Transient Pair Discovery using Unsupervised Clustering Methods

Jackson Moores, Michael Winders, Hailey Anton, Kevin Yang
QUANTT, Queen's University

March 2026

Contents

Abstract

Traditional pairs trading strategies rely on cointegration to identify asset pairs that share a long-run equilibrium, but this assumption fails to capture the transient, short-lived correlations that frequently emerge and dissolve across equity markets. This paper presents a clustering-driven framework for pairs trading that uses unsupervised learning to discover these transient relationships. We engineer nine features per asset, capturing volatility, market and sector beta, momentum, and regime shifts and apply OPTICS density-based clustering to hourly price data across 142 equities spanning five sectors. A noise-adjusted co-clustering frequency metric addresses OPTICS's high noise rate (approximately 58–68% across clustering methods) to accurately measure pair affinity. Rather than trading individual cluster formations, we use persistent co-clustering frequency above a threshold as a screening criterion to identify pairs whose transient relationships recur reliably, then validate these candidates through a five-test scored framework designed specifically for transient data, replacing the cointegration test which yields a 0% pass rate on short-lived relationships. Method validation on 40 semiconductor equities demonstrates a 5x lift in pass rate over randomly generated pairs, confirming that clustering detects real market structure. Scaling to the full universe, 3,643 pairs exceed the frequency threshold, of which 2,148 are deemed tradeable by the scoring framework, and a baseline daily z-score strategy yields 41% profitability across all tradeable pairs, rising to 64% among top-scoring pairs under a Kalman-filtered backtest with optimized parameters and realistic transaction costs. These results suggest that transient correlation detection via unsupervised clustering offers a viable alternative to cointegration-based pair discovery, particularly in markets where structural relationships are short-lived and regime-dependent.

1 Introduction

Pairs trading is a market-neutral strategy that exploits temporary mispricings between two statistically related assets. When the price spread between a pair deviates from its historical norm, a trader takes opposing positions in each asset, going long the undervalued one and short the overvalued one, and profits when the spread reverts to its mean. The strategy's appeal lies in its market neutrality: returns are driven by the relative relationship between the two assets rather than by the direction of the broader market.

The dominant framework for pairs trading since the late 1980s has been cointegration, introduced by Engle and Granger. Two assets are said to be cointegrated if a linear combination of their prices forms a stationary process, meaning the spread between them fluctuates around a stable long-run equilibrium. This framework has been widely adopted because it provides a rigorous statistical basis for pair selection: if two assets are cointegrated, deviations from equilibrium are by definition temporary, and mean reversion is guaranteed over a sufficiently long horizon.

However, the cointegration framework carries a strong implicit assumption: that the equilibrium relationship between two assets is permanent. In practice, many equity relationships are transient. Structural similarities between firms (shared exposure to a supply chain disruption, correlated earnings cycles, or temporary alignment in volatility regimes) can cause assets to move together for periods of hours to weeks before the relationship dissolves. These transient correlations represent real, exploitable structure, but they are invisible to cointegration tests, which require stationarity over the entire sample period. A pair that co-moves reliably for three weeks out of every month will fail a cointegration test because the relationship is not permanent, yet it may be highly tradeable.

This project proposes an alternative approach: using unsupervised clustering to discover pairs whose transient relationships recur with sufficient frequency to be traded. Rather than testing whether two assets share a permanent equilibrium, we ask a different question: when two assets are behaving similarly in feature space, how often does that happen? If the answer is "frequently and reliably," the pair is a candidate for trading, regardless of whether the relationship satisfies cointegration.

Our framework operates in two stages. In the first stage, we compute nine features per asset at each hourly timestamp, capturing short-term volatility, market and sector beta, momentum, relative strength, and regime shifts, and apply OPTICS density-based clustering to group assets exhibiting similar behavior. We then measure the noise-adjusted co-clustering frequency of every pair: the proportion of clusterable timestamps in which both assets appear in the same cluster. Pairs exceeding a persistence threshold enter a candidate registry. In the second stage, candidates are validated through a five-test scored framework that evaluates mean-reversion properties (stationarity, half-life, Hurst exponent, variance ratio, and rolling correlation stability) without requiring cointegration. Pairs scoring sufficiently high are backtested using a z-score mean reversion strategy with walk-forward validation.

During Phase 1 validation, we observed that many pairs did not simply co-cluster once and disappear. Instead, certain pairs exhibited recurring co-clustering patterns, repeatedly entering and exiting the same cluster across the sample period. This persistence suggested a secondary hypothesis: if two assets transiently co-cluster with sufficient regularity, the relationship may also be exploitable under a conventional daily mean-reversion framework, even without trading individual formation events in real time. The trading strategy evaluated in this paper tests this hypothesis, using hourly clustering as a pair screening mechanism and daily z-score mean reversion as the execution layer. Real-time detection and trading of individual clustering events as they form remains a primary direction for future work.

We develop and validate this framework in two phases. Phase 1 serves as a proof of concept on 40 semiconductor equities, where we demonstrate a 5x lift in validation pass rate over randomly generated pairs, confirming that the clustering pipeline identifies real market structure rather than statistical noise. Phase 2 scales the framework to 142 equities across five sectors: Technology, Healthcare, Energy, Financial Services, and Industrials, producing 3,643 candidate pairs, of which 2,148 are deemed tradeable. An enhanced backtest incorporating Kalman-filtered hedge ratios, optimized entry and exit thresholds, and realistic transaction costs yields a 64% profitability rate on top-scoring pairs out of sample.

Importantly, our framework operates in three distinct layers. First, hourly OPTICS clustering serves as a *pair selection* mechanism, identifying which pairs exhibit recurring transient co-movements. Second, a five-test validation framework confirms that selected pairs possess genuine mean-reverting statistical properties. Third, a separate daily z-score trading strategy executes trades on validated pairs using standard mean-reversion signals. The clustering does not generate trading signals directly; rather, it screens the universe to identify pairs whose transient relationships make them suitable candidates for profitable mean-reversion trading.

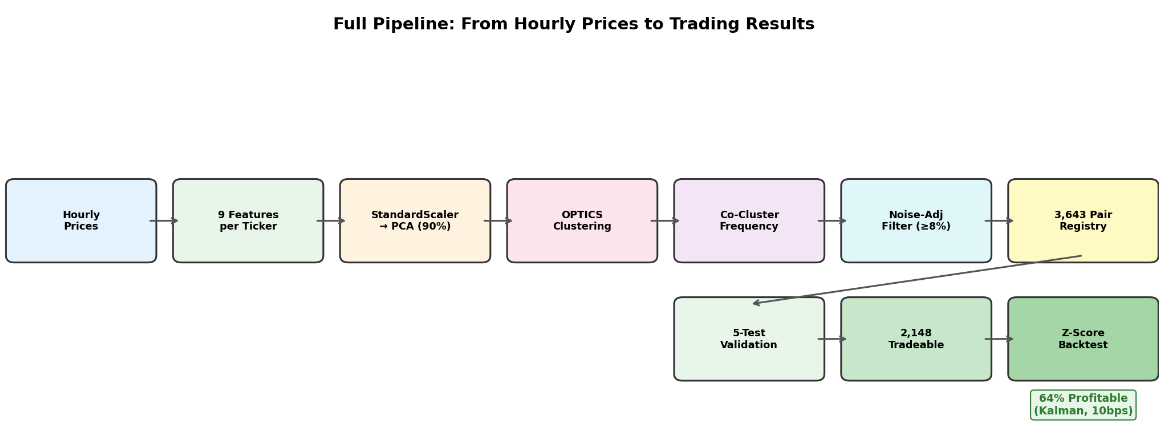

Figure 1: End-to-end pipeline overview. Hourly prices for 142 equities are transformed into 9 clustering features, clustered via OPTICS at each timestamp, and filtered by noise-adjusted co-clustering frequency ($\geq 8\%$). The resulting 3,643 candidate pairs are scored by a five-test validation framework, yielding 2,148 tradeable pairs. A z-score mean-reversion backtest with walk-forward validation produces the final profitability results (64% under Kalman hedging).

