per sector

Step 2: Trading (during the month)

1. Trade each pair according to the pair trading strategy (with the computed parameters)
2. Allocate $\frac{\text{value\_ptf}}{nb\_pairs}$ to each pair
3. At the end of the month, force the exit of all positions
4. Reinvest the portfolio (with or without compounding)

6 Live performance

6.1 Backtest on 2022–2025

The model was calibrated on 2018–2022, then tested on 2022–2025. The results show encouraging performance with two different interest-management approaches:

6.1.1 With interest compounding

The portfolio exhibits steady growth, with an average Sharpe ratio of **1.35** over the period. Annual performance is:

Table 1: Annual performance with interest compounding

Year	Final Value	Profitability	Max DD	Volatility	Sharpe
2022-2023	224 330	49.55 %	-4.54 %	4.64 %	2.40
2023-2024	168 630	12.42 %	-4.39 %	2.94 %	0.85
2024-2025	163 966	9.31 %	-1.37 %	2.05 %	0.79
Global	185 642	23.76 %	-3.43 %	3.21 %	1.35

The 2022–2023 period performed particularly well, positively impacting all indicators. The maximum drawdown remains very low (about -3.5% on average), highlighting the robustness of the strategy.

Monthly performance:

- 75 % positive months
- 25 % negative months
- Average monthly return: 1.76 %

6.1.2 Without interest compounding

The approach without reinvesting profits shows a slightly different risk–return profile:

Table 2: Annual performance without interest compounding

Year	Final Value	Profitability	Max DD	Volatility	Sharpe
2022-2023	214 340	42.89 %	-2.52 %	3.80 %	2.54
2023-2024	169 239	12.83 %	-3.03 %	2.59 %	1.00
2024-2025	164 500	9.67 %	-1.25 %	1.95 %	0.88
Global	182 693	21.80 %	-2.27 %	2.78 %	1.47

Risk is logically lower (maximum drawdown falls to around -2.3%), while total PnL decreases slightly. The Sharpe ratio improves to 1.47.

6.2 Strategy beta

A key indicator is the strategy's **beta**, measuring its sensitivity to movements in the overall market.

$$\beta = \frac{\mathrm{Corr}(R_{\text{strategy}}, R_{\text{S\&P500}})}{\mathrm{Var}(R_{\text{S\&P500}})} = 0.0049 \tag{15}$$

The estimated beta, close to zero, suggests very low exposure to the market, which is consistent with a market-neutral design.

7 Issues encountered and solutions

7.1 Prediction vs ranking

Problem: The linear regression used to predict PnL produced an R^2 close to 0 because a few extreme pairs dominated the model.

Solution: Switch to a **weighted ranking** approach, which is far less sensitive to outliers.

7.2 Asset weighting

Problem: A simple weighting scheme was initially used without taking volatility into account, which distorted the interpretation of the linear regression.

Solution: Implement inverse-volatility weighting, adjusting beta accordingly.

7.3 Past-performance bias

Problem: At the beginning of the backtest, large losses appeared after large gains, which degraded the Sharpe ratio.

Solution: Introduce a penalty for pairs that performed well in the previous month, preventing the ranking from being biased toward recent winners.

8 Areas for improvement

8.1 Integration of macroeconomic factors

The pair trading strategy currently relies solely on mathematical indicators. Adding a penalty related to macroeconomic attention (pairs where one of the stocks is under strong media focus) could improve robustness by avoiding periods during which cointegration breaks down.

8.2 Leverage and long/short overlays

Around 75% of months are positive, but gains are not always very large. Using leverage and additional long/short overlays could increase returns, although this would also increase risk.

8.3 Stop loss and risk management

For now, no "large" monthly losses have been observed. However, if a pair significantly underperforms while representing a large share of the portfolio, this could wipe out monthly performance. A well-calibrated **stop loss** would be necessary for a real-world market implementation.

9 Conclusion

This project, initially designed to test cointegration between S&P500 assets, evolved into the construction of a full-fledged quantitative trading strategy. Although summarized here, the development process was long and rewarding: from initial exploration to validation on several years of data.

As a first real quantitative project, it provided an opportunity to learn many fundamental techniques: time series manipulation, statistical regression, strategy optimization, and risk management. Getting familiar with tools such as *yfinance*, *pandas*, *numpy*, and statistical tests (Dickey–Fuller, correlations) will certainly be useful for future projects.

Although theoretically promising after calibration, this strategy would require further improvements before a live deployment on financial markets. The robustness of the results and the low drawdown are encouraging signals to continue this work.

Looking forward to tackling the next challenge!