Pairs Trading

David Baines Josh Nielsen Peter Kinder

May 8, 2023

1 Problem Space

Given a machine learning algorithm's ability to effectively parse through voluminous, veracious, and variable dimensional data to derive effective data signals to make useful predictions, these powerful statistical methods have risen in popularity in quantitative finance, specifically in "stock picking." To effectively pick stocks for investment, the analyst must be able to separate the "signal from the noise" in the data, and financial data has a notoriously low signal-to-noise ratio given its extremely high dimensionality, deep sparsity, and variable feature sets (AQR Capital Management, 2019). They're able pick stocks effectively if they can adequately predict the price movement of the stock, typically in the short term, and will "long" (buy) stocks that they predict to go up in price in the future and "short" (sell) stocks they believe to go down in price in the future.

To achieve this goal, the analyst can group stocks together by common traits, both fundamental and technical in nature, and then trade them such that arbitrage profits are generated. (*Arbitrage* is the practice of buying two identical assets at different prices and profiting from their mispricing; it's the only theoretical "risk-free" investment opportunity that exists. While we and other papers use the term loosely, the idea is that these opportunities are ones ripe for investment exploitation.) Fundamental signals in these clusters would be measures in profitability, debt governance, industry classification, growth rates, or size; technical signals would be measures in past return behavior, trading volume, investor sentiment, analyst coverage, and trading spreads.

In this paper, we extend a stock selection framework proposed by Han, He & Toh (2022), wherein clustering algorithms are used to form these groups of stocks using both fundamental and technical features, after which a portfolio of stocks is created which is comprised of stocks which are materially deviating from the general behavior of the group; the assumption being that they will "regress to the mean" and re-assimilate into the group's general behavior in the future and thus an arbitrage opportunity is available. In a general sense, this strategy is called "pairs trading" because the analyst is assuming that two stocks which are fundamentally similar are going to converge in the future if they are currently diverged. In this paper, it is this strategy that we explore and then augment with a new clustering algorithm and portfolio creation optimization via hyperparameter tuning.

2 Approach

The approach we took for this project was largely based on a recent paper published in the *European Journal of Operational Research* entitled "Pairs Trading via Unsupervised Learning" by Han, He & Toh (2022). The paper explored the application of clustering to identify "pairs" of stocks which would form the foundation of a trading strategy. Specifically, the methods of k-means, agglomerative and density-based spatial clustering of applications with noise (DBSCAN) clustering were used to identify the clusters from which the pairs would be selected.

For the k-means clustering, the k-means++ algorithm was used to determine the k centroids, and outliers were removed if they were farther from their respective centroids than the median distance of nearest neighbors for the entire set. Additionally, a l_2 norm was the method of distance measurement for the k-means. Lastly, values of 5, 10, 50, 100, 500, 1000, and 1500 for k were tested.

For the agglomerative clustering, the maximum distance for clusters to be merged, known as the linkage distance, was the hyperparameter chosen instead of the number of clusters hyperparameter. To determine the linkage distance for data points, the average linkage was used and a l_1 norm was the method of distance measurement. To determine the maximum linkage distance that served as the hyperparameter, the percentile of the distances of nearest neighbors for the entire set were used. Lastly, the values of $0.1, \ldots, 0.9$ for percentiles for the maximum linkage distance were tested.

For the DBSCAN, a distance metric was also used as the sole hyperparameter. However, unlike k-means and agglomerative clustering, instead of a percentile of the just the nearest neighbors, a percentile of the average a certain number of nearest neighbors for each data point. This certain number, termed $MinPts$, was the natural logarithm of the total number of data points and was recommend by Ester et al. (2019), who proposed DBSCAN. Furthermore, a l_1 norm was the method of distance measurement used for DBSCAN. Lastly, the values of $0.1, \ldots, 0.9$ for percentiles of this distance metric were tested.

Once the clusters were determined, the stocks in each cluster were paired based on their previous month's return. The highest returning stock was paired with the lowest returning stock, the second highest returning stock was paired with the second lowest returning stock, and so on. Then, the standard deviation of all of the differences in monthly returns for each of the pairs and all of the clusters was used as a threshold. If the difference between a pair was greater than the computed standard deviation, then the pair was kept. All of the pairs that were kept were put into an equally balanced portfolio where the stocks that had high returns in the previous month would be shorted, and stocks that had low returns in the previous month would be long. All of the data used to make this determination was from the previous month (time $t-1$) and before, and current month's returns were used to calculate how the stocks performed at time t.

So far, what has been described in the approach was what we replicated in our approach. However, our approach added three significant changes:

- We explored hierarchical density-based spatial clustering of applications with noise (HDBSCAN)
- We considered different standard deviation threshold for portfolio formation
- We split the data into a train and test set

The HDBSCAN approach that we used "extends DBSCAN by converting it into a hierarchical clustering algorithm, and then using a technique to extract a flat clustering based in the stability of clusters" (McInnes, Astels & Healy, 2016). It has a similar hyper parameter as DBSCAN, and we utilized the same method used for DBSCAN to determine it.

For the standard deviation threshold, we considered different scalar values to multiply the standard deviation by. The intent was to see if being more or less inclusive in our portfolio construction has a significant effect on the results. The values that we tested for a scalar multiple of standard deviation were were 0.845, 1, 1.285, 1.645, 1.96, 2.325.

Finally, we split the data into a training and test set to determine the optimal hyperparameters for each clustering method, and their respective optimal standard deviation scalar multiple. From our understanding of the paper we replicated, their optimal hyperparameters were selected based on the entire dataset, which we believe would introduce look-ahead bias. We instead selected the hyperparameters based on the training set and evaluated the models' robustness on the test set. Lastly, the hyperparameters were chosen based on the average annualized Sharpe ratio of the resulting portfolios, which is essentially a measure of returns discounted by the standard deviation of the portfolio (Appendix: Figures 11, 12, & 13).

3 Data

For this project, we evaluated three dataset based with the intent to assess how the feature set may influence the results. The first dataset that we used was our best attempt to recreate the dataset used in the paper that we based our project on, which will be referred to as the Complete Dataset. The data set contained 48 return features and 78 firm characteristics (Appendix: Figure 4) chosen from Green, Hand, & Zhang (2017). The portfolio returns generated from the cluster were all in time t, the return features were from time $t-1$ to $t-48$ and the firm characteristics were from time $t-1$. The i-month return feature at the end of month $t-1$ is defined as the cumulative return from month $t-i$ to $t-2$ for $i > 1$

and as the previous one-month return for $i = 1$:

$$mom_i = r_{t-1}, \qquad i = 1,$$
$$mom_i = \prod_{j=t-i}^{t-2} (r_j + 1) - 1, \qquad i \in 2, \ldots, 48,$$

where r_j denotes the return in month j.