2 Feature Engineering

Nine features were engineered from hourly price data to capture short-term dynamics relevant to transient pair identification. Features were computed at two time scales: a short window of 50 hours (approximately one trading week) and a medium window of 147 hours (approximately one month). Three regime shift features, defined as the normalized difference between short and medium windows, capture directional changes in market behaviour.

Table 1: Feature Descriptions and Parameters

Feature	Window	Description
Returns	1h	Percentage returns
Vol_Short	50h	Rolling standard deviation of returns
Beta_SPX_Short	50h	Rolling beta to S&P 500
Beta_Sector_Short	50h	Rolling beta to sector (leave-one-out)
RSI	70h	Relative Strength Index (Wilder's smoothing)
Momentum_5H	5h	5-hour price momentum
Vol_Regime_Shift	50h/147h	$(\sigma_{\text{short}} - \sigma_{\text{med}})/\sigma_{\text{med}}$
Beta_SPX_Regime_Shift	50h/147h	$(\beta^{\text{SPX}}_{\text{short}} - \beta^{\text{SPX}}_{\text{med}})/\beta^{\text{SPX}}_{\text{med}}$
Beta_Sector_Regime_Shift	50h/147h	$(\beta^{\text{sec}}_{\text{short}} - \beta^{\text{sec}}_{\text{med}})/\beta^{\text{sec}}_{\text{med}}$

Sector betas were computed using a leave-one-out methodology: when calculating a ticker's beta to its sector, the ticker itself was excluded from the sector average to avoid self-correlation. RSI used Wilder's exponential smoothing with $\alpha = 1/70$, providing a smoother signal than standard RSI implementations.

The three regime shift indicators are computed as the normalized difference between short and medium windows:

$$\text{Regime\_Shift}_t = \frac{\text{metric}_{\text{short},t} - \text{metric}_{\text{medium},t}}{\text{metric}_{\text{medium},t}} \tag{1}$$

These features detect when a stock's recent behaviour diverges from its medium-term norm, the key signal for identifying transient clustering events. In total, 12 features are computed per asset (including the three medium-window metrics: Vol_Medium, Beta_SPX_Medium, and Beta_Sector_Medium at 147h), but only the 9 short-window and regime shift features are used for clustering. The medium-window features serve solely as denominators for computing regime shifts.

All nine clustering features were standardized using `StandardScaler` and then reduced via PCA retaining 90% of variance, typically compressing the 9 features to 3–4 principal components. This preprocessing step removes scale differences between features and reduces dimensionality while preserving the dominant sources of variation in the data.

3 Phase 1: Semiconductor Proof-of-Concept

Before scaling to a cross-sector universe, the clustering framework was validated on a controlled 40-ticker semiconductor universe. Semiconductors were chosen for their strong intra-industry correlations and homogeneous factor exposures, providing a favourable environment to test whether unsupervised clustering could detect tradeable transient relationships.

The validation methodology used a backward-looking three-window structure designed to test whether cluster formations predicted future mean-reverting behaviour:

- **Execution Lag** (2 hours), a realistic delay between observing a cluster formation and initiating a trade.
- **Calibration Window** (20 hours), used to estimate the hedge ratio via OLS regression on the pair's spread.
- **Exploitation Window** (40 hours), the out-of-sample period where mean-reversion diagnostics (AR(1) coefficient, lag-1 autocorrelation, half-life, and bounce rate) were recorded.

Of 657 formation events tested across OPTICS cluster outputs, 26 passed all five validation criteria: return correlation > 0.70, spread coefficient of variation < 0.03, half-life < 8 hours, hedge ratio drift < 0.20, and presence of a tradeable signal, yielding a 4.0% pass rate, compared to 0.8% for randomly generated pairs (a 5x lift). This result confirms that OPTICS clustering identifies pairs with genuine short-term mean-reverting properties rather than statistical noise, validating the method before applying it to a larger universe.

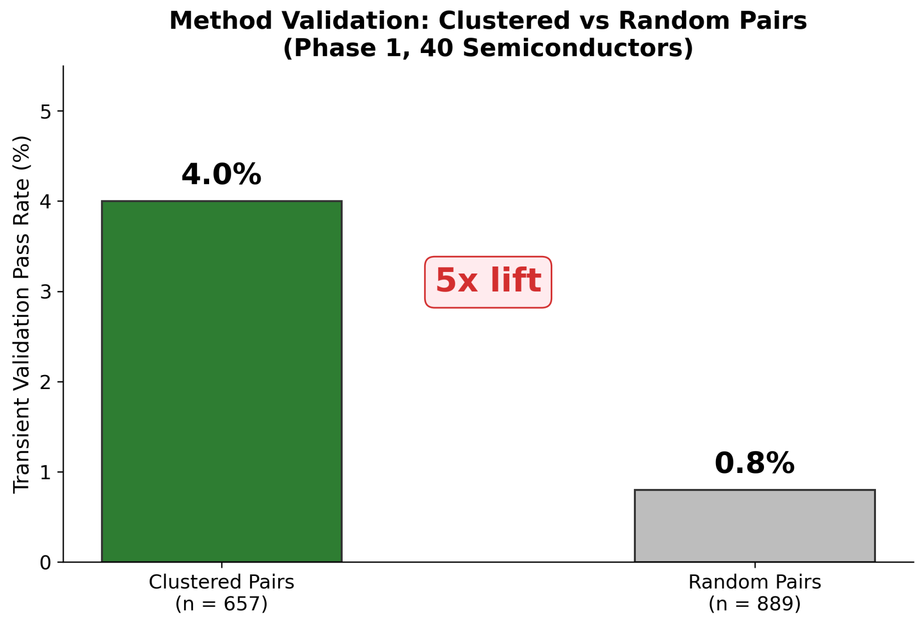

Figure 2: Phase 1 transient validation results. Clustered pairs achieve a 4.0% pass rate across 657 formation events, compared to 0.8% for 889 randomly generated pairs, a 5× lift. This confirms that OPTICS clustering identifies pairs with genuine short-term mean-reverting properties rather than statistical artifacts.

Classical cointegration testing was also applied to the semiconductor pairs: 0 of 6 stable pairs passed the Engle/Granger test, confirming that these transient relationships are invisible to methods requiring permanent equilibrium. This failure directly motivated the development of the five-test validation framework used in Phase 2.

During this phase, we also identified an error in the initial implementation. The earlier version estimated hedge ratios on the full dataset and evaluated performance over the same period, which overstated results and produced 23/23 profitable pairs. After introducing a proper calibration window with a strictly separate out-of-sample test period, profitability declined to 14/26 pairs. This reinforced the need for a clear separation between calibration and testing periods.

4 Clustering Comparison

4.1 K-Means Clustering

K-means clustering is a sorting algorithm that divides data points into k distinct, non-overlapping groups based on the closest cluster centroid using Euclidean distance. The optimal k value can be determined using various methods, such as the elbow test, which plots different k values against the Within-Cluster Sum of Squares (WCSS). This produces a graph with an "elbow" or bend, which represents the optimal number of centroid groups for clustering. K-means clustering is very prone to being affected by noise and outliers because it considers every single data point and attempts to position a centroid to best accommodate them.

For this project, a K-means clustering algorithm was applied to the nine-feature representation: returns, short-term volatility (50h), beta to S&P (50h), beta to sector (50h), RSI (70h), 5-hour momentum, and three regime shift indicators (volatility, S&P beta, and sector beta regime shifts computed as the normalized difference between short and medium windows). The idea is to identify pairs or groups of stocks that constantly overlap within the same clusters for these various factors, and potentially identify a trading relationship that could be validated using the statistical testing framework described in Section 7.

