EMPIRICALLY IDENTIFIED INDUSTRY CLASSIFICATION

A Dissertation

presented to

the Faculty of the Graduate School

at the University of Missouri-Columbia

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

---

by

MICHAEL GIBBS

Dr. Dan French, Dissertation Supervisor

MAY 2016

The undersigned, appointed by the dean of the Graduate School, have examined the dissertation entitled

EMPIRICALLY IDENTIFIED INDUSTRY CLASSIFICATION

presented by Michael Gibbs,

a candidate for the degree of Ph.D. of Business Administration, and hereby certify that, in their opinion, it is worthy of acceptance.

Professor Dan French

Professor Michael O'Doherty

Professor Kuntara Pukthuanthong

Professor Inder Khurana

To my family, friends, and especially my wife

-for their support

Acknowledgements

I would like to thank numerous people for their help on my dissertation. I would like to sincerely thank Professor Dan French for chairing my dissertation and the many hours he has spent to ensure I produced a quality product. I would like to thank Professors Mike O'Doherty and Kuntara Pukthuanthong for sitting on my committee and all of the valuable feedback. I would like to thank Professor Inder Khurana for sitting on my committee as an outside member. Additionally I would like to acknowledge all of the professors in the Finance Department for feedback and suggestions. Outside of the University of Missouri, I thank the faculty of Oklahoma State University, Binghamton University, participants of the South West Finance Association, and especially Professor Andrew Lynch of the University of Mississippi.

Table of Contents

ACKNOWLEDGEMENTS..... ii

LIST OF FIGURES v

LIST OF TABLES..... vii

LIST OF ABBREVIATIONS..... viii

ABSTRACT..... ix

Chapter 1. CLASSIFICATION OF FIRMS INTO INDUSTRIES USING MARKET DATA…….1

Classification Systems…….3

Method……...…..6

Model……...9

Estimating the Model……..…13

Classification Methods……...……23

Regression Methods……...29

Empirical Evaluation of Selection Methods……..……32

Results……..……..35

Robustness……...……..38

Conclusions…….……..39

Chapter 2. MEASURING DIVERSIFICATION LEVEL THROUGH STOCK AND INDUSTRY RETURN CORRELATION……..…….……58

Introduction…….…..58

Literature Review…….…….……60

Empirically Identifying Industries.....66

Data sources and sample creation.....69

Calculation of Diversification Level.....71

Causes of Time Trends.....80

Conclusions.....91

Chapter 3. EMPIRICALLY IDENTIFYING INDUSTRY RETURNS……..112

Introduction.....112

Empirical Method.....112

Data sources and sample creation.....114

Methods and Results.....114

Conclusions.....121

Reverences.....140

VITA.....150

LIST OF FIGURES

Chapter 2

Figure 1-A: Absolute Error Across All Specification of each Method……42

Figure 1-B: Percent Error Across All Specification of each Method……43

Figure 2-A: Absolute Error Across Time Horizon……44

Figure 2-B: Percent Error Across Time Horizon……..……45

Figure 3-A: Absolute Error Across Targets……..……46

Figure 3-B: Percent Error Across Targets……..…47

Figure 4-A: Absolute Error Across Variance……48

Figure 4-B: Percent Error Across Variance……..……..…49

Figure 5-A: Squared Error Across All Specification of each Method……..…..…50

Figure 5-B: Squared Error Across Time Horizon……..……51

Figure 5-C: Squared Error Across Targets……52

Figure 5-D: Squared Error Across Variance…….……53

Figure 6-A: Error Across Methods, The Worst Performance Case……..…54

Figure 6-B: Percent Error Across Methods, The Worst Performance Case…..…55

Figure 6-C: Error Across Methods, The Best Performance Case…….……56

Figure 6-D: Percent Error Across Methods, The Best Performance Case……57

Chapter 3

Figure 7: Diversification Level Through Time……104

Figure 8-A: Diversification Premium Through Time By Market to Book……..105

Figure 8-B: Diversification Premium Through Time By Tobin's Q……..…106

Figure 8-C: Diversification Premium Through Time By Fama MacBeth MB…107

Figure 8-D: Diversification Premium Through Time By Fama MacBeth TQ…108

Figure 8-E: Diversification Premium Through Time By Portfolio Returns……109

Figure 9-A: Statistically Significant Diversification Premium Through Time MB……...……110

Figure 9-B: Statistically Significant Diversification Premium Through Time TQ……..……..111

APPENDIX I

Figure A1: Diversification Premium Mapped Against Credit Spread MB…144

Figure A2: Diversification Premium Mapped Against Credit Spread TQ……145

LIST OF TABLES

Chapter 3

Table 1: Summary Statistics on Diversification Level: January 1975 to December 2013

Table 2: Fama-MacBeth Regressions of Valuation Proxies on Measures of Diversification

Table 3: Risk Factor Explanation of Diversification Discount

Table 4: Comparison of Results

Table 5: Explanations for Time Varying Diversification Premium

Chapter 4

Table 6: SIC Code Classification System

Table 7: Implied Idiosyncratic Volatility by Classification

Table 8: Correlations Across Iteration

Table 9: Industry Summary Statistics

Table 10: Accuracy of Classification System Using Highest R-Squared

Table 11: Consistency of Industries

Table 12: Comparability of Identification Method

Table 13: Sample of Companies from Each Computed Industry

APPENDIX I

Table A1: The effect of Macro Variables on Diversification Through Time

LIST OF ABBREVIATIONS

SIC standard industrial code assigned by classifying agency

TQ Tobin's Q, a commonly accepted measure of firm value

MB Market to Book, a commonly accepted measure of firm value

SEC Security and Exchange Commission

EDGAR Electronic Data Gathering, Analysis, and Retrieval System

FAS Financial Accounting Standards

ABSTRACT

This study examines return based correlations between industry returns and firm returns to create more objective and comparable industry classifications. In my first essay I model a market with firms that invest in one or more categories of assets. Firms that invest in assets with similar return correlations are grouped into categories that are comparable to industry groups in the standard scheme of classifying firms into industries based on offering a common product or service. Because these categories are based on objective rather than subjective criteria, use of these categories by investors might have advantages when using industry information to make investment decisions and construct portfolios. I also derive estimable equations to measure firms' exposures to category risk thereby identifying in which category or categories the firm belongs, and we use simulation to explore the efficacy of three different estimation methods.

In my second essay I evaluate the question does the number of industry exposures (i.e. diversification level) affect corporate value. I find an unconditional diversification premium. However, there is substantial time series variation in the relation between diversification and valuation. This variation is able to reconcile many of the conflicting conclusions in the prior literature.

In my third essay I perform empirical tests to determine whether industry returns can be refined by applying an iterative regression of firm returns on industry returns to create returns than are more inter-correlated but also more orthogonal to other industries' returns. I find strong evidence that an iterative process of return generation provides benefits to researchers as well as practitioners.

Chapter I: Classification of Firms into Industries Using Market Data
Classification of Firms into Industries Using Market Data

Introduction

The Purpose of this research is to investigate the efficacy of using statistical properties of firms to classify them into "industry" groupings rather than the traditional method that bases classification on having firms in the group sell similar products or services. We develop a model of a market composed of firms in one or more industries based on the correlations of their asset portfolios. We then propose and simulate a method for extracting industry category returns from a market and categorizing firms into those industries.

Use of industry classifications is a standard tool among practicing investment managers who use industry classifications as means to aid in a number of investment decisions such as diversification and hedging. For example, a portfolio manager whose goal is to construct a portfolio that is broadly diversified across a market might choose several companies from each of a number of different industries as a method for achieving such diversification. Another manager might be interested in investing in General Motors as a result of a valuation analysis but might hold back if the fund is already invested in another automobile manufacturing company such as Ford.

Industry classification systems currently in use are generally subjective in nature and based on the definition of the "industry" as a set of firms offering mutual products or services. Firms ascertain the primary industry within which they reside and identify this

industry in their financial reports. There are two important issues that can reduce the effectiveness of this procedure. First is that firms may invest in assets from more than one industry. While they may opt to identify secondary industries in which they operate, they are required to identify only the primary industry. Second, firms self-identify their industry and may not correctly identify the industry that generates the largest portion of their returns.

While these industry classifications do provide a useful means to assist in investment decisions, the driving factor that determines the effectiveness of diversification or hedging is not simply membership in a group of firms selling a common product but rather the statistical relationships among the returns of those member firms. We propose observed commonality of firms within an industry group is a result of similarity of the returns on firms' component assets. Therefore, the correlations between firm returns should provide an alternative and potentially more effective method of categorizing firms into common groups compared to the traditional method.

We construct a model of a market composed of assets, each of which will have some degree of commonality with the other assets in the market. These commonalities will result in asset returns that have varying levels of correlations with the remaining market assets. A given asset may have a high correlation with one asset and a low correlation with another, and assets can be grouped into categories ("industries") such that the assets in a given category have correlations with the other assets in that category that are higher than their correlations with assets external to the category. That is, assets

in any category may still be correlated with assets in other categories but at a lower level than with their own category.

Firms in our model market concentrate their holdings of assets in one or more categories, and the result is that firms holding assets in the same category will have returns correlated with the returns on other firms holding assets in the same category. Note that there may be "pure" firms holding assets in only one category but also "diversified" firms holding assets in multiple categories. The problem is that firms self-report a single primary industry but when they hold assets in multiple categories, their firm returns are "polluted" with the returns on their non-primary category assets. These "polluted" returns are used to compute the industry returns even though they contain components from non-industry asset categories. Firms categorized into one industry may actually have greater correlations with firms in one of the other industries. However, identifying cases of such industry misclassification would not be straightforward because the calculated industry returns are polluted with returns from other industry groups.

Because firms investing in a portfolio of multiple asset categories results in a market with firms that have a principal correlation with one asset category but also with correlations to multiple categories, a procedure is needed that uses observed market returns to identify asset categories and to classify firms into those categories. Ideally the procedure will not only identify the primary and secondary industries of a firm but also the percentages each contributes to firm returns. The final purpose of our paper is to propose a method of identifying the asset categories that comprise a firm's asset portfolio and to demonstrate the method with a simulation.

Classification Systems

The primary data source for classification of firms into industries is the firm's financial report. To comply with FAS 131 (which superseded FAS 14 and now covered in FASB ASC 280), firms that are internally organized into operating segments must report certain financial items (such as revenues, interest expense, depreciation expense, profit or loss, and assets) for segments that comprise ten percent or more of the firm's revenues, profits, or assets. Reported segments may aggregate two or more operating segments that have similarities in products, production, customer type, and regulatory environment. Companies must report a sufficient number of segments to cover at least 75 percent of revenues. Firms with no internal segment organization must classify their revenue by similar product or service and report those numbers in a fashion similar to firms organized into segments. When firms report their segment data, they identify the segment by name and description and may identify a corresponding industry classification from one of the systems such as SIC or NAICS.

A number of industry classification systems are currently in use, each with its own benefits as well as shortcomings. Personnel in these systems create classification definitions and may also analyze company financial reports and assign primary and possibly one or more secondary industry classifications to firms. One consequence of this is that each a given firm may not fall into the same industry classification by all systems. For example, one system might assign the primary industry using the segment with the greatest contribution to firm revenue while another might use assets. This disagreement

between classification systems can even have an impact on the results of financial research that uses industry classifications as demonstrated by Kahle and Walkling (1996).

Seven of the more common systems are SIC, NAICS, ISIC, BITS, GICS, ICB, and TRBC, summarized in Appendix I. Of these, SIC, NAICS, ISIC, and BITS are constructs of governmental agencies or partnerships while GICS, ICB, and TRBC are private for-profit organizations. Probably the most widely used is the SIC.

The government-sponsored systems tend to rely on cash-flow data from segment reporting while the privately maintained classification systems tend to include additional properties such as production, customers, etc. The privately created classification systems appear to allocate more importance to how industry classifications affect investors than do governmental classifications systems. Research in the academic community tends to focus on classifications reported in the CRSP (uses SIC assigned by S&P) and Compustat (SIC and NAICS) databases. The SEC in the EDGAR database organizes firms by a primary SIC. The SEC uses these only for assigning SEC review personnel to the financial report, and the firms themselves identify and report this SIC to the SEC.

The concept behind industry classification is to group firms with similar risk exposures together, and existing methods group firms based on the product or service provided by the firm. There is lack of a classification system that identifies a purely objective and quantitative evaluation of investor perception of "industry" exposure. We believe that by offering a classification system that concerns itself only with the consequences of exposure, i.e. investor perception of importance, we can provide a meaningful contribution to the understanding of asset pricing. Industries are among the

most common control variables seen in the finance literature. Often we see studies focus on returns of a firm and generally they will, at least for robustness, control for the industry returns. By concentrating on the largest asset group without consideration for the percent of returns that are affected by the systematic exposure this group experiences, researchers may introduce bias to their study. For example we can consider the event study, where researchers consider the effect an event has on stock returns. Firms that have a high exposure to the reported asset classification would see a more representative effect of the event on stock prices, while those that have only a small portion of their returns affected by the reported industry (but have a large portion affected by other industries) would show having a magnified industry adjusted return (without a bias toward the returns being positive or negative). The bias may lead us to falsely reject results through the additional noise that occurs from incomplete industry adjustment and therefore deflated t-statistics.

Method

Most commonly accepted methods of classification seek to provide information to investors to allow them to assess the risks associated with potential investments. Often this is done through comparison of common business areas and consumer demand. We propose that the systematic risk associated with industry exposure is the correlation of portfolios of underlying assets and macro-economic changes, whether those assets be physical, contractual or abstract, such as human capital. When firms hold common assets we contend that the value of those assets will have similar correlation with the state of the market. Said differently, in an efficient market there are expected returns for each

individual asset and those expected returns are constant across firms. It is logical to assume that firms that operate similarly, will sell similar products, have similar customers, hold similar assets, and as a result have similar asset driven returns. If the expected cash flows to those assets change then the expected returns will change for all firms holding those assets. The degree to which the expected cash flows change will be based on the percent of total assets that the underlying asset makes up. The result is that firm returns will update to represent the change in expected cash flows and therefore expected returns of that particular asset. For example, if a firm holds assets in construction equipment or machinery then we would expect as expectations about the future of construction improve that investors would increase their valuation of that equipment. Clearly the amount of construction ongoing at any point in time is related to the economy as a whole, however there is no reason to suspect that the correlation is perfect. In this example construction represents a sub-sector of the economy that affects the value of a specific group of assets much differently than assets in general.