The second dataset was a scaled down version of this, and essentially the same as the paper recreation except for 40 firm characteristics (Appendix: Figure 5) instead of 78. The second dataset will be referred to as the Reduced Dataset. The third dataset was based on work done by Chen and Zimmerman (2021) and had the same return characteristics and 87 firm characteristics (Appendix: Figures 6 & 7). The third dataset will be referred to as the CZ Dataset.

Additionally, the data was generated every month for the sample period (time $t-1$) from December 1979 to November 2020. The out-of-sample period (time t) is from January 1980 to December 2020. The data was also normalized using cross-sectional means and standard deviations. Furthermore, PCA was performed for feature selection and to alleviate the curse of dimensionality. The number of principal components kept was such that 99% of the variance was explained. Lastly, we performed a 77.5 / 22.5 train test split, with 31 years of data ranging from January 1980 to December 2010 in the training set and 10 years of data ranging from January 2011 to December 2020 in the test set.

4 Results

For the Complete Dataset, the HDBSCAN and DBSCAN algorithms outperformed the Agglomerative and k-means algorithms in the training set based on annualized return and Sharpe ratio. However, the Agglomerative algorithm demonstrated the best out-of-sample performance with the highest annualized Sharpe ratio (4.118) and the lowest maximum drawdown (-0.008).

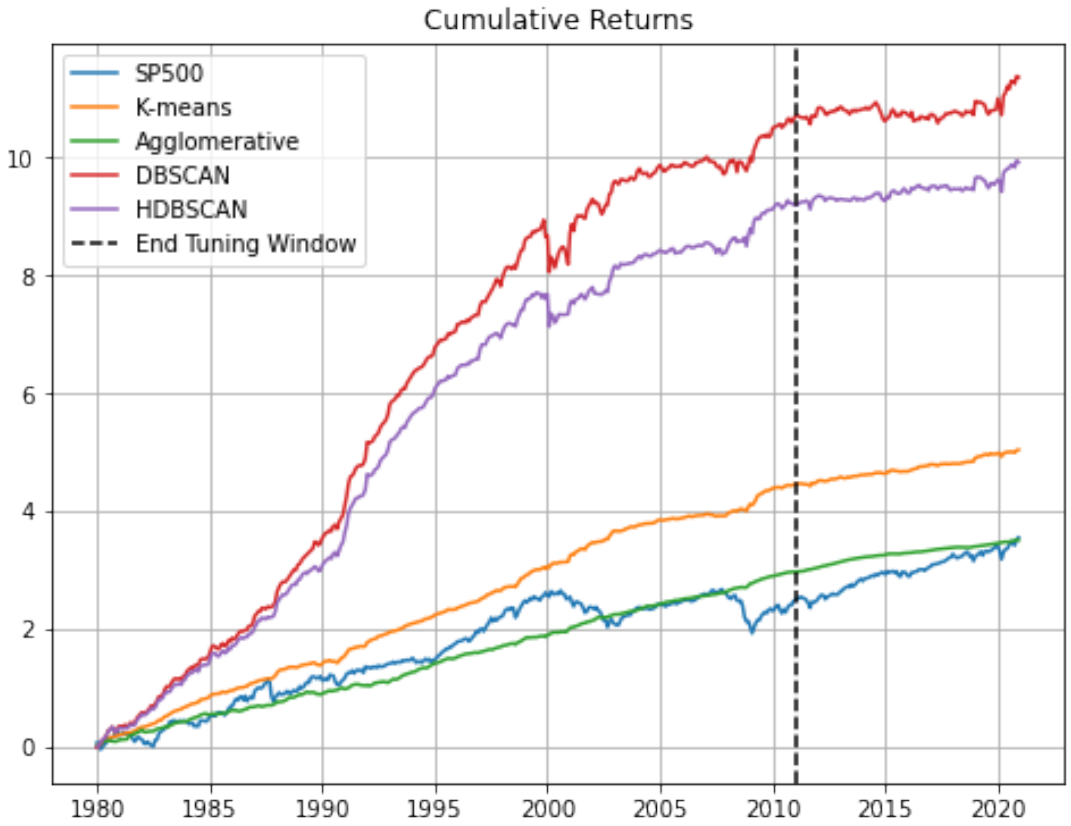


Figure 1: Complete Dataset Cumulative Returns

Table 1: Complete Dataset Training Data

	SP500	Agglomerative	k-means	HDBSCAN	DBSCAN
Monthly Return Mean	0.007648	0.008045	0.012130	0.026697	0.031735
Monthly Return Std	0.044965	0.010984	0.015914	0.056687	0.075582
Annualized Return	0.095736	0.100931	0.155671	0.371853	0.454850
Annualized Return Std	0.155765	0.038051	0.055127	0.196370	0.261825
Annualized Sharpe Ratio	0.614617	2.652510	2.823861	1.893633	1.737232
Maximum Drawdown	-0.525559	-0.047799	-0.045201	-0.438868	-0.587296

Table 2: Complete Dataset Testing Data

	SP500	Agglomerative	k-means	HDBSCAN	DBSCAN
Monthly Return Mean	0.009863	0.004530	0.005075	0.007419	0.008311
Monthly Return Std	0.039042	0.003907	0.015286	0.056584	0.070004
Annualized Return	0.124986	0.055732	0.062628	0.092749	0.104416
Annualized Return Std	0.135247	0.013533	0.052953	0.196011	0.242500
Annualized Sharpe Ratio	0.924137	4.118097	1.182708	0.473183	0.430581
Maximum Drawdown	-0.200011	-0.007535	-0.057256	-0.221200	-0.297706

In the case of the Reduced Dataset, the HDBSCAN algorithm exhibited the highest in-sample annualized Sharpe ratio (3.273), while the k-means algorithm achieved the highest annualized return (0.508). However, the out-of-sample performance favored the HDBSCAN algorithm with the highest Sharpe ratio (2.134) and the lowest maximum drawdown (-0.023).

