



STEVENS INSTITUTE OF TECHNOLOGY  
MASTER OF SCIENCE IN FINANCIAL ENGINEERING  
FE 800 - SPECIAL RESEARCH PROBLEMS

---

**Learned Sectors**  
A FUNDAMENTALS-DRIVEN SECTOR RECLASSIFICATION PROJECT

---

May 16, 2019

*Authors:*

Rukmal WEERAWARANA  
Yiyi ZHU  
Yuzhen HE

*Supervisors:*

Thomas LONON  
Ionut FLORESCU  
Papa NDIAYE  
Dragos BOZDOG

This work is licensed under a Creative Commons “Attribution-ShareAlike 3.0 Unported” license.



---

The underlying source code is licensed under the MIT License:

Copyright (c) 2019 Rukmal Weerawarana, Yiyi Zhu, Yuzhen He

Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the "Software"), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions:

The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Software.

THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.

## Abstract

Market sectors play a key role in the efficient flow of capital through the modern Global economy. Their use is widespread across the entire spectrum of market participants and observers, ranging from Governments using them to better regulate industry, to retail investors gaining exposure to particular segments of the economy through exchange traded funds tracking sector indices. We analyze existing sectorization heuristics, and observe that the most popular - the GICS (which informs the S&P 500), and the NAICS (published by the U.S. Government) are not entirely quantitatively driven, but rather appear to be highly subjective and rooted in dogma.

We examined alternative approaches to market sectorization, and found that returns-based methods were inherently flawed due to the significant bias of existing classifications on the structure of correlation distributions of the returns. Following this, we inspected determinants of firm value that would be intrinsically descriptive of the economic operating domain of a company. Building on inferences from analysis of the capital structure irrelevance principle and the Modigliani-Miller theoretic universe conditions, we postulate that corporation fundamentals - particularly those components specific to the Modigliani-Miller universe conditions - would be optimal descriptors of the true economic domain of operation of a company. Fundamentals data from Form 10-K for 15 features were downloaded for 362 companies in the S&P 500, forming the feature space on which train our classification model.

To this end, we developed a new, objective data-driven sector classification heuristic, based on a HCA algorithm. We utilized this novel heuristic to generate a set of potential candidate learned sector universes, by varying the linkage method of the HCA algorithm (testing SLINK, CLINK, ALC, and WARD linkage methods), and the number of resulting sectors derived from the model (ranging from 5 to 19), resulting in a total of 60 candidate learned sector universes.

We then introduce *reIndexer*, a backtest-driven sector universe evaluation research tool, to rank the candidate sector universes produced by our learned sector classification heuristic. *reIndexer* backtests portfolios of synthetic exchange traded funds, constructed based on the specifications of a candidate sector universe. The backtest period was from January 1<sup>st</sup> 2012 to December 31<sup>st</sup> 2017, tracking the evolution of the portfolio daily. The backtest results of each classification universe are then evaluated against each other, to derive a de facto *rank* for the candidate sector universes. This rank was utilized to identify the risk-adjusted return optimal learned sector universe as being the universe generated under CLINK (i.e. complete) linkage, with 17 sectors.

Finally, we evaluate our risk-adjusted return optimal learned sector against the benchmark classification heuristic, the GICS S&P 500 Classification. *reIndexer* was used again to backtest the GICS classification universe against the optimal (complete linkage; 17 sectors) learned sector universe. We found that our learned sector universe portfolio outperformed the benchmark with respect to both absolute portfolio value, and the risk-adjusted return of the portfolio over the backtest period.

We conclude that we fully explored the scope of our thesis statement, and addressed our specific research goals through the successful development of a fundamentals-driven Learned Sector classification heuristic with a superior risk-diversification profile than the status quo classification heuristic.

# Contents

<b>Abstract</b>	<b>i</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Applications of Market Sectors . . . . .	1
1.2 Status Quo . . . . .	2
1.2.1 GICS - Global Industry Classification Standard . . . . .	2
1.2.2 ICB - Industry Classification Benchmark . . . . .	2
1.3 Key Limitations . . . . .	3
<b>2 Research Goals</b>	<b>4</b>
2.1 Thesis Statement . . . . .	4
2.2 Specific Research Goals . . . . .	5
<b>3 Literature Review</b>	<b>6</b>
3.1 Existing Heuristic Evaluation . . . . .	6
3.2 Alternative Approaches to Market Sectorization . . . . .	7
3.3 Relationship between Economic Sectors and Fundamentals Data . . . . .	8
<b>4 Model Data</b>	<b>9</b>
4.1 Fundamentals Data Overview . . . . .	9
4.2 Feature Selection . . . . .	9
4.3 Benchmark Sector Universe . . . . .	10
<b>5 Learning Methods Survey</b>	<b>11</b>
5.1 Evaluation Criteria . . . . .	11
5.2 Candidate Learning Methods . . . . .	11
5.2.1 $K$ -means Clustering . . . . .	11
5.2.2 Support Vector Classifier . . . . .	12
5.2.3 Hierarchical Cluster Analysis . . . . .	12
<b>6 Hierarchical Clustering Model</b>	<b>13</b>
6.1 HCA Overview . . . . .	13
6.1.1 Distance Metric . . . . .	13
6.1.2 Linkage Method . . . . .	14
6.2 Unsupervised Learning Approach . . . . .	14
6.3 Learned Sector Universe Search Space . . . . .	15
<b>7 Candidate Universe Ranking</b>	<b>16</b>
7.1 Sample Learned Sector Universe . . . . .	16
7.2 reIndexer . . . . .	17
7.2.1 Synthetic ETF Formulation . . . . .	17
7.2.2 Efficient Portfolio Optimization . . . . .	18
7.2.3 Software Architecture Overview . . . . .	20
7.3 Performance Evaluation Metrics . . . . .	21
7.3.1 SETF Restructuring Turnover . . . . .	21
7.3.2 Portfolio Rebalancing Turnover . . . . .	21

7.3.3	Portfolio Return . . . . .	22
7.3.4	Sharpe Ratio . . . . .	22
7.4	Backtest Configuration . . . . .	22
<b>8</b>	<b>Optimal Sector Universes</b>	<b>23</b>
8.1	Complete Backtest Results . . . . .	23
8.2	Backtest Results Analysis . . . . .	23
8.2.1	Minimum Cumulative Turnover . . . . .	24
8.2.2	Maximum Absolute Portfolio Value . . . . .	24
8.3	Risk-Adjusted Return Optimal Universe . . . . .	25
<b>9</b>	<b>Benchmark Comparison</b>	<b>27</b>
9.1	Comparison Overview . . . . .	27
9.2	Performance Metric Comparison . . . . .	27
9.2.1	Cumulative Turnover Comparison . . . . .	27
9.2.2	Absolute Portfolio Value Comparison . . . . .	29
9.2.3	Risk-Adjusted Return Comparison . . . . .	29
9.3	Qualitative Comparison . . . . .	29
<b>10</b>	<b>Conclusion</b>	<b>31</b>
10.1	Research Goal 1 . . . . .	31
10.2	Research Goal 2 . . . . .	31
10.3	Research Goal 3 . . . . .	32
<b>11</b>	<b>Future Work</b>	<b>33</b>
11.1	HCA Model Tuning . . . . .	33
11.2	Varied ETF Construction Heuristics . . . . .	33
11.3	Temporal Variation of Sector Assignments . . . . .	33
11.4	Existing Sectorization Scheme Ranking . . . . .	33
<b>A</b>	<b>Backtest Visualization</b>	<b>34</b>
	SETF Restructuring Turnover . . . . .	35
	Portfolio Rebalancing Turnover . . . . .	36
	Portfolio Return . . . . .	37
	Sharpe Ratio . . . . .	38
<b>B</b>	<b>Optimal Learned Sector Universe</b>	<b>39</b>
B.1	Optimal Learned Sector Asset Distribution . . . . .	39
B.2	Optimal Learned Sector Universe Dataset . . . . .	40
<b>C</b>	<b>Backtest Portfolio Weights</b>	<b>44</b>
	Optimal Learned Sector Universe Portfolio . . . . .	45
	Benchmark Sector Universe Portfolio . . . . .	46
	<b>Bibliography</b>	<b>47</b>
	<b>Glossary</b>	<b>50</b>



# Chapter 1

## Introduction

The United States today is home to approximately 20,000 publicly traded corporations. Despite only a small minority capturing the public eye on a regular basis, they all contribute to the foundation on which the modern Global economy is built. As postulated by nearly all economic theory, the efficient flow of capital and information through these markets is necessary for a healthy economy.

To this end, market sectors and the practice of sectorization have been an integral component of healthy markets, both in the United States and around the world. Market Sectors - in their ideal form - group together corporations of similar business function and economic operating arena for easier regulation, management, investment, etc. A related practice to that of market sectorization is Market Segmentation, the practice of dividing a market into subgroups of consumers (i.e. *segments*).

The evolution of Market Segments and Market Sectors have historically been extremely useful metrics for gauging the development of the economy. There are four generally accepted stages of evolution in market segmentation; fragmentation, unification, segmentation, and hyper-segmentation.<sup>1</sup> The United States economy - being the archetype on which this four-stage heuristic was built - developed through these four stages over the course of the last two centuries. Being a decidedly *hyper-segmented* market today, there is a marked shift toward ever more narrow market segments. This shift has notably been amplified by the enable of hyper-targeted marketing and product delivery by technologies such as the smartphone.

### 1.1 Applications of Market Sectors

The United States Government began classifying companies into segments and sub-groups with the introduction of the Standard Industrial Classification (hereafter *SIC*) system in 1937.<sup>2</sup> Certainly, the effective classification of companies is a key prerequisite to scalable monitoring and governance. The sheer scope of the modern economy guarantees the necessity of such classifications; the current scope of the economy spans - indisputably - all facets of Human culture. This gargantuan scope demands specialization, which - in turn - demands organization; motivating the need for market sectors.

Credit rating is the practice of evaluating the risk of a prospective counterparty in a transaction. This metric is integral to risk management, and relies on evaluating the probability that a candidate counterparty to a transaction will not default on their obligation. In addition to the idiosyncratic forces affecting any given corporation, its risks are often decomposed to market factors and - increasingly - market sector factors. This means that a positive outlook on a specific market sector would imply a more positive outlook for the constituent corporations composing that sector, underscoring the importance of appropriate and accurate sector assignment. The significance of accurate sector assignments is reaffirmed by the fact that all of the *Big 3* credit rating agencies cite market sector rating as a key component in determining credit ratings.<sup>3,4,5</sup>

---

<sup>1</sup> Tedlow 1996

<sup>2</sup> Office of Statistical Standards - Bureau of the Budget 1957

<sup>3</sup> Standard & Poor's Rating Services 2014

<sup>4</sup> Hill et al. 2016

<sup>5</sup> Fitch Ratings 2019

Finally, another major application of asset sectors is to provide investors with targeted exposure to specific segments of the the market. It is a well-known corollary of Modern Portfolio Theory that diversification provides savings an enhanced risk-return portfolio for any given basket of assets.<sup>6</sup> This, combined with the excellent cost savings provided by modern Exchange Traded Funds (hereafter *ETFs*), has led to a rapid proliferation of these products in the Financial System today.

## 1.2 Status Quo

In the United States today, there are myriad sectorization taxonomies (hereafter *sector universes*). To limit the scope of this analysis, we will focus on two of the three most popular sector classification systems<sup>7</sup>; the GICS, and the ICB.

### 1.2.1 GICS - Global Industry Classification Standard

The GICS (Global Industry Classification Standard)<sup>8</sup> is published by MSCI (Morgan Stanley Capital International) and S&P (Standard & Poor's). This classification is arguably the most widely used sector universe in the United States, and provides the basis for the popular S&P 500 Market Sectors and Market Sector ETFs used in popular financial analysis resources.

Corporations are divided into four different categories, each in increasing order of specificity. The first classification is its *sector* (the most general), followed by the *industry group*, *industry*, and finally *sub-industry*, the most specific. The hierarchical taxonomy of these sector classifications are displayed in Figure 1.1<sup>9</sup>.

Additionally, the GICS methodology specification indicates that sector assignments and assignment updates are made primarily based on three factors; the primary source of revenue, earnings and market perception, and finally - in the case of a new company - information derived from the company prospectus.<sup>10</sup>

*Note:* A key change in this sector classification taxonomy was made in the Fall of last year (September 2018).<sup>11</sup> Specifically, the previously-labeled *Telecommunications Services* sector was broadened and renamed to *Communication Services*. Company sector assignment changes were also made commensurate to the name and scope change:

- Media companies were moved from *Consumer Discretionary* to *Communication Services*
- Internet services companies were moved from *Information Technology* to *Communication Services*
- E-Commerce companies were moved from *Information Technology* to *Consumer Discretionary*



**Figure 1.1:** Overview of the GICS sector classification universe.

### 1.2.2 ICB - Industry Classification Benchmark

The ICB (Industry Classification Benchmark)<sup>12</sup> is published by FTSE International (previously jointly owned by Dow Jones and FTSE). Similarly to the GICS, the ICB also classifies corporations into four increasingly specific categories; *industry* (the most general), *supersector*, *sector*, and finally *subsector*, the most specific. A visualization of the ICB taxonomy is reproduced in Figure 1.2<sup>13</sup>.

Similar to the GICS, ICB too utilizes three main criteria when classifying companies into specific sectors, and other categories. They are; the primary source of revenue, description in annual filings, and - in the case of a new company - information derived from the company prospectus, or regulatory filing descriptions (company-supplied).

<sup>6</sup> Markowitz 1952

<sup>7</sup> Fidelity Investments 2019

<sup>8</sup> MSCI - Morgan Stanley Capital International 2019

<sup>9</sup> *ibid.*

<sup>10</sup> S&P Global Market Intelligence 2018

<sup>11</sup> MSCI Research 2018

<sup>12</sup> FTSE International Limited 2019

<sup>13</sup> *ibid.*



### 1.3 Key Limitations

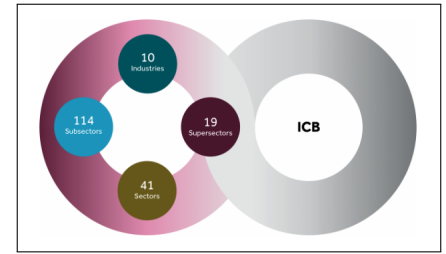
In this section, we analyze the GICS and ICB sector classification schemes described above through the lens of their limitations. We then utilize these key limitations to inform our research goals.

As discussed above, both GICS and ICB utilize information from company prospectuses to determine an initial sector assignment. Unfortunately, this inherently implies that the initial classification is not based on an objective criteria, and is highly subject to the initial vision of the authors of the company prospectus. This is in stark contrast to utilizing a company-specific quantifiable metric, and relies on accurate reporting in the initial prospectus; a document whose authors are highly incentivized to inflate in grandiosity.

Furthermore, the initial sector groups and constituent sectors in both classification schemes are defined based on a qualitative analysis of the economy, as opposed to a quantitatively-driven process. Similarly, the number of sectors in the market is also arbitrary, and not based on a quantifiable or objective metric.

Additionally, as implied by the reorganization of the GICS classification scheme in 2018, there appears to be no strong objective criteria governing the creation and deletion of new sectors. The companies reassigned during this reorganization were not new, and imply that the underlying newly created sector existed before its recognition by the GICS scheme.

As evidenced by the short list of limitations outlined in this section, the status quo of market sector classifications are far from perfect. To this end, we hope to develop a new *Learned Sectors* scheme, addressing each of the limitations described above.



**Figure 1.2:** Overview of the MSCI sector classification universe.

## Chapter 2

# Research Goals

In this section, we outline our overarching thesis statement. Additionally, we also isolate specific research goals based on this thesis statement, which we will address in sequence throughout the report. Through tackling each of the stated research goals, we hope to address the full scope of our thesis statement.

### 2.1 Thesis Statement



#### Thesis Statement

Utilize relationships in the idiosyncratic characteristics of corporations to inform a fundamentals-driven, non-subjective sector classification framework.

The thesis statement above encapsulates - at a very high level - the key issues encountered in existing sector classification heuristics, and how we plan to address these limitations. We believe that - given the objectivity of our classification - our *Learned Sectors* will provide a better basis for natural economic diversification. This is due to the fact that the underlying division of sectors, and assignment of corporations into those sectors will be objectively and quantitatively driven, rather than subjectively and qualitatively driven as is the status quo.

To combat the primary issue of the previously discussed classification heuristics (their lack of objectivity, particularly for newly classified companies), we seek to restrict our input data to the classification algorithm to reduce complexity. We believe that this restriction sufficiently limits the scope of our investigation, while also providing us with a quantifiable, objective measure of a capital structure, which - we believe - will reflect underlying economic function.

Given the constraint on our input data - motivated by end goal of increasing objectivity and reducing subjective involvement in the classification of the corporations - we also postulate that we will use data driven algorithms to derive potential classifications. That is, we plan to use entirely Unsupervised Learning methods, which do not require the definition of a *cost function*. This lack of a cost function - in addition to reducing complexity of the project - also removes another aspect of potential bias in the classification of the companies.

However, a shortcoming of this approach is that we will have a clustering algorithm parameterized by some set of arbitrary parameters, which will map a set of potential corporations to a set of potential market sectors. In keeping with the spirit of objectivity, we cannot arbitrarily assign values to the parameters, and thus must derive a method for *ranking* our potential sector universes. Note that this ranking cannot be on an objective scale, but rather would be a relative ranking comparing each candidate universe to its peers, thus maintaining objectivity of the ranking.

Finally, we hope to evaluate our sector classification against a benchmark sector universe; the *GICS S&P 500 Sector Classification* (hereafter *the benchmark*). To do this, we will use the same metric(s) utilized in ranking the candidate sector universes, and maintain any specific methodology used to compute those rankings between the *best* Learned Sector Universe, and the benchmark. If our initial hypothesis is correct, the superior risk diversification benefit inherent to our fundamentals-driven sector divisions will lead our Learned Sector Universe to outperform the benchmark with respect to the evaluation metric.

## 2.2 Specific Research Goals

Here, we encapsulate the gist of the previous discussion of our thesis statement in a collection of specific research goals. We will then address each of these research goals in sequence through the rest of the report, thereby fully exploring the scope of our thesis statement in the process.

Number	Description
RG-1	Utilize data-driven algorithms to derive a truly objective classification heuristic.
RG-2	Rank candidate sector universes against each other using entirely objective criteria.
RG-3	Evaluate our risk-adjusted return optimal sector universe against the benchmark.

**Table 2.1:** *Specific, itemized research goals of the Learned Sectors project.*

## Chapter 3

# Literature Review

Having identified concrete research goals designed to fully explore the scope of our thesis statement, we explored existing research in the field of market segmentation.

Due to their many applications and widespread use across the world, the problem of market sectorization has been approached through myriad lenses. For example, some authors have focused on sub-dividing an already established (i.e. dogmatic) sectors, while others have focused on the specific learning algorithms that may be successfully applied to the task of hierarchical decomposition of a set of related entities.