Jensen and Meckling's (1976) seminal work holds that the firm is merely a "nexus of contracts," which implies that the value of those underlying contracts is the value of the firm. As any of those contracts' values change we would expect to see the expectations about future value changes as well. The result is that total return of the firm should be completely measured by the sum of the values of each individual contract. Edward Thurlow was quoted saying "*Did you ever expect a corporation to have a conscience, when it has no soul to be damned, and no body to be kicked*" implying that a firm's decisions (and therefore its performance) will be the result of mechanics. Those

mechanics could be contracts, or optimal decision making given assets under control. Because assets have measurable value based on the cash flows that can be derived from owning the assets, the price that an investor is willing to pay for the cash flows from said asset is efficient. If two firms had the same asset but the value were lower for one than for the other, either the market would price down the firm with a lower asset driven cash flow or that company would sell the asset to the company with the higher asset driven cash flow, the result is that the expected cash flow would be the same across both firms that continue holding the underlying asset. Because of the market correction mechanism for pricing of underlying assets, we can suggest that assets represent the systematic portion of the stock price, and so the systematic value of the firm is rather a nexus of assets, where the changes in systematic returns are a result of the weighted systematic changes in the value of each individual asset. Any additional changes in firm value (not explainable by assets) are a result of non-systematic value. Additionally, human capital is little different from other assets given that similar products, services, customers, or innovations will make use of similar talent pools and skill sets. One of the most common inspirational claims made by corporations is the value they place on human capital. Goldman Sachs for example, a firm that specializes in valuing assets is quoted on its website as sayings "*Our people are our greatest asset – we say it often and with good reason. It is only with the determination and dedication of our people that we can serve our clients, generate long-term value for our shareholders and contribute to the broader public."* There are innumerable skill sets that provide greater or lesser value given different economic states (for example oil fracking) and these intangible assets will also

experience systematic value changes as those underlying states changes. Whether we consider physical assets, human capital, or legal contracts, we have reason to believe not only that there is a correlation between state of the economy (or state of a segment of the economy) and the value investors place upon those assets, but also, different assets will have different systematic exposure to different sectors of the state of the economy.

We propose that the summation of the individual asset values represents the systematic value of the firm and the summation of the individual assets' systematic risk is equal to the total systematic risk experienced by the firm. Total firm systematic risk can be measured as correlation between a firm's performance and the economy as a whole, and individual assets' systematic risk is measurable by the correlation between the firm's returns and the returns of sub-sectors of the economy. Through the identification of correlations between the sub-sector returns and firm returns we can identify common asset holdings and make comparisons about industry similarities of companies in a consistent way so that investors can compare two firms of similar beta loadings.

In the following section we derive a model to prove that the total systematic risk can be decomposed into a number of correlated systematic risks that can be compared across firms to identify how firms should be classified into common industries. The following section identifies methods to empirically identify these industries. Lastly we perform simulations to determine the accuracy of the empirical identifications.

Model

Define assets as physical items that have value based on current and future expected net cash flows or the legal rights to such cash flows. Let the economy (market)

be composed of a set of assets and the return on each asset n be denoted A_n. The return on the entire market portfolio of assets is therefore equal to

$$M = \sum_{n=1}^{N} a_n A_n \tag{1}$$

where M is the return on the market portfolio of assets, N is the total number of assets in the market, and a_n is the weight of asset n in the market. (Appendix II provides a listing of all equations.) Each weight a_n is the portion of the total market value represented by the value of asset n, where V_j is the market value of asset j.

$$a_n = \frac{V_n}{\sum_{j=1}^{N} V_j} \tag{2}$$

Allow assets to be classified into categories, and denote the return on asset j in category i as $A_{i,j}$. The criterion for inclusion of an asset in a category is all assets that meet

$$\begin{aligned} &\rho\left(A_{i,j}, A_{i,g}\right) > r_i > \rho\left(A_{i,j}, A_{k,h}\right) \\ &\quad \text{for all } i = 1 \ldots I \\ &\qquad j = 1 \ldots J_i \\ &\qquad g = 1 \ldots J_i,\ g \neq j \\ &\qquad k = 1 \ldots I,\ k \neq i \\ &\qquad h = 1 \ldots H_k \end{aligned} \tag{3}$$

where r_i is the cutoff correlation for category i, and I is the number of asset categories, J_i is the number of assets in category i, and H_k is the number of assets in category k. In words, an asset is included in category i when the correlation of the return on that asset j in category i ($A_{i,j}$) with the return on any other same-category asset ($A_{i,g}$) is greater than

a target cutoff correlation which is greater than the return correlation of the asset with any out-of-category-i asset ($A_{k,h}$). Further, let all assets fall into one and only one category.

The return on a category is

$$C_i = \sum_{j=1}^{J_i} c_{i,j} A_{i,j} \tag{4}$$

where C_i is the return on category i and $c_{i,j}$ is the weight of asset j in asset category i. The return on the market portfolio (1) can also be expressed in terms of categorized assets

$$M = \sum_{i=1}^{I} m_i C_i = \sum_{i=1}^{I} \sum_{j=1}^{J_i} a_{i,j} A_{i,j} \tag{5}$$

where m_i is the weight of category i in the total market asset portfolio.

Allow firms to exist that, for simplicity, are unleveraged. Ownership in these firms is represented by shares owned by investors in the market, and firms invest in a portfolio of assets. Some firms will specialize in only one category of assets while others may diversify into mutiple categories. The return on an individual firm is

$$F = \sum_{i=1}^{I} \sum_{j=1}^{J_i} w_{i,j} A_{i,j} \tag{6}$$

Where F is the return on an individual firm and $w_{i,j}$ is the weight in the firm's portfolio of asset j from category i. Firms concentrating in only one category n will have $w_{i,j}$ equal to zero for all $i \neq n$.

Individual asset returns are observable only by the firms who own and operate them, and firms report their asset level returns on a regular periodic basis. Market investors therefore can regularly observe firm returns but only periodically observe asset returns. Furthermore, firms must disclose the principal category of asset they invest in,

but they may or may not disclose other categories they own. Additionally, firms may or may not disclose the percentage of each category comprising their portfolio. As a result, investors will often only know the principal category that a firm is invested in, which may not even represent a majority of the firm. For example, suppose a firm owns assets in categories automotive, music, and electronic instruments, and each represents 40, 35, and 25 percent of the firm respectively. The firm appears to investors as an automotive firm even though this category represents only 40% of the firm's value.

The challenge represented by such a market is that it is not possible to know the returns on categories. Categorizing firms by their principal asset and computing the mean results in only an estimate of the category's true return. Furthermore, we can only assume that these categories are on average correct if we assume that there is not a structural correlation across all firms $asset_i$ given that they own $asset_j$. Taking an example of a firm that sells services such as cable or telecommunications. The firm needs to provide every employee a vehicle and tools even though they have little intention of being in the vehicle or tool industry. Because their human capital component is always paired with the vehicle and tools component, identification of the human capital asset alone becomes problematic. Hence, we must rely on the combination of assets being correlated with other the combination of assets of other firms that operate in the same. Logically, if the value of the tools and vehicles is a significant portion of the firm value (or ability to be profitable) then as the value of the underlying necessary assets changes, so will the value investors are willing to pay to have access to the cash flows derived from those assets.

Assume the Capital Asset Pricing Model (CAPM). The expected return on any individual firm is

$$\tilde{F}_n = r + \beta_{n_i}\left(\tilde{M} - r\right) \tag{7}$$

where $\left(\tilde{M} - r\right)$ represents the excess return of the entire market portfolio, which is the weighted average returns of each individual asset in the market.

In terms of categories, from (5) we can state the expected (excess) market return as

$$\tilde{M} - r = \sum_{i=1}^{I} m_i\left(\tilde{C}_i - r\right) \tag{8}$$

Where C_i-r is the return on the category portfolio and the m_i is weight the category caries in the market portfolio. We substitute (8) into (7) in order to represent the systematic return of the firm as a weighted summation of its individual asset returns, this yields

$$\tilde{F}_n = r + \beta_n \sum_{i=1}^{I} m_i\left(\tilde{C}_i - r\right) \tag{9}$$

Moving β_n within the summation gives

$$\tilde{F}_n = r + \sum_{i=1}^{I} \beta_n m_i\left(\tilde{C}_i - r\right) \tag{10}$$

Algebraically, while (10) is equivalent to (9), there is a subtle difference in concept. The CAPM states that the return on any asset is a function of only the return on the market portfolio, and the asset's β_n fully captures that linear relationship. However, once we move β_n within the summation, there is no necessity that it be identical for every category *i*, only that the weighted sum of the individual $\beta_{n,i}$ be equal to the single CAPM β_n, or

$$\beta_n = \sum_{i=1}^{I} m_i \beta_{n,i} \tag{11}$$

Implying that if the summation of $\beta_{n,i}$ is equal to its CAPM β_n then the returns of the firm must be equal to the summation of the returns of its assets.

Substituting (11) into (10) yields

$$\tilde{F}_n = r + \sum_{i=1}^{I}(m_i(C_i - r)(\sum_{i=1}^{I} m_i \beta_{n,i}))$$

which after moving summations can be rewritten as

$$F_n = r + \sum_{i=1}^{I}\sum_{j=1}^{I} m_i(C_i - r) m_j \beta_{n,j} \quad (12)$$

Equation (12) provides a model showing that the expected return on an individual firm is the sum of the product of weighted expected return on each category *i* and beta risk coefficient of the firm's return with that category return. Because the weight of each category in which the firm holds no investments is zero, only those categories in which the firms operates is relevant to its expected return. That is, the expected return on a firm is a function only of the weighted expected returns on categories in which the firm invests its assets.

Estimating the model

Jensen (1969) shows that we can recast the CAPM from expectations to realized returns in order to estimate model coefficients using a historical return series. The beta can be decomposed into industry level components, i.e. $\beta_{n,i} = \sum_{j=1}^{I} \beta_{n,j} m_j$, and so modifying (17) in such a fashion gives

$$(F_n - r) = \hat{\alpha} + \sum_{i=1}^{I} m_i \hat{\beta}_{n,i} (C_i - r) + \varepsilon_n \quad (13)$$

where $\hat{\alpha}$ and $\hat{\beta}_{n,i}$ are estimates of the intercept and industry level betas respectively. However, because we are using the model to identify the categories in which the firm invests (those categories in which $m_i > 0$), the only estimation that (13) can provide is the weighted beta,

$$(F_n - r) = \alpha + \left(\sum_{i=1}^{I} \sum_{j=1}^{I} \beta_{n,j} m_i m_j (C_i - r) \right) + \varepsilon_n \tag{14}$$

where $\hat{\beta}_{n,i}\hat{m}_j$ is the estimated weighted beta. Therefore, assuming that $\hat{\beta}_{n,i} \neq 0$ for all categories, any estimated $\hat{\beta}_{n,i}\hat{m}_j$ greater than zero signifies that $m_i > 0$; that is, the firm has a significant investment in asset category *i*. Using (13) we cannot estimate the *extent* to which the firm invests in category *i*, only *whether* it invests or not.

Unfortunately there are other issues with the estimation of the model. Each category's returns can be, and likely are, correlated with some or all of the returns of the remaining categories. That is,

$$\rho_{C_i, C_j} > 0 \text{ for all } i, j = \{1 \ldots I\} \tag{15}$$

Where all industries are inter-correlated to some degree. To mitigate this problem, use the CAPM to model specific category returns using the market model.

$$\tilde{C}_i = r + \beta_{C_i} \left(\tilde{M} - r \right) \tag{16}$$

and to estimate that relationship empirically we move the risk free rate to the left and create the standard regression

$$\left(C_i - r \right) = \alpha + \hat{\beta}_{C_i} \left(M - r \right) + \varepsilon_{C_i} \tag{17}$$

Substituting (17) into (14) gives

$$(F_n\text{-}r)=\alpha+\sum_{i=1}^{I}\sum_{j=1}^{I}\beta_{n,j}m_i m_j(\alpha+\beta_{C_i}(M-r)+\varepsilon_{c_i})+\varepsilon_n \tag{18}$$

where *i* corresponds to firm level returns and *j* denotes category level returns, $\beta_{n,j}m_i m_j=\beta_{I,C}$ which is the industry level beta, and $\beta_{I,C}$ comes from $(C_i-r)=\beta_{C_i}(M-r)+\varepsilon_{C_i}$. Where (18) is quantitatively the same as (13) is allows us to decompose market return into category return.

Consider the composition of (19),

$$(\mathrm{F}_n\text{-r})=\alpha+\sum_{i=1}^{I}\sum_{j=1}^{J}\underbrace{\beta_{n,j}m_i m_j}_{\beta_{i,C}}\underbrace{(\alpha+\beta_{C_i}(M-r)+\varepsilon_{c_i})}_{C_i-r}+\varepsilon_n \tag{19}$$

Note that

$$(C_i-r)=\alpha+\beta_{ci}(M-r)+\varepsilon_n \tag{20}$$

and

$$M-r=\sum_{i=1}^{I}m_i\left(C_i-r\right) \tag{21}$$

That is, (20) is equal to the estimated beta loading on the market, and (21) is equal to the market return. To examine a firm for holdings of assets in category *j*, we can estimate equation (19) by estimating its two portions (equations 20 and 21). A resulting estimate of $\beta_{\mathrm{i,C}}$ for any category *i* that is not significantly different from zero then m = 0, or in other words, the firm does not hold assets within this category.