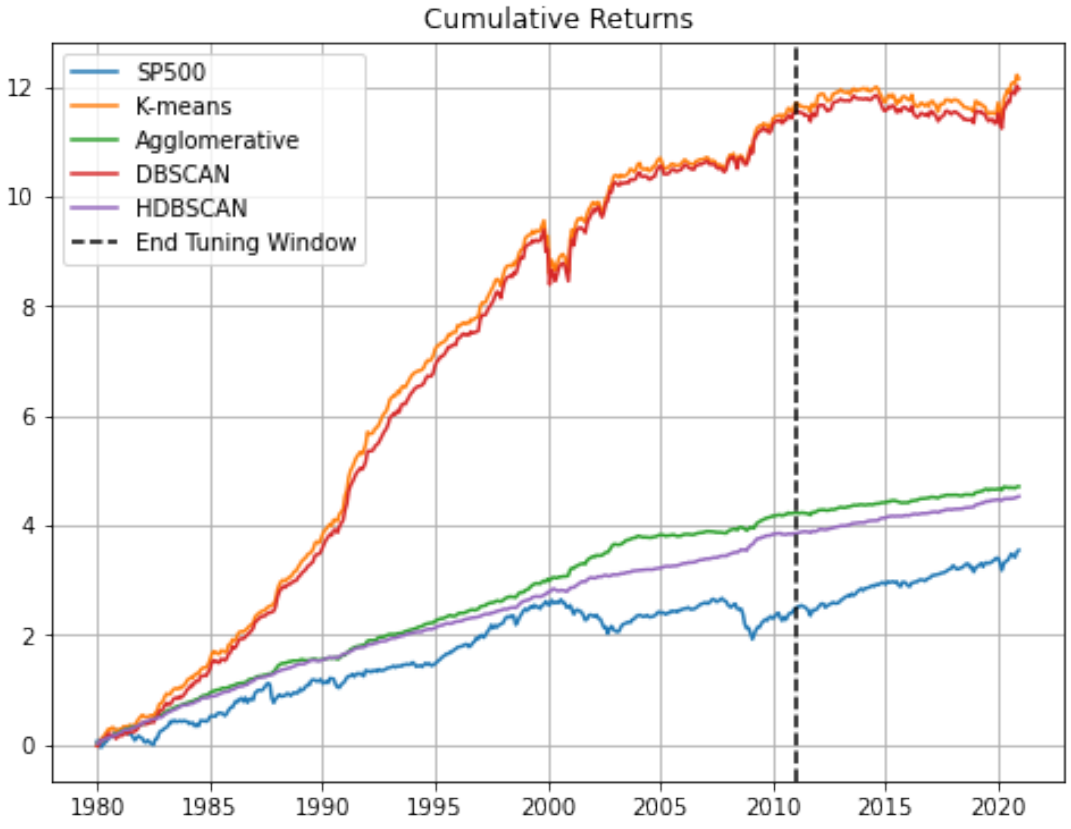


Figure 2: Reduced Dataset Cumulative Returns

Table 3: Reduced Dataset Training Data

	SP500	Agglomerative	k-means	HDBSCAN	DBSCAN
Monthly Return Mean	0.007648	0.011538	0.034851	0.010468	0.034517
Monthly Return Std	0.044965	0.016042	0.082017	0.011738	0.080876
Annualized Return Mean	0.095736	0.147587	0.508469	0.133109	0.502620
Annualized Return Std	0.155765	0.055570	0.284116	0.040663	0.280162
Annualized Sharpe Ratio	0.614617	2.655883	1.789654	3.273472	1.794033
Maximum Drawdown	-0.525559	-0.061990	-0.621369	-0.052673	-0.631209

Table 4: Reduced Dataset Testing Data

	SP500	Agglomerative	k-means	HDBSCAN	DBSCAN
Monthly Return Mean	0.009863	0.004099	0.007110	0.005678	0.006138
Monthly Return Std	0.039042	0.013035	0.075117	0.009512	0.073128
Annualized Return Mean	0.124986	0.050306	0.088738	0.070307	0.076188
Annualized Return Std	0.135247	0.045156	0.260212	0.032951	0.253323
Annualized Sharpe Ratio	0.924137	1.114070	0.341021	2.133641	0.300756
Maximum Drawdown	-0.200011	-0.041841	-0.435755	-0.022939	-0.453244

For the CZ Dataset, the k-means algorithm showed the best in-sample performance with the highest annualized return (0.228) and Sharpe ratio (2.966). In the testing set, the Agglomerative algorithm performed better with a higher annualized return (0.122) and Sharpe ratio (1.313) compared to other unsupervised learning methods.

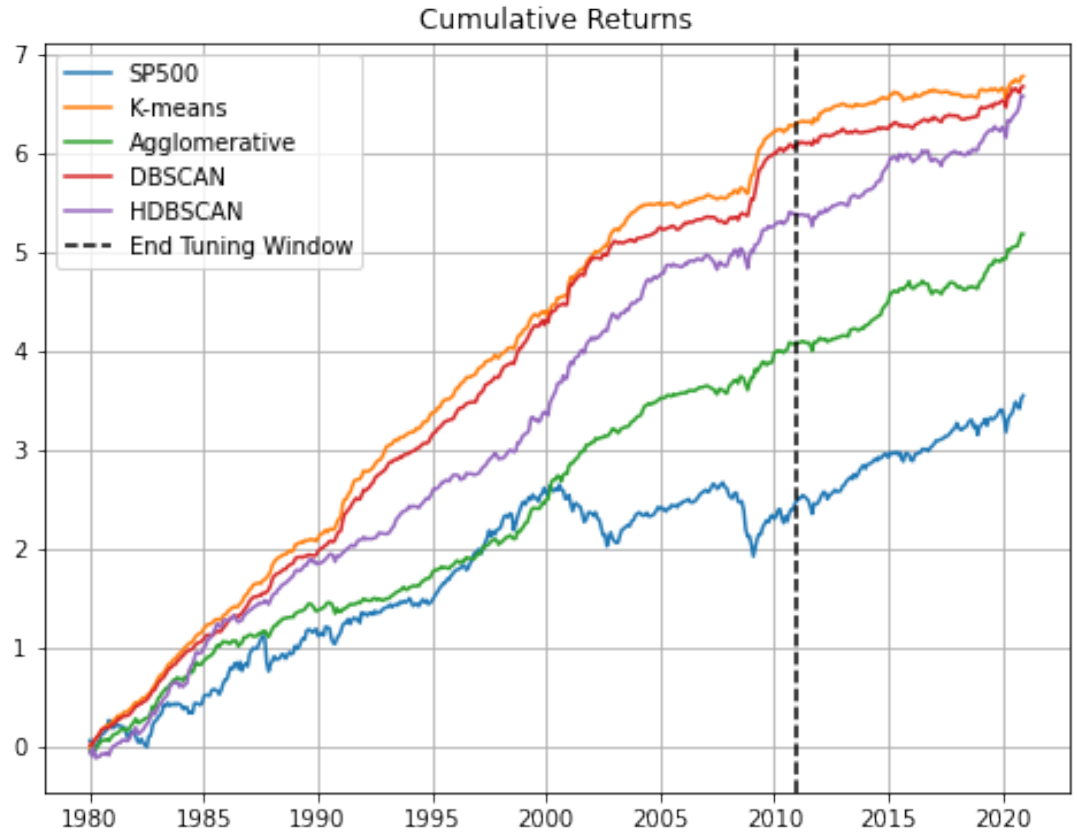


Figure 3: CZ Dataset Cumulative Returns

Table 5: CZ Datset Training Data

	SP500	Agglomerative	K-means	HDBSCAN	DBSCAN
Monthly Return Mean	0.007648	0.011287	0.017278	0.015024	0.016649
Monthly Return Std	0.044965	0.024302	0.022212	0.030948	0.022428
Annualized Return Mean	0.095736	0.144178	0.228220	0.195961	0.219130
Annualized Return Std	0.155765	0.084186	0.076945	0.107206	0.077693
Annualized Sharpe Ratio	0.614617	1.712616	2.966034	1.827883	2.820473
Maximum Drawdown	-0.525559	-0.126524	-0.057303	-0.180620	-0.066829