4.2 DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a density-based clustering algorithm that groups points based on the density of their local neighborhoods rather than distance to centroids. The algorithm requires two parameters: ϵ (eps), which defines the radius of a point's neighborhood, and `min_samples`, the minimum number of points required within that radius to form a dense region. A point is classified as a *core point* if it has at least `min_samples` neighbors within its ϵ-neighborhood. Clusters are formed by connecting chains of density-reachable core points, while points that cannot be reached from any core point are labeled as noise. Unlike K-means, DBSCAN does not require specifying the number of clusters in advance and can discover clusters of arbitrary shape, making it well-suited for detecting non-spherical groupings in feature space.

In this project, DBSCAN was applied to the same nine-feature representation used across all clustering methods, with `min_samples` set to 3 and an adaptive ϵ selection strategy. For each hourly timestamp, the algorithm computes a k-distance graph and evaluates candidate ϵ values at the 50th, 60th, 70th, and 80th percentiles, selecting the value that maximizes cluster count while maintaining a moderate noise rate (15–50%). This adaptive approach addresses DBSCAN's well-known sensitivity to the ϵ parameter. Applied to the semiconductor universe, DBSCAN produced an average of 1.47 clusters per timestamp with a 28% noise rate, and achieved the highest out-of-sample stability among the three algorithms with a train-test correlation of 0.840.

4.3 OPTICS

OPTICS (Ordering Points To Identify the Clustering Structure) extends DBSCAN by eliminating the need for a fixed ϵ parameter. Instead of producing a flat partition, OPTICS constructs a reachability plot that captures the clustering hierarchy across all density levels. Each point is assigned a *reachability distance*, the minimum radius needed to density-connect it to the growing cluster ordering. Valleys in the reachability plot correspond to dense clusters, while peaks indicate transitions between clusters or noise. Clusters are extracted using a steepness parameter ξ, which identifies significant drops in reachability distance. This hierarchical approach allows OPTICS to simultaneously detect clusters at different density scales, making it particularly effective when the data contains regions of varying density.

For this project, OPTICS was configured with `min_samples` $= 3$, $\xi = 0.05$, and a minimum cluster size of 3, applied to the same PCA-reduced nine-feature space. OPTICS produced an average of 3.36 clusters per timestamp, more than double DBSCAN's 1.47, reflecting its ability to resolve finer-grained density structure. However, this came at the cost of a substantially higher noise rate of 58.46%, as OPTICS applies stricter density requirements and only assigns points to clusters when they fall within clearly defined reachability valleys. Despite the higher noise rate, OPTICS identified the most transient pairs of any algorithm (681 vs. 549 for K-means and 474 for DBSCAN), suggesting it is the most sensitive to short-lived clustering relationships. Its out-of-sample correlation of 0.665, while lower than DBSCAN (0.840), still confirms that the discovered patterns are statistically meaningful rather than artifacts of noise.

4.4 Comparative Analysis

All three clustering algorithms were evaluated on the Phase 1 semiconductor universe (40 equities, 1,579 hourly timestamps) using identical nine-feature inputs and PCA preprocessing. Table 2 summarizes the key performance metrics.

The three algorithms exhibit a clear trade-off between sensitivity and stability. OPTICS identified the most transient pairs (681) and actionable formation events (20,242), but at the cost of the highest noise rate (58.5%) and lowest out-of-sample stability ($r = 0.665$). DBSCAN occupied the opposite end of this spectrum, producing fewer but more stable clusters with the highest out-of-sample correlation ($r = 0.840$). K-means provided a middle ground across all metrics, with moderate noise (10%) and good stability ($r = 0.735$).

Table 2: Clustering Algorithm Performance Comparison

Metric	K-Means	DBSCAN	OPTICS
Valid Windows	1,578 (99.9%)	1,578 (99.9%)	1,364 (96.1%)
Avg. Clusters/Timestamp	3.09	1.47	3.36
Avg. Noise Rate	10.0%	28.0%	58.5%
Actionable Formations	18,829	17,507	20,242
Transient Pairs Found	549	474	681
Stable Candidate Pairs	231	306	6
OOS Correlation	0.735	0.840	0.665

Despite these differences, all three algorithms converged on the same eight consensus pairs among their respective top-20 most frequently co-clustered pairs: ADI/NXPI, ADI/SWKS, ADI/TXN, AMAT/KLAC, AMAT/LRCX, KLAC/LRCX, NXPI/TXN, and QRVO/SWKS. This convergence across different clustering approaches (centroid-based, flat density-based, and hierarchical density-based) suggests these pairs represent genuine structural relationships rather than algorithmic artifacts.

Based on average ranking across all metrics, OPTICS was selected as the primary clustering algorithm for Phase 2 cross-sector scaling (average rank 1.50 vs. 2.00 for K-means and 2.50 for DBSCAN). Its higher sensitivity to transient relationships aligned with the project's core objective of detecting short-lived co-movements, and its hierarchical density approach naturally accommodates the varying density structures expected when scaling from 40 to 142 equities across five sectors.

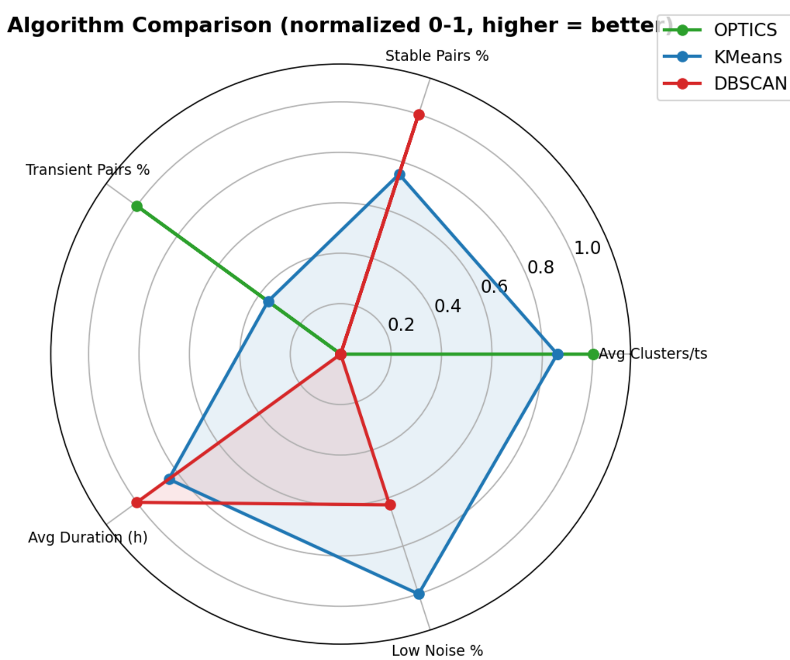

Figure 3: Normalized radar chart comparing K-Means, DBSCAN, and OPTICS across five clustering performance metrics. OPTICS dominates transient pair discovery while K-Means achieves the lowest noise rate.

5 Asset Selection

The cross-sector universe was constructed through a three-stage screening pipeline designed to select liquid, institutionally investable equities suitable for hourly-frequency pairs trading.

Stage 1: Fundamental Screen. Using the `yfscreen` library for programmatic Yahoo Finance screening, equities were filtered by: market capitalization above $2 billion, average daily trading volume exceeding 5 million shares, share price above $5, positive trailing EBITDA, and US-domiciled listings only. These thresholds ensure sufficient liquidity for realistic trade execution and exclude micro-cap stocks with unreliable price data.

Stage 2: Sector Selection. Screens were executed separately for five major GICS sectors: Technology, Healthcare, Energy, Financial Services, and Industrials. Running the screen on a sector-by-sector basis prevents overrepresentation of a single industry and ensures adequate cross-sectional diversity for clustering analysis. The screened results from each sector were then pooled into a combined universe, allowing both intra-sector and cross-sector pair discovery.