We often are concerned with not only the industries that a firm's assets are in, but also the amount of assets that are in each industry. We have shown that an industry can be empirically identified by a decomposition of the CAPM regression, and the natural

extension is that for each industry to exist in the model, that is $m_i \neq 0$, it must improve the portion of returns explained.

$$\forall \beta_i \in \beta_{i=1}^{N} w_i$$

That is for every industry beta to exist in the CAPM beta, a portion of the sum of squared errors must be reduced by including that industry where $m_i > 0$. The full model is estimated as (21), and the portion of returns that are unexplained is defined as the sum of squares of the model (22)

$$\tilde{F}_n = r + \sum_{i=1}^{I} \beta_n \left(\tilde{C}_i - r \right) + \varepsilon \tag{22}$$

$$(\tilde{F}_n - \overline{F}_n)^2 \tag{23}$$

Where $\tilde{F}_n$ is the firm's returns and $\overline{F}_n$ is the predicted value of the firm's returns. In theory following (3), we assume a small amount of cross-industry correlation, however in reality this becomes an empirical question.

$$\begin{gathered} \rho\left(A_{i,j}, A_{i,g}\right) > r_i > \rho\left(A_{i,j}, A_{k,h}\right) \\ \text{for all } i = 1 \ldots I \\ j = 1 \ldots J_i \\ g = 1 \ldots J_i,\ g \neq j \\ k = 1 \ldots I,\ k \neq i \\ h = 1 \ldots H_k \end{gathered} \tag{18}$$

If we assume that no cross-industry correlation exists then we can evaluate the weights of each industry as

$$1-\frac{SS_{\mathrm{Res}X_i}}{SS_{Total}}$$

$$\textit{from} \qquad (24)$$

$$F_n - r = \alpha + \beta_i\left(\tilde{C}_i - r\right) + \varepsilon$$

Where $SS_{\mathrm{Res}X_i}$ is the residuals squared from a regression of industry X_1 on firm level returns, and SS_{Total} is the sum of squares for the firm level returns. However when industries are correlated, we need to isolate the unique explanatory power of each industry. In (25) we remove the correlated information of $C_{j,\ j=1 \text{ to I but does not equal i}}$, from C_i.

$$\left(\tilde{C}_i - r\right) = \alpha + \sum_{\substack{j=1 \\ i \neq j}}^{I} \beta_j\left(\tilde{C}_j - r\right) + \varepsilon_{i,j} \qquad (25)$$

And we also remove the information from firm level returns that can be explained by any of the other industries in order to isolate the explanatory power of the industry of interest.

$$F_n - r = \alpha + \sum_{\substack{j=1 \\ i \neq j}}^{I} \beta_j\left(\tilde{C}_j - r\right) + \varepsilon_{j,F_n} \qquad (26)$$

$$\text{for each industry i against all other industries j}$$

$$\varepsilon_{j,F_n} = \alpha + \varepsilon_{i,j} + \varepsilon_{F_n,i} \qquad (27)$$

That is the portion of Industry J that cannot be explained empirically by any other industry; from a regression standpoint the left hand is equal to the residuals of the regression of firm returns on all industries other than the industry of interest (j) while the right hand side is the residuals of the regression of industry j on all other industries. Now we can determine the unique information of firm returns explained by the unique information from the industry whose weight we wish to know (27). The natural question

that arise is whether we wish to know the weight of an industry with respect to all assets of the firm, or with respect to the assets that fall within one of the N categories of commonality. Said differently, do we consider weight to be with respect to the explainable portion of F_n (systematic), or with respect to the total variation of F_n (systematic and non-systematic)? If we want the former, we standardize the sum explained variation from the regression of residuals (27) by the sum of explained variation from the full model (22).

$$\tilde{F}_n = r + \sum_{i=1}^{I} \beta_n \left(\tilde{C}_i - r \right) + \varepsilon \tag{22}$$

$$w_j = 1 - \frac{\left(\varepsilon_{j,F_n} - \alpha - \varepsilon_{i,j} \right)^2}{\left(\tilde{F}_n - r + \sum_{i=1}^{I} \beta_n \left(\tilde{C}_i - r \right) \right)^2} \tag{28}$$

If we want the latter, then we instead standardize by the total variation with $\tilde{F}_n$ (23).

$$(\tilde{F}_n - \overline{F}_n)^2 \tag{23}$$

$$w_j = 1 - \frac{\left(\varepsilon_{j,F_n} - \alpha - \varepsilon_{i,j} \right)^2}{\left(\tilde{F}_n - \overline{F}_n \right)^2} \tag{29}$$

The result from equation 29 (or 28) is intuitively very similar to a coefficient of partial determination. We can consider w_j to be the m_i term for the weight of $\beta_{n,j}$ from the equation 20[1].

Finally, given the large amount of literature devoted to the diversification level of the firm (Lang and Stulz (1994), Berger and Ofek (1995), Rajan, Serveas and Zingales (2000), Villalonga (2004)), it may be beneficial to be able to differentiate between firms that have large weights of assets in a few asset groups and those that have small weights of assets in many different asset classes. As an example, consider two firms, one of which has 80% of its assets explained by these classes of systematically correlated assets, and are identified by two classes, while another has only 40% of its assets explained by the systematically correlated asset classes but spread across 10 different portfolios, is one more diversified than the other? In order to compare diversification levels, we develop a geometrically weighted diversification level. This is calculated as the geometric average of the individual class weights, where each class will explain between 0 and 100% of asset returns for each firm and the amount of asset returns explained by all classes will range from 0 to 100%.

$$\overline{m} = \left(\prod_{i=1}^{I} (1+m_i) \right) \left(\frac{SS_{\mathrm{Res}X_{i=1\to n}}}{SS_{Total}} \right) - 1 \qquad (30)$$

The first function of the geometric mean weight is a quotient function of one plus each individual class weight, and the second function represents the non-systematic portion of

[1] $(F_n\text{-}r) = \alpha + \sum_{i=1}^{I}\sum_{j=1}^{I} \beta_{n,j} m_i m_j \underbrace{(\alpha + \beta_{ci}(M - r) + \varepsilon_{ci})}_{C_i - r} + \varepsilon_n$ $\beta_{i,C}$

assets, which is the portion of assets not explained by any of the 1 to n asset classes. Intuitively the outcome is pleasing given that the natural bounds are between 0 and Euler's number minus 1 (appx. 1.718) regardless of the number of sectors possible, however, equation (30) needs to be altered if we want to eliminate any weight being given to idiosyncratic assets. We can simply remove the second term and have

$$\overline{m} = \left(\prod_{i=1}^{I} (1 + m_i) \right) - 1 \tag{31}$$

Which is intuitively proportional to equation (30) given that (30) is calculated from equation (28), which also is unconcerned with the non-systematic portion of returns, and that 31 is calculated from equation (29) where both (29) and (31) do consider the non-systematic returns. In order to scale $\overline{m}$ s.t. it ranges between 0 and 100% diversified, we drop the -1 and perform a natural log on equation (30) or (31) and $\overline{m}$ appears as (32) or (33).

$$\overline{m} = \ln\left(\left(\prod_{i=1}^{I} (1 + m_i) \right) \left(\frac{SS_{\mathrm{Res}X_{i=1 \to n}}}{SS_{Total}} \right) \right) \tag{32}$$

Or

$$\overline{m} = \ln\left(\prod_{i=1}^{I} (1 + m_i) \right) \tag{33}$$

The final functions (32 and 33) define a geometrically weighted average of diversification level.

Finally, approximating industries though regressions allows us to objectively compare not only the diversification level but also how common a diversification is. Said

differently, how common is it to be in industry$_j$ given that you are industry$_i$. We would not be surprised to find that a firm operating within the armaments industry is also in the steel industry, but we would be surprised if we found that a firm involved in the armaments industry were also in the fast-food industry. The method proposed below allows us to separate these observations into those we find typical and those that are atypical.

Preserving our objectivity, we can rely upon empirical partitioning to define typical vs. atypical diversification. This can be performed by calculating the conditional probability of a firm being in industry$_j$ given that you are industry$_i$.

$$\begin{matrix} p(I_1 \mid I_1) & \cdots & p(I_M \mid I_i) \\ \vdots & \ddots & \vdots \\ p(I_j \mid I_N) & \cdots & p(I_M \mid I_N) \end{matrix} \tag{34}$$

$p(I_j \mid I_i)$ is conditional probability of being in industry$_j$ given that you are in industry$_i$. Additionally this matrix can be broken down into binaries around an arbitrary cutoff (say the median conditional probability of the overall matrix) to determine the level of the firm's typical and atypical diversification level.

$$\sum_{i=1}^{M}\sum_{j=1}^{N}\begin{pmatrix} p(I_1 \mid I_1) & \cdots & p(I_M \mid I_i) \\ \vdots & \ddots & \vdots \\ p(I_j \mid I_N) & \cdots & p(I_M \mid I_N) \end{pmatrix}\begin{pmatrix} I_1 & \cdots & I_M \\ \vdots & \ddots & \vdots \\ I_N & \cdots & I_{MN} \end{pmatrix} \tag{35}$$

$p(I_M \mid I_N)$ takes a value of 1 if over the cutoff (1 if under the cutoff when you are looking for atypical diversification), or 0 otherwise, and I_1 through I_M (I_N) is the binary for whether the firm is in the industry. By summing the quotient we have a number spans 0 to N that describes how many industries the firm is exposed to and whether the industry

is common (or un-common), given the matrix of all other industries the firm is exposed to.

Classification Methods

As a matter of practice, the investment services industry separates companies into categories based loosely on the correlations of firms within the category. They do this not necessarily by selecting firms with high correlations but often by selecting firms that share common characteristics. Asset return correlations are likely greater between assets that have common characteristics, so the investment services industry's method of categorizing firms is consistent with (3)[2], the is inter-correlation corollary, but not identical. These characteristics include but are not limited to features such as similar products, customer profile, raw material inputs, labor needs, exposure to risk factors, and sources of financing. Grouping companies into categories based on these characteristics forms what is typically termed in the investment industry "sectors", which are then further subdivided into smaller categories called "industries."

Classification of firms into industries is formalized in the U.S. by two classification systems. The Bureau of the Census maintains the North American Industry Classification System (NAICS), and U.S. Office of Management of the Budget publishes the Standard Industrial Classification (SIC) codes. Additionally, MSCI maintains GICS and Compustat Segment data reports the classification required by the SEC regulation

[2] an asset is included in category *i* when the correlation of the return on that asset *j* in category *i* with the return on any other same-category asset is greater than a target cutoff correlation which is greater than the return correlation of the asset with any out-of-category-*i* asset.

FASB 14. The list of classification systems is robust, however these are the most established.

To test equation (20), i.e., identify a firm's membership in one or more categories (industries), we first must classify firms into sectors; therefore it is necessary to choose an appropriate grouping system. There are numerous systems commonly used both in academia as well as in the market place, but there is no clear consensus that any particular system is uniformly superior. In general we can consider each system to fall somewhere along the spectrum between purely qualitative and purely quantitative (purely subjective to purely objective). Additionally each method places different value on different points of the production line. Compustat Segment data, for example, is qualitative in that it is determined by the management of each firm based on internal decision making. SIC, and NAICS are determined by governmental agencies based on the largest cash flows of the firm. TRBC, ICB, and GICS are determined by third party market analyst through comparisons of product and service consumption rather than production or internal concerns. Given that none of the above mentioned classification systems is perfect for every situation, researchers must decide whether a qualitative or a quantitative approach is more appropriate for their research question. While motivation in classifying these firms is certainly relevant to which approach is optimal, shortfalls certainly face both methods. A qualitative approach, for example, that proposed by Ahn, Conrad, and Dittmar (2009), would focus on the correlation of firms within each industry. Specifically, sectors would be designed to maximize correlation of firm level returns within each category while maximizing orthogonality of returns across all other sectors.

$$Max\left(\rho\left(A_{i,j},A_{i,g}\right)\right)\left|Min\left(\rho\left(A_{i,j},A_{k,h}\right)\right)\right.$$

for all $i=1\ldots I$

$j=1\ldots J_i$

$g=1\ldots J_i, g\neq j$

$k=1\ldots I, k\neq i$

$h=1\ldots H_k$

This is performed through an iterative process of taking all N firms in the population and initially breaking them down into *N*/2 portfolios, based on similarity of returns. Next they iterate again and assign each of *N*/2 portfolios into *N*/2 new portfolios based on the optimal similarity of returns, conditional on the minimal cross-portfolio similarity of returns. This process is iterated as many times as it takes to reduce the sample of N firms into the desired number of portfolios, or sectors.

The objectivity of this method is pleasing for those researchers not interested in subjective perceptions of the firm but rather require a market wide consensus. However the method is empirically difficult and faces a few intuitive problems. The first issue with the quantitative method is commonality of returns over a given time period could be spurious. To our knowledge, no literature suggests that returns themselves should be persistent, but rather that future returns should be stochastic in nature. If we assume that returns are stochastic and the sorting is performed across returns, then there is little reason to believe that a firm would fall into the same portfolio in the future as it does today. On the contrast, the underpinning logic of our method (to some degree) expects a certain persistence of industries. It may be possible that the method proposed by Ahn, Conrad, and Dittmar (2009) is picking up fundamentals that are similar across firms within each portfolio, and are therefore persistent and consistent with our empirical

selection method, however an evaluation of the common characteristics would be time intensive and subjective itself. In an efficient market world, we can speculate that these N portfolios are based on a sorting across the CAPM market loading, or Beta. Since we expect return to risk to be fairly linear across the investment opportunity set, it seems that ACD's method would identify portfolios based on common deviations from the market portfolio rather than based on fundamentals related to underlying assets. If this is the case then there may be essential methodological break between ACD and our decomposition; if their method identifies deviations from the market portfolio then firms are being sorted into bins based on their idiosyncratic returns rather than systematic, the theory for our classification assumes that the bins represent the decomposition of the firm's market beta into sub-market (industry) loadings. Furthermore, the literature has still not come to a consensus whether the firm's market loading itself can change through time. Our method is intended as a decomposition of the CAPM model into the industry components that collectively make up the market portfolio rather than an identification of external pricing factors therefore the purely quantitative approach may be inappropriate. Quantitative classification systems might provide benefit when a researcher's goal is to identify common non-systematic characteristics across firms, but it is unclear how this method would be superior to performing principle component analysis or factor analysis when we wish to identify industries.

The second concern with the purely quantitative methods such as ACD, factor analysis, or principle component analysis is more fundamental. Even if the sorting works perfectly, we are still unable to identify how the assets of firms in the same industry are

related intuitively. In order to make use of the information provided, for decision making of managers, we need to understand why these firms move together. The only method we can identify is visually examining the assets of the firms placed within each bin to qualitatively identify common characteristics, but even then we may find this method fails to have asset similarities that are prevalent enough (or persistent enough) to identify asset commonality. A concern even with the comparison of assets is that the correlations could be spurious. If we compare twenty assets we could randomly find that one of them is statistically correlated without being certain that they will be correlated in the following period.

While intuitive information can be gained from the quantitative identification methods, it is unclear that this is the optimal sorting procedure to identify industries, even though it may be appropriate for identifying the number of risk factors a firm is affected by.