Table 6: CZ Dataset Testing Data

	SP500	Agglomerative	K-means	HDBSCAN	DBSCAN
Monthly Return Mean	0.009863	0.009607	0.004274	0.010560	0.005361
Monthly Return Std	0.039042	0.026719	0.021695	0.034950	0.022375
Annualized Return	0.124986	0.121571	0.052517	0.134347	0.066259
Annualized Return Std	0.135247	0.092558	0.075154	0.121070	0.077511
Annualized Sharpe Ratio	0.924137	1.313452	0.698799	1.109665	0.854839
Maximum Drawdown	-0.200011	-0.126272	-0.108360	-0.143657	-0.066047

The results indicate that the choice of clustering algorithm and dataset features has a significant impact on the performance of pairs trading strategies. While HDBSCAN and DBSCAN showed promising in-sample results in some cases, the Agglomerative algorithm consistently demonstrated better out-of-sample performance. These findings suggest that the Agglomerative clustering algorithm is more robust and generalizable for pairs trading strategies when compared to other unsupervised learning methods considered in this study.

5 Discussion

What we find in our study are three main takeaways: 1) the clustering algorithms in a general sense perform well in-sample but their performance degrades out-of-sample, 2) the performance of the portfolios are captured in the tails of the expected returns, and 3) the firm-level characteristics are certainly not uniform in the information they contain relating to predicting returns.

For the first point, every clustering algorithm we tested performed materially worse out-of-sample than in-sample, even when tuning the standard deviation threshold during portfolio construction on performance (annualized Sharpe ratio). Theory would indicate this is due to overfitting, however our training set contains large business cycle events such as the Savings and Loans crisis in the 1980s, the Dot-com Bubble collapse in the early 2000s, and the Global Financial Crisis in 2008 and 2009. It is possible that our models actually perform worse in "boom" economies, which was the case during our testing period.

Secondly, the higher the standard deviation threshold we use in the portfolio construction procedure in time $t-1$, the higher the return of the long-short portfolio becomes in time t. This implies that stocks in the tails of the expected return distribution contains the largest mispricings, while also containing the largest volatility. This makes intuitive sense, but reinforces the fact that the intuition behind the pairs strategy, at least how it was deployed here, is empirically sound.

Finally, our experiments with the three different data sources show that forward-looking return information (prediction signal) is higher in certain firm-level characteristics than others. We know this because, even when we add more features to the data set, the performance of the models both in- and out-of-sample does not improve. In fact, using properly imputed data from the firm-level characteristics (Reduced Dataset) is better for predictive modeling, even if its dimensionality is much less than that of the larger, more feature-rich data sets (CZ and Combined Datasets).

References

[1] AQR Capital Management. *Machine learning: Why finance is different.* AQR Capital Management. (2019, June 7). https://www.aqr.com/Learning-Center/Machine-Learning/Machine-Learning-Why-Finance-Is-Different

[2] Chen, Andrew Y. and Zimmermann, Tom, *Open Source Cross-Sectional Asset Pricing* (June, 2021). FEDS Working Paper No. 2021-37. http://dx.doi.org/10.17016/FEDS.2021.037

[3] Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). *A density-based algorithm for discovering clusters in large spatial databases with noise.* In kdd (Vol. 96, No. 34, pp. 226-231).

[4] Green, J., Hand, J. R., & Zhang, X. F. (2017). *The characteristics that provide independent information about average U.S. monthly stock returns.* The Review of Financial Studies, 30(12), 4389-4436. https://doi.org/10.1093/rfs/hhx019

[5] Han, C., He, Z., & Toh, A. J. (2023). *Pairs Trading Via Unsupervised Learning.* European Journal of Operational Research, 307(2), 929–947. https://doi.org/10.1016/j.ejor.2022.09.041

[6] McInnes, L., Astels, S., & Healy, J. (2016). *How HDBSCAN works. How HDBSCAN Works.* https://hdbscan.readthedocs.io/en/latest/how_hdbscan_works.html

[7] Sarmento, S. M., & Horta, N. (2018). *A Machine Learning based Pairs Trading Strategy.* Springer International Publishing.

Appendix

Acronym	Firm characteristic	Acronym	Firm characteristic
absacc	Absolute accruals	invest	Capital expenditures and inventory
acc	Working capital accruals	IPO	New equity issue
aeavol	Abnormal earnings announcement volume	lev	Leverage
age	# years since first Compustat coverage	lgr	Growth in long-term debt
agr	Asset growth	maxret	Maximum daily return
baspread	Bid-ask spread	ms	Financial statement score
beta	Beta	mve	Size
betasq	Beta squared	mve_ia	Industry-adjusted size
bm	Book-to-market	nincr	Number of earnings increases
bm_ia	Industry-adjusted book to market	operprof	Operating profitability
cash	Cash holdings	pchcapx_ia	Industry adjusted % change in capital expenditures
cashdebt	Cash flow to debt	pchcurrat	% change in current ratio
cashpr	Cash productivity	pchdepr	% change in depreciation
cfp	Cash flow to price ratio	pchgm_pchsale	% change in gross margin - % change in sales
cfp_ia	Industry-adjusted cash flow to price ratio	pchquick	% change in quick ratio
chatoia	Industry-adjusted change in asset turnover	pchsale_pchrect	% change in sales - % change in A/R
chcsho	Change in shares outstanding	pctacc	Percent accruals
chempia	Industry-adjusted change in employees	pricedelay	Price delay
chinv	Change in inventory	ps	Financial statements score
chmom	Change in 6-month momentum	quick	Quick ratio
chpmia	Industry-adjusted change in profit margin	rd	R&D increase
chtx	Change in tax expense	retvol	Return volatility
cinvest	Corporate investment	roaq	Return on assets
convind	Convertible debt indicator	roeq	Return on equity
currat	Current ratio	roic	Return on invested capital
depr	Depreciation / PP&E	rsup	Revenue surprise
divi	Dividend initiation	salecash	Sales to cash
divo	Dividend omission	salerec	Sales to receivables
dolvol	Dollar trading volume	securedind	Secured debt indicator
dy	Dividend to price	sgr	Sales growth
ear	Earnings announcement return	sin	Sin stocks
egr	Growth in common shareholder equity	SP	Sales to price
ep	Earnings to price	std_dolvol	Volatility of liquidity (dollar trading volume)
gma	Gross profitability	std_turn	Volatility of liquidity (share turnover)
herf	Industry sales concentration	sue	Unexpected quarterly earnings
hire	Employee growth rate	tang	Debt capacity/firm tangibility
idiovol	Idiosyncratic return volatility	tb	Tax income to book income
ill	Illiquidity	turn	Share turnover
indmom	Industry momentum	zerotrade	Zero trading days