To best navigate the large corpus of research that is relevant to our research goals, we divide the presentation of our literature review into three sections:

- **Existing Heuristic Evaluation:** Evaluating the existing dominant sectorization heuristics.
- **Alternative Approaches to Market Sectorization:** Exploration of unorthodox approaches to market sectorization.
- **Relationship between Economic Sectors and Fundamentals Data:** Analysis of the relationship between company fundamentals data and their business function.

### 3.1 Existing Heuristic Evaluation

Originally established in the United States in 1937, the SIC<sup>1</sup> is a system for classifying industries with a four-digit code. Due to its abundant use in industry, the SIC Classification system has been widely used as an instrument in published Finance and Accounting Research. In 1997, the North American Industry Classification System<sup>2</sup> (hereafter *NAICS*) was adopted as an alternative to the SIC by various Government agencies, and is often cited interchangeably with the SIC in certain research. The key difference between the two heuristics is that NAICS is production-oriented, whereas the SIC is market-oriented.<sup>3</sup>

In *The Impact of Industry Classification Schemes on Financial Research*,<sup>4</sup> the author evaluates the usage of existing classification heuristics in Finance and Accounting Research published in major research journals. The author finds that approximately 30% of all research published in the top 3 Finance, and top 2 Accounting journals utilize industry classification systems. Given this relatively abundant usage, it is extremely concerning that the underlying heuristic itself is not entirely objective or quantitatively derived, as discussed in Section 1.3.

They are mainly used for sample restriction (34%), comparable company selection (31%), and detection of industry effects (12%). Under the reasonable assumption that Finance and Accounting Research is utilized when publishers create or update classification heuristics, this behavior of widespread use in existing work may be indicative of a feedback pattern, where existing structural dogmas of prior classification heuristics are implicitly reimposed on new systems. Additionally, the author also discovers that approximately 45% of all corporations change their industry over time based on the SIC Classification, and

<sup>1</sup> Office of Statistical Standards - Bureau of the Budget 1957

<sup>2</sup> United States Office of Management and Budget 1997

<sup>3</sup> Economic Classification Policy Committee (ECPC) 2007

<sup>4</sup> Weiner 2005

20% based on the GICS<sup>5</sup> industry classifications. This result highlights the lack of temporal stability prevalent in popular classification heuristics.

Despite this apparent lack of temporal stability of assignment, in her paper *Structural change and industrial classification*,<sup>6</sup> the author evaluates the impact of the slow rate of change of existing heuristics with respect to the addition and deletion of new and emergent industry groups to sector classification taxonomies. The author recognizes the fact that existing heuristics provide an incalculable resource to researchers. Due to this, she also infers that the co-dependence of researchers and classification publication agencies have led to existing classification schemes becoming de facto descriptors of economic industries, as opposed to the other way around. The author then performs an empirical analysis of the classification of highly innovative firms providing products and services in gaming devices, packaging, filtration, photonics, imaging, biomedical research, and fabless semiconductor design. Through her analysis, she finds significant vertical disintegration in existing classification heuristics.

The performance of NAICS and the GICS S&P 500 Classification heuristics are evaluated from a quantitative perspective in *A comparison of industry classification schemes: A large sample study*.<sup>7</sup> The authors perform individual linear regressions of a selection of fundamentals and earnings data of companies in the S&P Composite 1500 index against the sector assignments implied by the NAICS and GICS heuristics (among others). The various linear regressions are compared through the lens of an adjusted  $R^2$ -derived metric. The results indicate that the GICS heuristic performed best, but the maximum adjusted  $R^2$ -derived metric (realized on the monthly returns vs. GICS sector linear regression) was only 13.59%. Clearly, this is an extremely sub-optimal result.

## 3.2 Alternative Approaches to Market Sectorization

While not directly applied to the specific research problem of Market Sectorization, *Correlation Structure and Evolution of World Stock Markets: Evidence from Pearson and Partial Correlation-Based Networks*<sup>8</sup> provides excellent insight into the correlation structure of returns in the more generalized global economic environment. The authors analyzed daily price indices of 57 stock markets from 2005 to 2014, and inspected the distributions of the Pearson and Partial correlations between pairs of stock markets. In addition to affirming Economic theory through the confirmation that correlations between markets increase substantially during crisis, they also found that large groups of correlated markets exist based on their geographic location.

The authors' results confirm that the existence of pre-determined groupings of assets significantly affects the correlation distribution of those assets over time. Treating the geographic location of markets as a proxy for a generalized pre-existing group, we can extrapolate these effects to the more localized United States market. This generalization suggests that the existing sector groups would have a significant effect on the historical returns of companies in a given sector. This in turn implies that the usage of historical asset returns would introduce bias from existing sector groups to a new heuristic.

*Marketing segmentation using support vector clustering*<sup>9</sup> explores the application of a support vector clustering (a permutation of the support vector machine) to a relatively low-dimensional marketing dataset to derive clusters. The support vector clustering method is parameterized with a cluster count, and a random initialization of cluster centroids. Additionally, support vector clustering does not guarantee cluster assignment for all data points, and outliers remain unclassified. This approach is then compared to a  $K$ -means clustering and self-organizing feature map (SOFM) method, and is found to perform better based on a mean and standard error index evaluation. Despite appearing to be a promising approach in the authors' sample case study, the support vector clustering algorithm's cluster count parameterization, and its treatment of outliers do not make it suitable for the problem of market sectorization.

The authors of *A purchase-based market segmentation methodology*<sup>10</sup> apply a genetic algorithm to cluster transactional purchase data from a set of customers, with the end goal of training an RFM (Recency, Frequency, Monetary Value) model. The genetic algorithm, along with a cost function to assess the fidelity of fit, is used to segment customers into unique clusters based on their purchasing data. The iterative and stochastic behavior of the genetic algorithm ensures that the resulting cluster assignments are extremely stable, while also being non-variant with respect to centroid initialization. However, as with the previously discussed support vector clustering method, this approach is hindered by the required prior specification of a cluster count, as well as not being hierarchical in nature.

<sup>5</sup> MSCI - Morgan Stanley Capital International 2019

<sup>6</sup> Hicks 2011

<sup>7</sup> Hrazdil et al. 2013

<sup>8</sup> Wang et al. 2018

<sup>9</sup> Huang et al. 2007

<sup>10</sup> Tsai et al. 2004

### 3.3 Relationship between Economic Sectors and Fundamentals Data

*The determinants of capital structure in transitional economies*<sup>11</sup> provides an in-depth quantitative analysis of the alignment of traditional optimal capital-structure dogma against the real-world behavior of companies in transitional economies. The results suggest that while some traditional capital structure theories are indeed applicable to transitional economies, a large portion of capital structures are not well described by these traditional theories. Rather, the author finds that disparities in legal systems, shareholder power and demographics, and corporate governance provide a significantly better frame of explanation for the variance observed in capital structure.

*Determinants of capital structure of Chinese-listed companies*<sup>12</sup> analyzes the capital structures of corporations in China, providing a much better proxy for the large developed United States economy. The results presented by the authors echo that of Delcours, asserting that traditional theory does not fully describe the distribution of capital structures in China. Furthermore, the authors also allude to myriad other factors affecting capital structure, similar to Delcours. This work confirms that the dynamics of the determinants of capital structures observed in transitional economies are portable to larger, more established economies.

As highlighted above in Section 3.1, existing sectorization heuristics do not exhibit strong temporal stability. Thus, to avoid overfitting against the changing dynamics of a market when designing our new sectorization heuristic, we postulate that it would be beneficial to treat the market as constantly transitional. Under this assumption, the findings of Delcours can be applied to our prospective heuristic to great effect. The author's findings would suggest that we focus on determinants of the factors listed above to best capture the idiosyncratic dynamics of a given company, as opposed to focusing on traditional metrics of performance, such as asset returns.

Under the assumptions of the Modigliani-Miller theoretic universe (no taxes, bankruptcy costs, agency costs, and asymmetric information), the capital structure irrelevance principle<sup>13</sup> postulates that - in an efficient market - the value of a firm is unaffected by how that firm is financed. However, given that all of the conditions of the theoretic universe are violated in the real world, this leads to the profound realization that capital structure is the single most important determinant of firm value.<sup>14</sup> Based on the observation that firm value is derived from a company's intrinsic economic domain of operation, the capital structure irrelevance principle - in conjunction with the violation of Modigliani-Miller universe assumptions - implies that capital structure is governed by the idiosyncrasies of the true economic domain of a company.

Given that the exact quantified magnitude of violation of each of the Modigliani-Miller theoretic universe conditions are inherently specific to a given economic segment, we postulate that corporation fundamentals reflective of capital structure - particularly those components specific to the Modigliani-Miller universe conditions - would be the optimal descriptors of the true underlying economic domain of a company.

---

<sup>11</sup> Delcours 2007

<sup>12</sup> Chen 2004

<sup>13</sup> Modigliani et al. 1958

<sup>14</sup> Vernimmen et al. 2005

## Chapter 4

# Model Data

In this section, we describe the data sources used in our project, and identify specific features to be used in our sector classification heuristic. Additionally, we also describe the benchmark sector classification universe that we will use to evaluate our final results. This section begins the discussion of our first research goal, RG-1.

RG-1	Utilize data-driven algorithms to derive a truly objective classification heuristic.
------	--

### 4.1 Fundamentals Data Overview

In the previous section (see page 8), we explored the effect of the violation of the Modigliani-Miller theoretic universe conditions on the capital structure irrelevance principle, in conjunction with observations of the dynamics of the determinants of capital structure in transitional and established economies. The logical corollary of this analysis is that fundamentals data reflective of capital structure - particularly those specific to the Modigliani-Miller universe conditions - are optimal descriptors of the economic domain of a company.

Based on this conclusion, we identified earnings data from Form 10-K<sup>1</sup> filings to be our model input data. This data was retrieved for 362 companies in the S&P 500 Index, for every year from 2010 to 2017, from the Compustat Database<sup>2</sup>, via the Wharton Research Data Services<sup>3</sup> Cloud (hereafter *WRDS*).

### 4.2 Feature Selection

Given the variability of earnings reports, we identified 15 specific features from the annual Balance Sheet, Income Statement, and Statement of Cash Flows guaranteed to exist for all companies in our dataset. In addition to being common across all companies, they were also isolated on the basis of being related to, or direct arguments of, the capital structure of the company.

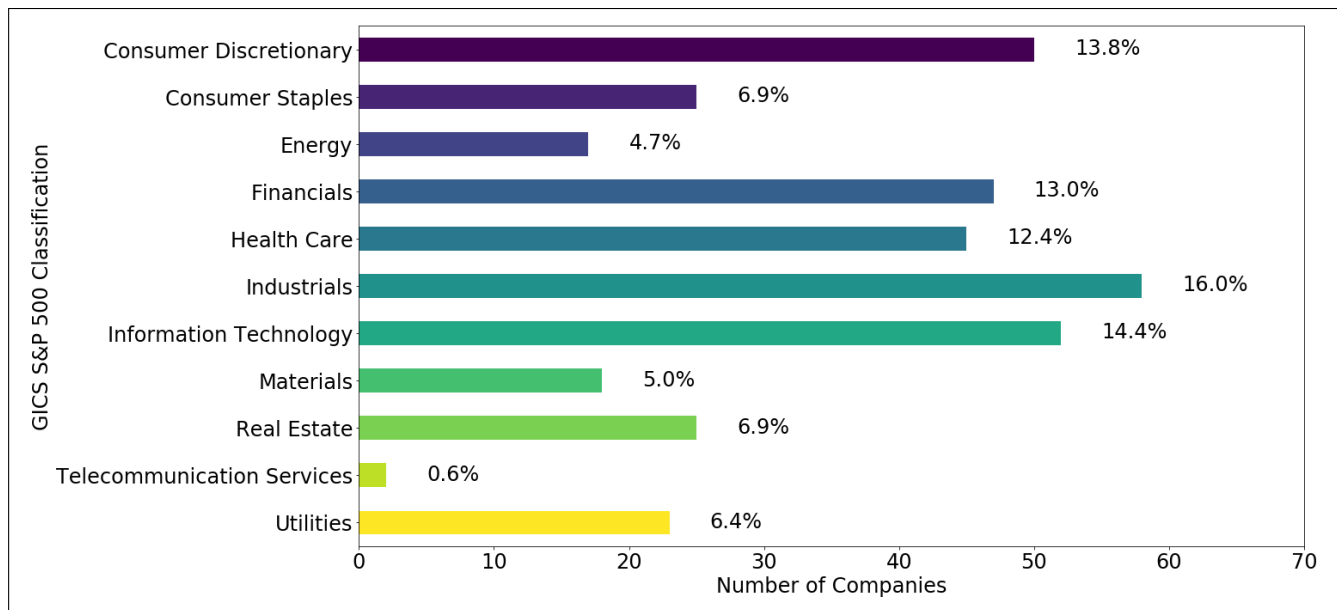
Total Assets	Cash & Equivalents	Receivables
Inventories	Sales	Cost of Goods Sold
Gross Profit	Operating Cash Flow	Operating Income
Depreciation, Depletion & Amortization	Interest Expense	Non-Operating Income/Expense
Income Taxes	Advertising Expense	Research & Development Expense

**Table 4.1:** Selected model input data features from Form 10-K for each company.

<sup>1</sup> U.S. Securities and Exchange Commission 2019

<sup>2</sup> S&P Global Market Intelligence 2019

<sup>3</sup> The Wharton School 1993



**Figure 4.1:** Distribution of input data companies ( $n = 362$ ) across sectors in the benchmark universe (i.e. GICS S&P 500 Classification).

### 4.3 Benchmark Sector Universe

To evaluate our final learned sector universe and fully address RG-3, we identified the GICS S&P 500 Classification<sup>4</sup> (hereafter *benchmark universe*) to be our benchmark. Unfortunately, the complete dataset of sector assignments our benchmark universe is proprietary. Due to this, we were unable to collate historical sector assignments, and were limited to the latest sector assignments for companies in our input data space.

Due to the disparity in temporal alignment between our data, we decided to utilize only the latest available data for our learned sector universe evolutions. That is - unless stated otherwise - we only utilized learned sector assignments implied by the 2017 10-K Form data for the remainder of this project.

The distribution of the 362 companies in our input data across various sectors in the benchmark universe is displayed in Figure 4.1.

<sup>4</sup> MSCI - Morgan Stanley Capital International 2019



## Chapter 5

# Learning Methods Survey

In this chapter, we outline a set of ideal characteristics and desired behavior of an ideal candidate classification algorithm, and conduct a survey of potential unsupervised learning methods. We then evaluate each of these surveyed methods against our selection criteria, and determine the optimal method with which to proceed.

### 5.1 Evaluation Criteria

Despite being not entirely objective, existing classification heuristics have a range of desirable behavior that we would wish to replicate with our candidate clustering algorithm. Additionally, we would want to replicate this behavior while also maintaining objectivity and stability in our new heuristic by utilizing a highly nonparametric learning method.

In particular, we would be extremely interested in preserving the nested hierarchical clustering behavior of the current schemes. That is, to be able to classify a market into sectors, and in turn those sectors into subsectors. Furthermore, it would be desirable to be able to determine these nested sub-sectors in the context of the greater market, rather than in isolated analysis of a particular sector.

Additionally, we would also like to vary the number of resulting sectors of our algorithm while maintaining stability. That is, if we were to request two sectors from a heuristic that was initially resulting in four sectors, the two sectors would be some combination of the initial four sectors, as opposed to an entirely new segmentation profile. This behavior is reflective of the real world, where economic sectors often exhibit nesting, as opposed to independent clustering.

Finally, as per RG-1 (see Section 2.2), we are extremely motivated to design a heuristic that is either entirely non-parametric, or parameterized with highly objective, quantitatively derived criteria. In addition to preserving mathematical objectivity of our results, a nonparametric approach would ensure that no personal biases - either explicit or implicit - are introduced to the final learned sectors.

### 5.2 Candidate Learning Methods

Given the required behavior outlined above, we evaluated three major families of clustering algorithms. We empirically evaluate each clustering technique through the lens of the requirements outlined above.

#### 5.2.1 $K$ -means Clustering

$K$ -Means Clustering is a method of partitioning  $n$ -dimensional data into a set of  $K$  distinct clusters. The basic algorithm is outlined below<sup>1</sup>:

---

<sup>1</sup> Lloyd 1982

Let  $C_1, C_2, \dots, C_K = \text{Set of } K \text{ possible clusters}$

Let  $W(C_k) = \text{Measure of pairwise difference of observations in a cluster}$

Let  $x_{ij} = j^{\text{th}}$  feature in cluster  $i$  with coordinates  $x$

$$\Rightarrow W(C_k) = \frac{1}{|C_k|} \sum_{i, i' \in C_k} \sum_{j=1}^p (x_{ij} - x_{i'j})^2$$

$$\Rightarrow K\text{-Means Clusters} = \underset{C_1, \dots, C_K}{\text{minimize}} \sum_{k=1}^K W(C_k) = \underset{C_1, \dots, C_K}{\text{minimize}} \sum_{k=1}^K \frac{1}{|C_k|} \sum_{i, i' \in C_k} \sum_{j=1}^p (x_{ij} - x_{i'j})^2$$

Notice that in the algorithm outlined above, the  $K$ -means clustering process requires two sets of parameters at initialization. First, it requires the number of target clusters,  $K$ , as well as a set of random initializations for cluster centroids,  $\frac{1}{|C_k|} \sum_{i \in C_k} x_{ij}$ . This high level of parameterization, coupled with the clear lack of congruity of assignment across varying values of  $K$  make this family of algorithms poorly suited to the task of sector classification, as per the constraints detailed above.

### 5.2.2 Support Vector Classifier

The support vector classifier is based on the notion of finding a set of hyperplanes in a higher dimensional feature space that optimally divides a set of data into classes. Data is mapped to a higher dimensional space to ensure orthogonal hyperplanes in the divisions of the clusters. The support vector classifier objective function is outlined below<sup>2</sup>:

$$\begin{aligned} \underset{R, a, \alpha}{\text{minimize}} \quad & R^2 - \sum_i \alpha_i (R^2 - \|x_i - a\|^2) \\ \text{subject to} \quad & \alpha_i \geq 0 \\ & (R^2 - \|x_i - a\|^2) = 0 \quad \forall i \text{ (KKT Condition)} \end{aligned}$$

As indicated by the literature, and through inspection of the objective function, it is clear that the support vector classifier is parameterized on the kernel used for optimization, as well as the specific loss function employed during model training. Furthermore, this model also optimizes to a fixed number of sectors, as opposed to a dynamic number. Therefore, this method too is inappropriate as per the evaluation criteria.