An alternative approach is to qualitatively identify which firms should be classified together through an examination of assets and common characteristics of each firm, and assign them to sectors based on these similarities. This method also faces several difficulties such as subjectivity, noise, and labor intensity.

Because it is infeasible for any one person to evaluate every asset owned by every firm, we are forced to rely, at least for preliminary sorting, on disclosures that place a firm into one of a very large number of possible categories, whether these categorizations are performed by management (such as the Compustat Segment Data) an external agent (such as TCRB, ICB or GICS) is determined by the researcher's use of industry classifications. SIC codes are probably the most commonly used measure of industry

classification and they range from 1-9,999, which can clearly become un-meaningful given the very large number of possible assignments. Often researchers will simply drop the first two digits of the SIC code in order to have a more manageable number of industries. As an alternative Fama and French (1997) develop an algorithm to place these sic codes into specific industries based on common characteristics. Although the two methods are comparable in number, around 48 base industries, F&F provide arguments to convince the reader that their method avoids many of the problems that are faced by SIC codes alone. Whether we consider two digit SIC codes or F&F industry classifications, we still face with a number of problems such as subjectivity, and that classification based on major assets ignores minor assets. The benefit of this subjective and noisy method is that the underlying securities are sorted based, at least loosely, on fundamental assets, which we believe are likely to be persistent.

Even though all classification systems are noisy due to assets being miss-classified into the same bin as the largest asset held, on average the mean returns for each bin should be representative of the actual returns of the largest assets. Clearly the returns of these portfolios are not perfectly correlated with the returns of the hypothetical portfolio (which is made up only of the largest asset) over any arbitrarily short time horizon, but given an appropriate time series of data the returns of the actual portfolio should be strongly correlated with the hypothetical portfolio. The difference between these two portfolios is simply noise, so long as there is not an endogenous relationship between the non-largest assets.

Regression Methods

After choosing a portfolio sorting method, we need to decide upon the most appropriate type of regression to apply. The techniques that seem most natural for identifying industries fall into two types; OLS and selection criterion methods. An OLS regression is appealing because it does not require as many subjective decisions on the part of the tester, however it also faces problems such as multi-collinearity. A large concern is inter-correlation of industry level returns. While this concern may minimalized contemporaneously for quantitative sector assignments, it is likely an issue across qualitative assignments. The correlation of returns of sectors will likely arise because we have made the assumption that firm assignment to sectors is based on a large subset of assets, but not all assets need to be related to the assigned sector. Said differently, although a firm may be largely involved with a particular sector, it could also have portions of its assets fall into a secondary, or tertiary classification, and if this occurs we should expect to see that the returns of the major sector will be correlated with those of the second or third by the degree to which the second and third set of assets (which are unfortunately assigned to the primary sector) are correlated with other sectors. In fact both SIC and NAICS explicitly consider only the largest source of revenue for determining classification and this source need not make up a majority. Although OLS is unbiased in its estimates, if correlated returns are prevalent, OLS may not be able to differentiate between correlation of firm level return and industry level returns especially when we use shorter time horizons.

Another concern is the performance of our proposed methodology during periods of high volatility or ambiguous industries. Time periods such as extremely inflated market volatility (e.g. 1987, 2008) could introduce additional noise and damage both qualitative and quantitative methods of classification. The multi-collinearity between industries would likely increase dramatically. Additionally, merger waves such as was seen in the late eighties would decrease the weight that the largest asset has in the portfolio, especially if the waves are diversifying mergers. Care should be taken when evaluating any measure of industry classification during time periods of high noise.

Selection criterion alternatives to OLS provide at least partial solutions to the problem of semi-correlated returns, but require judgment calls by the experimentalist. Two mainstream methods are stepwise regressions and LASSO regressions. A stepwise regression selects which variables (when given a list of possible variables) should be in the model through an automatic process of addition and removal of variables until there is no more benefit to the model (the way that benefit is defined may change the structure of a stepwise regression). A LASSO regression (Least Absolute Shrinkage and Selection Operator) minimizes the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant (Tibshirani (1996)). More intuitively, the LASSO shrinks the coefficients of variables that are less instrumental in explaining the dependent variable until there are two types of parameter estimates; zero and non-zero. The reduction of non-relevant variables from the model helps to alleviate the concerns that arise with an OLS regression, since this process will generally remove variables that are overly correlated but require specification decisions such as entrance and exit

significance levels for the model (stepwise) or the constant level (LASSO). OLS also faces subjectivity such as choosing what statistical significance levels are required to reject or fail to reject the null hypothesis, but these decisions are widely accepted within the literature. Additionally, the selection criterion of both stepwise and LASSO procedures will strongly impact not only which sectors are significant, but also the number of sectors.

Another concern that faces all methods is whether a firm should have negative industry exposures? It is not perfectly clear what the intuition of a negative exposure is, does this imply that a firm is in the identified industry, does it measure customers and supplier effects, or is it meaningless and merely a noise created by the methodology?

There are a variety of methods that can be applied to identify industry exposure and it is unclear that there should be an absolute prior for which method is optimal. Rather, we suggest that the selection method be decided with respect to the goals of each specific study.

When there are many unknowns the empirical model will often be unreliable. Within the context of industry identification we do not know the true number of segments in the model and therefore it is difficult to determine whether assumptions about the model are correct. One assumption of particular interest is the probability of appearing in the model, $p(\theta)$. In context of industry identification, how do we know whether our model is picking up the true number of industries? From simulations we can see that while the true number may not be perfect, firms tend to fall at the correct part of the spectrum (from very few industries to many industries) even if the measured number is

not correct. For some researchers this ranking may not be good enough, so they may wish to consider a Bayesian approach where a prior $p(\theta|y)$ is approximated by taking realizations of the probability of which an industry is in the model $p(\theta)$, and then condition their model on those approximations. The natural problem with such an application is that Bayes requires a joint distribution on observable y and the parameters θ, that is $p(\theta|y)= p(y|\theta)\,p(\theta)$.

When identifying industries we generally will apply a normal linear regression, $y=x\beta+\varepsilon$ where $\varepsilon \sim N(0,\sigma^2)$ and $y \sim N(y|x\beta,\sigma^2)$. When σ^2 is unknown, an approximation from realizations of data may be appropriate. This can often be done through a Markov chain Monte Carlo (MCMC) (Gelfand and Smith (1990)). We are of course still faced with the Bayes' requirement of an observable joint distribution across which to apply our burn-in sample. This can be achieved through simulating returns of firms from a known number of industry level returns. The simulated returns give us the opportunity to calibrate our MCMC to more accurately identify industry exposure in terms of exact numbers.

To summarize, we use three different methods of estimation: single-pass OLS, stepwise, and LASSO regression.

Empirical Evaluation of Selection Methods

A problem that researchers face when trying to identify a true model is that often the true model is unknown. In order to determine the strengths and weaknesses of each selection method we need to know how many industries a firm is in, otherwise it would

be difficult to ascertain the successes and failures of each model. By assumption we know that this information itself is non-transparent and so any comparison of methods would not be beneficial since we cannot determine the accuracy without assuming the true model. To circumvent this problem we simulate firm level returns from random drawings of industry returns so that we can see which industries a firm is in and what the weights of each industry are. Through these simulations we hope to gain insight into which method is most accurately able to sort through the noise that arises because sector level returns are calculated from industry classifications based on plurality cash flows rather than individual assets.

We assign a simulated firm's category (industry) by a random drawing of between 1 and 48 potential industries from the Fama and French industry dataset (Fama and French (1997)) and equal weight is provided to each industry so that the number of industries is I. For simplicity we assign the weight for each industry to be 1/I. We choose the industry classification of Fama and French because of its widespread acceptance among the social science community, to date which has been cited nearly 5000 times. Although Fama and French provide many different datasets that sort the total number of SIC codes into larger or smaller numbers of industries, we choose the most broad dataset (48 industries) in order to err on the side of type II error. Each observation may have anywhere between one and forty eight industries. To determine how many industries will be in each simulated firm, We apply a standard normal distribution random drawing, where the mean is the number of industries we wish to have in the simulation and the standard deviation is one. We simulate returns over one half year, one year, two years, and three

years of daily returns. Randomly distributed error terms are added to the weighted industry returns at low levels (.05 percent daily), medium levels (.5 percent daily), high levels (1.5 percent daily) and very high levels (2.5 percent daily). These error terms are intended to span times of very low to very high market volatility. The error term is normally distributed around zero and is applied to each daily return such that the daily return will be equal to:

$$\mathrm{Re}\,t_{daily} = \left(\sum_{1}^{N} R_i * \frac{1}{N} * \rho(i) \right) + N \square (0, \sigma)$$

where rho(i) is a binary for whether or not industry i is in the model, and a normally distributed error term with a mean of zero and a standard deviation of sigma is added to the daily returns of the hypothetical firm.

We regress the simulated returns on the 48 industry returns using each of the three methods discussed, OLS, stepwise, and LASSO, across a range of specifications.

$$F - r = I_1 - r + \cdots + I_{48} - r$$

Specifications include time horizon, noise level (or volatility), and number of industries in the "true" model (target), defined as the mean from the random drawing from a standard normal distribution.

We evaluate the correctness of each regression based on the number of industries that are in the true model versus the number that are in the empirically predicted model. This process is iterated 100 times for each specification and the results are recorded in appendix 2. Our 288 simulations covers all combinations of the method, time horizon, volatility, and target number of industries. We use the output to provide insight into the

accuracy that each method provides, but also to show the sensitivity of each method to alterations in the specifications.

Results

The results of our simulations are reported by both absolute error and percentage error. Absolute error is the absolute value of the difference between the number of industries in the true model and the number of industries in the measured model. The percentage error is the absolute error divided by the number of industries in the true model. Appendix 3 reports the results from each specification such that side by side comparisons can be made of OLS, stepwise, and LASSO regressions, however for friendliness to the reader we report the results of the simulations in graphical form, averaging the error for all specifications across an individual criterion at a time. We evaluate accuracy across horizon, target, and noise. Because we use a random drawing for the target number of industries the mean number of industries across methods could be slightly different. A direct comparison of absolute error is not precise, instead both absolute and percent error should be considered.

All Specfications

Figure 1-A (1-B) graphs the average error (percent error) of all specifications for each of the three methods. On average OLS performs the worst, either considering absolute error or percent error. LASSO and stepwise both appear to be approximately the same in absolute error, but LASSO outperforms on average, when we compare percent error. While Figure 1-A (1-B) are able to show general performance of the three measures, differing research questions may be able to more accurately predict levels of

noise in their data, appropriate time horizon, or estimations of the number industries in the true mean. Because of the differing motivations it is also interesting to evaluate whether different priors or specifications may yield different accuracies cross-sectionally across the three methods.

Time Horizons

Time horizon is one of the three comparisons we measure. We expect that as the number of data points increases, regression measures such as OLS should increase in accuracy, and we should see OLS appears as a much more logical candidate. Alternatively, with selection criterion methods, the difference should be less significant given that the final model need not have the full set of all independent variables (in our case 48 industries) across which to share their power. Figure 2-A (and 2-B) reports average error for each method across different time horizons, that is how much past data we are using to measure empirically determined industries. The time horizons are 126, 252, 504, and 756 trading days. In both figures OLS again performs the worst by a pretty large margin. LASSO performs better as time goes on, but there is no clear trend in the accuracy provided by stepwise with longer horizons. Across absolute error stepwise performs better in short time horizons than LASSO, however in percentage error LASSO always outperforms regardless of the horizon.

Mean Number of Industries

The mean number of industries is the second of our three comparisons because we suspect that the larger the number of industries, the weaker the empirical determination of industries will work. In general, it is unclear what the true number of industries should

be. Villalonga (2004) reports that Compustat Segment data lists only 28.6% of firms being multi-segment, and that the BITS data set reports only 36.2% of firms being multi-segment and 79.4% of firms being multi-business unit. Figure 3-A (and 3-B) reports the average error across target level, i.e. the number of industries that appear in the true model. As we would expect, the more industries that appear in the true model the more inaccuracies we see, regardless of the empirical method of testing that model. Loss of empirical power is likely the reason for decreasing accuracy across increasing target. That being said, within small targets, LASSO tends to outperform, but in large targets stepwise appears to perform better. For simulations with the largest numbers of industries (mean of 24), both LASSO and OLS have about 40% more error in absolute terms than stepwise under this simulation specification. In small targets (mean of 1) LASSO appears to consistently get within one industry of the true industry. In percentage error, there appears to be virtually no difference of large targets among any of the methods.

Noise

Noise is the third category that we compare our method's accuracy across because it is important to understand whether the methodology is likely to fail in different economic states such as the high volatility experienced during 1987 and 2008. In figure 4-A (and 4-B) we compare the average error across noise levels. It is interesting to note that noise levels appear to be less important than other factors such as target level. This is good news given our concerns of multi-collinearity. In general, being consistent with previous figures, OLS performs the worst while stepwise and LASSO compete for best. In absolute errors stepwise and LASSO appear about the same while LASSO outperforms

in some noise levels and stepwise outperforms in others. When we compare percentage error, LASSO consistently outperforms. In fact stepwise has nearly half again as high of error as does LASSO, and OLS has nearly twice the error as LASSO. As we would expect there is a fairly monotonic increase in error as the noise level increases, however it is perhaps, not as high as we would expect. For a numerical example, across all simulations that have a mean number of one true industries, LASSO identifies about 0.5 to 1.5 industries on average, while OLS would identify between (approximately) 0 and 2.15. At high levels of noise OLS is only a little worse, identifying, approximately, a range of 0-2.25 industries, whereas LASSO increases to the approximate range of 0.25-1.75 industries.

Robustness

For a robustness test, we report graphs of squared error to determine whether any of the measures perform significantly worse at extreme observations. Alternatively readers can explore appendix B to evaluate specifically where any extreme observations occur. Figures 5-A through 5-D graph the squared error across each specification to check whether outliers appear to drastically change the results of the simulations. We find that the results appear to change little. OLS still performs worse except in an extreme minority of specifications; LASSO on the other hand appears to perform a little worse than stepwise but often the difference is un-meaningful. Extreme observations do not seem to be much more persistent across any of the methods.