Note: This table lists the 78 firm characteristics used in the study. Readers are referred to Green et al. (2017) and the references therein for the exact definitions.

Figure 4: Complete Dataset Firm Characteristics

Acronym	Firm characteristic	Acronym	Firm characteristic
absacc	Absolute accruals	invest	Capital expenditures and inventory
acc	Working capital accruals	~~ipo~~	~~New equity issue~~
aeavol	Abnormal earnings announcement volume	lev	Leverage
~~age~~	~~# years since first Compustat coverage~~	~~lgr~~	~~Growth in long-term debt~~
agr	Asset growth	maxret	Maximum daily return
baspread	Bid-ask spread	~~ms~~	~~Financial statement score~~
beta	Beta	~~mve~~	~~Size~~
betasq	Beta squared	~~mve_ia~~	~~Industry-adjusted size~~
bm	Book-to-market	~~nincr~~	~~Number of earnings increases~~
~~bm_ia~~	~~Industry-adjusted book to market~~	~~operprof~~	~~Operating profitability~~
~~cash~~	~~Cash holdings~~	~~pchcapx_ia~~	~~Industry-adjusted % change in capital expenditures~~
cashdebt	Cash flow to debt	pchcurrat	% change in current ratio
cashpr	Cash productivity	pchdepr	% change in depreciation
cfp	Cash flow to price ratio	pchgm_pchsale	% change in gross margin - % change in sales
~~cfp_ia~~	~~Industry-adjusted cash flow to price ratio~~	pchquick	% change in quick ratio
~~chatoia~~	~~Industry-adjusted change in asset turnover~~	~~pchsale_pchrect~~	~~% change in sales - % change in A/R~~
~~chcsho~~	~~Change in shares outstanding~~	pctacc	Percent accruals
~~chempia~~	~~Industry-adjusted change in employees~~	~~pricedelay~~	~~Price delay~~
chinv	Change in inventory	~~ps~~	~~Financial statements score~~
~~chmom~~	~~Change in 6-month momentum~~	~~quick~~	~~Quick ratio~~
~~chpmia~~	~~Industry-adjusted change in profit margin~~	~~rd~~	~~R&D increase~~
chtx	Change in tax expense	~~retvol~~	~~Return volatility~~
~~cinvest~~	~~Corporate investment~~	roaq	Return on assets
~~convind~~	~~Convertible debt indicator~~	~~roeq~~	~~Return on equity~~
currat	Current ratio	roic	Return on invested capital
depr	Depreciation / PP&E	rsup	Revenue surprise
divi	Dividend initiation	salecash	Sales to cash
divo	Dividend omission	salerec	Sales to receivables
dolvol	Dollar trading volume	securedind	Secured debt indicator
~~dy~~	~~Dividend to price~~	sgr	Sales growth
~~ear~~	~~Earnings announcement return~~	~~sin~~	~~Sin stocks~~
~~egr~~	~~Growth in common shareholder equity~~	SP	Sales to price
~~ep~~	~~Earnings to price~~	~~std_dolvol~~	~~Volatility of liquidity (dollar trading~~ volume)
~~gma~~	~~Gross profitability~~	~~std_turn~~	~~Volatility of liquidity (share turnover~~)
herf	Industry sales concentration	~~sue~~	~~Unexpected quarterly earnings~~
hire	Employee growth rate	~~tang~~	~~Debt capacity/firm tangibility~~
idiovol	Idiosyncratic return volatility	~~tb~~	~~Tax income to book income~~
ill	Illiquidity	~~turn~~	~~Share turnover~~
indmom	Industry momentum	zerotrade	Zero trading days

Note: This table lists the 78 firm characteristics used in the study. Readers are referred to Green et al. (2017) and the references therein for the exact definitions.