Stage 3: Data Quality Filter. Hourly price data was fetched for all candidates passing the fundamental screen. Tickers were removed if they had less than 80% price coverage over the sample period or more than 5% zero-volume hours, as gaps in hourly data would distort clustering and spread calculations.

The final universe comprised 142 tickers: 59 Technology, 27 Energy, 23 Healthcare, 18 Industrials, and 15 Financial Services. This produced $\binom{142}{2} = 10{,}011$ unique pairwise combinations, of which 3,643 exceeded the 8% noise-adjusted co-clustering frequency threshold.

6 Validation

6.1 Co-integration Failure

Our initial validation followed the traditional pairs trading framework based on cointegration. After clustering candidate pairs, we applied the Engle-Granger approach by estimating the long-run regression and testing the residual spread for stationarity using the Augmented Dickey-Fuller (ADF) test. Rejection of the null hypothesis indicates stationary residuals and therefore supports cointegration.

In Phase 1, cointegration tests on 40 semiconductor equities produced a 0% pass rate (0 of 6 stable pair candidates passed). Phase 2 abandoned cointegration testing entirely, proceeding directly to the transient-aware five-test framework.

This revealed a structural inconsistency. Cointegration assumes a long-run equilibrium relationship, while our clustering method identifies relationships that exist over short time frames. In hourly equity data, stable long-term relationships are uncommon. As a result, the validation framework required a revision.

6.2 Transient Correlations

We shifted our validation method from cointegration to short-term mean reversion, known as transient correlation. Validation became event-driven. Once a cluster formed, a three-window structure was applied (see Section 3 for specific duration parameters):

- **Execution Lag:** realistic trading delay
- **Calibration Window:** hedge ratio estimation
- **Exploitation Window:** spread evaluation

Within the exploitation window, we measured:

- AR(1) coefficient (ϕ)
- Lag-1 autocorrelation
- Half-life of mean reversion
- Bounce rate from z-score extremes

The AR(1) coefficient was estimated by applying:

$$x_t = c + \phi x_{t-1} + \epsilon_t \tag{2}$$

where x_t denotes the spread at time t.

The half-life was estimated by regressing the first difference of the spread on its lagged level:

$$\Delta s_t = \alpha + \beta s_{t-1} + \epsilon_t \tag{3}$$

The half-life of mean reversion was then calculated as:

$$t_{1/2} = -\frac{\ln(2)}{\beta}, \qquad \beta < 0 \tag{4}$$

The goal was no longer to identify long-term stationarity, but to ensure that mean reversion occurred quickly enough to support trade execution within the clustering window.

6.3 Five-Test Validation Framework

While the transient correlation tests were effective for evaluating individual cluster events, they were not enough on their own to rank thousands of pairs. Therefore, we introduced a five-test validation framework. The score ranges from 0 to 5, and a pair scoring 3 or higher is interpreted as tradeable:

Test	**Criterion**
ADF	$p < 0.10$
Half-Life	5–60 days
Hurst Exponent	$H < 0.5$
Lo-MacKinlay Variance Ratio	Reject at 10%
Rolling Correlation	Stability > 0.5

Each test targets a distinct aspect of mean-reversion suitability:

- **ADF** ($p < 0.10$): Tests whether the spread is stationary, a prerequisite for mean-reversion trading.
- **Half-Life** (5–60 days): Measures how quickly the spread reverts to its mean. Too slow ties up capital; too fast suggests noise.
- **Hurst Exponent** ($H < 0.5$): Confirms anti-persistent, mean-reverting behaviour in the spread time series.
- **Lo-MacKinlay Variance Ratio** (reject at 10%): Rejects the random walk hypothesis in favour of mean-reversion.
- **Rolling Correlation** (stability > 0.5): Ensures the pair relationship is stable over time, not a one-time coincidence.

This framework replaced binary cointegration with a broader assessment of statistical strength across several reinforcing tests. The pass rates for tradability increased from 0% under strict cointegration to 59.0%.

6.4 Persistence and Transience

Certain pairs displayed clustering patterns that re-formed over time. Although continuous cointegration failed, these pairs consistently showed short-term mean-reverting behaviour. This suggests that long-run profitability arises not from a permanent equilibrium, but from recurring temporary equilibria. For example, semiconductor pairs like KLAC/LRCX may co-move for several hours during a sector-wide catalyst, dissolve as idiosyncratic factors dominate, then re-form days later during the next shared event. The trading opportunity lies not in any single formation, but in the statistical regularity of these recurrences.

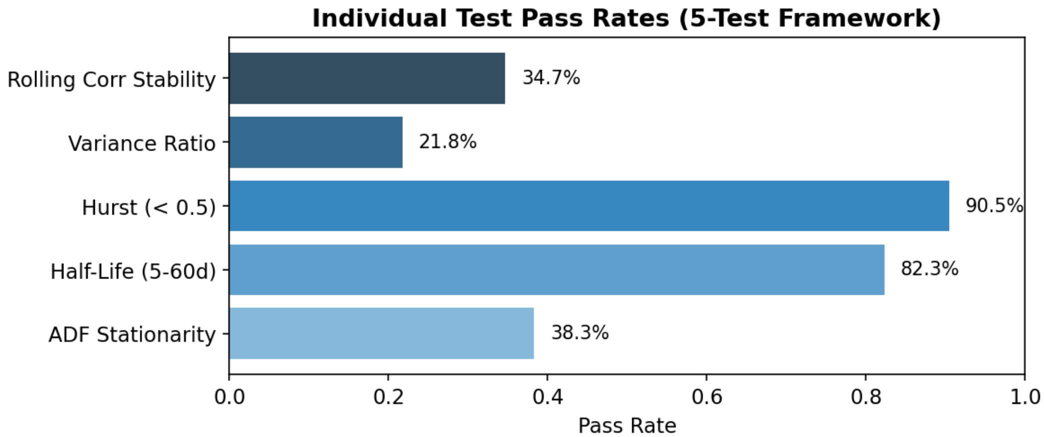

Figure 4: Individual pass rates for each of the five validation tests. Hurst exponent (90.5%) and half-life (82.3%) pass most frequently, while variance ratio (21.8%) and rolling correlation stability (34.7%) act as the primary filters.

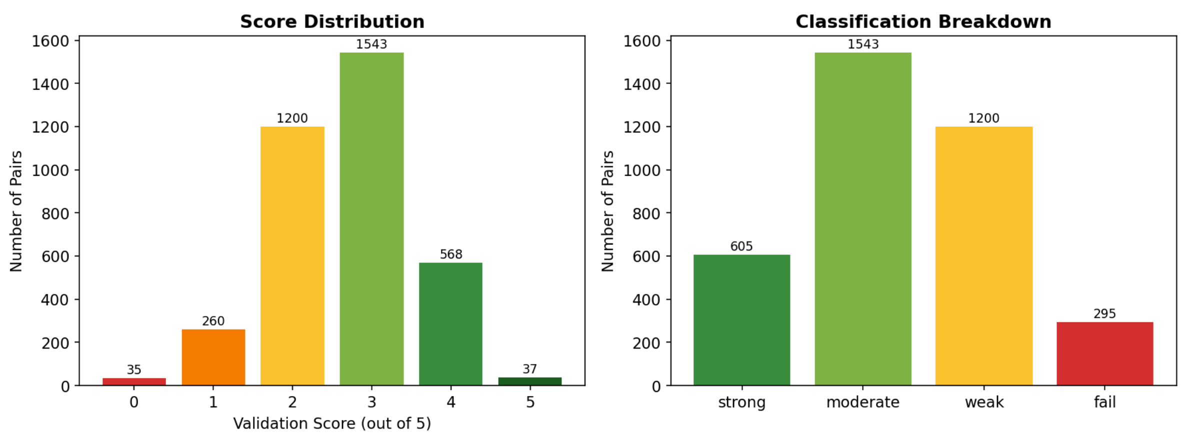

Figure 5: Distribution of validation scores across all 3,643 candidate pairs (left) and classification breakdown (right). Pairs scoring 3 or higher are deemed tradeable, yielding 605 strong and 1,543 moderate candidates.