Finally we evaluate the two extreme specifications, that is, the least informative and most informative specifications in terms of accuracy of empirical determination of

industries. The specification that results in the least accuracy is, as we would generally expect, the high volatility, a large mean industries, and the shortest regression horizon. Additionally, as we would expect the most informative is low volatility, low mean industries, and a long horizon. Figures 6-A and 6-B report the error and squared error for the worst case scenario while Figures 6-C and 6-D report the error and squared error for the best case scenario. Uniformly across all four figures, LASSO outperforms each of the other methods, but OLS and stepwise alternate in accuracy across specifications. In the worst case scenarios OLS outperforms stepwise but in the best case scenarios stepwise outperforms OLS.

Conclusions

In order to not bias the evaluation of methods by calibrating them specifically to minimize error, we run very generic stepwise and LASSO regressions. Both methods can be calibrated given a set of priors about the data, hence within a simulation framework it would be quiet easy to design a specification model that (because we designed the true model) could result in very high accuracy. Some of the simulation results are probably a byproduct of this precaution. For example, in several simulations (especially those across target) we will generally see that the farther the sample goes from a normal prior of a few significant industries, the worse the LASSO regression works. In general, it appears that LASSO is on average the most accurate method of identifying industries and OLS is the worst. Additionally, for sample selection models such as stepwise and LASSO, accuracy can likely be increased through the typical Bayesian method of burning observations to determine the priors of data. This is the process of calibrating the selection criterion

based on observable data in the past. Specific requirements and priors of the data could lead other researchers to choose an alternate method. Each researcher should choose their regression technique based on the needs of their project as well as the data availability and priors that can be justified. The reported simulation results should be considered to be conservative because we strived to err on the side of type II errors.

Overall the results suggest that empirical determination of industries may not be optimal for all research questions (dependent on the researcher's restrictions) but does provide a valuable alternative to current industry classification systems. In general, empirically determined industries perform well from the perspective of firms that have a larger number of true industries will empirically identify more industries than those that have lower number of true industries, even if the method cannot peg the exact number of true industries perfectly. The empirical identification of industries provides four benefits to researchers

i. allows researchers to view investor perspective of industry involvement
ii. allows researchers to determine weights of each industry a firm is exposed to
iii. allows a comprehensive view of the diversification level by comparing not only the industries but the weight each industry makes up
iv. allows researchers to compare diversifications that are typical against those that are atypical, conditional on other firms diversification into each industry

The four benefits provided by the discussed methodology contribute to the finance literature greatly. Nearly all areas of financial research consider or control for industries and our method allows a more complete picture of how this can be done.

Figure 1-A: Absolute Error Across All Specification of each Method

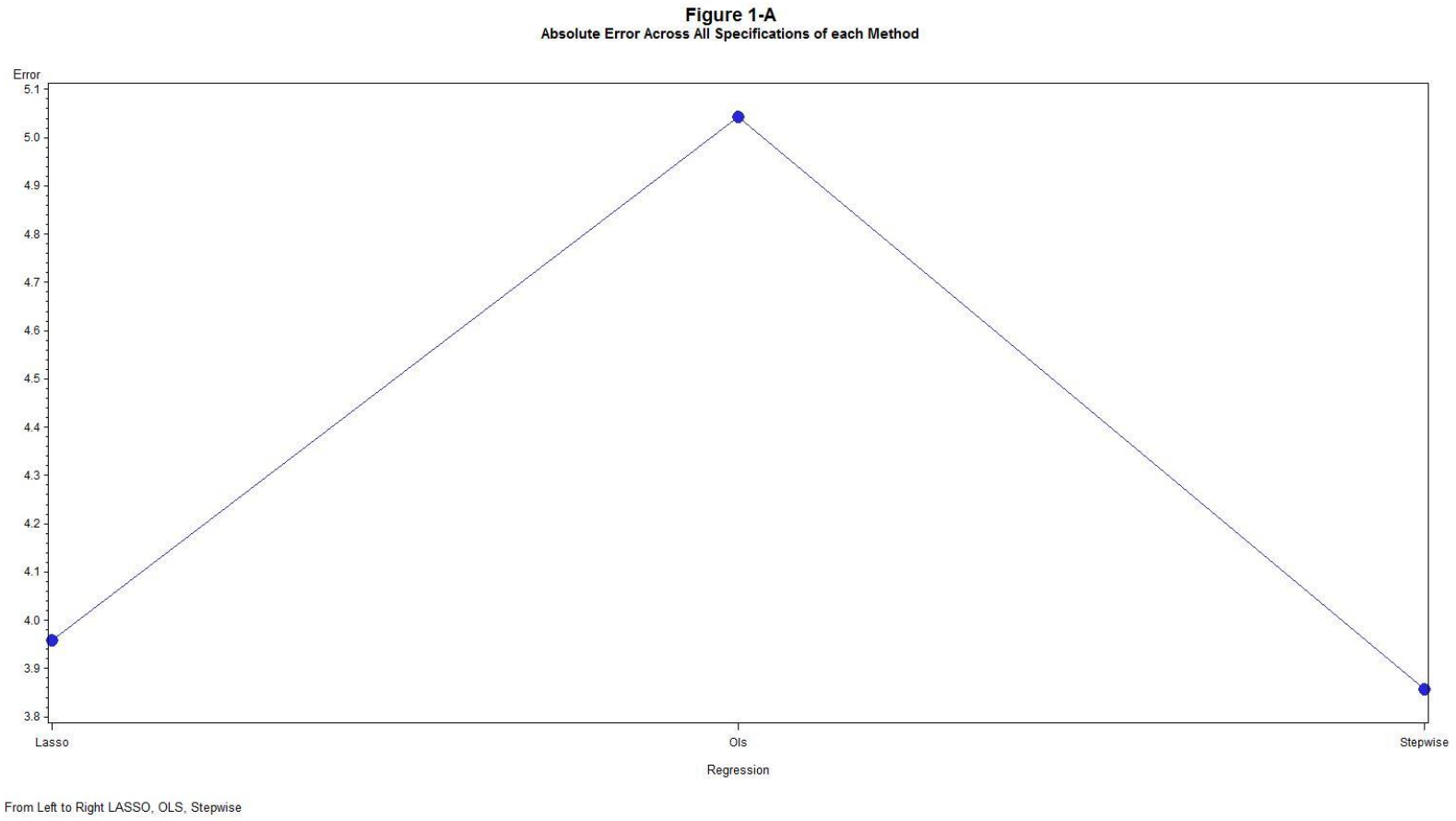


Figure 1-B: Percent Error Across All Specification of each Method

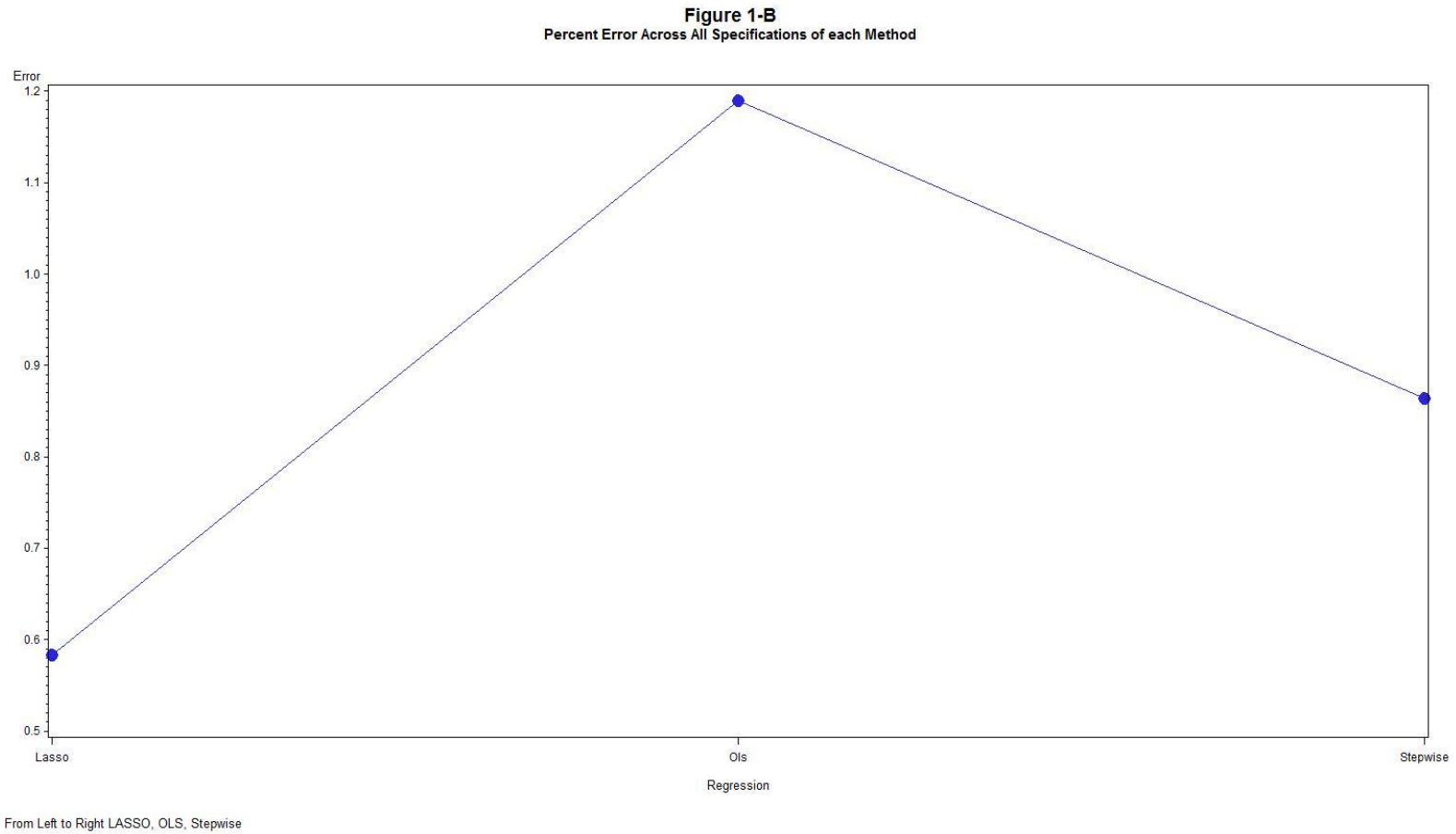
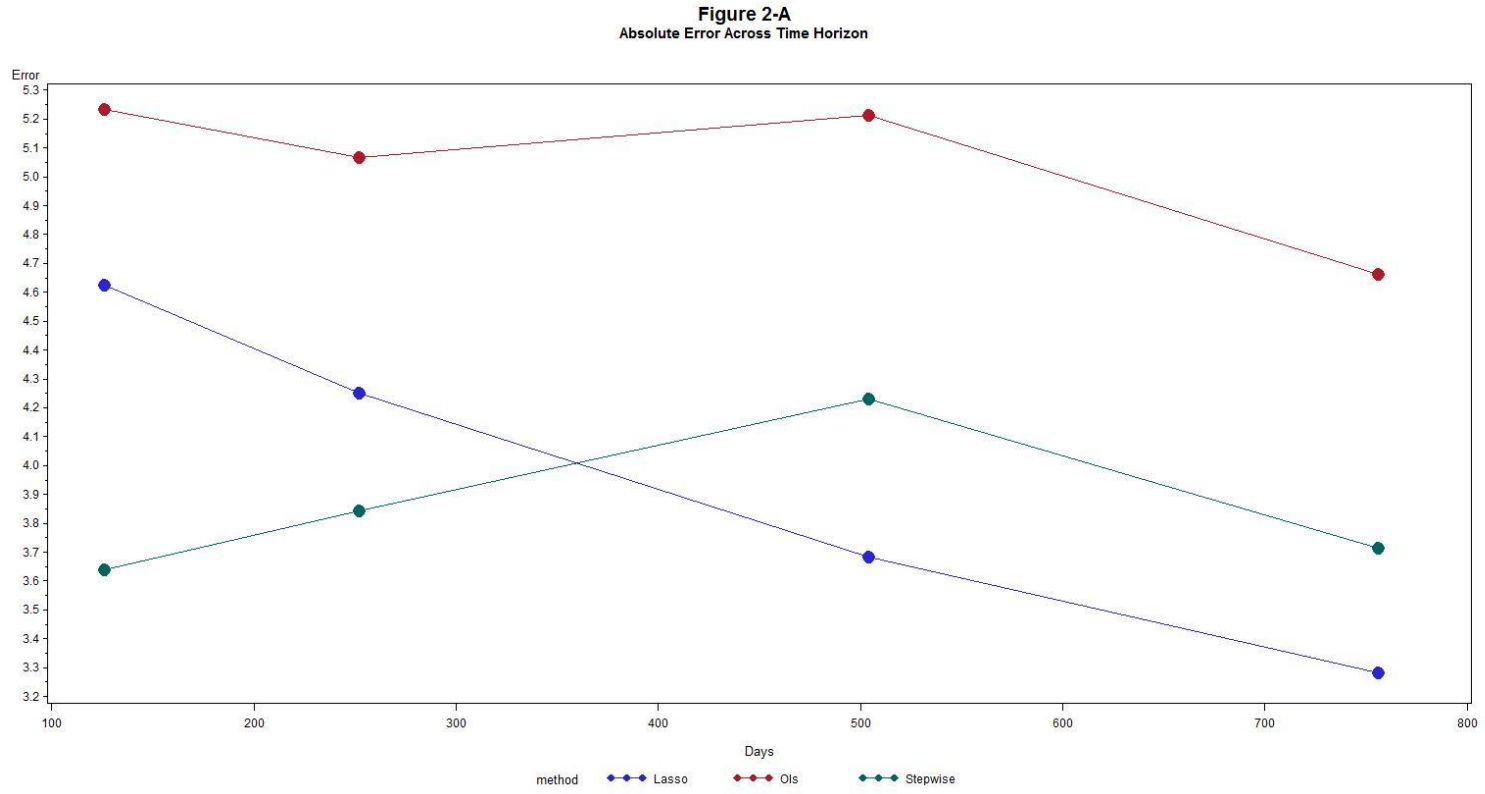
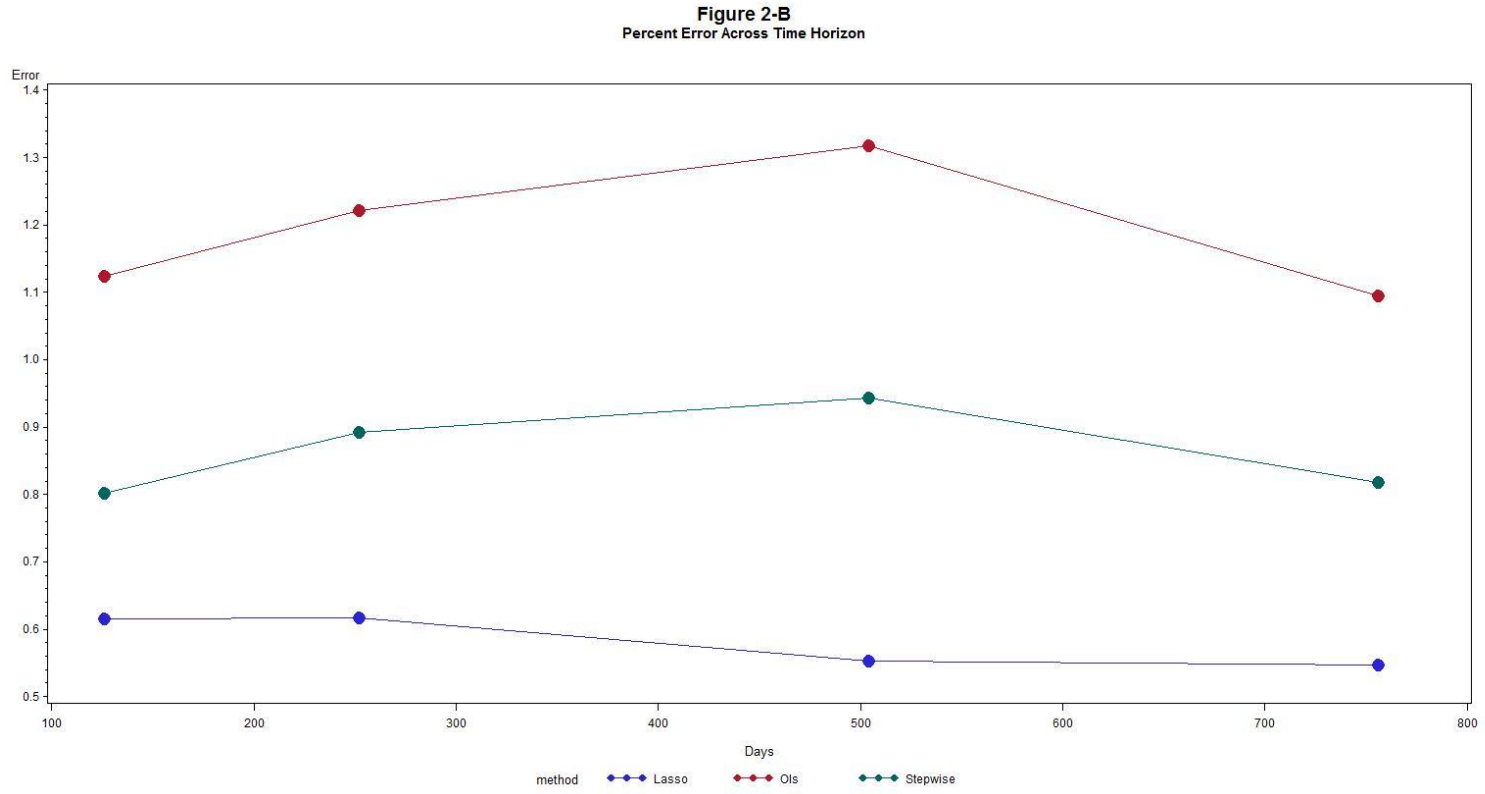