Figure 5: Reduced Dataset Firm Characteristics

Acronym	Description	Publication Year	Reference	% missing
Accruals	Accruals	1996	Sloan (1996)	0.50
AssetGrowth	Asset growth	2008	Cooper et al. (2008)	0.00
Beta	CAPM beta	1973	Fama and MacBeth (1973)	0.00
BetaFP	Frazzini-Pedersen Beta	2014	Frazzini and Pedersen (2014)	6.82
BetaTailRisk	Tail risk beta	2014	Kelly and Jiang (2014)	34.65
BidAskSpread	Bid-ask spread	1986	Amihud and Mendelson (1986)	7.33
BMdec	Book to market using December ME	1992	Fama and French (1992)	0.00
BookLeverage	Book leverage (annual)	1992	Fama and French (1992)	0.00
Cash	Cash to assets	2012	Palazzo (2012)	28.91
CashProd	Cash Productivity	2009	Chandrashekar and Rao (2009)	9.73
CBOperProf	Cash-based operating profitability	2016	Ball et al. (2016)	29.60
CF	Cash flow to market	1994	Lakonishok et al. (1994)	9.10
cfp	Operating Cash flows to price	2004	Desai et al. (2004)	14.34
ChEQ	Growth in book equity	2010	Lockwood and Prombutr (2010)	3.64
ChInv	Inventory Growth	2002	Thomas and Zhang (2002)	0.00
ChInvIA	Change in capital inv (ind adj)	1998	Abarbanell and Bushee (1998)	11.45
CompEquIss	Composite equity issuance	2006	Daniel and Titman (2006)	37.48
CompositeDebtIssuance	Composite debt issuance	2008	Lyandres et al. (2008)	40.26
Coskewness	Coskewness	2000	Harvey and Siddique (2000)	0.00
DelCOA	Change in current operating assets	2005	Richardson et al. (2005)	0.00
DelCOL	Change in current operating liabilities	2005	Richardson et al. (2005)	0.50
DelFINL	Change in financial liabilities	2005	Richardson et al. (2005)	0.79
DelLTI	Change in long-term investment	2005	Richardson et al. (2005)	0.00
DelNetFin	Change in net financial assets	2005	Richardson et al. (2005)	0.79
EarningsSurprise	Earnings Surprise	1984	Foster et al. (1984)	19.17
EBM	Enterprise component of BM	2007	Penman et al. (2007)	9.67
EntMult	Enterprise Multiple	2011	Loughran and Wellman (2011)	27.24
EP	Earnings-to-Price Ratio	1977	Basu (1977)	35.05
EquityDuration	Equity Duration	2004	Dechow et al. (2004)	1.83
GP	gross profits / total assets	2013	Novy-Marx (2013)	19.42
grcapx	Change in capex (two years)	2006	Anderson and Garcia-Feijoo (2006)	18.78
GrLTNOA	Growth in long term operating assets	2003	Fairfield et al. (2003)	2.64
GrSaleToGrInv	Sales growth over inventory growth	1998	Abarbanell and Bushee (1998)	23.15
Herf	Industry concentration (sales)	2006	Hou and Robinson (2006)	16.75
High52	52 week high	2004	George and Hwang (2004)	0.00
hire	Employment growth	2014	Belo et al. (2014)	1.53
IdioRisk	Idiosyncratic risk	2006	Ang et al. (2006)	0.00
Illiquidity	Amihud's illiquidity	2002	Amihud (2002)	4.72
IndMom	Industry Momentum	1999	Moskowitz and Grinblatt (1999)	9.10
IntMom	Intermediate Momentum	2012	Novy-Marx (2012)	9.25
Investment	Investment to revenue	2004	Titman et al. (2004)	25.73
InvestPPEInv	change in ppe and inv/assets	2008	Lyandres et al. (2008)	12.02
InvGrowth	Inventory Growth	2012	Belo and Lin (2012)	40.76
Leverage	Market leverage	1988	Bhandari (1988)	9.33
LRreversal	Long-run reversal	1985	De Bondt and Thaler (1985)	16.00
MaxRet	Maximum return over month	2010	Bali et al. (2011)	0.00
MeanRankRevGrowth	Revenue Growth Rank	1994	Lakonishok et al. (1994)	35.49
Mom12m	Momentum (12 month)	1993	Jegadeesh and Titman (1993)	9.28
Mom12mOffSeason	Momentum without the seasonal part	2008	Heston and Sadka (2008)	9.15
Mom6m	Momentum (6 month)	1993	Jegadeesh and Titman (1993)	9.18
MomOffSeason	Off season long-term reversal	2008	Heston and Sadka (2008)	9.69
MomOffSeason06YrPlus	Off season reversal years 6 to 10	2008	Heston and Sadka (2008)	31.66
MomSeason	Return seasonality years 2 to 5	2008	Heston and Sadka (2008)	9.68
MomSeason06YrPlus	Return seasonality years 6 to 10	2008	Heston and Sadka (2008)	31.53
MomSeasonShort	Return seasonality last year	2008	Heston and Sadka (2008)	9.18
MRreversal	Medium-run reversal	1985	De Bondt and Thaler (1985)	9.32
NetDebtFinance	Net debt financing	2006	Bradshaw et al. (2006)	10.75
NetEquityFinance	Net equity financing	2006	Bradshaw et al. (2006)	0.92
NOA	Net Operating Assets	2004	Hirshleifer et al. (2004)	0.42
OperProf	operating profits / book equity	2006	Fama and French (2006)	56.49

Figure 6: CZ Dataset Firm Characteristics

OPLeverage	Operating leverage	2010	Novy-Marx (2011)	0.20
PriceDelayRsq	Price delay r square	2005	Hou and Moskowitz (2005)	2.43
PriceDelaySlope	Price delay coeff	2005	Hou and Moskowitz (2005)	2.43
PriceDelayTstat	Price delay SE adjusted	2005	Hou and Moskowitz (2005)	2.71
RDS	Real dirty surplus	2011	Landsman et al. (2011)	6.63
ResidualMomentum	Momentum based on FF3 residuals	2011	Blitz et al. (2011)	13.33
ReturnSkew	Return skewness	2016	Bali et al. (2016)	0.59
roaq	Return on assets (qtrly)	2010	Balakrishnan et al. (2010)	13.83
RoE	net income / book equity	1996	Haugen and Baker (1996)	0.01
ShareIss1Y	Share issuance (1 year)	2008	Pontiff and Woodgate (2008)	9.31
Size	Size	1981	Banz (1981)	0.00
SP	Sales-to-price	1996	Barbee Jr et al. (1996)	9.30
STreversal	Short term reversal	1989	Jegadeesh (1990)	0.00
Tax	Taxable income to income	2004	Lev and Nissim (2004)	11.58
TotalAccruals	Total accruals	2005	Richardson et al. (2005)	4.97
TrendFactor	Trend Factor	2016	Han et al. (2016)	52.96
VarCF	Cash-flow to price variance	1996	Haugen and Baker (1996)	15.51
VolMkt	Volume to market equity	1996	Haugen and Baker (1996)	4.07
VolSD	Volume Variance	2001	Chordia et al. (2001)	5.93
VolumeTrend	Volume Trend	1996	Haugen and Baker (1996)	10.48
XFIN	Net external financing	2006	Bradshaw et al. (2006)	11.43
zerotrade	Days with zero trades	2006	Liu (2006)	4.00

Figure 7: CZ Dataset Firm Characteristics (Conitnued)