### 5.2.3 Hierarchical Cluster Analysis

Hierarchical clustering is a greedy algorithm which seeks to build clusters following either an agglomerative, or a divisive approach.<sup>3</sup> Agglomerative clustering is *bottom-up*, with each observation starting in its own cluster, whereas divisive is *top-down*, with all observations starting in one cluster and splits performed recursively at each level. The clusters output by this algorithm are determined by two model settings; the distance metric (i.e. the algorithm for computation of pairwise distance between observations), and the linkage method, which specifies the algorithm governing the dissimilarity of entire sets, as a function of the pairwise distances of observations in those sets.

This method has the distinct advantage of being entirely additively hierarchical, with groups being nested as described in the evaluation criteria. Furthermore, this method is entirely nonparametric, with the sole exception being the choice of linkage and distance metric. Additionally, it is extremely stable with varying sector counts. This is a direct result of the greedy nature of the algorithm, as it does not recompute the hierarchy each time a new cluster arity is extracted, but rather just changes the level of extraction from the same hierarchy.

As per the evaluation of the different families of learning methods detailed in this chapter, we selected Hierarchical Clustering to be the basis of our Learned Sectors classification heuristic.

<sup>2</sup> Ben-Hur et al. 2001

<sup>3</sup> Ward et al. 1963

## Chapter 6

# Hierarchical Clustering Model

In the previous chapter, we evaluated major families of learning methods and identified the Hierarchical Clustering Analysis (hereafter *HCA*) algorithm to be the method best aligned with the research goals of the project. Here we outline the specifics of our approach to applying HCA to our model input data, and build a search space of candidate universes to be evaluated, fully addressing RG-1.

### 6.1 HCA Overview

As outlined in Section 5.2.3, hierarchical clustering is a greedy learning algorithm which seeks to construct a hierarchy of clusters. The greedy nature of this algorithm results in extremely high computational complexity for any given model fit, but is extremely stable in its solution. Furthermore, it is an  $\mathcal{O}(1)$  complexity operation to extract classifications of varying arity due to the persistent hierarchical nature of the algorithm.

One of the main requirements of our heuristic is the ability to create sector universes with varying numbers of sectors. Because of this, we elected to utilize an Agglomerative approach to clustering. That is, we utilize a *bottom-up* HCA model, where each company begins in its own sector, with larger clusters derived at each successive step of the tree by merging existing pairs of clusters.

Any given HCA algorithm tree is parameterized by two distinct settings; the distance metric, and the linkage method. To best understand the potential candidate universes that may be generated by this HCA-driven classification heuristic, we analyzed each of these model settings in turn:

#### 6.1.1 Distance Metric

The distance metric is the measure of the distance between pairs of observations. This setting primarily affects the shape of the clusters. Due to the fact that our model input data is exclusively in monetary units (i.e. United States Dollars), we do not intend to transform the existing metric of wealth reflected by the dollar value measurement. Thus, we chose to use the  $\ell^2$  (i.e. Euclidean) distance metric for our heuristic.

Let  $\mathbf{p}, \mathbf{q}$  = Cartesian coordinates  $\mathbf{p} = (p_1, \dots, p_n)$  and  $\mathbf{q} = (q_1, \dots, q_n)$  where  $\{\mathbf{p}, \mathbf{q}\} \in \mathbb{R}^{n \times 2}$

Let  $dist(\mathbf{p}, \mathbf{q}) = \ell^2$  (i.e. Euclidean) distance between points  $\mathbf{p}$  and  $\mathbf{q}$

$$\Rightarrow dist(\mathbf{p}, \mathbf{q}) = dist(\mathbf{q}, \mathbf{p}) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \dots + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^n (p_i - q_i)^2}$$

### 6.1.2 Linkage Method

The second setting governing the behavior of the HCA algorithm is the selection of a linkage method. The linkage is a measure of distance between sets of observations as a function of the pairwise distances between observations. There are four major linkage method choices that we evaluate in our HCA model:

Let  $A, B, C, X, Y = \text{Sets (i.e. clusters) of observations}$

Let  $C = X \cup Y$

Let  $N = |A| + |X| + |Y|$ , where  $|\alpha| = \text{Cardinality}(\alpha)$

**Single Linkage<sup>1</sup>:**

$$d_{\text{SLINK}}(A, B) = \min \text{dist}(\mathbf{a}, \mathbf{b}) \forall \{\mathbf{a}, \mathbf{b} : \mathbf{a} \in A, \mathbf{b} \in B\}$$

**Complete Linkage<sup>2</sup>:**

$$d_{\text{CLINK}}(A, B) = \max \text{dist}(\mathbf{a}, \mathbf{b}) \forall \{\mathbf{a}, \mathbf{b} : \mathbf{a} \in A, \mathbf{b} \in B\}$$

**Average Linkage<sup>3</sup>:**

$$d_{\text{ALC}}(A, B) = \frac{1}{|A| \cdot |B|} \sum_{\mathbf{a} \in A} \sum_{\mathbf{b} \in B} \text{dist}(\mathbf{a}, \mathbf{b})$$

**Ward (variance minimization) Linkage<sup>4</sup>:**

$$d_{\text{WARD}}(C, A) = \sqrt{\frac{|A| + |X|}{N} d_{\text{WARD}}(A, X)^2 + \frac{|A| + |Y|}{N} d_{\text{WARD}}(A, Y)^2 - \frac{|A|}{N} d_{\text{WARD}}(X, Y)^2}$$

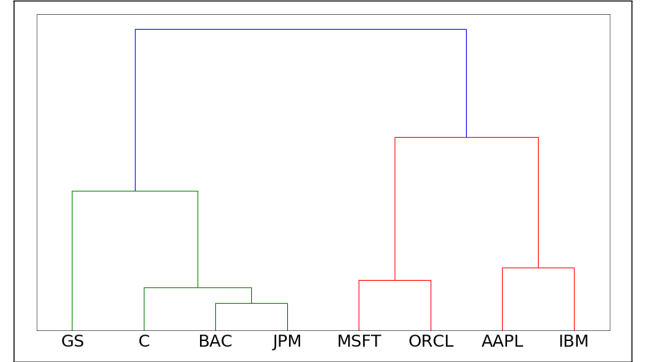
## 6.2 Unsupervised Learning Approach

As per RG-1 (see Section 2.2), we seek to create an entirely objective classification heuristic. Logically, this implies that we utilize an entirely nonparametric approach when designing the classification heuristic. However, as discussed above, the HCA algorithm is parameterized by both the distance metric, and the linkage method (in addition to a posterior selection of the number of sectors).

To work around the semi-supervised nature of the learning method, we elected to utilize HCA to build a search space of potential candidate sector universes. Following this, we will address our second research goal, RG-2, to rank these candidate sector universes against each other to determine the optimal learned sector classification.

Note that this search-space generation varies the sector count and linkage method parameters of the HCA model, but not the distance metric. This is due to the fact that we wish to preserve the monotonic and geometric difference of magnitudes of wealth implied by the dollar values of our input data.

Figure 6.1 is a dendrogram of a sample HCA model generated on a subset of the model input data. Despite having a very high time complexity for model training, the HCA model thrives in its ability to extract sector classifications of varying arity from a HCA model. Its ability to perform this action in constant time complexity greatly enhanced our ability to generate a large search space of candidate learned sector universes.



**Figure 6.1:** *Dendrogram of a sample hierarchical clustering model result.*

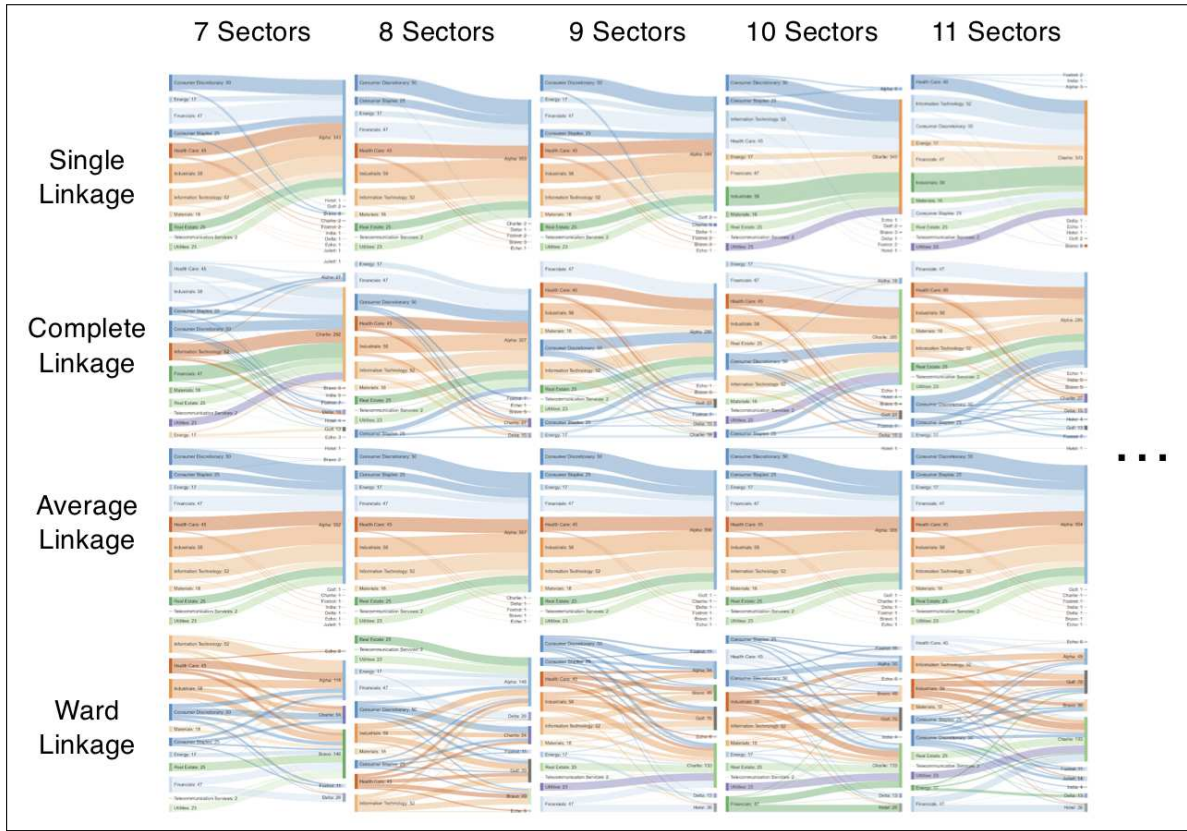
<sup>1</sup> Sibson 1973

<sup>2</sup> Defays 1977

<sup>3</sup> Seifoddini 1989

<sup>4</sup> Ward et al. 1963

### 6.3 Learned Sector Universe Search Space



**Figure 6.2:** Candidate learned sectors partial search space visualization.

With the end goal of building a comprehensive search space of candidate learned sector classifications, we generated HCA models parameterized with each of the linkage methods, and then isolated sector classifications for varying numbers of sectors. Specifically, we varied the number of sectors in our search space universes with  $N = \{5, 6, \dots, 19\}$  for each of the four linkage methods, for a total of 60 candidate learned sector universes. A subset of our search space is visualized in Figure 6.2.

The HCA models were generated iteratively using the model input data discussed in Section 4. We utilized the built-in Hierarchical Clustering Module in *Scikit-learn*<sup>5</sup> to generate the models, and saved them in specially formatted CSV files (publicly available on the *reIndexer* website<sup>6</sup>) for later ingestion by the ranking system we developed to identify the optimal learned sector classification universe.

<sup>5</sup> Pedregosa et al. 2011

<sup>6</sup> Weerawarana 2019

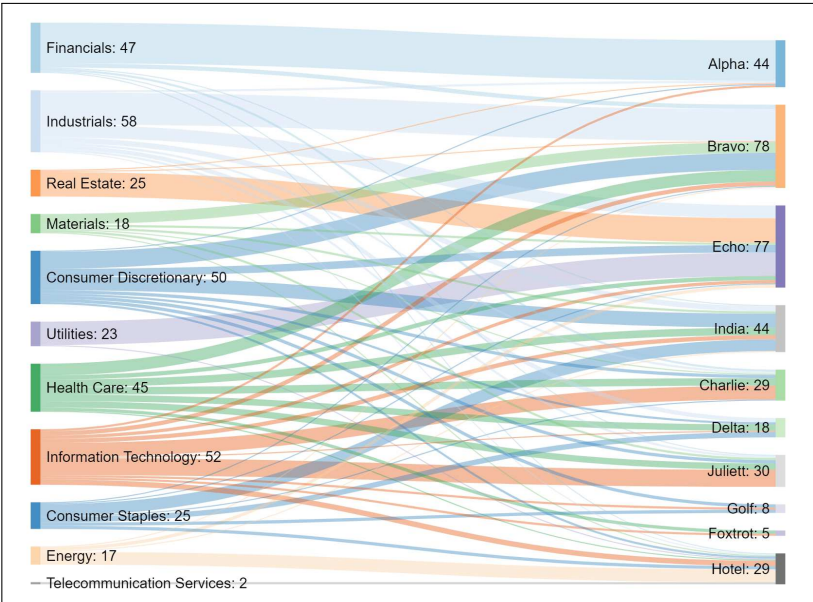
# Chapter 7

## Candidate Universe Ranking

Following the successful derivation of an objective sector classification heuristic (addressing RG-1), the next step is to derive a methodology to rank our sectors against each other, to isolate the optimal sector/sectors, without imposing any subjective criteria on the selection. Thus, this section begins the discussion of addressing the second research goal, RG-2 (see Section 2.2):

RG-2	Rank candidate sector universes against each other using entirely objective criteria.
------	---

### 7.1 Sample Learned Sector Universe



**Figure 7.1:** Sample learned sector universe - Ward Linkage; 2010 Data; 10 Sectors.

congruity between the benchmark classification and the learned sector universe heuristic. In the example, the only sector that can be considered remotely similar to a benchmark sector would be learned sector *Alpha* to benchmark sector *Financials*.

Other sectors however, appear largely broken up and dispersed when comparing their benchmark sector to the new learned sector universe. Particularly noticeable examples of this include the benchmark sectors *Health Care*, *Information Technology*, and *Consumer Discretionary*. Due to the fundamentals-driven nature of our classification heuristic, this result is not entirely

To best motivate the approach we decided to use when comparing and ranking candidate sector universes, it is worthwhile to discuss an example.

Figure 7.1 is a Sankey Diagram, representing the new learned sector assignments of various corporations from our search space by means of comparison to the benchmark. The left-hand side of the diagram represents the original benchmark sectors, and their constituent assets, while the right hand side represents the new learned sector assignment of the same assets.

As evidenced by the diagram, there appears to be significant transitory behavior of corporations across sectors when comparing the benchmark sectors to the learned sectors. Additionally, there also appears to be a large amount of mixing between the sectors, with very few sectors appearing to be preserved between the sector universes. This implies a stark lack of

surprising; the benchmark sectors that appear to have the most dispersion in the learned sector universe are ones which are increasingly integral to regular business, regardless of sector; particularly *Information Technology*.

As illustrated in this section with a single example, there is extremely little congruity between the benchmark sector classification and the learned sector classification. This pattern can be observed across a larger set of learned sector universes in Figure 6.2, a partial visualization of the learned sector search space.

Due to this fact, it would be extremely difficult to perform a sector-by-sector analysis across sector universes as a means for comparison. There is obvious difficulty in matching sectors across universes (as illustrated by the example in Figure 7.1) without introducing significant bias to the comparison metric. Furthermore, there is an additional issue presented by the fact that the number of sectors in two given candidate learned sector universes may not be identical (let alone the number of constituent corporations in a given sector), thus completely prohibiting a sector-by-sector analysis.

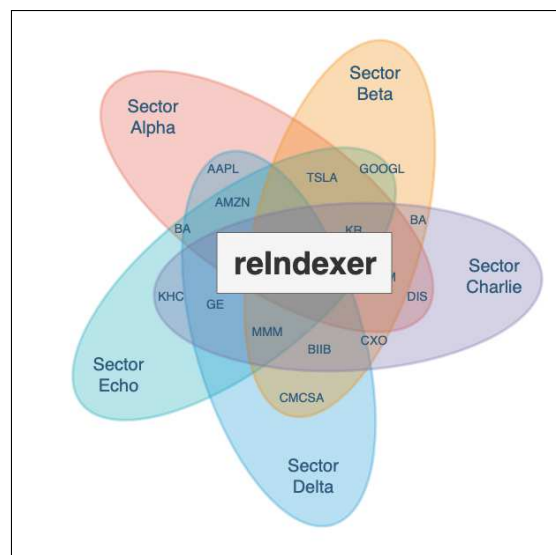
To combat this issue, we decided to evaluate the sector universes as a whole, and then analyze universe-level metrics to rank the candidate learned sectors. To this end, we developed *reIndexer*, which is discussed at length in the following section.

## 7.2 reIndexer

*reIndexer*<sup>1</sup> is an open-source research tool for backtest-driven evaluation of different sector universes, using a system of Synthetic ETFs (hereafter *SETF*), and efficient portfolios of those SETFs. *reIndexer* was designed and implemented to solve the problem discussed in the previous section; namely, the fact that we cannot perform a sector-by-sector comparison, and thus must compare learned sectors at the universe-level.

Built on Quantopian's *Zipline Pythonic Algorithmic Trading Library*<sup>2</sup>, *reIndexer* is fully API-compatible with the Quantopian suite of analytics tools, including *Alphalens*<sup>3</sup> and *Pyfolio*<sup>4</sup>. While these libraries are not directly used in this project, the modular design of *reIndexer* make it an extremely powerful platform that is highly extensible in scope and functionality.

*reIndexer* was designed under the hypothesis that the universe-level statistics of a given sectorization scheme would provide a superior metric for comparison across sector universes, compared to a sector-by-sector analysis. To this end - at a very high level - it provides a level of abstraction between single-asset trades, and constructed SETFs (whose composition is provided by the user), to simulate and record the performance of a portfolio of these SETFs over a predefined backtesting window.



**Figure 7.2:** The *reIndexer* backtest-driven sector universe evaluation research tool.

### 7.2.1 Synthetic ETF Formulation

As the assets prescribed by the classification heuristic are not traded in the real market, it is not possible to get historical ETF prices directly from Zipline's engine. *reIndexer* works around this fact by implementing a layer between the portfolio optimization, and the assets, maintaining a hypothetical SETF. To compute efficient portfolios, and to treat the SETFs as unique assets from the point of view of the portfolio optimization engine, all that is necessary is a historical chain of prices for a given asset, over a specified lookback period.

To this end, *reIndexer* currently implements a price-weighted synthetic ETF. The pythonic implementation of this ETF is based on a highly flexible and reproducible template, ensuring that different types of ETFs (market-weighted, etc.) can be easily implemented and used with *reIndexer*. For this project, price-weighted SETFs were used due to the fact that Zipline does not have historical market capitalization data for certain firms in its asset universe.