Figure 2-A: Absolute Error Across Time Horizon



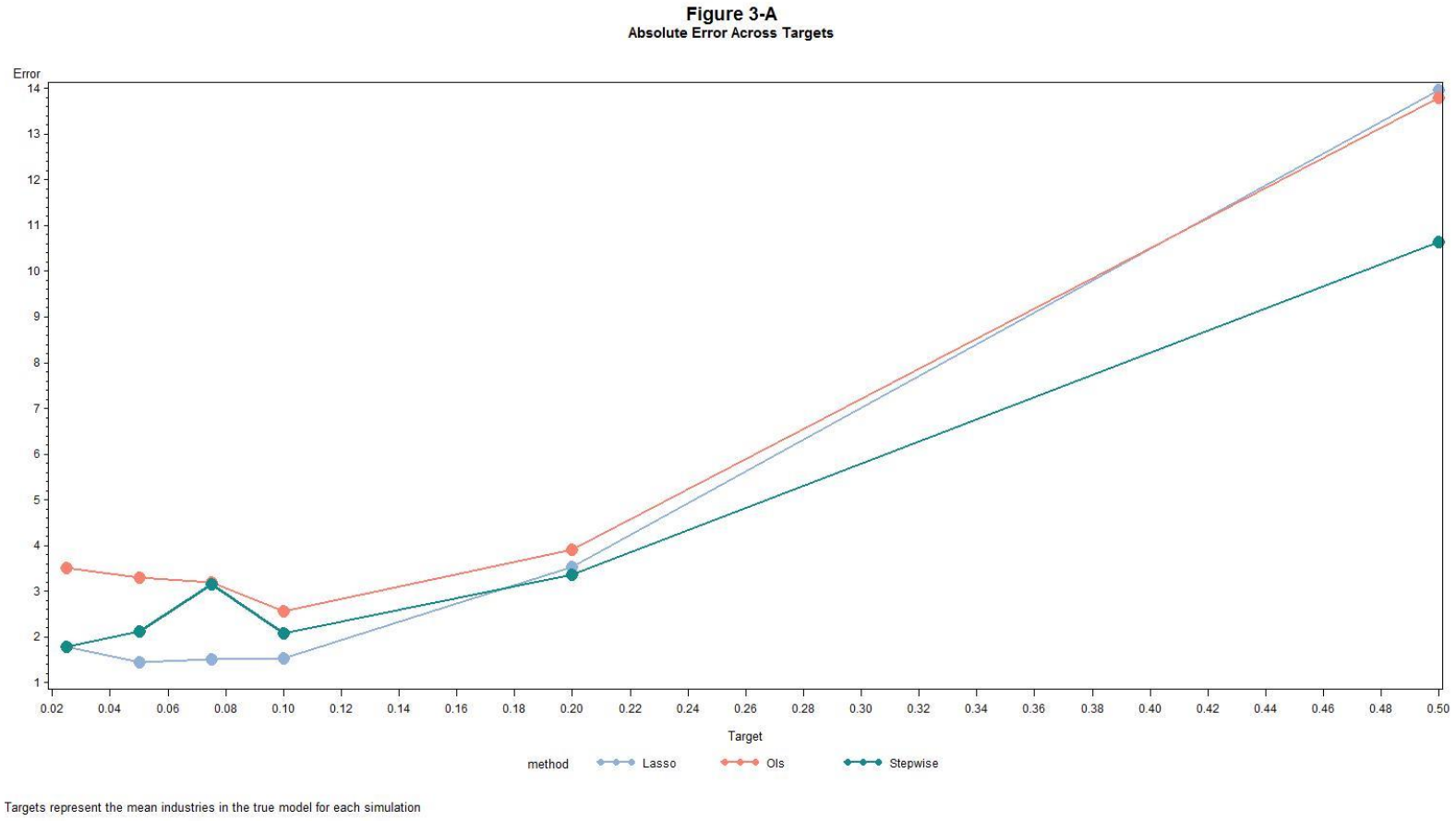
Time Horizon is the number of days of returns used in the regression

Figure 2-B: Percent Error Across Time Horizon



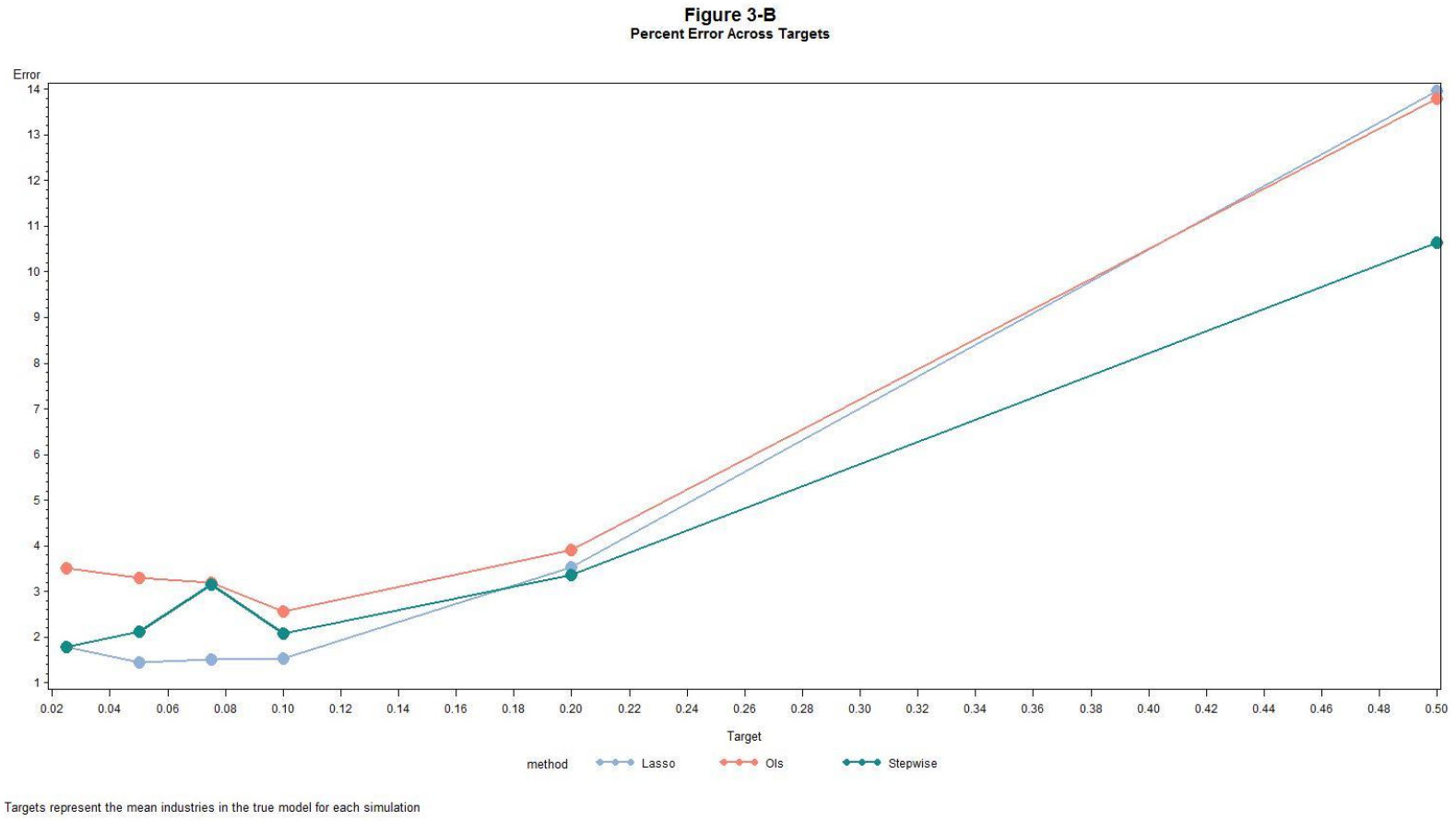
Time Horizon is the number of days of returns used in the regression

Figure 3-A: Absolute Error Across Targets



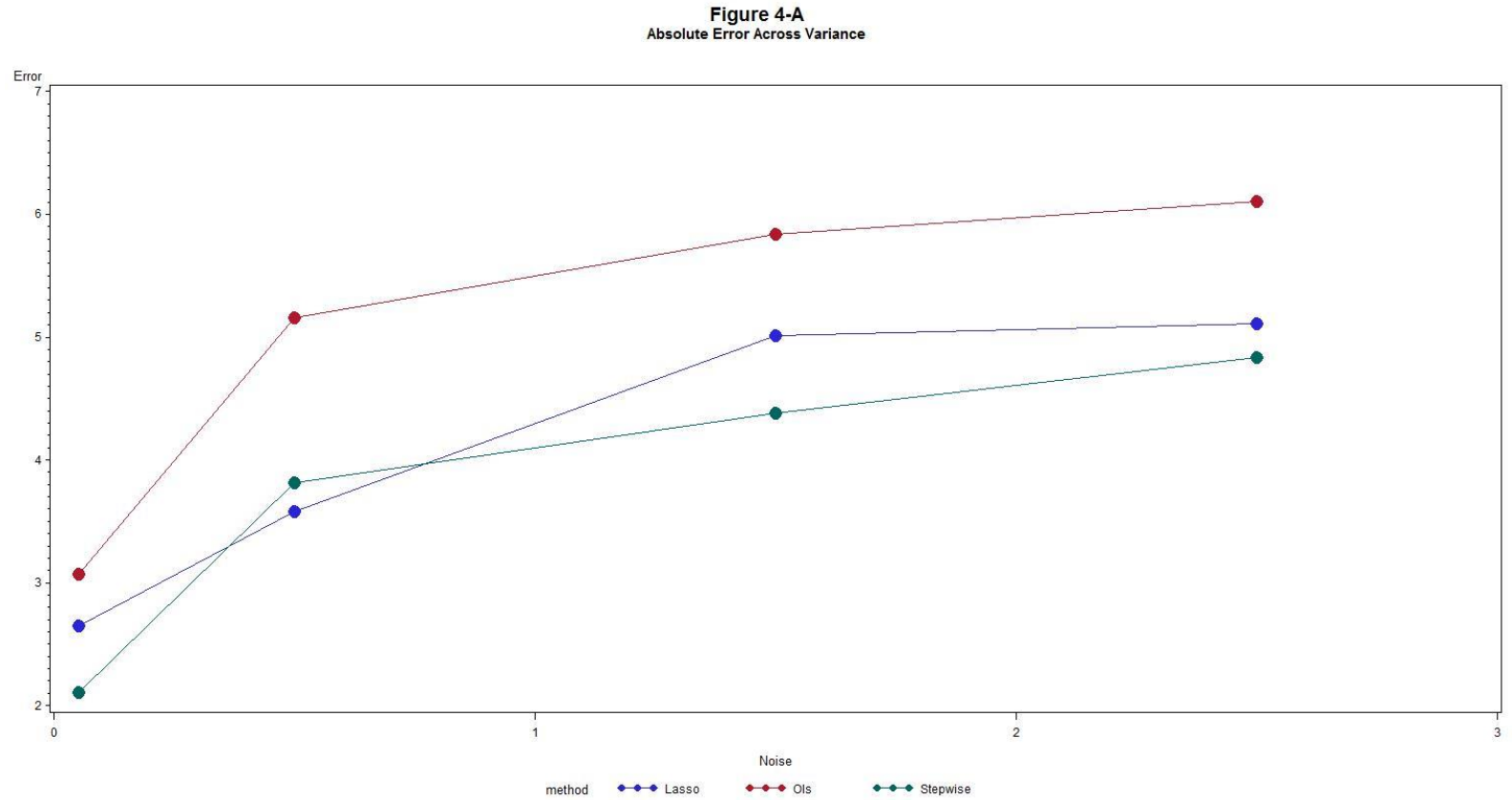
Targets represent the mean industries in the true model for each simulation