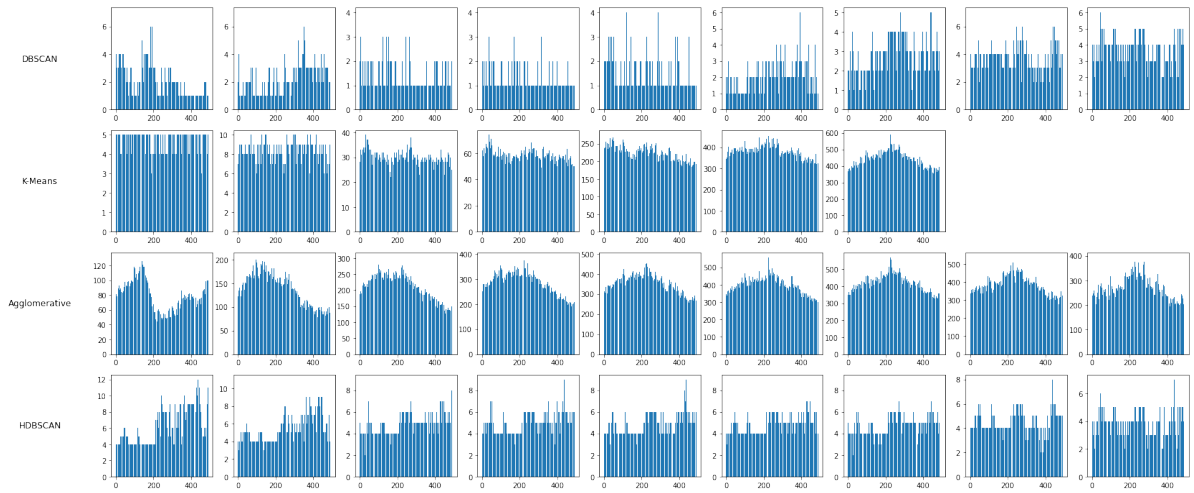


Figure 8: Complete Data Cluster Counts per Clustering Methodology. Each column represents the clustering algorithm's tuning parameter. For Agglomerative, DBSCAN, HDBSCAN we used $\epsilon = [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]$. For K-Means we used $K = [5, 10, 50, 100, 500, 1000, 1500]$.

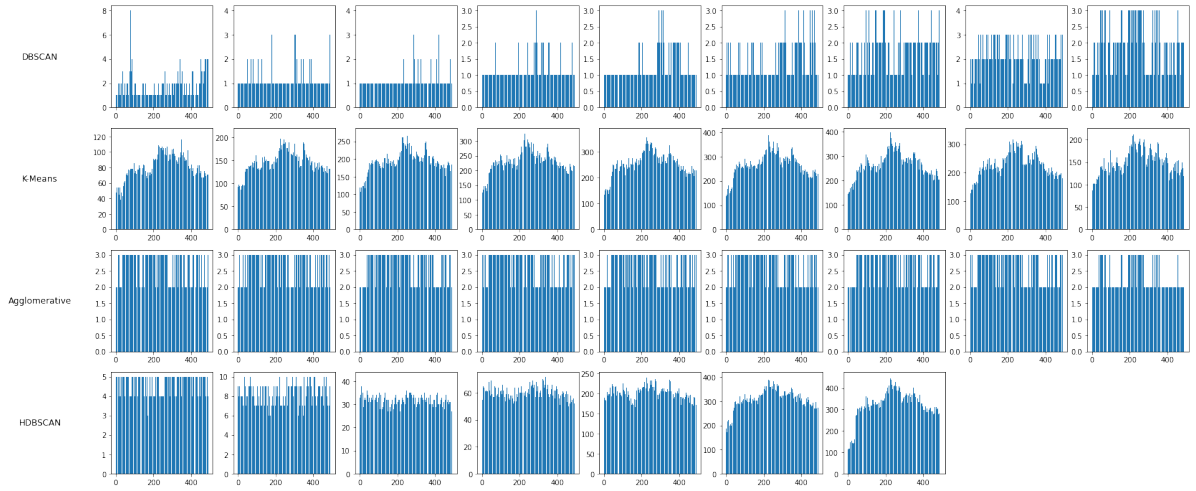


Figure 9: OAP Data Cluster Counts per Clustering Methodology. Each column represents the clustering algorithm's tuning parameter. For Agglomerative, DBSCAN, HDBSCAN we used $\epsilon = [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]$. For K-Means we used $K = [5, 10, 50, 100, 500, 1000, 1500]$.

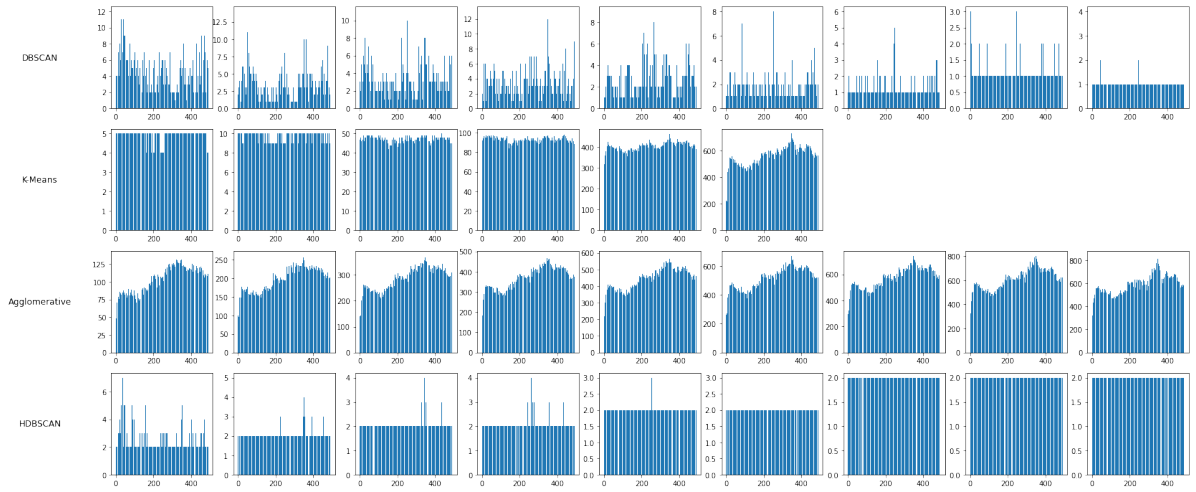
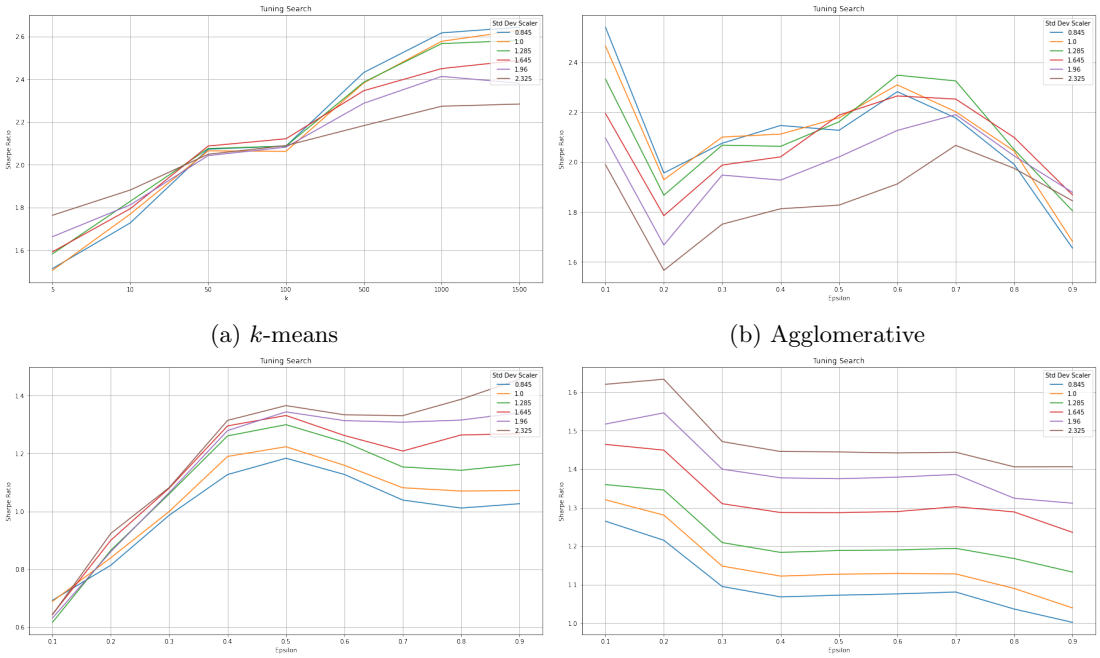