The mathematical formulation of the price-weighted SETF is outlined below:

<sup>1</sup> Weerawarana 2019

<sup>2</sup> Quantopian Inc. 2019c

<sup>3</sup> Quantopian Inc. 2019a

<sup>4</sup> Quantopian Inc. 2019b



Let  $\mathbf{S}_\alpha$  = Assets in sample sector  $\alpha$

Let  $\mathbf{P}_\alpha$  = Prices of assets in  $\mathbf{S}_\alpha$

Let  $\mathbf{w}_\alpha$  = Weights of assets in  $\mathbf{S}_\alpha$

$$\Rightarrow \mathbf{w}_\alpha = \left[ \frac{P_i}{\sum \mathbf{P}_\alpha} \forall P_i \in \mathbf{P}_\alpha \right]^\top$$

This process of recomputing the weights of the constituent assets in a SETF is referred to as **SETF Restructuring**.

Suppose that this SETF restructuring process occurs at every time step  $\{r_t, r_{t+1}, \dots, r_{t+n}\}$ .

Additionally, let there be additional time steps in-between the restructuring process times,  $\{r_{t+\delta t}, r_{t+2\delta t}, \dots, r_{t+m\delta t}\}$ , such that  $r_t < \{r_{t+\delta t}, \dots, r_{t+m\delta t}\} \leq r_{t+1}$ .

At each of these intermediate timesteps, the price-weighted SETF will have a price equal to the dot product of the sector asset weights computed at the immediately preceding SETF restructuring time  $r_t$ ,  $\mathbf{w}_{\alpha, r_t}$  and the prices of the constituent assets at the intermediate timestep,  $\mathbf{P}_{\alpha, r_{t+m\delta t}}$ .

Let  $\Pi_{\alpha, \tau}$  = Price of sector SETF  $\alpha$  at time  $\tau$

$$\Rightarrow \Pi_{\alpha, \tau_i} = \mathbf{w}_{\alpha, r_t} \cdot \mathbf{P}_{\alpha, \tau_i} \forall \tau_i \in \{r_{t+\delta t}, r_{t+2\delta t}, \dots, r_{t+m\delta t}\}$$

reIndexer provides an extremely flexible interface to specify SETF restructuring trigger times,  $\{r_t, r_{t+1}, \dots, r_{t+n}\}$ . The restructure can be triggered on any specific day of any specific week of the month (eg: third Friday of each month), or simply the first trading day of each month. Additionally, it also handles intelligent rebalancing rollovers, in the event that the specified rebalancing date trigger is a holiday with respect to the configured trading calendar.

## 7.2.2 Efficient Portfolio Optimization

Following the construction of the SETFs, reIndexer builds an efficient portfolio, treating each of the SETFs as distinct, unitary assets. As our goal with the project is to assess the benefit of fundamentals-driven objective sector classifications through the lens of risk diversification, reIndexer currently implements a backtest of a Global Minimum Variance Portfolio, not allowing short-sales.

However, in a similar fashion to the price-weighted SETF, the pythonic implementation of this portfolio is highly generalized in the reIndexer source code, and can be easily reconfigurable to work with myriad different portfolio configurations. reIndexer provides functionality to retrieve a Matrix of historical SETF prices for a given lookback window, with SETF prices being correctly computed using the formulation described above.

Following this, reIndexer computes the correlation matrix of returns using the historical prices over a specific lookback period, and then performs the non-convex optimization necessary to compute SETF weights in the Global Minimum Variance Portfolio. A sequential quadratic programming solver from the Python library *SciPy*<sup>5</sup> was used to perform the optimization.

The mathematical formulation for this process is outlined below (note that notation is preserved from the preceding section):

<sup>5</sup> Oliphant 2007



Let  $\Omega$  = Set of sectors in the candidate universe,  $\Omega_i \in \{\Omega_1, \Omega_2, \dots, \Omega_n\}$

Let  $\Pi_{\Omega_i}$  = Historical log price vector of sector SETF  $\Omega_i$

Let  $\Sigma$  = Covariance matrix of historical log-returns of SETFs

Let  $\omega$  = Vector of SETF weights in the Global Minimum Variance Portfolio

$$\Sigma = \begin{bmatrix} \mathbb{E}[(\Pi_{\Omega_1} - \mathbb{E}[\Pi_{\Omega_1}])^2] & \mathbb{E}[(\Pi_{\Omega_1} - \mathbb{E}[\Pi_{\Omega_1}])(\Pi_{\Omega_2} - \mathbb{E}[\Pi_{\Omega_2}])] & \cdots & \mathbb{E}[(\Pi_{\Omega_1} - \mathbb{E}[\Pi_{\Omega_1}])(\Pi_{\Omega_n} - \mathbb{E}[\Pi_{\Omega_n}])] \\ \mathbb{E}[(\Pi_{\Omega_2} - \mathbb{E}[\Pi_{\Omega_2}])(\Pi_{\Omega_1} - \mathbb{E}[\Pi_{\Omega_1}])] & \mathbb{E}[(\Pi_{\Omega_2} - \mathbb{E}[\Pi_{\Omega_2}])^2] & \cdots & \mathbb{E}[(\Pi_{\Omega_2} - \mathbb{E}[\Pi_{\Omega_2}])(\Pi_{\Omega_n} - \mathbb{E}[\Pi_{\Omega_n}])] \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}[(\Pi_{\Omega_n} - \mathbb{E}[\Pi_{\Omega_n}])(\Pi_{\Omega_1} - \mathbb{E}[\Pi_{\Omega_1}])] & \mathbb{E}[(\Pi_{\Omega_n} - \mathbb{E}[\Pi_{\Omega_n}])(\Pi_{\Omega_2} - \mathbb{E}[\Pi_{\Omega_2}])] & \cdots & \mathbb{E}[(\Pi_{\Omega_n} - \mathbb{E}[\Pi_{\Omega_n}])^2] \end{bmatrix}$$

$\therefore$  Global minimum variance portfolio weights are determined by solving the non-convex optimization:

$$\begin{aligned} & \underset{\omega}{\text{minimize}} && \omega^\top \Sigma \omega \\ & \text{subject to} && \mathbf{1}^\top \cdot \omega = 1 \\ & && \omega_i \geq 0 \forall \omega_i \in \omega \end{aligned}$$

This process of recomputing the weights of the SETFs in the portfolio is referred to as **Portfolio Rebalancing**.

Similar to the SETF restructuring process, the portfolio rebalancing process too occurs at discrete, user-specified time intervals. Additionally, as with the restructuring process, portfolio weights from the preceding timestep are used to compute the value of the portfolio at each intermediate timestep. This computation is outlined below:

Let  $\omega_\tau$  = SETF portfolio weights at time  $\tau$

Let  $\pi_{\Omega, \tau}$  = Value of portfolio under sector universe  $\Omega$  at time  $\tau$

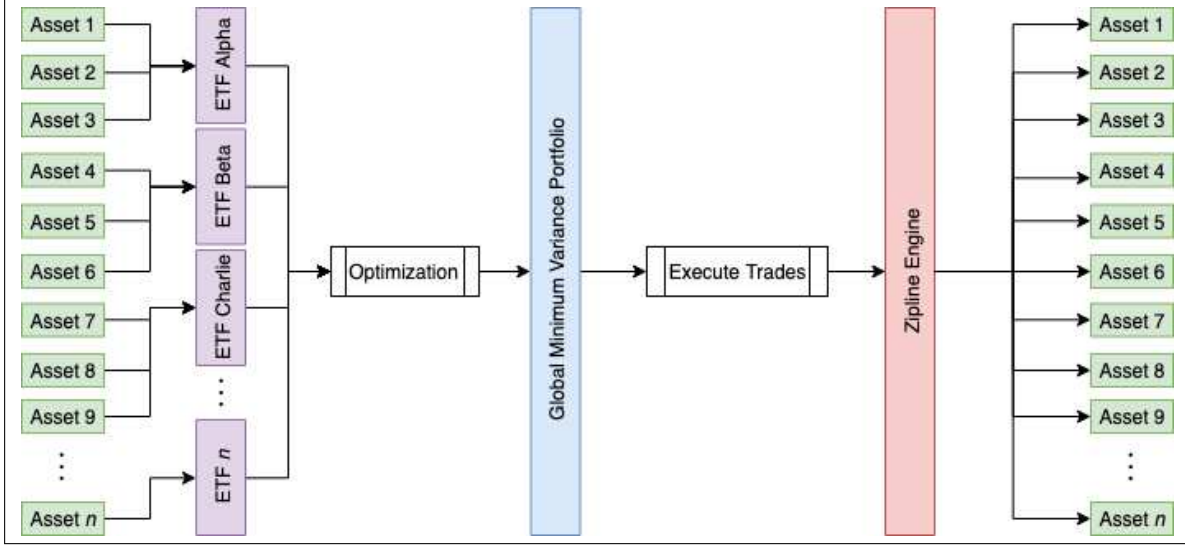
$$\Rightarrow \forall \{i, j, \tau\} : \{\{i \in \Omega\}; \{j \in \mathcal{S}_i\}; \{r_t < \tau \leq r_{t+m\delta t} < r_{t+1}\}\}$$

$$\begin{aligned} \therefore \pi_{\Omega, \tau} &= \Pi_{\Omega} \cdot \omega \\ &= \sum_{i \in \Omega} \Pi_{\Omega_i, \tau} \cdot \omega_{i, \tau} \\ &= \sum_{i \in \Omega} w_{i, \tau} \cdot P_{i, \tau} \cdot \omega_{i, \tau} \\ &= \sum_{i \in \Omega} \sum_{j \in \mathcal{S}_i} w_{j, \tau} \cdot P_{j, \tau} \cdot \omega_{i, \tau} \end{aligned}$$

In addition to the portfolio value, other metrics are also recorded at each timestep. These include open positions, the Sharpe ratio, the information ratio, and myriad other portfolio statistics. In this project, we do not utilize the full battery of statistics provided by Zipline (over 30 individual statistics in total), but we maintain the capability to generate these statistics for future expansion of reIndexer, and to maintain API compatibility with the Quantopian suite of analysis tools.

### 7.2.3 Software Architecture Overview

In this section, we describe the software architecture of reIndexer. Figure 7.3 is the architecture diagram of the system, describing the logical flow of the system, from individual asset statistics on the left, to final trades executed with the Zipline engine on the right.



**Figure 7.3:** *reIndexer architecture overview diagram.*

As seen in the diagram, SETFs are constructed from the bare asset statistics on the left, and are treated as unitary assets. Upon computation of SETF historical price chains (outlined above), the Global Minimum Variance Portfolio (with no short-sales) is computed by solving the non-convex optimization, also outlined above.

Both of the intervals of computation for these distinct determinants of portfolio value, the asset weights in the SETF, and the SETF weights in the portfolio, are computed at discrete, user-configurable intervals and do not necessarily have to occur at the same time. This behavior allows to replicate the real market to a high degree of accuracy, as most popular ETFs are rebalanced on the third Friday of every month, whereas retail investor portfolios are typically rebalanced on the first trading day of each month. Furthermore, Zipline also maintains the full historical order book for each asset, thus providing realistic price drift effects when large orders are placed.

After both computations are performed, individual trades are passed to the Zipline layer of the system (highlighted in red). These trades are only executed on the days of either SETF restructuring (i.e.  $\mathbf{w}$  update) or portfolio rebalancing (i.e.  $\boldsymbol{\omega}$  update). This provides us with a significant performance boost, as Zipline is optimized to record historical portfolio performance extremely quickly when trade execution is not required.

As Zipline does not see the individual SETFs (they are maintained internally by reIndexer), it is necessary to compute the individual weight for a given asset when a trade is to be executed. Due to the fact that the portfolio of SETFs can be considered a portfolio of portfolios, which both have asset weights that sum to 1, the weight of an individual asset in the larger portfolio of SETFs is simply the product of its weight in the SETF, and the weight of the SETF in the global minimum variance portfolio. This computation is outlined below:

Let  $\Theta$  = Set of all assets in the simulation

Let  $\mathbf{w}_i(\theta)$  = Weight of asset  $\theta$  in SETF  $i$

Let  $\boldsymbol{\omega}(\theta, i)$  = Weight of SETF  $i$  in the Global Minimum Variance Portfolio

Let  $\gamma_\theta$  = True weight of asset  $\theta$  in the larger Zipline portfolio

$$\therefore \gamma_\theta = \mathbf{w}_i(\theta) \cdot \boldsymbol{\omega}(\theta, i) \forall \{\theta, i\} : \{\{\theta \in \Theta\}; \{i \in \Omega\}\}$$

## 7.3 Performance Evaluation Metrics

To fully address RG-2 (see Section 2.2), we selected a set of objective criteria to record over the duration of each backtest, for each candidate learned sector universe. These metrics were selected to evaluate specific, but varied attributes of learned sector universe SETF portfolios, and will be evaluated independently to identify the optimal learned sector universe with respect to each metric.

### 7.3.1 SETF Restructuring Turnover

As described above, a SETF event is the recomputation of constituent asset weights in a synthetic ETF, and is a user-configurable triggered event. In the financial markets today, sector ETFs are typically created and sold by financial institutions. This abundance of ETFs provides increased liquidity to the market, while also reducing the barrier for entry to retail investors to gain exposure to specific sectors.

A key cost of creating and holding these ETFs for financial institutions is the fee incurred during ETF restructuring. Due to their enhanced status in the market, large financial institutions do not pay traditional commission fees charged to retail investors when executing trades on a stock exchange.

However, we believe it is valid to assume that their cost of trading would be proportional to the asset turnover of the component assets at each of these restructuring times. Therefore, we chose to record the component SETF asset restructuring turnover at each time a restructure is triggered, and utilize it as a proxy for judging the cost to financial institutions that would be incentivized to create ETFs of these sectors if they were indeed real. The mathematical formulation of this turnover for a single restructuring event is outlined below:

Let  $r_t$  and  $r_{t+1}$  = Times SETF restructures are triggered  
 Let  $\mathbf{w}_\tau$  = Vector of underlying asset weights in the SETF at time  $\tau$   
 Let  $\mathbf{P}_\tau$  = Vector of underlying SETF asset prices at time  $\tau$

$$\Rightarrow \text{SETF Restructuring Turnover} = \left| \mathbf{w}_{r_{t+1}}^\top - \mathbf{w}_{r_t}^\top \right| \cdot \mathbf{P}_{r_{t+1}}$$

### 7.3.2 Portfolio Rebalancing Turnover

Similar to how we treat the SETF restructuring turnover as a proxy for the cost to a financial institution to create the SETFs in the real market, we treat the portfolio rebalancing turnover as a proxy for the cost to a retail investor holding the SETFs in a portfolio.

To avoid introducing specific cost bias to our evaluation, we simply record the turnover of each SETF at each portfolio rebalance event. The mathematical formulation of this turnover for a single rebalancing event is outlined below:

Let  $\tau_t$  and  $\tau_{t+1}$  = Times portfolio rebalancing events are triggered  
 Let  $\boldsymbol{\omega}_t$  = Vector of SETF weights in the portfolio at time  $t$   
 Let  $\boldsymbol{\pi}_t$  = Vector of prices of SETFs in the portfolio at time  $t$

$$\Rightarrow \text{Portfolio Restructuring Turnover} = \left| \boldsymbol{\omega}_{\tau_{t+1}}^\top - \boldsymbol{\omega}_{\tau_t}^\top \right| \cdot \boldsymbol{\pi}_{\tau_{t+1}}$$

### 7.3.3 Portfolio Return

We also record the overall portfolio return over time. The computation of portfolio value at each timestep is discussed in depth in the previous section.

### 7.3.4 Sharpe Ratio

To best evaluate our initial hypothesis that our learned sector classification heuristic - based on objective fundamentals data - would provide better economically diversified sector classifications, we must test the diversification benefit of our learned sector universes, relative to the benchmark. Similarly, we also utilize this metric to isolate the optimal learned sector universe portfolio with respect to this diversification benefit metric.

As we are already computing the global minimum variance portfolio with no short sales, the total variance of the portfolio is minimized. Due to this fact, we can treat the risk-adjusted return (i.e. the Sharpe Ratio) as a metric for quantifying the diversification benefit of each learned sector classification universe with respect to a single unit of risk. The mathematical formulation of risk-adjusted return captured by reIndexer is reproduced below (parameterized by the annualized portfolio return,  $R_p$  and the portfolio variance  $\sigma_p$ ):

$$\text{Sharpe Ratio} = \frac{R_p - r_f}{\sigma_p}$$

## 7.4 Backtest Configuration

A backtest was performed on all 60 candidate learned sector universes with the following configuration:

Configuration Parameter	Setting
Start Date	January 1, 2012
End Date	December 31, 2017
SETF Restructure Trigger	Third Friday of each month
Portfolio Rebalance Trigger	First trading day of each month
Starting Capital	\$10,000,000,000
Backtest Frequency	Daily

**Table 7.1:** *Learned sector universe backtest configuration parameters.*

*Note: A high initial capital base is necessary to account for the fact that Zipline does not allow for fractional asset trades. Using a large capital base mitigates the rounding effect caused by this behavior.*

## Chapter 8

# Optimal Sector Universes

In this chapter, we utilize the backtesting system outlined in the previous chapter, we performed historical universe-level analysis of each of our candidate learned sectors. To reiterate, we plan to use the SETF restructuring turnover, the portfolio rebalancing turnover, the portfolio value, and the Sharpe ratio to rank our candidate learned sectors against each other.

Due to the magnitude of data output by the reIndexer backtesting system, all calculations in this section were performed in the cloud, on Google Research Colaboratory.<sup>1</sup> Data was loaded dynamically from its output location on Google Drive, directly into Google Colaboratory. Following this, individual output files were opened sequentially to extract the necessary longitudinal data dimension for the candidate learned sector universe, with all data being kept in-memory on Google Cloud Platform servers for increased efficiency during analysis.<sup>2</sup>

## 8.1 Complete Backtest Results

The complete results of the backtest are published online, and are available on the reIndexer website.<sup>3</sup>

Due to the magnitude of data generated by reIndexer, it is impractical to display quantitative summaries for each sector. Rather, we plotted graphs to visualize the progression of each of the risk metrics discussed in Section 7.3. The graphs were plotted for the full backtesting window outlined in Section 7.4.

- Cumulative SETF Restructuring Turnover (Figure A.1)
- Cumulative Portfolio Rebalancing Turnover (Figure A.2)
- Portfolio Value (Figure A.3)
- Rolling Sharpe Ratio (Figure A.4)

See Appendix A on Page 34 for full-sized plots of each of the graphs.

## 8.2 Backtest Results Analysis

We computed and recorded each of the Performance Evaluation Metrics outlined in Section 7.3 for each time step across all 60 candidate learned sector universes.

Following this, we computationally extract the optimally performing learned sector universe for each of the performance metrics. As both turnover metrics track a cost, we isolated the learned sector universe with the minimum cumulative value at the end of the simulation. Conversely, we isolated the learned sector universe that yielded the maximum value for both the portfolio value, and the (average) rolling Sharpe ratio.