Figure 3-B: Percent Error Across Targets



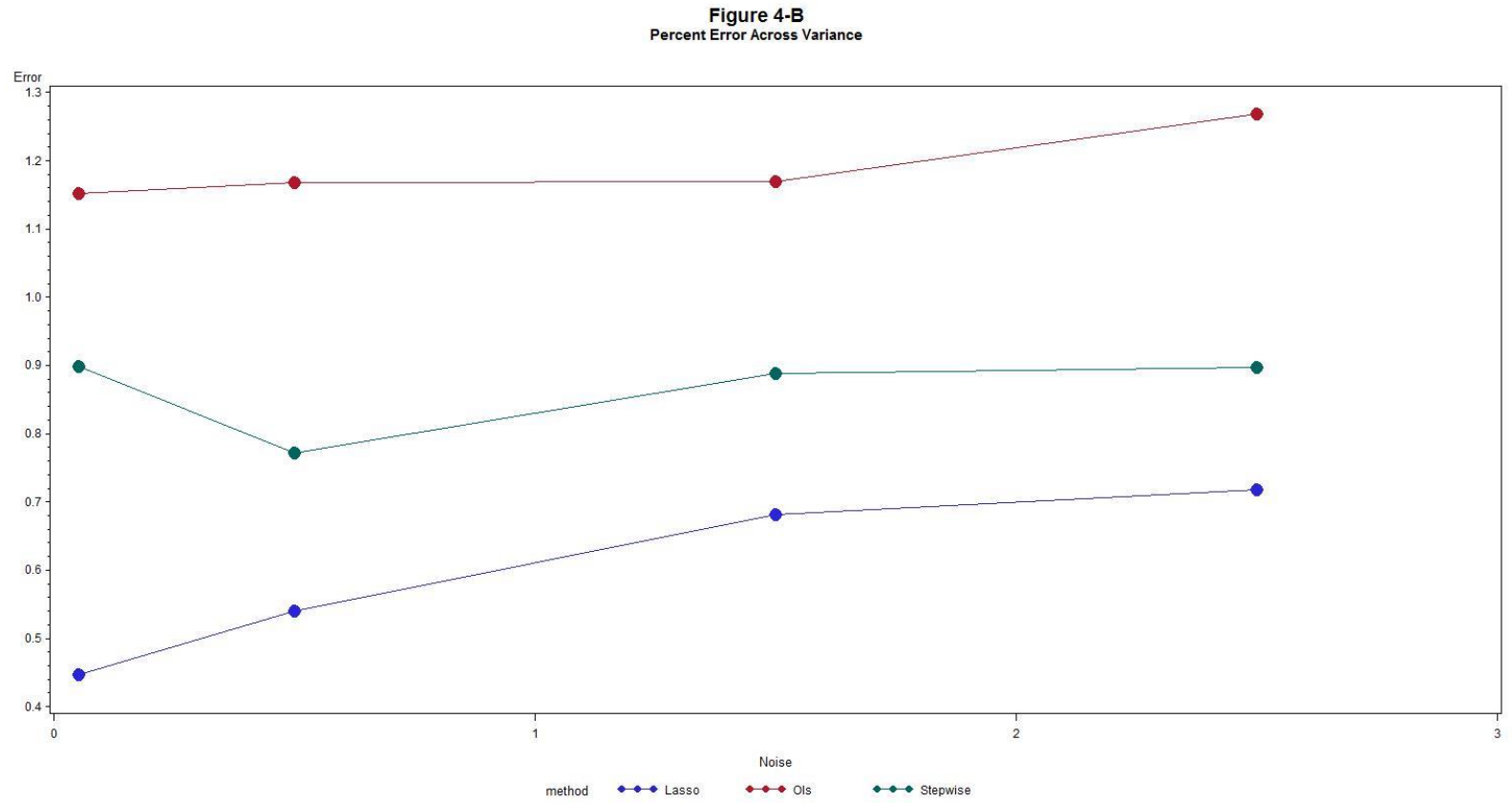
Targets represent the mean industries in the true model for each simulation

Figure 4-A: Absolute Error Across Variance



Variance is the amount of error added to the daily returns of each simulated stock

Figure 4-B: Percent Error Across Variance



Variance is the amount of error added to the daily returns of each simulated stock

Figure 5-A: Squared Error Across All Specification of each Method

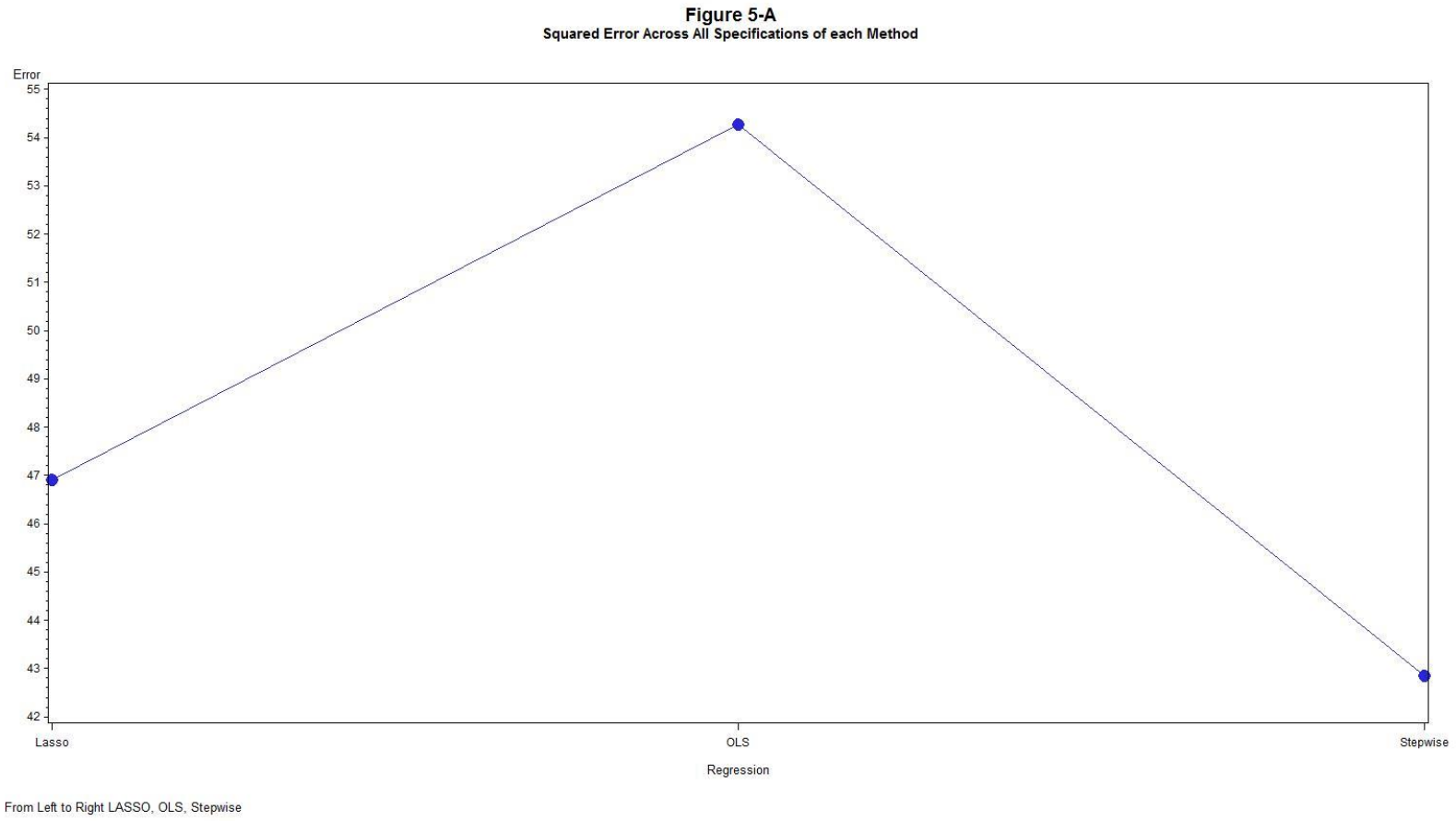
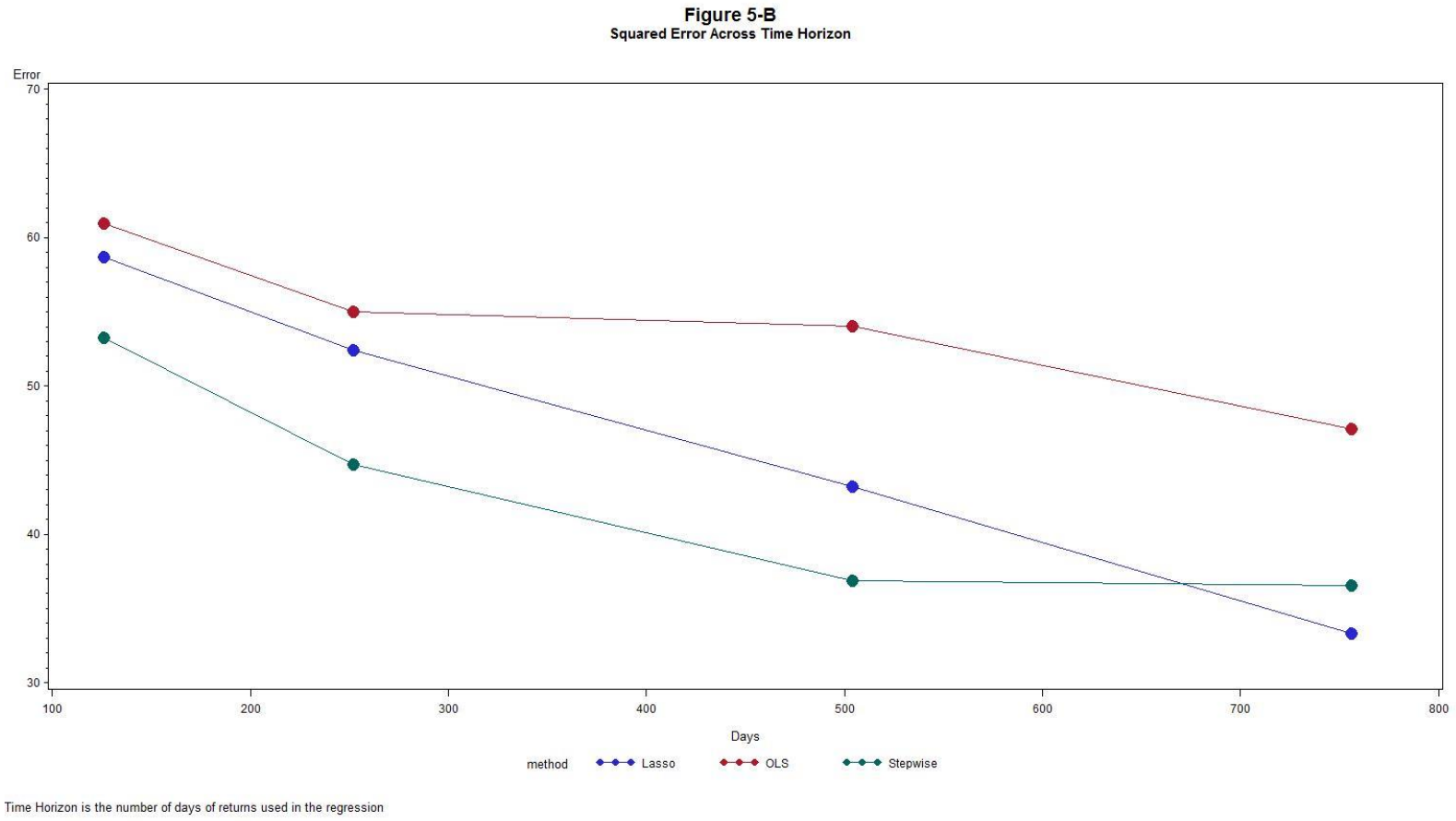
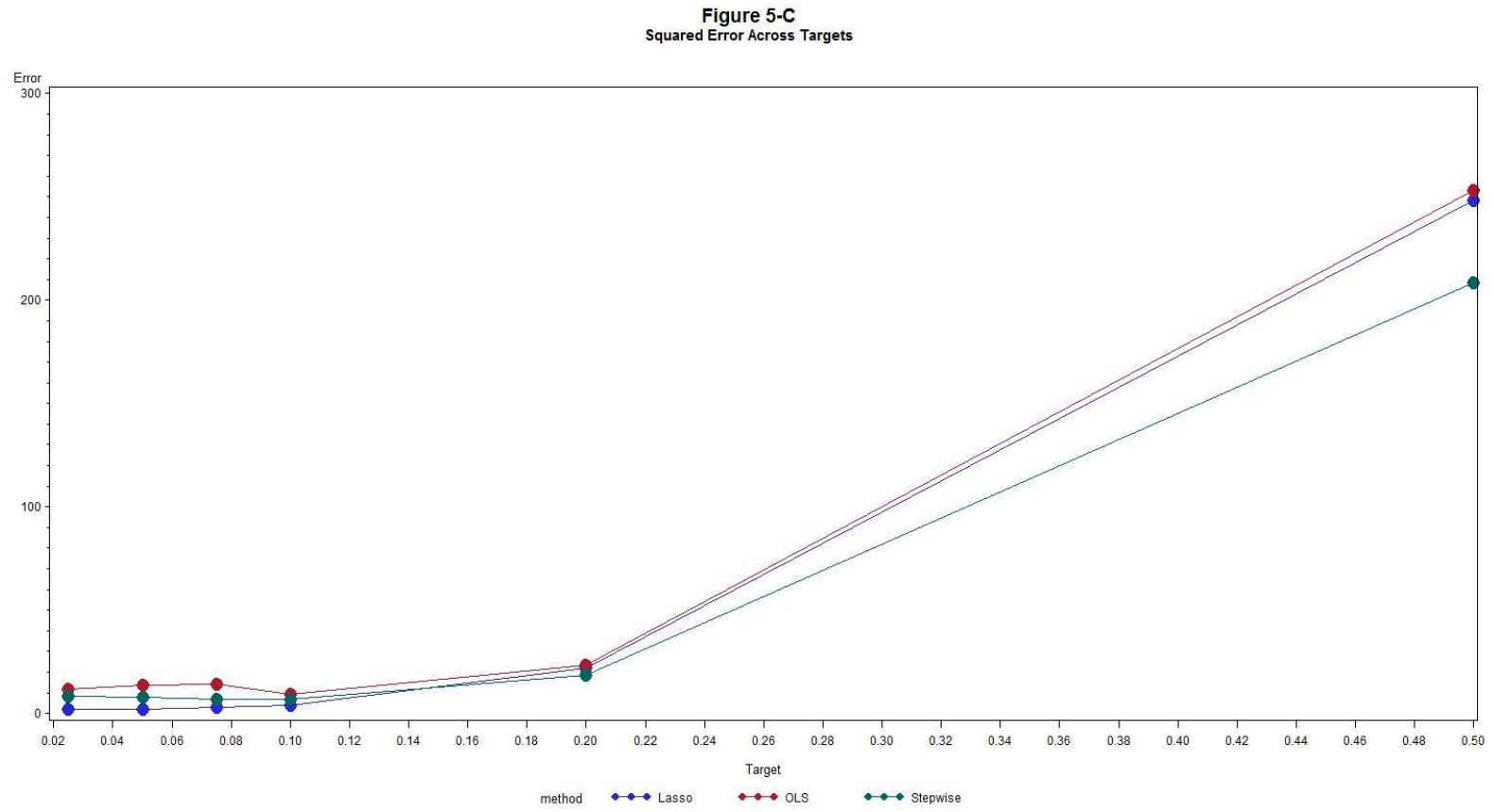