OPTICS cluster formations lasted an average of 19.86 hours before dissolving, compared to 25.39 hours for K-Means and 26.42 hours for DBSCAN. The median formation duration across all algorithms was just 1 hour, as most formations are very brief, and the higher means are driven by a fat right tail of long-lived events. For example, the pair QRVO/SWKS co-clusters at a 38.3% noise-adjusted frequency, meaning the pair repeatedly enters and exits cluster membership rather than maintaining a permanent relationship. This transient-but-recurring pattern is precisely what the framework is designed to exploit.

7 Trading Strategy

Our project tested three strategies to trade valid pairs identified by the clustering pipeline. The tested strategies include transient event validation, classical pairs trading, and enhanced mean-reversion with Kalman filtering. After testing each strategy we concluded that the Enhanced Mean-Reversion with Kalman Filtering was the best option to implement into our framework.

7.1 Transient Event Validation

Transient Event Validation is not a deployable trading strategy but rather a method validation layer, designed to verify that the clustering engine identifies economically meaningful short-term relationships rather than statistical noise. It operates on hourly data and evaluates individual co-clustering formation events. Each co-clustering formation event is tested using a strict three-window structure consisting of a short execution lag, a calibration window for ordinary least squares (OLS) hedge ratio estimation, and a forward exploitation window for out-of-sample testing. Z-score entry signals are triggered at ± 2 standard deviations with exits at ± 0.5, and five validation criteria (correlation strength, spread stability, half-life of mean reversion, hedge ratio drift, and signal presence) must all be satisfied. While no transaction costs are applied, this framework provides strong statistical evidence that transient regime alignment produces genuine, short-lived mean-reversion opportunities.

7.2 Classical Pairs Trading

Classical Pairs Trading benchmarks the discovered pairs against traditional cointegration-based methodology using daily data. Price histories are divided into a 67% calibration and 33% out-of-sample split, with the hedge ratio estimated via OLS on calibration data only and held fixed during testing to avoid look-ahead bias. Pairs must satisfy three classical criteria (Engle-Granger cointegration at $p < 0.05$ on daily price levels, tradeable half-life, and a Hurst exponent below 0.5) to formally pass. Note that this cointegration threshold differs from the ADF test in the five-test framework (Section 6.3), which tests spread stationarity at $p < 0.10$ on a different data input. Although no pairs met all strict cointegration requirements, many near-miss pairs still generated positive out-of-sample performance, confirming that the clustering engine captures meaningful mean-reversion even when long-run equilibrium relationships are absent. This strategy therefore serves as a theoretical benchmark rather than a final production system.

7.3 Enhanced Mean-Reversion with Kalman Filtering

This strategy introduces the scored five-test validation framework, pair-specific optimization of z-score entry, exit, and lookback parameters through grid search on calibration data, and a Kalman-filtered hedge ratio. The Kalman filter exponentially weights recent observations, allowing the hedge ratio to adapt to structural drift while still extracting a terminal calibration estimate that is fixed in the out-of-sample period to prevent look-ahead bias. Unlike the earlier strategies, realistic transaction costs of 10 basis points per round-trip trade are incorporated. By combining adaptive hedging, optimized signal thresholds, and realistic cost modeling, this enhanced framework provides the most robust and deployable implementation.

7.4 Signal Generation

Trading signals are generated from the rolling z-score of the spread:

$$z_t = \frac{s_t - \bar{s}_t}{\hat{\sigma}_t} \tag{5}$$

where $\bar{s}_t$ and $\hat{\sigma}_t$ are the rolling mean and standard deviation over a lookback window. Entry occurs when $|z| \geq 2.0$ (long the spread if $z \leq -2.0$, short if $z \geq 2.0$), and positions are closed when $|z| \leq 0.5$. These serve as the baseline defaults; the enhanced and Kalman strategies optimize these parameters on a per-pair basis via grid search, as described in Section 7.6.

7.5 Hedge Ratio Estimation

Two hedge ratio methods were compared[1]:

- **OLS (Baseline):** The hedge ratio is estimated by ordinary least squares regression of P^A on P^B over the calibration window. This produces a static β applied throughout the out-of-sample period.

[1] A Total Least Squares (TLS) hedge ratio estimator was also implemented in the codebase but was not included in the final backtest comparison, as the Kalman filter consistently outperformed both OLS and TLS on calibration data.

- **Kalman Filter (Enhanced):** A Kalman filter is run on calibration data to estimate a time-varying hedge ratio. The terminal β from the calibration period is then frozen and applied to the out-of-sample period, avoiding look-ahead bias while capturing the most recent relationship dynamics.

We ultimately implemented the Kalman-filtered hedge ratio because it provides a more realistic and adaptive estimate of the relationship between paired stocks. Unlike OLS, which assigns equal weight to all historical observations and produces a static average hedge ratio, the Kalman filter exponentially weights recent data more heavily, allowing it to respond to structural drift and evolving market conditions. This results in a tighter and more accurately hedged spread at the start of the out-of-sample period. Importantly, only the terminal calibration beta is used and then fixed during testing, preventing look-ahead bias while still benefiting from the filter's adaptive properties. This balance between responsiveness and methodological rigor makes the Kalman approach better suited for live deployment in dynamic markets.

7.6 Backtest Design

Each pair's daily price history was split into a 67% calibration window and 33% out-of-sample (OOS) evaluation window. The hedge ratio was estimated exclusively on calibration data. The baseline strategy used static parameters: $z_{\text{entry}} = 2.0$, $z_{\text{exit}} = 0.5$, and a lookback window of 20 days, with no transaction costs.

The enhanced backtest introduced grid-searched z-score parameters optimizing over entry thresholds $z_{\text{entry}} \in [1.0, 2.5]$, exit thresholds $z_{\text{exit}} \in [0.0, 0.75]$, and lookback windows $\in [10, 15, 20]$ days. Realistic transaction costs of 10 basis points per round-trip trade were applied.

Walk-forward validation employed 5 rolling calibration/OOS splits, each using 80% of the data as the window size. Each split produced an independent OOS evaluation, and results were aggregated to assess robustness across different market regimes.

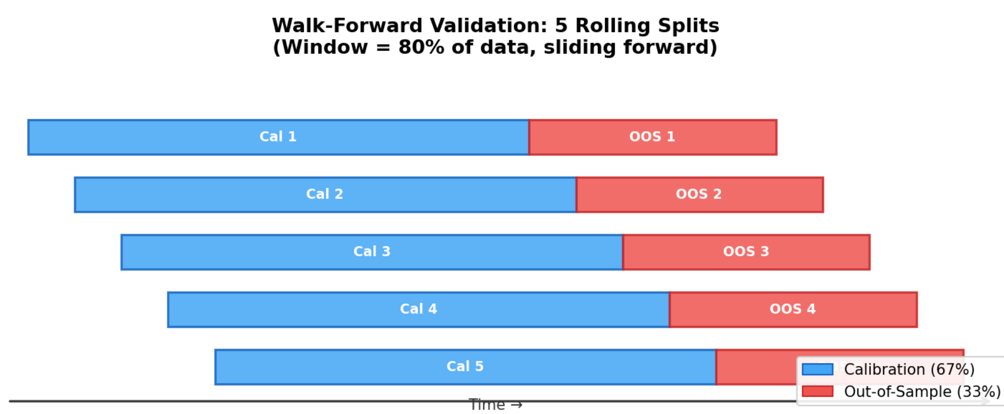

Figure 6: Walk-forward validation structure. The dataset is divided into 5 rolling splits, each using 80% of observations. Within each split, the first 67% serves as the calibration window (hedge ratio estimation and z-score parameter optimization) and the remaining 33% as the out-of-sample evaluation period.