---

<sup>1</sup> Google Research 2019

<sup>2</sup> Weerawarana et al. 2019b

<sup>3</sup> Weerawarana 2019

We now discuss the optimally performing learned sector universe with respect to the turnover and maximum absolute portfolio value metrics. Coincidentally, the minimum cumulative SETF restructuring turnover, and the minimum portfolio rebalancing turnover are both achieved by the same learned sector universe. These two performance metrics will be discussed together.

### 8.2.1 Minimum Cumulative Turnover

As discussed in Section 7.3, the cumulative turnover, for both SETF restructuring, and portfolio rebalancing, are designed to be a proxy for the cost of issuing, and holding these ETFs, respectively. The learned sector that performed best with respect to the cumulative turnover metrics has the following configuration:

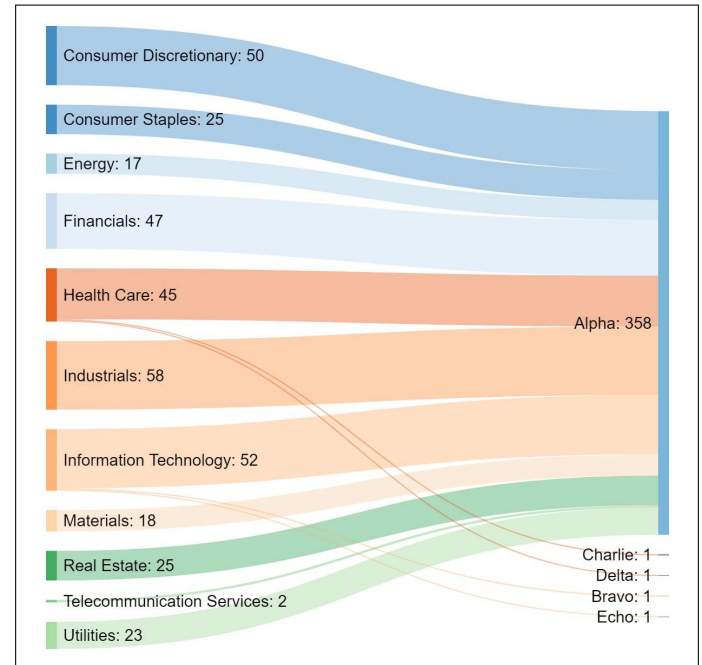
**Single Linkage; 5 Sectors**

The cost proxies each measure the effective cost of the SETF for entirely different segments of the Financial Lifecycle of this hypothetical product. The SETF restructuring turnover is a proxy for the cost that would be incurred by an ETF issuer (i.e. an institutional investor), whereas the portfolio rebalancing turnover is a proxy for the cost incurred by a holder of the ETF (i.e. a retail investor).

Upon closer inspection of the Sankey diagram corresponding to this sector in Figure 8.1, it is clear why it has superior minimum turnover, through the lens of both SETF restructuring, and portfolio rebalancing. The presence of a single large sector containing most of the assets (all of the assets with the exception of 4 in the case of Figure 8.1) implies that the sector universe would inherently have a significant advantage over its counterparts.

This single-sector pooling behavior would imply its SETF restructuring fee would be 0 in perpetuity for the sectors in which there is only a single asset. As single-asset sectors represent 80% of the sectors in this universe, this result is not surprising. This advantage is conferred in a similar fashion to the portfolio restructuring fee, as single-asset sector SETFs are not likely to be held in large quantities, relative to the singular vast sector SETF (Sector *Alpha*, in this example).

Unfortunately, it seems that the single linkage method results in a significant *pooling* effect, where the vast majority of assets are relegated to a single sector, with the remaining sectors being relative extremely small; containing just one asset each in the case of Figure 8.1. This behavior under the single linkage method is apparent from both the underlying data, as well as the partial search space visualization in Figure 6.2.



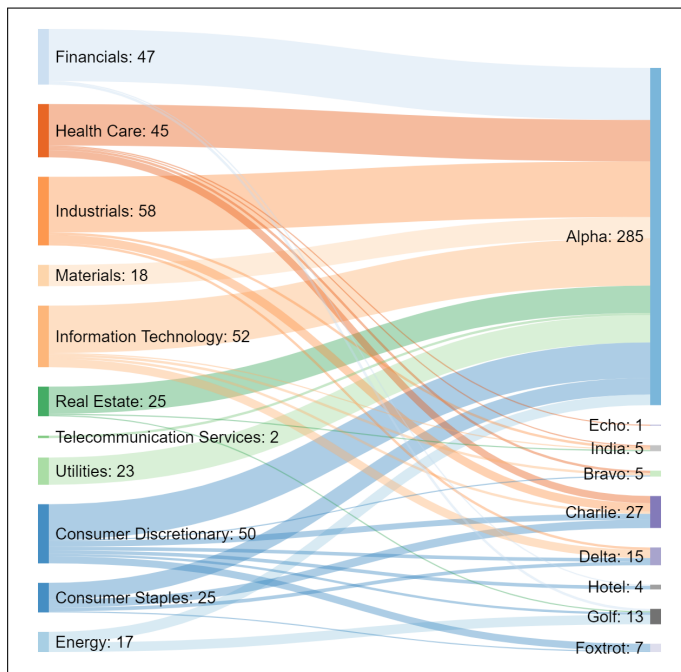
**Figure 8.1:** Learned sector universe with minimum cumulative SETF restructuring, and portfolio rebalancing turnover.

### 8.2.2 Maximum Absolute Portfolio Value

Unlike the previous performance metric, evaluating the historical learned sector universe portfolio value is not a proxy for an external cost. Rather, it is a candid assessment of the historical performance of the learned sector universes. While this does not directly address any of our research goals, it is an extremely important performance metric, and would be a key determinant of the success of any given learned sector universe in the real market. The learned sector that had the maximum absolute portfolio value over the lookback period (outlined in Section 7.4) has the following configuration:

**Complete Linkage; 9 Sectors**

The absolute portfolio value tracks the value of the portfolio throughout the backtesting period, utilizing the nested value computation outlined in Section 7.2.2. While not being a direct proxy of risk diversification, if viewed through the lens of potentially improved economic asset grouping, this learned sector universe is extremely encouraging. If the corollary that better capital market performance is correlated with improved economic sector performance, this result would imply that the fundamentals-driven classification heuristic does indeed provide a beneficial measure of diversification.



**Figure 8.2:** *Learned sector universe with maximum absolute portfolio value.*

Unlike the previous learned sector universe under the Single Linkage Method, the learned sector outlined in Figure 8.2 appears to have significantly less pooling. Despite this however, assets are still highly concentrated in a single extremely large sector, with other sector sizes (with respect to the number of constituent companies) being significantly smaller.

An interesting observation of this sector is that there is significant dispersion from the original benchmark classification. Omitting the sector *Alpha* due to its gargantuan size, sectors *Charlie* and *Delta* both have numerous components from highly diverse original benchmark sectors. Sector *Charlie* appears to have a high number of assets from the *Health Care* sector, as well as the *Industrials* and *Consumer Discretionary* sectors. Similarly, sector *Delta* has a large number of constituent assets classified as *Information Technology*, *Consumer Staples*, and *Consumer Discretionary* under the benchmark classification.

As with the sample sector classification discussed in Section 7.1, there appears to be high levels of dispersion of assets in sectors which are increasingly requisite to doing business. This is particularly evident (again, ignoring sector *Alpha*) through the dispersion of the *Information Technology*, *Health Care*, and *Consumer Discretionary* sectors.

### 8.3 Risk-Adjusted Return Optimal Universe

Finally, to isolate the sector with the maximum rolling annualized Sharpe ratio, we computed the mean rolling Sharpe ratio across the longitudinal temporal axis, and compared each of the learned sectors. Following this comparison, we determined that the learned sector that had the maximum average rolling Sharpe Ratio over the lookback period has the following configuration:

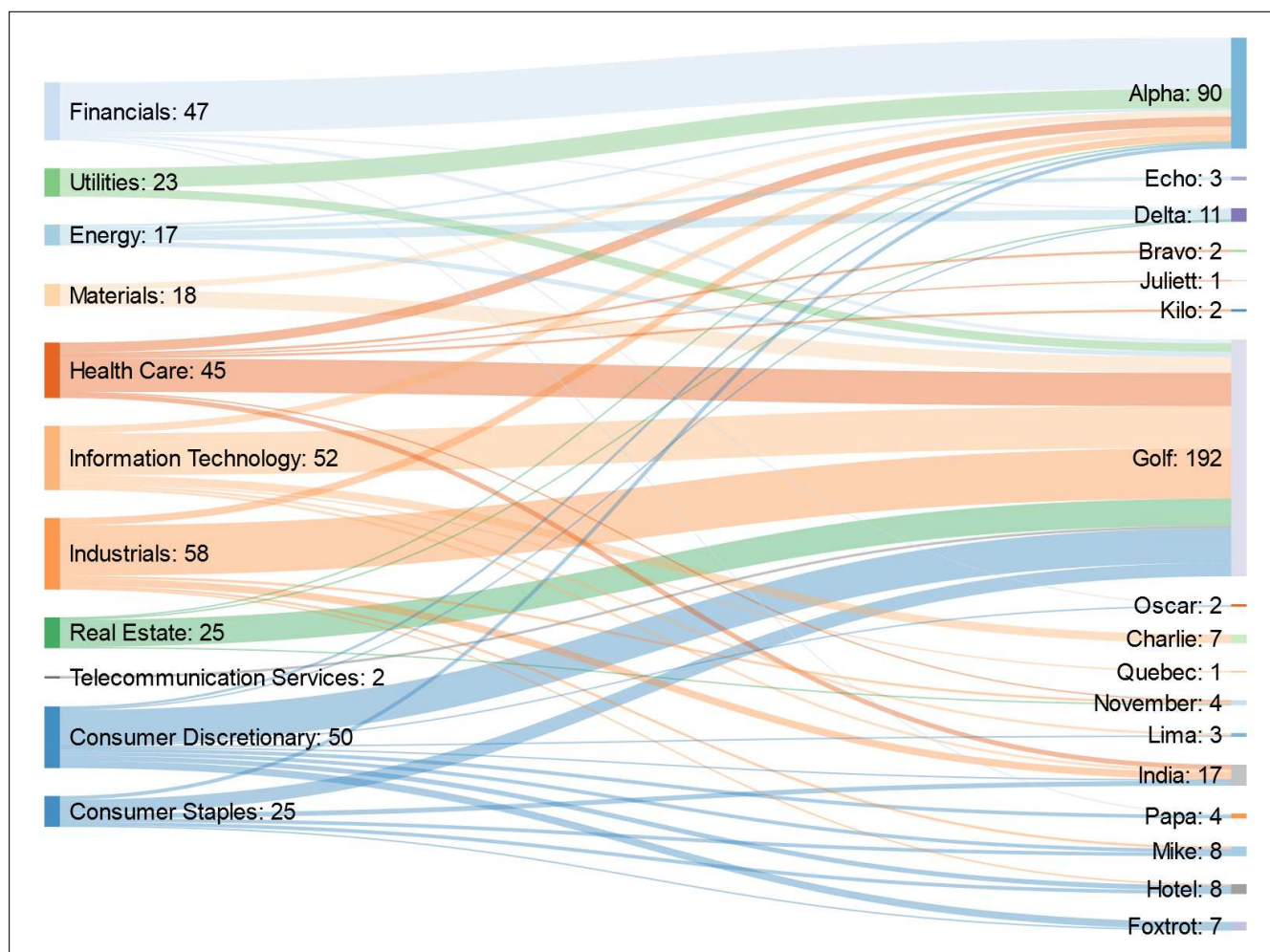
**Complete Linkage; 17 Sectors**

The Sankey diagram corresponding to the risk-adjusted return optimal learned sector is displayed in Figure 8.3. As indicated by the diagram, this learned sector universe has even less pooling behavior than both of the previously discussed learned sector universes (Figure 8.1 and Figure 8.2), despite still having 2 relatively large sectors, *Alpha*, and *Golf*, when compared to the rest.

As this learned sector universe provides the best risk-adjusted return on an already variance-minimized Global Minimum Variance Portfolio comparison, it would imply that this configuration provides the best risk diversification profile out of all of the candidate learned sector universes. In addition to the obviously highly prevalent dispersion from the benchmark classification to the new learned sectors, the size of each of the sectors is also considerably more even compared to its peers. This would imply a better portfolio-level diversification profile, as poorly performing sectors can be underweight during times of low implied risk-adjusted return. This is a key caveat of the pooling behavior particularly prevalent under the Single Linkage heuristic.

This lack of transitivity, combined with the implication that a higher risk-adjusted return implies a better risk-diversification profile would suggest that the learned sector universe heuristic is providing high levels of economic diversification, which appears to be at odds with the benchmark classification.

The full dataset of the risk-adjusted return optimal learned sector universe is reproduced in Appendix B.



**Figure 8.3:** *Learned sector universe with maximum rolling Sharpe ratio.*



# Chapter 9

## Benchmark Comparison

We now have a data-driven methodology for deriving learned sector universes (addressing RG-1), and have developed a objective criteria-driven ranking methodology to compare the learned sector universes against each other, addressing RG-2. The final step is to evaluate our objectively-identified risk-adjusted return optimal learned sector universe (see Section 8.3) against the benchmark classification. Thus, this section addresses the third and final research goal, RG-3 (see Section 2.2).

RG-3	Evaluate our risk-adjusted return optimal sector universe against the benchmark.
------	--

### 9.1 Comparison Overview

To preserve the impartial basis for comparison developed and maintained throughout this report, we isolated sector assignments for our benchmark sector universe, the *GICS S&P 500 Classification*. Unfortunately, as mentioned previously, we were only able to isolate transverse sector assignments for the year 2019, and were unable to access historical sector assignments, thus making a truly longitudinal comparison of our learned sector to the benchmark impossible.

To mitigate this issue, we compared the latest learned sector as implied by our clustering algorithm to the benchmark classification. To maintain the consistency of our analysis, we utilized reIndexer (see Section 7.2) to model SETFs of the benchmark, and to perform a backtest using the same configuration as was used for the candidate learned sector ranking (see Section 7.4). Similarly, we utilized the same performance metrics as were used to compare the candidate learned sector universes (see Section 7.3) to compare the risk-adjusted return optimal learned sector universe to the benchmark.

### 9.2 Performance Metric Comparison

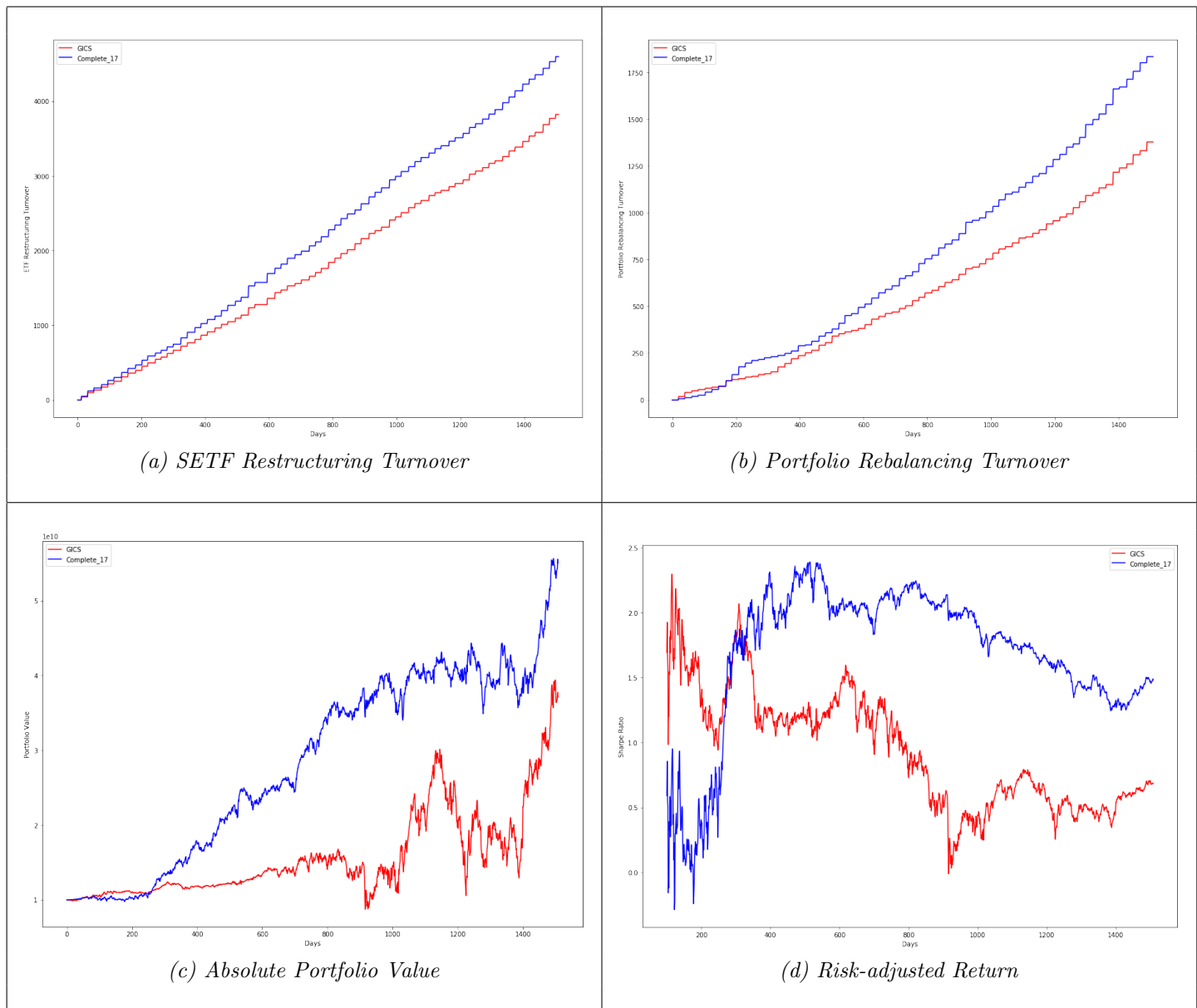
Figure 9.1 contains four panels, (a) through (d), with each displaying one of the four performance metrics outlined in Section 7.3. To best decompose the results of the comparison with the *GICS S&P 500 Classification* benchmark, we will analyze each of the performance metric comparisons in turn.

#### 9.2.1 Cumulative Turnover Comparison

Panels (a) and (b) in Figure 9.1 plot the cumulative turnover of SETF restructuring, and portfolio rebalancing, respectively. Recall from the previous section that for these turnover metrics, lower is better. As the red line represents the benchmark, it is apparent that the risk-adjusted return optimal learned sector did not outperform the benchmark with respect to both the SETF restructuring turnover, and the portfolio rebalancing turnover.

This phenomenon may be explained through analysis of the Sankey Diagram of the risk-adjusted return learned sector universe in Figure 8.3. It indicates a significant proliferation of component assets in 2 large sectors, with a large number of smaller sectors. Due to the fact that larger sectors have a higher notional value, and thus imply higher turnover when bought or sold, it is not surprising that the portfolio rebalancing turnover is higher for the risk-adjusted return optimal learned sector universe, compared to the more uniformly distributed benchmark universe.