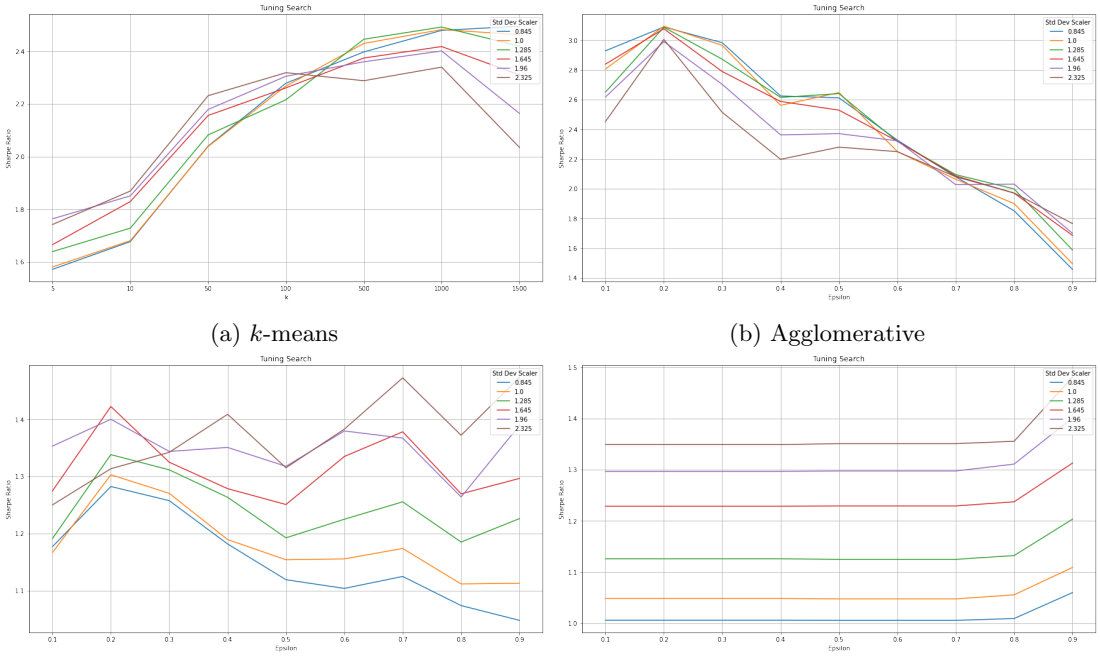
Figure 10: Reduced Data Cluster Counts per Clustering Methodology. Each column represents the clustering algorithm's tuning parameter. For Agglomerative, DBSCAN, HDBSCAN we used $\epsilon = [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]$. For K-Means we used $K = [5, 10, 50, 100, 500, 1000, 1500]$.



(a) k-means (b) Agglomerative

(c) DBSCAN (d) HDBSCAN

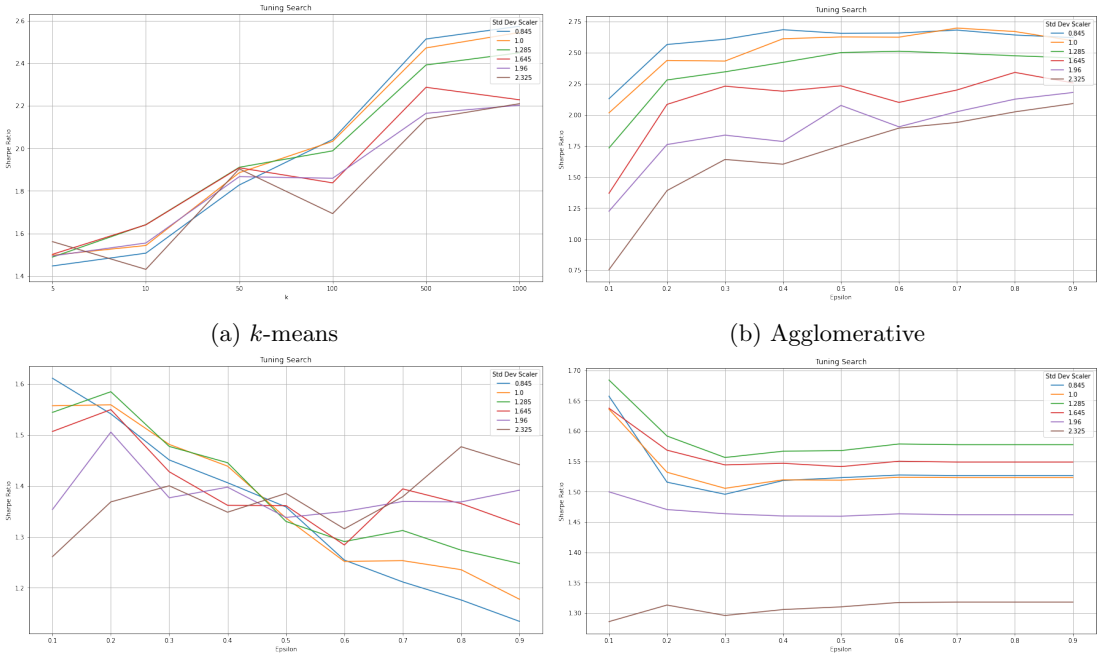
Figure 11: Complete Dataset Hyperparameter Tuning



(a) k-means (b) Agglomerative

(c) DBSCAN (d) HDBSCAN

Figure 12: Reduced Dataset Hyperparameter Tuning



(a) k-means

(b) Agglomerative

(c) DBSCAN

(d) HDBSCAN

Figure 13: CZ Dataset Hyperparameter Tuning