8 Results

8.1 Phase 1: Semiconductor Proof of Concept

OPTICS clustering on 40 semiconductor equities produced 20,242 cluster formation events across 776 unique pairs. The top pairs by noise-adjusted co-clustering frequency were QRVO/SWKS (0.383), KLAC/LRCX (0.364), AMAT/LRCX (0.361), and AMAT/KLAC (0.326). Eight consensus pairs appeared in the top 20 across all three clustering algorithms (OPTICS, K-Means, and DBSCAN): ADI/NXPI, ADI/SWKS, ADI/TXN, AMAT/KLAC, AMAT/LRCX, KLAC/LRCX, NXPI/TXN, and QRVO/SWKS.

To validate that clustering identifies real market structure rather than statistical noise, we applied backward-looking transient event validation. Of 657 formation events tested, 26 passed all validation criteria (4.0% pass rate), compared to a 0.8% pass rate for randomly generated pairs, a 5x lift. Passed events exhibited an average simulated P&L of $1.94 and an average win rate of 76.9%, confirming that the clustering pipeline detects genuine short-term relationships.

For trading, 26 near-miss pairs from the Phase 1 semiconductor universe (meeting 2 of 3 classical criteria: half-life, Hurst exponent, but rarely cointegration) were backtested out of sample. A z-score mean reversion strategy produced 14 profitable pairs (54%), with a top Sharpe ratio of 3.55.

8.2 Phase 2: Cross-Sector Scaling

Scaling to 142 equities across five sectors, the OPTICS clustering pipeline identified 3,643 pairs exceeding the 8% noise-adjusted co-clustering frequency threshold. The five-test scored validation framework classified these pairs as follows: 605 strong (16.6%), 1,543 moderate (42.3%), 1,200 weak (32.9%), and 295 fail (8.1%). In total, 2,148 pairs (59.0%) scored 3 or higher and were deemed tradeable.

Intra-sector pairs outperformed cross-sector pairs on average. Of the 1,442 intra-sector pairs, 925 were tradeable with an average score of 2.77 and a strong-pair rate of 19.6%. The 2,201 cross-sector pairs produced 1,223 tradeable candidates with an average score of 2.62 and a strong-pair rate of 14.7%. Technology dominated in raw pair count (880 intra-sector pairs), while Healthcare exhibited the highest pass rate at 19.2%.

Permutation testing confirmed statistical significance for 940 of 3,539 pairs tested (26.6%, $Z > 1.96$). The most statistically significant pairs by sector included CVX/XOM ($Z = 13.0$) in Energy, DAL/UAL ($Z = 11.5$) in Industrials, ASX/TSM ($Z = 13.1$) in Technology, and MRK/PFE ($Z = 9.9$) in Healthcare.

8.3 Noise-Adjusted Co-Clustering Frequency

A key methodological contribution of this work is the noise-adjusted co-clustering frequency metric. OPTICS assigns approximately 58% of hourly timestamps to noise, meaning points that fall outside any detected cluster. Naively dividing a pair's co-cluster count by the total number of timestamps therefore massively underestimates true clustering affinity, since the denominator includes timestamps where one or both tickers were classified as noise and thus had no opportunity to co-cluster.

The noise-adjusted frequency corrects for this by restricting the denominator to timestamps where both tickers were assigned to a non-noise cluster:

$$f_{\text{adj}}(A, B) = \frac{|\{t : A_t = B_t,\ A_t \neq \text{noise}\}|}{|\{t : A_t \neq \text{noise}\} \cap \{t : B_t \neq \text{noise}\}|} \tag{6}$$

where A_t and B_t denote the cluster assignments of tickers A and B at timestamp t. Pairs exceeding an 8% noise-adjusted frequency threshold were retained as candidates, yielding 3,643 pairs from the Phase 2 cross-sector universe.

8.4 Backtest Performance

The baseline backtest employed OLS hedge ratios estimated on a 67% calibration window, with z-score entry at $|z| \geq 2.0$, exit at $|z| \leq 0.5$, and a 20-day rolling lookback. Under this configuration, 41% of all 2,148 tradeable pairs were profitable out of sample. Restricting attention to the top 50 pairs by composite validation score raised the baseline profitability to 54%, with an average Sharpe ratio of 1.73.

Two enhancements were applied sequentially. The **Enhanced** backtest retained OLS hedge ratios but introduced grid-searched z-score parameters, optimizing entry thresholds (1.0–2.5), exit thresholds (0.0–0.75), and lookback windows (10–20 days), along with realistic transaction costs of 10 basis points per trade. Under this configuration, profitability rose to 57% of top pairs.

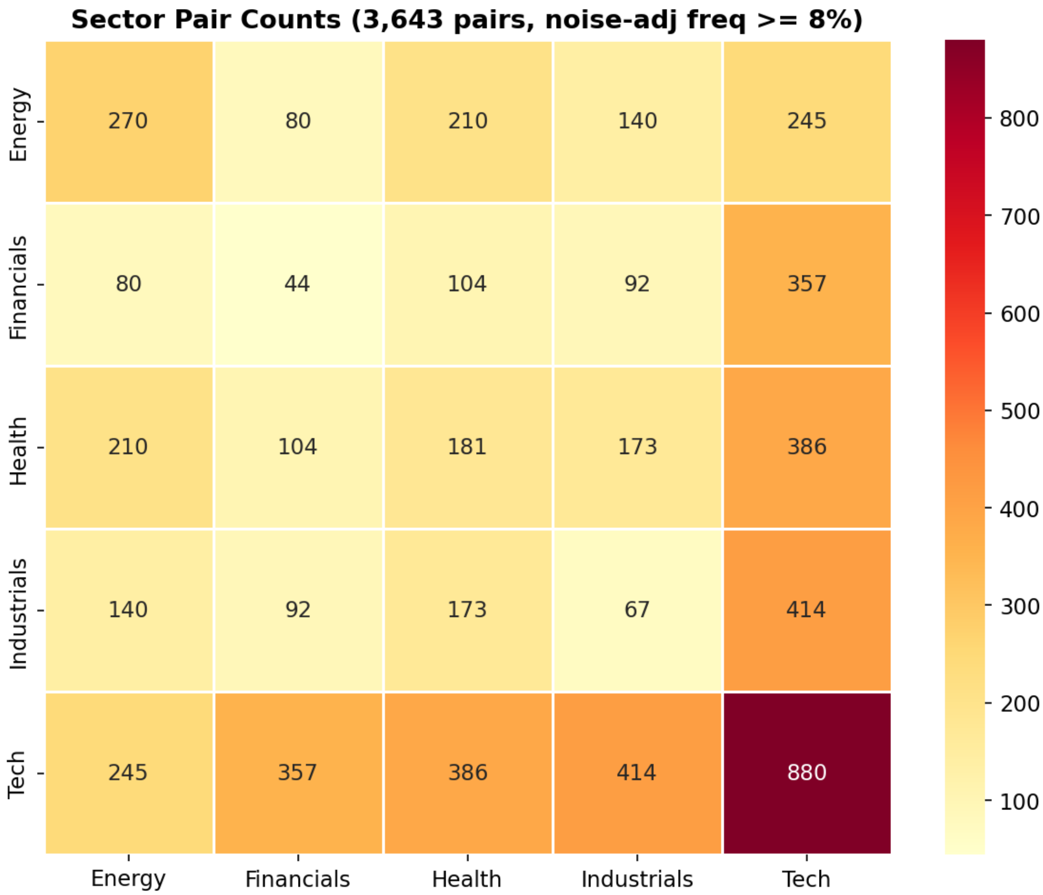

Figure 7: Heatmap of tradeable pair counts by sector pairing. Intra-sector pairs (diagonal) exhibit higher average validation scores, while Technology/Industrials and Energy/Healthcare cross-sector pairings also produce meaningful clusters.