Additionally, the larger individual sectors *Alpha* and *Golf* would also imply a higher rate of turnover during SETF restructuring. As a higher number of assets implies a more volatile total value, the extremely large sectors unique to the risk-adjusted return optimal learned sector universe would command a higher level of turnover during SETF restructuring, when compared to the more modestly sized sectors of the benchmark universe.



**Figure 9.1:** A comparison of sector universe performance metrics, with the *benchmark universe in red*, and the *risk-adjusted return optimal learned sector universe in blue*.

## 9.2.2 Absolute Portfolio Value Comparison

Panel (c) in Figure 9.1 is a comparison of the absolute portfolio value of both the risk-adjusted return optimal learned sector universe, and the benchmark universe. As indicated by the graph, the learned sector universe provides a significantly higher value at the terminus of the backtest, beating out the benchmark by nearly \$15,000,000,000 on a starting capital base of \$10,000,000,000 each, which translates to an outperformance of nearly 150%.

The progression of the portfolio over time for both the learned sectors universe and the benchmark universe indicate that the portfolio returns of the two sector universes are lightly correlated. This is to be expected, as the underlying base of investable assets is identical (by design) between the two sector universes. However, there does appear to be significantly less historical volatility in the returns of the learned sector universe portfolio compared to the benchmark portfolio.

This is particularly evident in the 750 - 1250 day interval in panel (c). This period shows that the portfolio value of the benchmark increased rapidly, at a significantly greater rate than its learned sector universe counterpart. However, at approximately the 1150 day mark, the benchmark suffers a extremely severe drop, losing nearly all of its gains of the preceding period.

Interestingly, the learned sector universe portfolio does not appear to fluctuate in value significantly (relative to the benchmark) during this period. This observation, coupled with the commensurate final rally in both sector universe portfolios near the end of the backtesting period suggests that the diversification profile of the learned sectors portfolio is *significantly* superior to that of the benchmark, resulting in not only a higher terminal portfolio value, but also significantly less volatility in reaching that value.

## 9.2.3 Risk-Adjusted Return Comparison

Figure 9.1 (d) is a comparison of the rolling risk-adjusted return (i.e. Sharpe Ratio) of the benchmark sector universe and the risk-adjusted return optimal learned sector universe. Given the results of the analysis of the absolute portfolio value comparison above, the outperformance of the learned sector universe relative to the benchmark universe is not a surprising result.

Continuing on the same line of analysis as the previous section, the Sharpe ratio of the benchmark sector universe performs extremely poorly during the interval of 750 - 1250 days discussed above. The negative effect of the increased volatility, despite a rally in the underlying portfolio is better reflected in the Sharpe ratio plot compared to the portfolio value plot of panel (c). In fact, the Sharpe ratio graph in panel (d) indicates that it was nearly more beneficial to own and hold the risk-free asset than the benchmark sector universe portfolio at approximately the 900 day mark, as the rolling Sharpe ratio of the portfolio briefly approaches 0. Despite rallying significantly during the interval, the Sharpe ratio of the benchmark sector universe never recovers, and does not approach the significantly higher value of the learned sector universe portfolio.

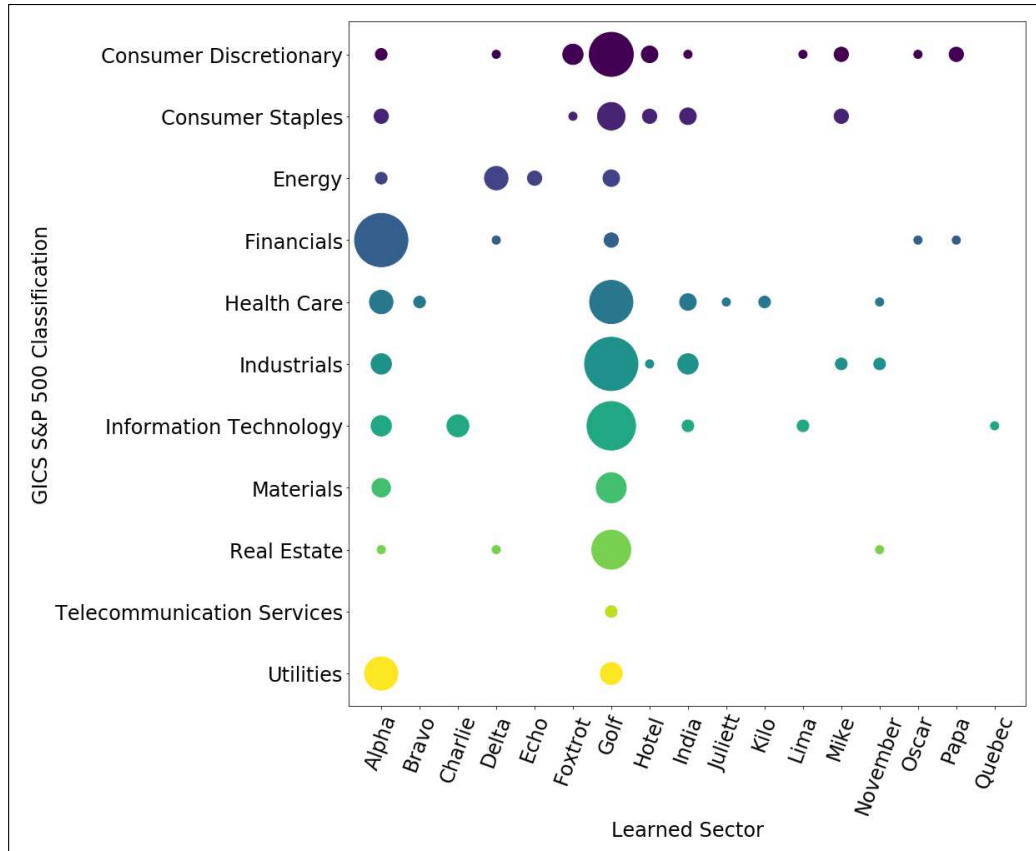
Additionally, despite rallying toward the end of the backtest, the Sharpe ratio plots indicate that the trend of the rolling risk-adjusted return for both sector universes was negative, with a much more smooth slope on the learned sector universe portfolio line. This indicates a lower *vol of vol* for the learned sector universe compared to the benchmark universe, which is further indication that the learned sector algorithm provides significantly better diversification benefits compared to the benchmark sector universe.

## 9.3 Qualitative Comparison

In this section, we attempt to conduct a more qualitatively-driven comparison and contrast of the risk-adjusted return optimal learned sector universe against the benchmark sector universe. Given the drastic difference in performance between the benchmark sector universe portfolio and the optimal learned sector universe portfolio, we believe that there is significant insight to be had by analyzing the composition of each of the sectors in the universe.

Analyzing each of the learned sectors in turn (from Figure 8.3 and Figure 9.2), it is clear that beyond the large sectors *Alpha* and *Golf*, a large majority of the remaining sectors are extremely small with respect to their numbers of component assets. Despite this however, the two large (major) sectors - as well as a selection of the smaller (i.e. minor) sectors - have an extremely high dispersion rate relative to the benchmark. That is, there doesn't seem to be a high level of congruence between the old and new sector assignments. This lack of agreement between the benchmark and learned sector universes is particularly apparent in the apparent lack of any direct transitional sector mappings in Figure 9.2.

Both major learned sectors *Alpha* and *Golf* comprise a large number of assets as their components. Particularly, it can be observed that learned sector *Alpha* contains a majority of the benchmark *Financials* sector, and the benchmark *Utilities* sector. Given that these sector assignments are derived from fundamentals data, this is a particularly interesting result, as both *Financials* and *Utilities* have become extremely risk-averse businesses over the last decade; *Financials* due to the Great Recession of 2008, and *Utilities* due to extensive capital damage incurred by the increased severity and number of Natural Disasters. This grouping indicates that the capital structure of these businesses are also becoming increasingly similar.



**Figure 9.2:** Sector assignment transitions between the benchmark sector classification universe and the risk-adjusted return optimal learned sector universe.


Considering learned sector *Golf*, it seems to be a mini-index within the original sector universe. From a component count perspective, it ingests a large amount of the benchmark *Consumer Discretionary* and *Consumer Staples* industries, as well as large swaths of industries that form the backbone of the US Economy as a whole; namely, the *Information Technology*, *Industrials*, and *Real Estate* sectors.

Appendix C contains stacked bar charts representing the level of investment in each of the sector SETFs for both the benchmark sector universe (see Figure C.2), and the risk-adjusted return optimal learned sector universe (see Figure C.1). Analyzing these graphs, in conjunction with the transition profile of sector assignments between the benchmark and optimal learned sector universe, it is clear that the learned sector *Golf* was not a strong performer. The benefit of containing a large cross-section of companies from myriad traditional sectors, combined with their poor performance during the years of 2014 and 2015 appears to be a key factor in the outperformance of the benchmark sector universe.

# Chapter 10

## Conclusion

In this section, we will reiterate our main findings, and relate them back to our specific research goals and thesis statement, initially outlined in Section 2.

**Thesis Statement**  
Utilize relationships in the idiosyncratic characteristics of corporations to inform a fundamentals-driven, non-subjective sector classification framework.

### 10.1 Research Goal 1

RG-1	Utilize data-driven algorithms to derive a truly objective classification heuristic.
------	--

- In the first portion of the report, we begin to address RG-1 goal by outlining our target data sources (see Section 4), the benchmark we plan to use for comparison, and how our specifically selected fields from our data sources relate to, and reinforce our research objective.
- Following this, we recognized that in order to maintain the level of objectivity enforced by RG-1, we would have to use an unsupervised learning method to determine our candidate learned sector universes. To this end, we conducted a survey of potential methodologies and identified Hierarchical Clustering as our target methodology in Section 5.
- In Section 6, we parameterized our HCA heuristic, and identified our search space consisting of 60 candidate learned sector universes. We then computed a set of candidate learned sector universes, fully addressing RG-1.

### 10.2 Research Goal 2

RG-2	Rank candidate sector universes against each other using entirely objective criteria.
------	---

- The second portion of the report was dedicated to addressing RG-2. This process began in Section 7, where we outlined the scope of RG-2. Following this, we introduced reIndexer, the backtest-driven sector universe evaluation research tool that powered the validation portion of our project (see Section 7.2).
- Next, we utilized reIndexer to rank our candidate learned sector universes, and computed a set of performance metrics (see Section 7.3) for each of our 60 candidate learned sector universes using reIndexer.
- Finally in Section 8, we utilized these performance metrics to identify the risk-adjusted return optimal learned sector universe (Complete Linkage; 17 Sectors), and therefore completing RG-2.

### 10.3 Research Goal 3

RG-3	Evaluate our risk-adjusted return optimal sector universe against the benchmark.
------	--

- The final research goal of this report was addressed in Section 9.
- We compared the risk-adjusted return optimal learned sector universe to the benchmark (i.e. *GICS S&P 500 Classification* sector universe utilizing reIndexer, and the same performance metrics used to rank the candidate learned sector universes.
- Our comparison showed that the benchmark sector universe provided a lower level of both SETF restructuring and portfolio rebalancing turnover compared to the risk-adjusted return optimal learned sector universe. However, the learned sector universe significantly outperformed the benchmark universe with respect to total portfolio return and rolling risk-adjusted return.
- Following this, we conducted a thorough quantitative and qualitative analysis of the reIndexer output for both the risk-adjusted return optimal learned sector universe, and the benchmark sector universe.
- We conclude that our risk-adjusted optimal learned sector universe does indeed provide a superior diversification profile compared to the benchmark universe, thus fully realizing RG-3.

Having addressed our specific research goals, RG-1, RG-2, and RG-3, we assert and affirm that we fully explored the scope of the thesis statement of our FE 800 Project in Spring Semester 2019.

# Chapter 11

## Future Work

In this section, we very briefly outline potential avenues for future research, building on the lessons learned during the course of this project.

### 11.1 HCA Model Tuning

Given the abundant pooling behavior (i.e. single large sector, and many single-asset sectors) of some of the hierarchical clustering models, it would be an extremely beneficial improvement to investigate methodologies to smooth the distribution of assets in the learned sectors.

### 11.2 Varied ETF Construction Heuristics

Currently, reIndexer creates and maintains price-weighted synthetic ETFs. However, a majority of market indexes today are market capitalization weighted, rather than price-weighted. A key improvement to reIndexer would be the implementation of market capitalization weighted SETFs, in addition to the current price-weighted SETF implementation.

### 11.3 Temporal Variation of Sector Assignments

As discussed in the report, we were unable to acquire historical sector assignment data for our benchmark sector universe, the *GICS S&P 500 Classification*. Due to this, we limited the scope of our sector ranking and benchmark evaluation to only use the latest fundamentals data we had available; 2017.

Given longitudinal sector assignment data for a collection of assets, reIndexer can be extended to be compatible with temporally varying sectors, increasing the overall accuracy of the evaluation system. This system would enable us to more accurately track metrics such as the SETF restructuring turnover over time, while also providing a more accurate assessment of the holding cost of SETFs to retail investors (i.e. portfolio rebalancing turnover).

### 11.4 Existing Sectorization Scheme Ranking

In addition to being an excellent tool for comparing hypothetical sector universes, reIndexer may also be used to compare existing sector classification schemes. That is, it may hypothetically be used to compare the performance of the GICS classification scheme against the FTSE classification scheme.

Similar to the analysis performed with the hypothetical sector universes, a *diversification ranking* of sorts of existing sector universes may be developed.

## Appendix A

# Backtest Visualization

### SETF Restructuring Turnover

See Figure A.1 on Page 35.

### Portfolio Rebalancing Turnover

See Figure A.2 on Page 36.

### Portfolio Return

See Figure A.3 on Page 37.

### Sharpe Ratio

See Figure A.4 on Page 38.







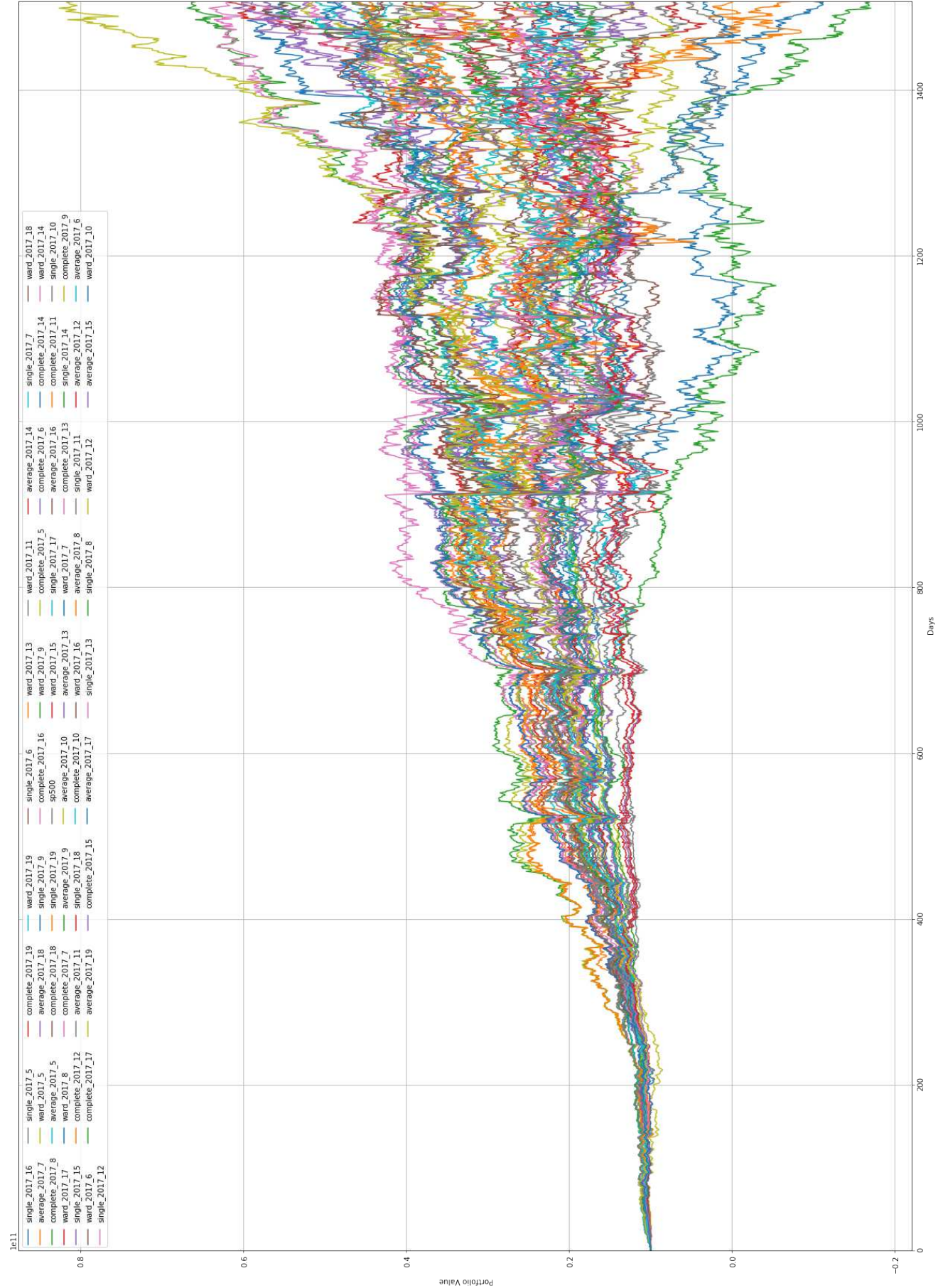
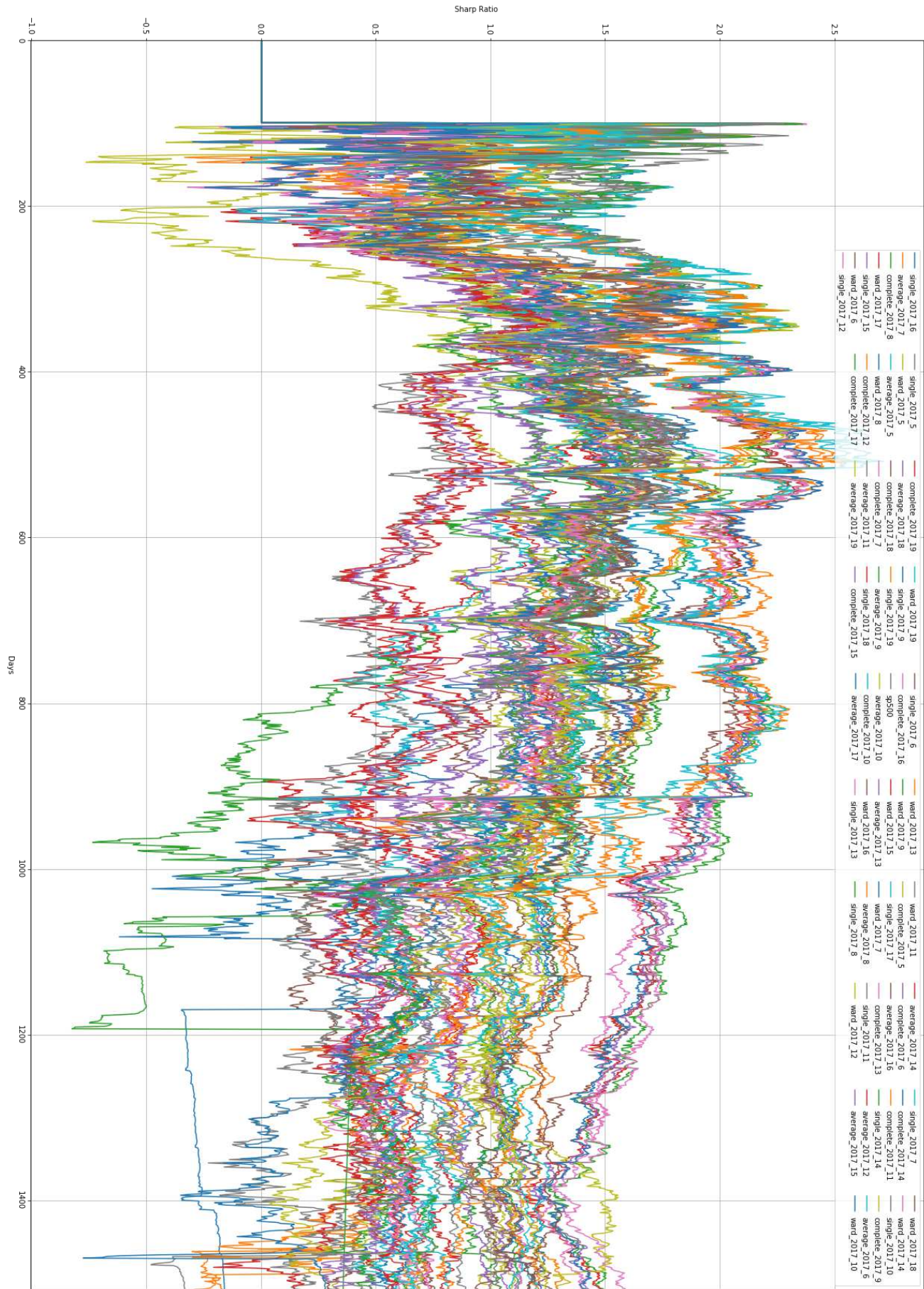


Figure A.3: Portfolio return for 60 candidate learned sector universes.