The **Kalman** backtest replaced the static OLS hedge ratio with a Kalman-filtered estimate. The filter was run on calibration data only and the terminal beta frozen for out-of-sample evaluation, eliminating look-ahead bias. Combined with the same optimized z-score parameters and 10 bps costs, this raised profitability to 64% of top pairs.

Walk-forward validation across five rolling calibration/out-of-sample splits confirmed robustness, with top performers including RUN/VNET (Sharpe 3.47), NVDA/ORCL (Sharpe 2.59), and ADT/NOW (Sharpe 2.56).

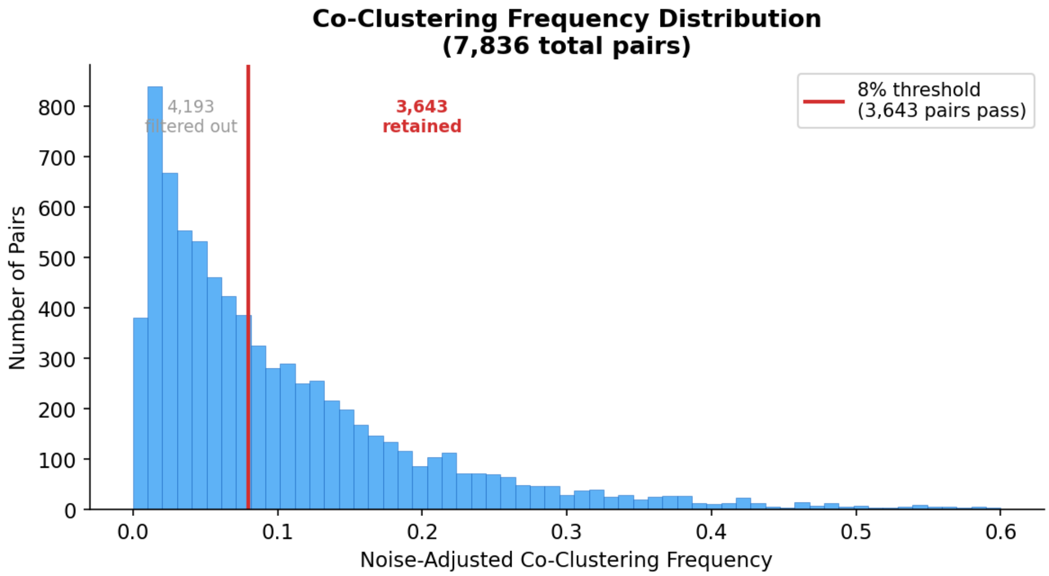

Figure 8: Distribution of noise-adjusted co-clustering frequency across the 7,836 pairs that co-clustered at least once. The 8% threshold (red line) separates the 3,643 candidate pairs (right of threshold) from the 4,193 pairs filtered out.

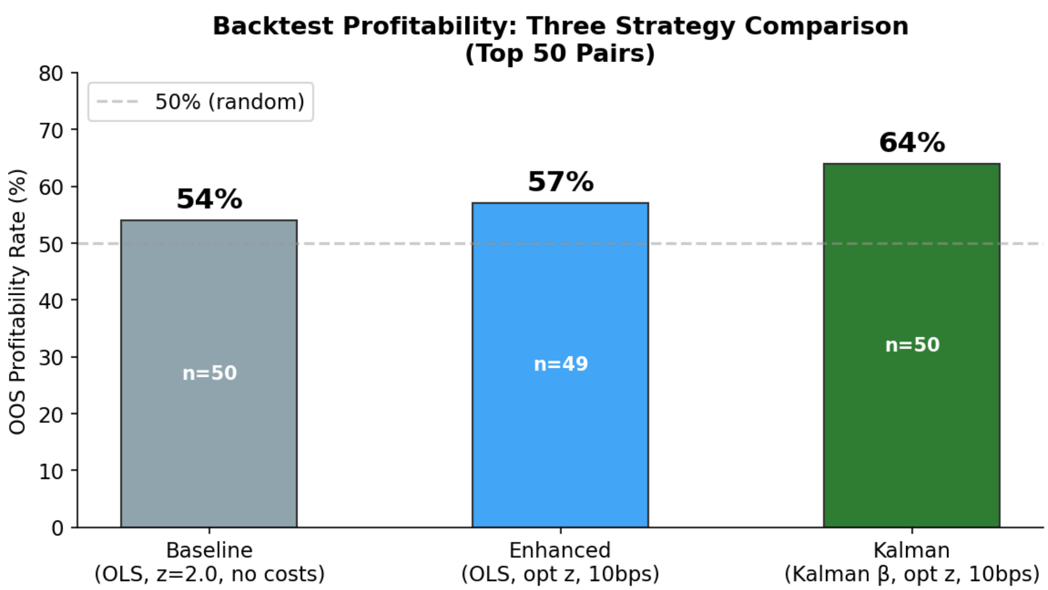

Figure 9: Profitability comparison across the three backtest strategies on the top 50 pairs by validation score. The baseline strategy (OLS hedge, static $z = 2.0$, no costs) achieves 54% profitability on this subset. The enhanced strategy (OLS hedge, optimized z-score parameters, 10 bps costs) reaches 57%. The Kalman strategy (Kalman-filtered hedge ratio with frozen terminal beta, optimized z-score, 10 bps costs) reaches 64%.

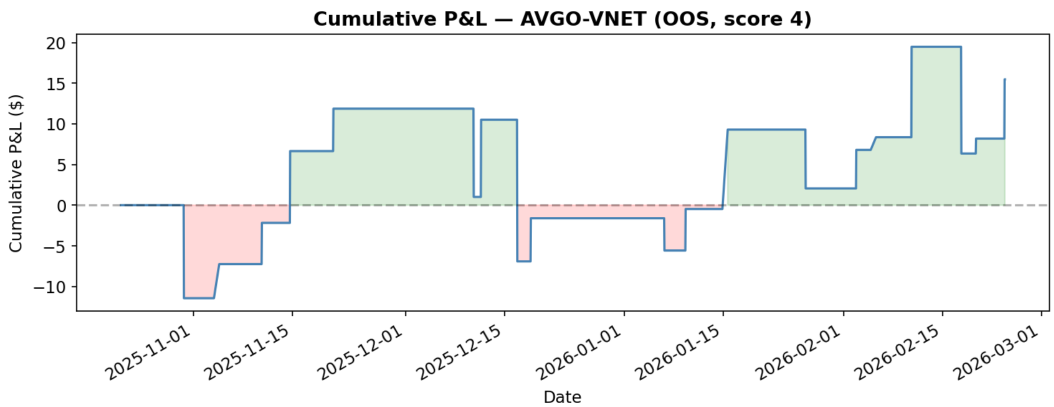

Figure 10: Cumulative profit and loss for the top-performing pair AVGO/VNET under the enhanced Kalman-filtered backtest. The strategy captures mean-reversion opportunities with consistent positive returns across the out-of-sample period.

9 Conclusions

This project set out to answer two questions: whether unsupervised clustering can detect real short-term relationships between equities, and whether those relationships can be traded profitably. The evidence supports affirmative answers to both.

The 5x lift in validation pass rate over randomly generated pairs in Phase 1 suggests that OPTICS clustering captures genuine market structure. This is not an artifact of the testing framework: the same validation criteria applied to random pairs yield a pass rate five times lower, confirming that persistent co-clustering identifies pairs with real mean-reverting properties during formation events.

The failure of cointegration testing (0% pass rate) on clustering-derived pairs is not a weakness of the framework but an expected outcome. Cointegration requires a permanent equilibrium relationship, which is fundamentally incompatible with transient correlations that form and dissolve over short time horizons. The five-test scored validation framework, designed specifically for transient data, proved far more appropriate, identifying 59% of candidate pairs as tradeable compared to 0% under strict cointegration.