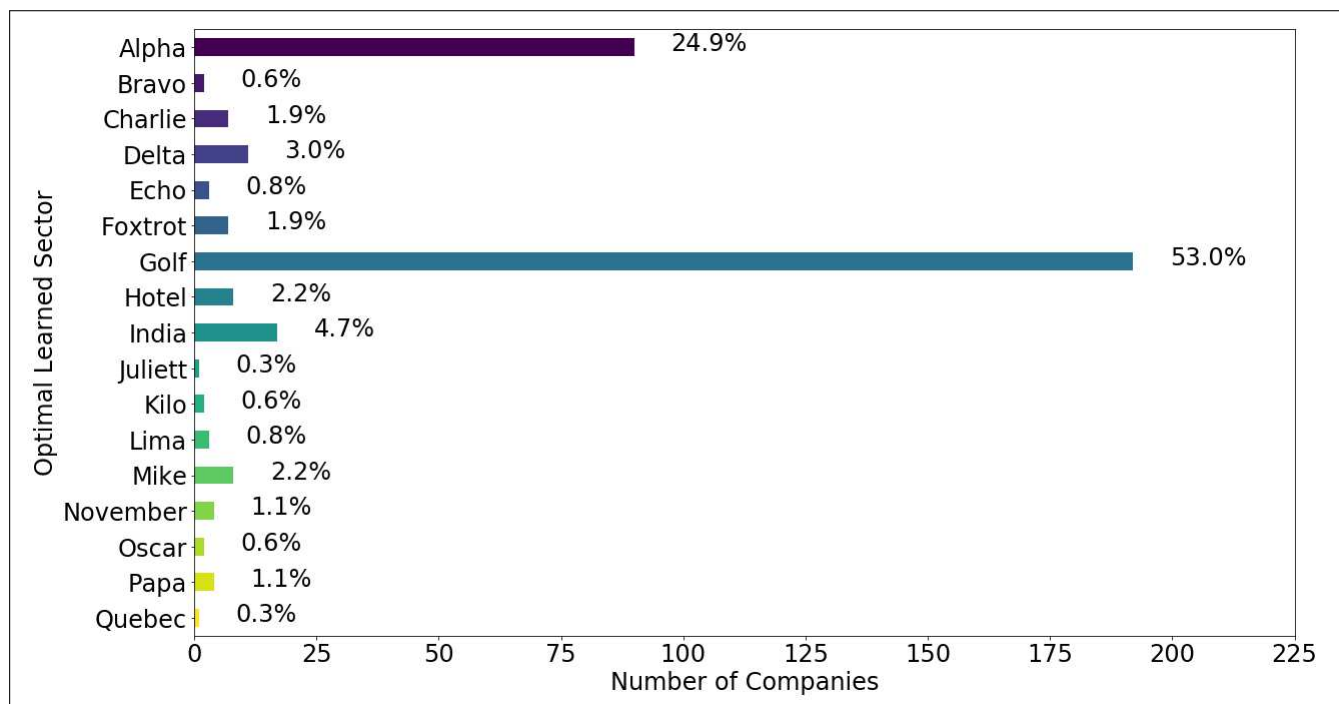




## Appendix B

# Optimal Learned Sector Universe

### B.1 Optimal Learned Sector Asset Distribution



**Figure B.1:** *Distribution of companies ( $n = 362$ ) across sectors in the risk-adjusted return optimal learned sector universe.*

## B.2 Optimal Learned Sector Universe Dataset

Ticker	GICS Sector	Optimal LS	Ticker	GICS Sector	Optimal LS
A	Health Care	Golf	BKNG	Consumer Discretionary	Foxtrot
AAP	Consumer Discretionary	Golf	BLL	Materials	Golf
AAPL	Information Technology	Golf	BRK.B	Financials	Alpha
ABC	Health Care	Kilo	BSX	Health Care	Golf
ABT	Health Care	Alpha	BWA	Consumer Discretionary	Golf
ACN	Information Technology	India	BXP	Real Estate	Golf
ADBE	Information Technology	Golf	C	Financials	Alpha
ADI	Information Technology	Alpha	CAH	Health Care	Kilo
ADM	Consumer Staples	India	CAT	Industrials	Golf
ADP	Information Technology	Alpha	CB	Financials	Alpha
ADS	Information Technology	Alpha	CBOE	Financials	Alpha
AEE	Utilities	Alpha	CCI	Real Estate	Golf
AEP	Utilities	Alpha	CCL	Consumer Discretionary	Golf
AES	Utilities	Golf	CDNS	Information Technology	Lima
AGN	Health Care	Alpha	CERN	Health Care	Golf
AIG	Financials	Alpha	CF	Materials	Golf
AJG	Financials	Alpha	CHD	Consumer Staples	Golf
AKAM	Information Technology	Golf	CHTR	Consumer Discretionary	Golf
ALGN	Health Care	Golf	CI	Health Care	Alpha
ALK	Industrials	Golf	CINF	Financials	Alpha
ALL	Financials	Alpha	CL	Consumer Staples	Mike
ALXN	Health Care	Alpha	CLX	Consumer Staples	Mike
AMAT	Information Technology	Golf	CMA	Financials	Alpha
AMD	Information Technology	Lima	CMCSA	Consumer Discretionary	Golf
AME	Industrials	Golf	CME	Financials	Alpha
AMG	Financials	Delta	CMG	Consumer Discretionary	India
AMGN	Health Care	Golf	CMI	Industrials	Golf
AMZN	Consumer Discretionary	Lima	CMS	Utilities	Golf
ANTM	Health Care	India	CNC	Health Care	India
AON	Financials	Alpha	CNP	Utilities	Golf
AOS	Industrials	Golf	COF	Financials	Alpha
APC	Energy	Delta	COL	Industrials	Alpha
APD	Materials	Golf	COO	Health Care	Golf
APH	Information Technology	Golf	COP	Energy	Delta
APTV	Consumer Discretionary	Golf	COST	Consumer Staples	Hotel
ARE	Real Estate	Golf	CPB	Consumer Staples	Golf
ATVI	Information Technology	Golf	CSCO	Information Technology	Golf
AVB	Real Estate	Golf	CSX	Industrials	Golf
AVGO	Information Technology	Golf	CTSH	Information Technology	Golf
AVY	Materials	Golf	CTXS	Information Technology	Golf
AWK	Utilities	Golf	CVS	Consumer Staples	India
AXP	Financials	Alpha	CVX	Energy	Delta
AYI	Industrials	Golf	CXO	Energy	Delta
AZO	Consumer Discretionary	Mike	DAL	Industrials	Golf
BA	Industrials	Golf	DE	Industrials	Alpha
BAC	Financials	Alpha	DFS	Financials	Alpha
BAX	Health Care	Golf	DGX	Health Care	Golf
BBT	Financials	Alpha	DHI	Consumer Discretionary	Papa
BBY	Consumer Discretionary	Hotel	DHR	Health Care	Alpha
BDX	Health Care	Alpha	DIS	Consumer Discretionary	Golf
BEN	Financials	Golf	DISCA	Consumer Discretionary	Golf
BIIB	Health Care	Golf	DISH	Consumer Discretionary	Delta
BK	Financials	Alpha	DLR	Real Estate	Golf

Ticker	GICS Sector	Optimal LS
DOV	Industrials	Golf
DRE	Real Estate	Golf
DTE	Utilities	Alpha
DUK	Utilities	Alpha
DVA	Health Care	Golf
DWDP	Materials	Alpha
EBAY	Information Technology	Golf
ECL	Materials	Golf
ED	Utilities	Alpha
EFX	Industrials	Golf
EIX	Utilities	Alpha
EL	Consumer Staples	Foxtrot
EMR	Industrials	Golf
EOG	Energy	Echo
EQIX	Real Estate	Golf
EQR	Real Estate	Golf
ES	Utilities	Alpha
ESS	Real Estate	Golf
ETFC	Financials	Alpha
ETN	Industrials	Golf
ETR	Utilities	Alpha
EW	Health Care	Golf
EXC	Utilities	Alpha
EXPD	Industrials	India
EXPE	Consumer Discretionary	Foxtrot
F	Consumer Discretionary	Golf
FAST	Industrials	Mike
FB	Information Technology	Charlie
FBHS	Industrials	Golf
FCX	Materials	Golf
FE	Utilities	Golf
FFIV	Information Technology	Charlie
FIS	Information Technology	Golf
FISV	Information Technology	Golf
FITB	Financials	Alpha
FLIR	Information Technology	Golf
FLR	Industrials	India
FLS	Industrials	Golf
FMC	Materials	Alpha
FOXA	Consumer Discretionary	Golf
FRT	Real Estate	Golf
GD	Industrials	Golf
GILD	Health Care	Golf
GLW	Information Technology	Golf
GM	Consumer Discretionary	Golf
GPC	Consumer Discretionary	Golf
GPN	Information Technology	Alpha
GRMN	Consumer Discretionary	Golf
GT	Consumer Discretionary	Golf
GWW	Industrials	Hotel
HAL	Energy	Golf
HAS	Consumer Discretionary	Foxtrot
HBAN	Financials	Alpha
HBI	Consumer Discretionary	Golf
HCA	Health Care	November
HCP	Real Estate	Golf

Ticker	GICS Sector	Optimal LS
HD	Consumer Discretionary	Mike
HES	Energy	Echo
HIG	Financials	Alpha
HII	Industrials	Golf
HOLX	Health Care	Golf
HON	Industrials	Golf
HP	Energy	Echo
HPQ	Information Technology	India
HRL	Consumer Staples	Golf
HRS	Information Technology	Golf
HSIC	Health Care	Golf
HST	Real Estate	Golf
HSY	Consumer Staples	Mike
HUM	Health Care	India
IBM	Information Technology	Golf
ICE	Financials	Alpha
IDXX	Health Care	Golf
ILMN	Health Care	Golf
INCY	Health Care	Bravo
INFO	Industrials	Alpha
INTC	Information Technology	Golf
INTU	Information Technology	Charlie
IP	Materials	Golf
IPG	Consumer Discretionary	Alpha
IPGP	Information Technology	Golf
IR	Industrials	Golf
IRM	Real Estate	Golf
ISRG	Health Care	Golf
IT	Information Technology	Alpha
ITW	Industrials	Golf
JBHT	Industrials	India
JCI	Industrials	Golf
JEC	Industrials	India
JNJ	Health Care	Golf
JNPR	Information Technology	Golf
JPM	Financials	Alpha
JWN	Consumer Discretionary	Hotel
K	Consumer Staples	Golf
KEY	Financials	Alpha
KIM	Real Estate	Golf
KLAC	Information Technology	Golf
KMB	Consumer Staples	Golf
KMI	Energy	Golf
KO	Consumer Staples	Golf
KSU	Industrials	Golf
LEG	Consumer Discretionary	Golf
LEN	Consumer Discretionary	Papa
LH	Health Care	Alpha
LKQ	Consumer Discretionary	Golf
LLL	Industrials	Golf
LLY	Health Care	Golf
LMT	Industrials	Golf
LNT	Utilities	Alpha
LOW	Consumer Discretionary	Hotel
LRCX	Information Technology	Golf
LUV	Industrials	Golf

Ticker	GICS Sector	Optimal LS
LYB	Materials	Golf
M	Consumer Discretionary	Golf
MA	Information Technology	Charlie
MAC	Real Estate	Golf
MAS	Industrials	Golf
MAT	Consumer Discretionary	Foxtrot
MCD	Consumer Discretionary	Golf
MDLZ	Consumer Staples	Alpha
MET	Financials	Alpha
MGM	Consumer Discretionary	Golf
MHK	Consumer Discretionary	Golf
MKC	Consumer Staples	Alpha
MMC	Financials	Golf
MMM	Industrials	Golf
MNST	Consumer Staples	Golf
MO	Consumer Staples	Golf
MOS	Materials	Alpha
MRO	Energy	Delta
MSFT	Information Technology	Golf
MSI	Information Technology	Golf
MTB	Financials	Alpha
MTD	Health Care	Golf
MU	Information Technology	Golf
NAVI	Financials	Alpha
NBL	Energy	Delta
NCLH	Consumer Discretionary	Golf
NDAQ	Financials	Alpha
NEE	Utilities	Golf
NEM	Materials	Golf
NFLX	Information Technology	Alpha
NI	Utilities	Alpha
NKTR	Health Care	Juliett
NLSN	Industrials	Golf
NOC	Industrials	Alpha
NOV	Energy	Alpha
NSC	Industrials	Golf
NTRS	Financials	Alpha
NUE	Materials	Golf
NVDA	Information Technology	Charlie
NWL	Consumer Discretionary	Golf
O	Real Estate	Golf
OMC	Consumer Discretionary	Alpha
ORLY	Consumer Discretionary	Mike
OXY	Energy	Delta
PBCT	Financials	Alpha
PCAR	Industrials	Alpha
PEP	Consumer Staples	Golf
PG	Consumer Staples	Golf
PH	Industrials	Golf
PHM	Consumer Discretionary	Papa
PKI	Health Care	Alpha
PLD	Real Estate	Delta
PM	Consumer Staples	Golf
PNR	Industrials	Golf
PNW	Utilities	Alpha
PPG	Materials	Golf

Ticker	GICS Sector	Optimal LS
PPL	Utilities	Golf
PRGO	Health Care	Golf
PSX	Energy	Delta
PWR	Industrials	India
QCOM	Information Technology	Golf
RCL	Consumer Discretionary	Golf
RE	Financials	Alpha
REG	Real Estate	Golf
REGN	Health Care	Golf
RF	Financials	Alpha
RHI	Industrials	Mike
RJF	Financials	Alpha
RMD	Health Care	Golf
ROK	Industrials	Golf
ROP	Industrials	Alpha
RSG	Industrials	Golf
RTN	Industrials	Golf
SBAC	Real Estate	November
SBUX	Consumer Discretionary	Oscar
SCHW	Financials	Papa
SEE	Materials	Golf
SHW	Materials	Golf
SIVB	Financials	Alpha
SLB	Energy	Alpha
SLG	Real Estate	Golf
SNA	Consumer Discretionary	Golf
SNPS	Information Technology	Golf
SO	Utilities	Alpha
SPG	Real Estate	Golf
SPGI	Financials	Golf
SRE	Utilities	Alpha
STI	Financials	Alpha
STT	Financials	Alpha
STX	Information Technology	Golf
SWK	Consumer Discretionary	Golf
SWKS	Information Technology	Charlie
SYK	Health Care	Golf
YYY	Consumer Staples	Hotel
T	Telecommunication Services	Golf
TAP	Consumer Staples	Alpha
TDG	Industrials	November
TEL	Information Technology	Golf
TPR	Consumer Discretionary	Golf
TRIP	Consumer Discretionary	Foxtrot
TROW	Financials	Oscar
TRV	Financials	Alpha
TSCO	Consumer Discretionary	Hotel
TSN	Consumer Staples	India
TSS	Information Technology	Golf
TXN	Information Technology	Charlie
TXT	Industrials	Golf
UAA	Consumer Discretionary	Foxtrot
UAL	Industrials	Golf
UDR	Real Estate	Alpha
UHS	Health Care	Golf
UNH	Health Care	India



<b>Ticker</b>	<b>GICS Sector</b>	<b>Optimal LS</b>
UNM	Financials	Alpha
UNP	Industrials	Golf
UPS	Industrials	India
URI	Industrials	November
USB	Financials	Alpha
UTX	Industrials	Golf
V	Information Technology	Golf
VAR	Health Care	Golf
VMC	Materials	Alpha
VNO	Real Estate	Golf
VRSK	Industrials	Golf
VRSN	Information Technology	Quebec
VRTX	Health Care	Bravo
VTR	Real Estate	Golf
VZ	Telecommunication Services	Golf
WBA	Consumer Staples	India
WDC	Information Technology	Golf
WEC	Utilities	Alpha
WFC	Financials	Alpha
WHR	Consumer Discretionary	Golf
WM	Industrials	Golf
WMB	Energy	Golf
WMT	Consumer Staples	Hotel
WRK	Materials	Alpha
WU	Information Technology	Golf
WY	Real Estate	Golf
XEC	Energy	Golf
XEL	Utilities	Alpha
XXRX	Information Technology	Golf
XYL	Industrials	Golf
ZION	Financials	Alpha
ZTS	Health Care	Golf

## Appendix C

# Backtest Portfolio Weights

### Optimal Learned Sector Universe Portfolio

See Figure C.1 on Page 45.

### Benchmark Sector Universe Portfolio

See Figure C.2 on Page 46.

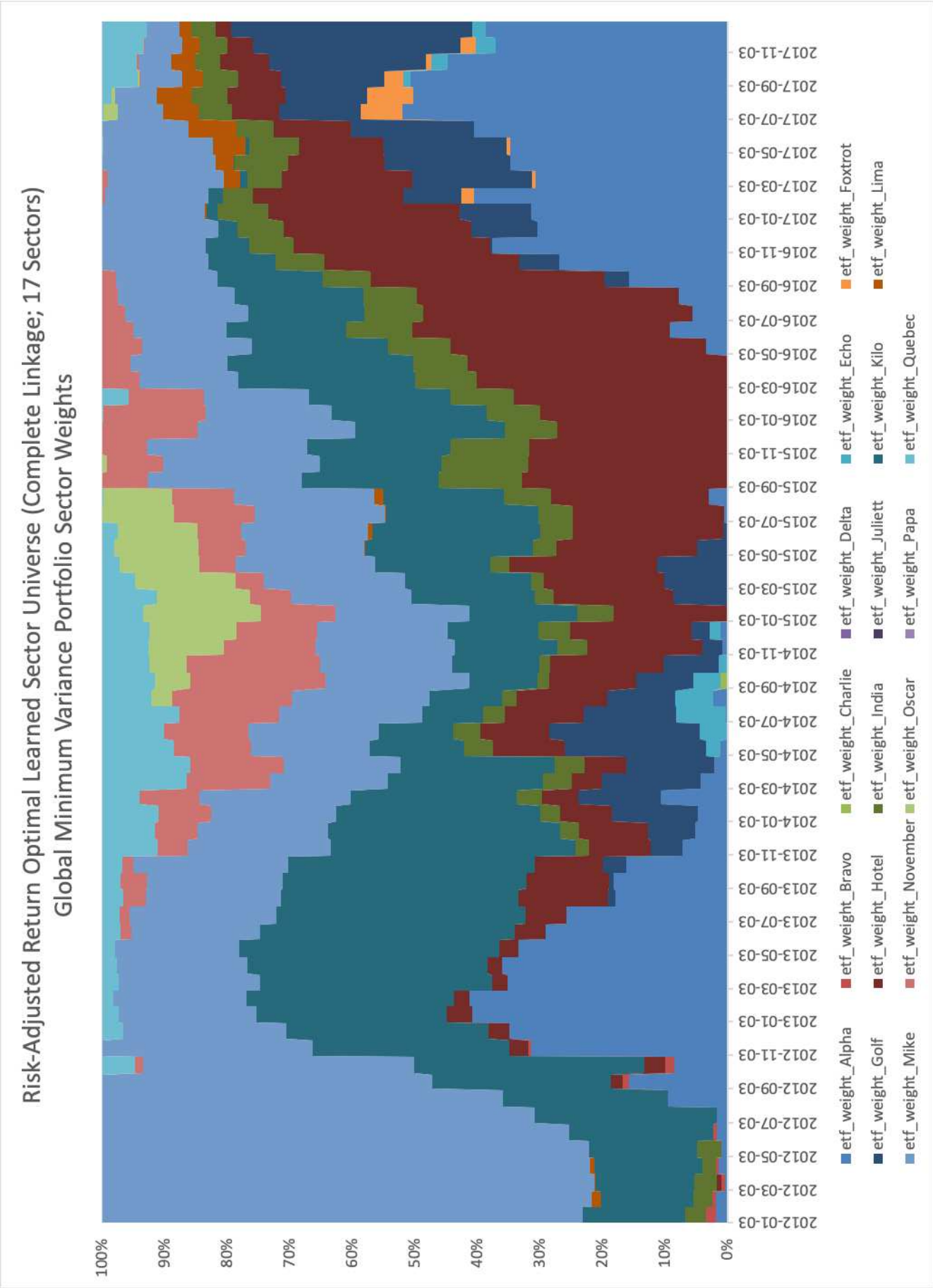
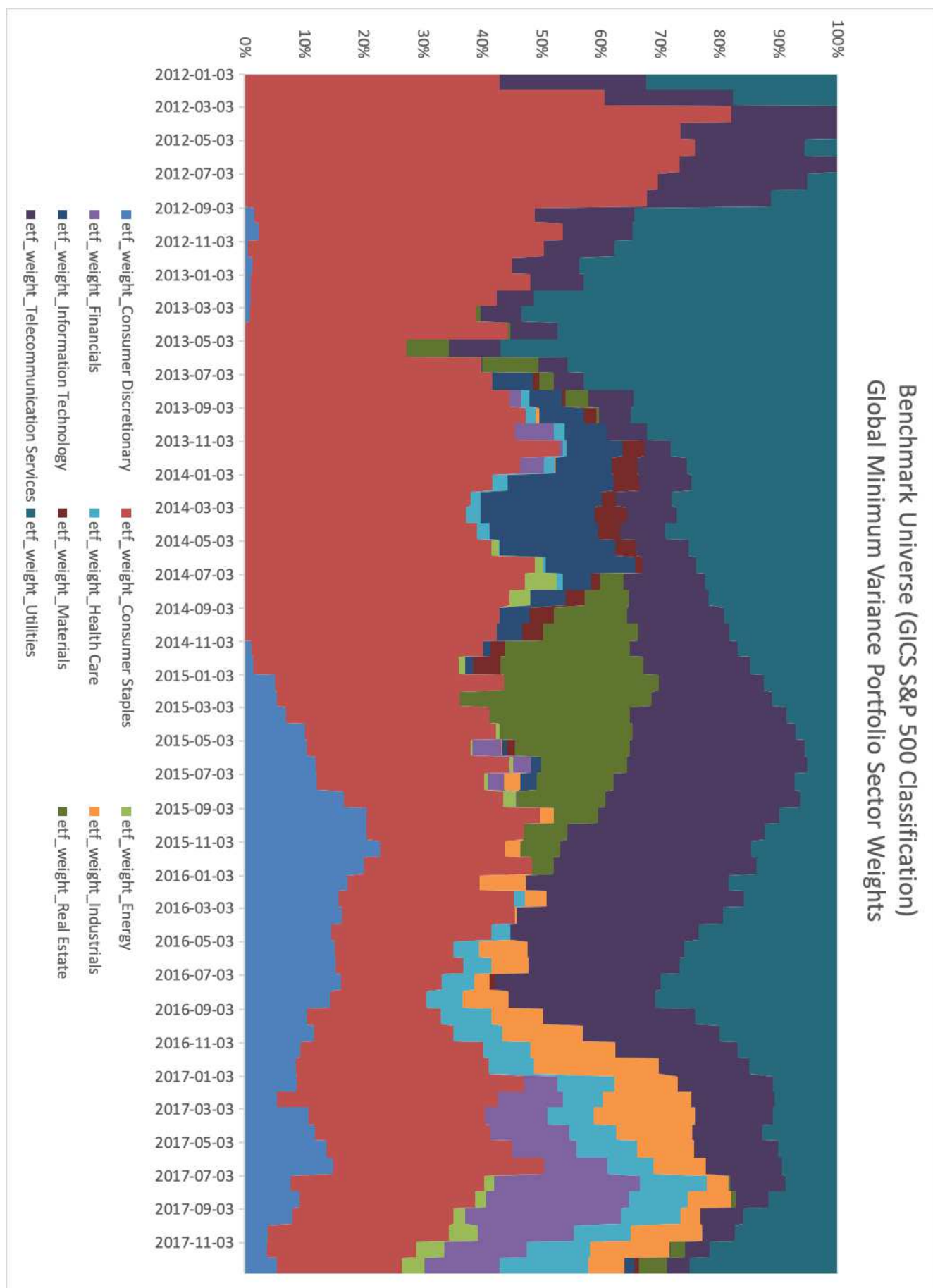


Figure C.1: Risk-adjusted return optimal learned sector universe portfolio assignment weights.



**Figure C.2:** Benchmark sector universe (i.e. GICS S&P 500 Classification) portfolio assignment weights.

# Bibliography

- Ben-Hur, Asa, David Horn, Hava T Siegelmann, and Vladimir Vapnik. 2001. "Support vector clustering." *Journal of machine learning research* 2 (Dec): 125–137.
- Chen, Jean J. 2004. "Determinants of capital structure of Chinese-listed companies." *Journal of Business Research* 57, no. 12 (December): 1341–1351. ISSN: 0148-2963. doi:10.1016/S0148-2963(03)00070-5. <https://www.sciencedirect.com/science/article/abs/pii/S0148296303000705>.
- Defays, D. 1977. "An efficient algorithm for a complete link method." *The Computer Journal* 20, no. 4 (April): 364–366. ISSN: 0010-4620. doi:10.1093/comjnl/20.4.364. <https://academic.oup.com/comjnl/article-lookup/doi/10.1093/comjnl/20.4.364>.
- Delcours, Natalya. 2007. "The determinants of capital structure in transitional economies." *International Review of Economics & Finance* 16, no. 3 (January): 400–415. ISSN: 1059-0560. doi:10.1016/J.IREF.2005.03.005. <https://www.sciencedirect.com/science/article/pii/S1059056006000086>.
- Economic Classification Policy Committee (ECPC). 2007. *2017 North American Industry Classification System (NAICS) Manual*. Technical report. Washington, D.C.: United States Office of Management and Budget. <https://www.census.gov/library/publications/2017/econ/2017-naics-manual.html>.
- Fidelity Investments. 2019. *Know Your Sectors and Industries*. <https://www.fidelity.com/learning-center/trading-investing/markets-sectors/know-your-sectors-and-industries>.
- Fitch Ratings. 2019. *Procedures and Methodologies for Determining Credit Ratings*. Technical report. New York.
- FTSE International Limited. 2019. *Industry Classification Benchmark (ICB) — FTSE Russell*. <https://www.ftserussell.com/financial-data/industry-classification-benchmark-icb>.
- Google Research. 2019. *Google Research Colaboratory*. Mountain View. <http://colab.research.google.com>.
- Hicks, Diana. 2011. "Structural change and industrial classification." *Structural Change and Economic Dynamics* 22 (2): 93–105. doi:10.1016/j.strueco.2011.03.001. <https://www.sciencedirect.com/science/article/abs/pii/S0954349X11000208>.
- Hill, Nick, Roland Auquier, Gregory Bauer, Michael Foley, Mark Lamonte, and Frederic Drevon. 2016. *Financial Institution Rating Methodology*. Technical report. New York: Moody's Investor Service. [www.moodys.com/methodologies](http://www.moodys.com/methodologies).
- Hrazdil, Karel, Kim Trottier, and Ray Zhang. 2013. "A comparison of industry classification schemes: A large sample study." *Economics Letters* 118:77–80. doi:10.1016/j.econlet.2012.09.022. [www.elsevier.com/locate/econlet](http://www.elsevier.com/locate/econlet).
- Huang, Jih-Jeng, Gwo-Hshiung Tzeng, and Chorng-Shyong Ong. 2007. "Marketing segmentation using support vector clustering." *Expert Systems with Applications* 32, no. 2 (February): 313–317. doi:10.1016/J.ESWA.2005.11.028. <https://www.sciencedirect.com/science/article/pii/S0957417405003404>.
- Lloyd, S. 1982. "Least squares quantization in PCM." *IEEE Transactions on Information Theory* 28, no. 2 (March): 129–137. ISSN: 0018-9448. doi:10.1109/TIT.1982.1056489. <http://ieeexplore.ieee.org/document/1056489/>.
- Markowitz, Harry. 1952. "Portfolio Selection." *The Journal of Finance* 7, no. 1 (March): 77–91. doi:10.1111/j.1540-6261.1952.tb01525.x. <http://doi.wiley.com/10.1111/j.1540-6261.1952.tb01525.x>.
- Modigliani, Franco, and Merton H. Miller. 1958. "The Cost of Capital, Corporation Finance and the Theory of Investment." *The American Economic Review* 48 (3): 261–297. doi:10.2307/1809766. <https://www.jstor.org/stable/1809766>.

- MSCI - Morgan Stanley Capital International. 2019. *GICS - Global Industry Classification Standard - MSCI*. <https://www.msci.com/gics>.
- MSCI Research. 2018. *Consultation on Implementation of 2018 GICS Changes in the MSCI Equity Indexes*. Technical report. New York. <https://www.msci.com/documents/1296102/8328554/GICS2018Consultation.pdf/0f246611-27f7-4126-b7f0-02a9255724d5>.
- Office of Statistical Standards - Bureau of the Budget. 1957. *History of the Standard Industrial Classification*. Technical report. Washington, D.C.: Executive Office of the President of the United States. <http://www.census.gov/epcd/www/sichist.htm> [4/5/2011 8:17:06AM].
- Oliphant, Travis E. 2007. "Python for Scientific Computing." *Computing in Science & Engineering* 9 (3): 10–20. ISSN: 1521-9615. doi:10.1109/MCSE.2007.58. <http://ieeexplore.ieee.org/document/4160250/>.
- Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. 2011. "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research* 12 (Oct): 2825–2830. <http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html>.
- Quantopian Inc. 2019a. *Alphalens - Performance Analysis of Predictive (alpha) Stock Factors*. Boston. <http://quantopian.github.io/alphalens/>.
- . 2019b. *Pyfolio - Portfolio and Risk Analytics in Python*. Boston. <https://quantopian.github.io/pyfolio/>.
- . 2019c. *Zipline - Pythonic Algorithmic Trading Library*. Boston. <http://www.zipline.io/>.
- Seifoddini, Hamid K. 1989. "Single linkage versus average linkage clustering in machine cells formation applications." *Computers & Industrial Engineering* 16 (3): 419–426. ISSN: 0360-8352.
- Sibson, R. 1973. "SLINK: An optimally efficient algorithm for the single-link cluster method." *The Computer Journal* 16, no. 1 (January): 30–34. ISSN: 0010-4620. doi:10.1093/comjnl/16.1.30. <https://academic.oup.com/comjnl/article-lookup/doi/10.1093/comjnl/16.1.30>.
- S&P Global Market Intelligence. 2018. *Global Industry Classification Standard - Methodology*. Technical report. New York. [https://www.spglobal.com/marketintelligence/en/documents/112727-gics-mapbook\\_2018\\_v3\\_letter\\_digitalspreads.pdf](https://www.spglobal.com/marketintelligence/en/documents/112727-gics-mapbook_2018_v3_letter_digitalspreads.pdf).
- . 2019. *Compustat Data from S&P Global Market Intelligence*. S&P Global, New York. <https://www.spglobal.com/marketintelligence/en/solutions/fundamental-data>.
- Standard & Poor's Rating Services. 2014. *Corporate Ratings Methodology*. Technical report. Sydney. <https://www.spratings.com/documents/20184/774196/Corporate+Ratings+Methodology.pdf>.
- Tedlow, Richard S. 1996. *New and improved : the story of mass marketing in America*. 483. Harvard Business School Press. ISBN: 9780875846729. <https://www.amazon.com/New-Improved-Story-Marketing-America/dp/0875846726>.
- The Wharton School. 1993. *Wharton Research Data Services*. Philadelphia. <https://wrds-web.wharton.upenn.edu/wrds/>.
- Tsai, C.-Y., and C.-C. Chiu. 2004. "A purchase-based market segmentation methodology." *Expert Systems with Applications* 27, no. 2 (August): 265–276. ISSN: 0957-4174. doi:10.1016/J.ESWA.2004.02.005. <https://www.sciencedirect.com/science/article/pii/S0957417404000132>.
- United States Office of Management and Budget. 1997. *North American Industry Classification System - 1987 Standard Industrial Classification Replacement Notice*. Technical report. Washington, D.C.: Executive Office of the President of the United States. [https://www.census.gov/eos/www/naics/federal\\_register\\_notices/notices/fr09ap97.pdf](https://www.census.gov/eos/www/naics/federal_register_notices/notices/fr09ap97.pdf).
- U.S. Securities and Exchange Commission. 2019. *Form 10-K*. <https://www.sec.gov/fast-answers/answers-form10khtml.html>.
- Vernimmen, Pierre, Pascal Quiry, Maurizio Dallocchio, Yann Le Fur, and Antonio Salvi. 2005. *Corporate Finance - Theory and Practice*. 1st ed. 1059. Chichester: John Wiley & Sons, Ltd. ISBN: 0-470-09225-4.
- Wang, Gang-Jin, Chi Xie, and H Eugene Stanley. 2018. "Correlation structure and evolution of world stock markets: Evidence from Pearson and partial correlation-based networks." *Computational Economics* 51 (3): 607–635. ISSN: 0927-7099.
- Ward, Joe H., and Jr. 1963. "Hierarchical Grouping to Optimize an Objective Function." *Journal of the American Statistical Association* 58, no. 301 (March): 236. ISSN: 01621459. doi:10.2307/2282967. <https://www.jstor.org/stable/2282967?origin=crossref>.

- Weerawarana, Rukmal. 2019. *reIndexer - A Research Tool for Backtest-Driven Evaluation of Different Sectorization Universes*. New York. <https://git.rukmal.me/reIndexer>.
- Weerawarana, Rukmal, Yiyi Zhu, and Yuzhen He. 2019a. *FE 800 Learned Sectors Report Source Code*. <http://git.rukmal.me/FE-800-Learned-Sectors-Report>.
- . 2019b. *Learned Sector Analytics*. New York. [https://colab.research.google.com/drive/1MvcyrTJW05tH09-bpFY46o6b2KW0\\_rm1](https://colab.research.google.com/drive/1MvcyrTJW05tH09-bpFY46o6b2KW0_rm1).
- Weiner, Christian. 2005. “The Impact of Industry Classification Schemes on Financial Research.” *SSRN Electronic Journal*. ISSN: 1556-5068. doi:10.2139/ssrn.871173. <http://sfb649.wiwi.hu-berlin.de>.

# Glossary

## Benchmark Universe

For this project, the benchmark sector classification universe is the *GICS S&P 500 Classification*.

## ETF

Exchange traded fund.

## GICS

Global Industry Classification System; informs the S&P 500 Sector Classifications.

## HCA

Hierarchical clustering analysis model.

## NAICS

North American Industry Classification System.

## Portfolio Rebalancing Turnover

The dollar-value change of constituent SETF turnover incurred when a portfolio of SETFs is rebalanced.

## reIndexer

Research tool for backtest-driven evaluation of different sectorization universes, using a system of synthetic ETFs, and efficient portfolios of those synthetic ETFs.

## Sector Universe

A specific sector classification taxonomy, such as the GICS (Global Industry Classification Standard), or the ICB (Industry Classification Benchmark).

## SETF

Synthetic Exchange Traded Fund; a hypothetical asset, used in backtesting simulations by the *reIndexer* research tool.

## SETF Restructuring Turnover

The dollar-value change of component asset turnover incurred when a synthetic ETF is restructured.

## SIC

Standard Industrial Classification.

## WRDS

Wharton Research Data Services.