The enhanced backtest demonstrates that careful strategy design meaningfully improves trading outcomes. The backtest results show a clear progression: the baseline strategy achieved 41% profitability across all 2,148 tradeable pairs. Restricting to the top 50 pairs by validation score, the enhanced strategy (optimized z-score parameters, 10 bps transaction costs) achieved 57% profitability, and the Kalman strategy (Kalman-filtered hedge ratios with terminal beta frozen to prevent look-ahead bias) further improved this to 64%. The improvement from Kalman filtering is particularly noteworthy: by estimating a time-varying hedge ratio on calibration data and freezing the terminal value for out-of-sample use, the approach captures dynamic relationships without introducing look-ahead bias.

Several limitations should be acknowledged. First, the 5x lift validation was conducted only in Phase 1 on semiconductor equities; Phase 2 did not repeat this backward-looking transient event analysis, so the method-validation evidence does not directly extend to the cross-sector universe. Second, all results are based on historical backtesting; a real-time implementation would require detecting cluster formations as they occur without the benefit of hindsight, introducing additional challenges around latency and look-ahead bias removal. Third, the enhanced backtest optimizes z-score parameters on calibration data, which may overfit to the specific calibration period despite walk-forward validation.

Future work should focus on three directions. First, the Phase 1 transient event validation showed that individual cluster formations can produce mean-reverting spreads with high win rates over short horizons (76.9% average win rate across passed events). A natural extension would be building a real-time clustering pipeline that detects these formations as they occur and generates trade signals before the relationship dissolves. Whether these results hold under live execution conditions, including latency, slippage, and partial fills, remains an open question, but the backward-looking evidence suggests it is worth investigating. Second, extending the transient event validation to the full cross-sector universe would help determine whether the 5x lift observed in semiconductors generalizes to other sectors. Third, paper trading or live simulation on the daily mean-reversion strategy would provide a more realistic assessment of the framework's profitability beyond historical backtesting. The framework presented here establishes that transient correlation detection via unsupervised clustering is a viable and promising alternative to cointegration-based pairs trading, particularly in equity markets where structural relationships are short-lived and regime-dependent.

References

[1] Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel, and Jörg Sander. Optics: Ordering points to identify the clustering structure. *ACM SIGMOD Record*, 28(2):49–60, 1999.

[2] David A. Dickey and Wayne A. Fuller. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 74(366):427–431, 1979.

[3] Robert F. Engle and Clive W. J. Granger. Co-integration and error correction: Representation, estimation, and testing. *Econometrica*, 55(2):251–276, 1987.

[4] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pages 226–231, 1996.

[5] Evan Gatev, William N. Goetzmann, and K. Geert Rouwenhorst. Pairs trading: Performance of a relative-value arbitrage rule. *The Review of Financial Studies*, 19(3):797–827, 2006.

[6] Vanshika Gupta, Vineet Kumar, Yuvraj Yuvraj, and Manoj Kumar. Optimized pair trading strategy using unsupervised machine learning. In *2023 IEEE 8th International Conference for Convergence in Technology (I2CT)*. IEEE, 2023. `https://ieeexplore.ieee.org/document/10126289/`.

[7] Eli Hadad, Sohail Hodarkar, Beakal Lemeneh, and Dennis Shasha. Machine learning-enhanced pairs trading. *Forecasting*, 6(2):434–455, 2024.

[8] Chulwoo Han, Zhaodong He, and Alenson Jun Wei Toh. Pairs trading via unsupervised learning. *European Journal of Operational Research*, 307(2):929–947, 2023.

[9] Hudson and Thames. ML based pairs selection — arbitragelab 1.0.0 documentation. Hudson and Thames ArbitrageLab Documentation, 2020. Accessed 2026. `https://hudson-and-thames-arbitragelab.readthedocs-hosted.com/en/latest/ml_approach/ml_based_pairs_selection.html`.

[10] Hudson and Thames. The comprehensive introduction to pairs trading. Hudson and Thames Research, 2021. Accessed 2026. `https://hudsonthames.org/definitive-guide-to-pairs-trading/`.

[11] Harold Edwin Hurst. Long-term storage capacity of reservoirs. *Transactions of the American Society of Civil Engineers*, 116:770–799, 1951.

[12] Rudolf E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35–45, 1960.

[13] Christopher Krauss. Statistical arbitrage pairs trading strategies: Review and outlook. *Journal of Economic Surveys*, 31(2):513–545, 2017.

[14] Bo Liu, Lo-Bin Chang, and Hélyette Geman. Intraday pairs trading strategies on high frequency data: The case of oil companies. *Quantitative Finance*, 17(1):87–100, 2017.

[15] Andrew W. Lo and A. Craig MacKinlay. Stock market prices do not follow random walks: Evidence from a simple specification test. *The Review of Financial Studies*, 1(1):41–66, 1988.

[16] James MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297, 1967.

[17] Rekhit Pachanekar. DBSCAN vs. K-means. QuantInsti Blog; reposted on IBKR Quant News, 2025. `https://www.interactivebrokers.com/campus/ibkr-quant-news/dbscan-vs-k-means/`.

[18] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[19] QuantConnect. PCA and pairs trading. QuantConnect Documentation, n.d. Accessed 2026. `https://www.quantconnect.com/docs/v2/research-environment/applying-research/pca-and-pairs-trading`.

[20] QuantStart. Backtesting an intraday mean reversion pairs strategy between SPY and IWM. QuantStart, 2021. Updated December 2021. Accessed 2026. `https://www.quantstart.com/articles/Backtesting-An-Intraday-Mean-Reversion-Pairs-Strategy-Between-SPY-And-IWM/`.

[21] QuantStart. Cointegrated augmented Dickey-Fuller test for pairs trading evaluation in R. QuantStart, n.d. Accessed 2026. `https://www.quantstart.com/articles/Cointegrated-Augmented-Dickey-Fuller-Test-for-Pairs-Trading-Evaluation-in-R/`.

[22] Simão Moraes Sarmento and Nuno Horta. Enhancing a pairs trading strategy with the application of machine learning. *Expert Systems with Applications*, 158:113490, 2020.

[23] statsmodels developers. statsmodels.tsa.stattools.adfuller — augmented Dickey-Fuller unit root test. statsmodels 0.14.6 documentation, 2026. Accessed 2026. `https://www.statsmodels.org/stable/generated/statsmodels.tsa.stattools.adfuller.html`.

[24] statsmodels developers. statsmodels.tsa.stattools.coint — engle-granger cointegration test. statsmodels 0.14.6 documentation, 2026. Accessed 2026. `https://www.statsmodels.org/stable/generated/statsmodels.tsa.stattools.coint.html`.

[25] Xing Tao. Kalman filter techniques and statistical arbitrage in China's futures market in Python. QuantInsti Blog (EPAT Project), 2021. `https://blog.quantinsti.com/kalman-filter-techniques-statistical-arbitrage-china-futures-market-python/`.

[26] Chainika Thakar. Augmented Dickey-Fuller (ADF) test for a pairs trading strategy. QuantInsti Blog, 2022. `https://blog.quantinsti.com/augmented-dickey-fuller-adf-test-for-a-pairs-trading-strategy/`.

[27] R. W. J. van der Have. Pairs trading using machine learning: An empirical study. Master's thesis, Erasmus School of Economics, Erasmus University Rotterdam, 2017.

[28] Ganapathy Vidyamurthy. *Pairs Trading: Quantitative Methods and Analysis*. John Wiley & Sons, 2004.

[29] Qiqin Zhou. Application of Black-Litterman Bayesian in statistical arbitrage. arXiv:2406.06706 [q-fin.CP], 2024. `https://arxiv.org/abs/2406.06706`.