Figure 5-B: Squared Error Across Time Horizon



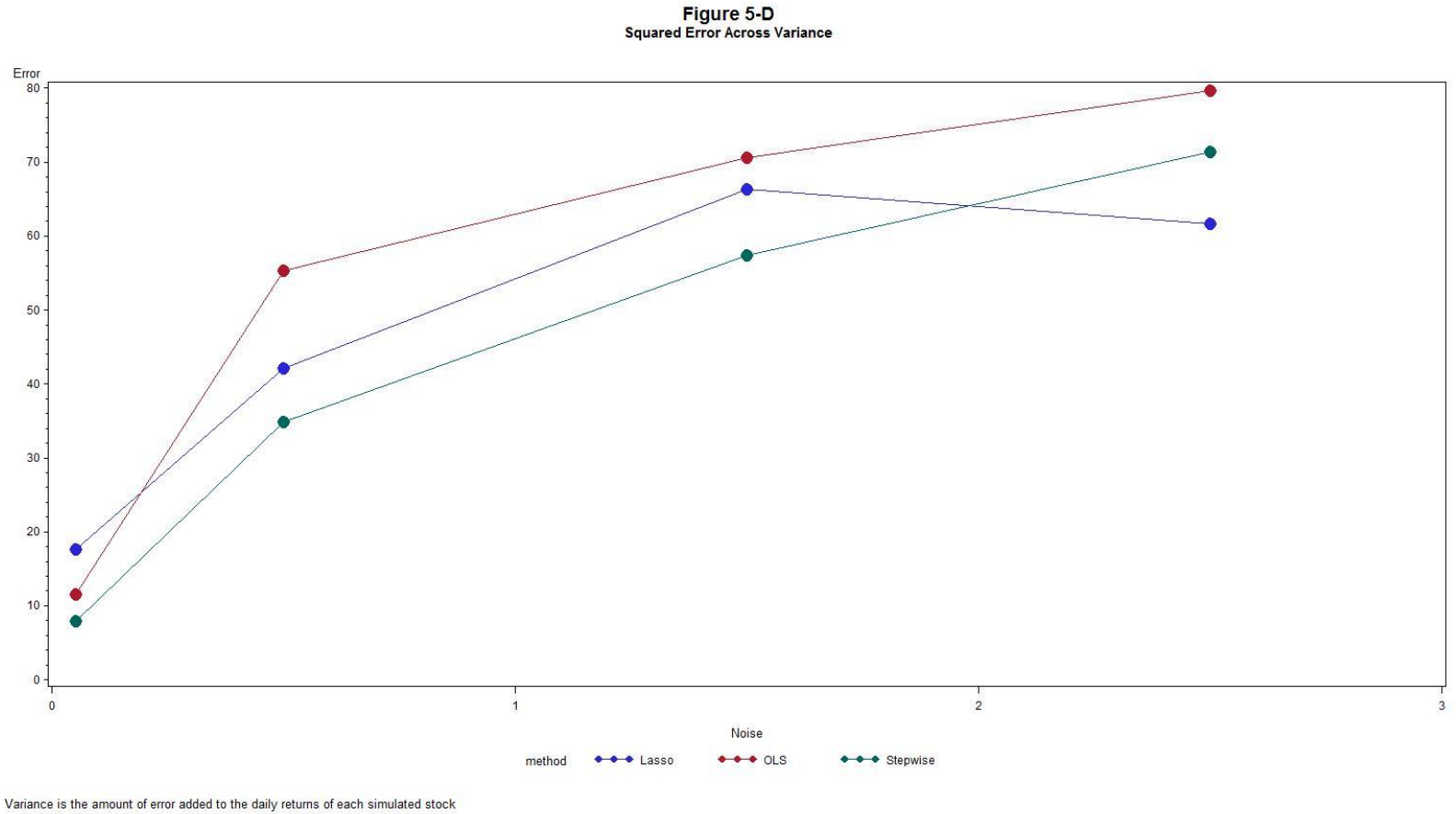
Time Horizon is the number of days of returns used in the regression

Figure 5-C: Squared Error Across Targets



Targets represent the mean industries in the true model for each simulation

Figure 5-D: Squared Error Across Variance



Variance is the amount of error added to the daily returns of each simulated stock

Figure 6-A: Error Across Methods, The Worst Performance Case

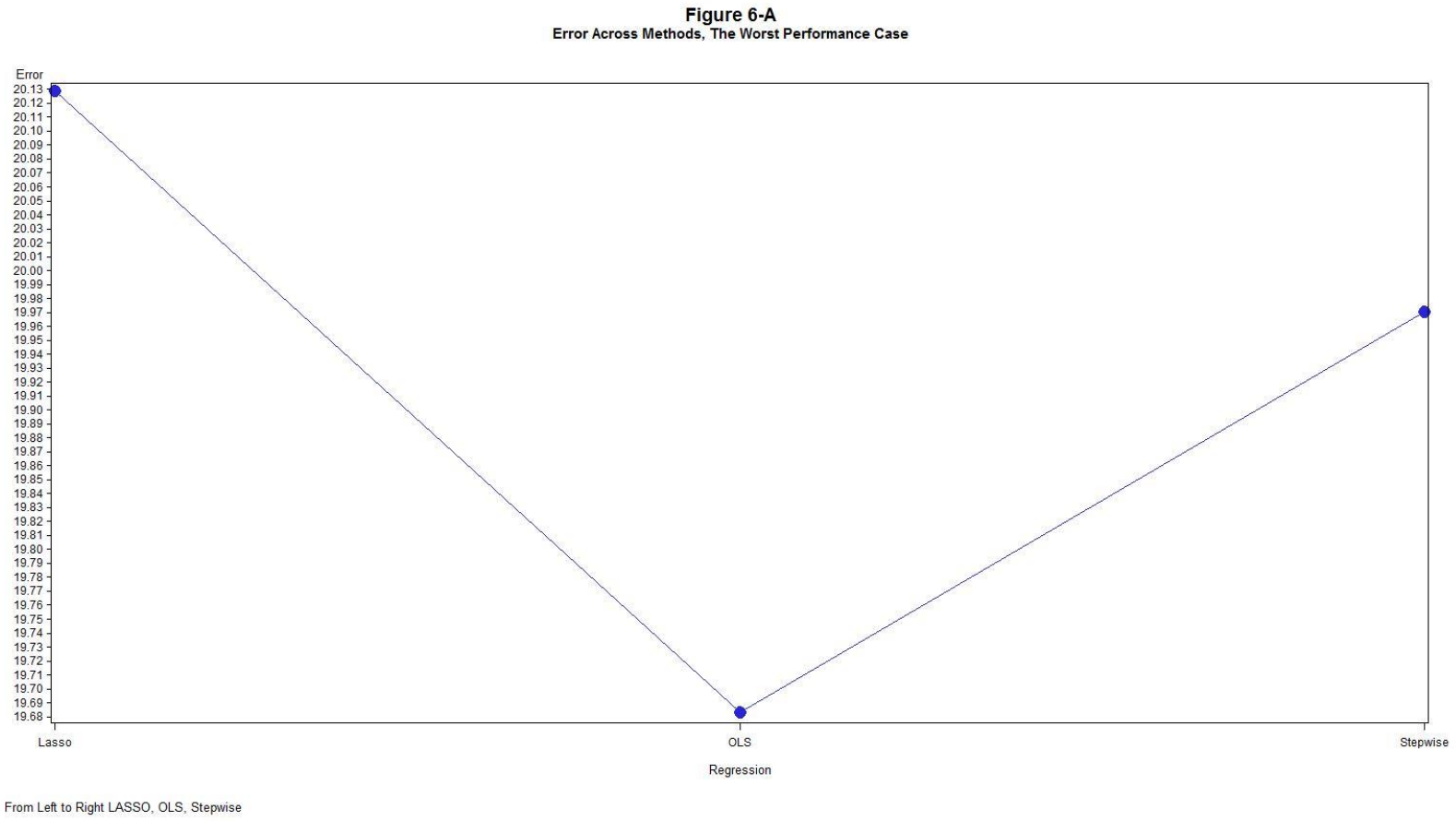


Figure 6-B: Percent Error Across Methods, The Worst Performance Case

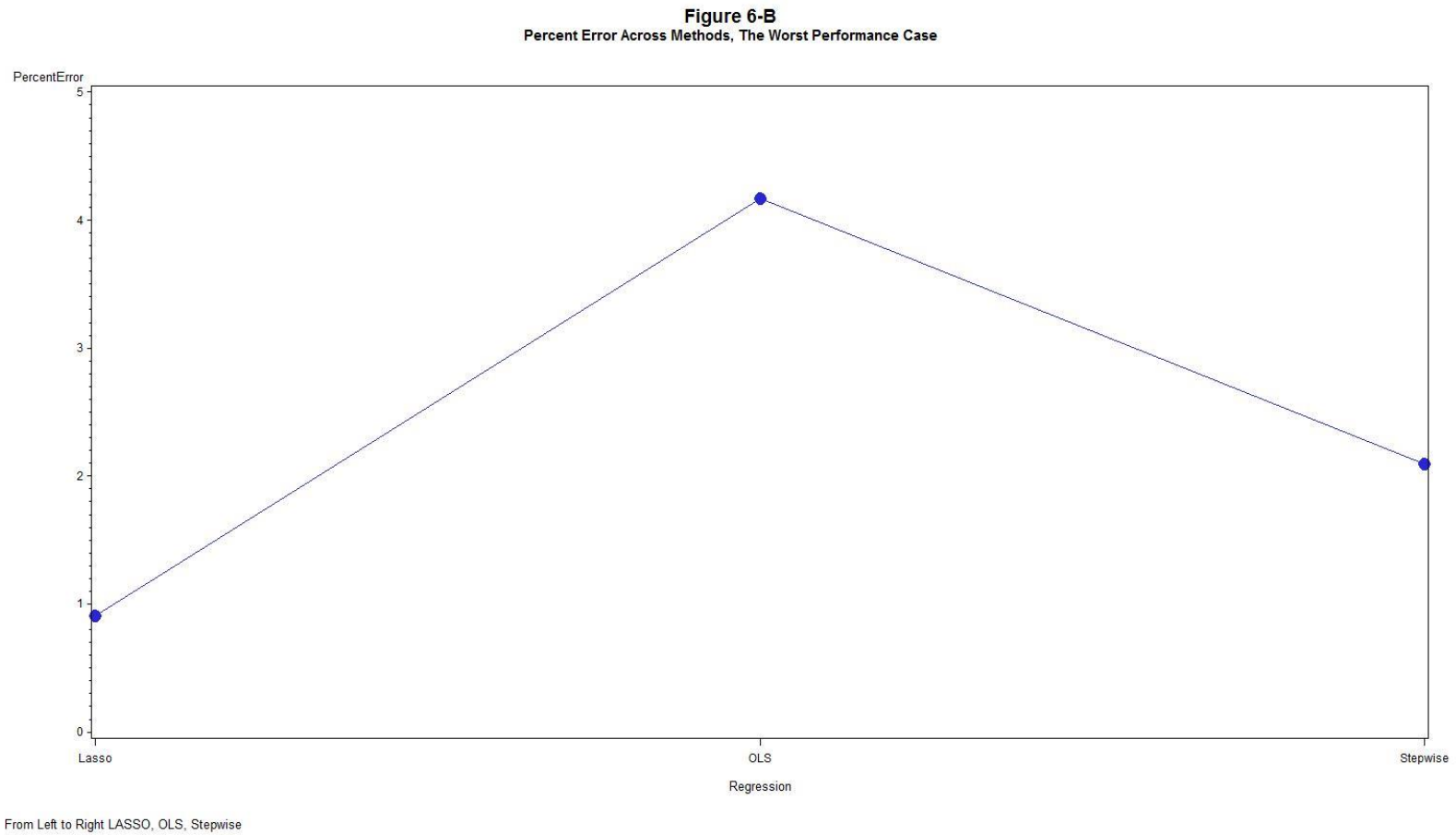


Figure 6-C: Error Across Methods, The Best Performance Case

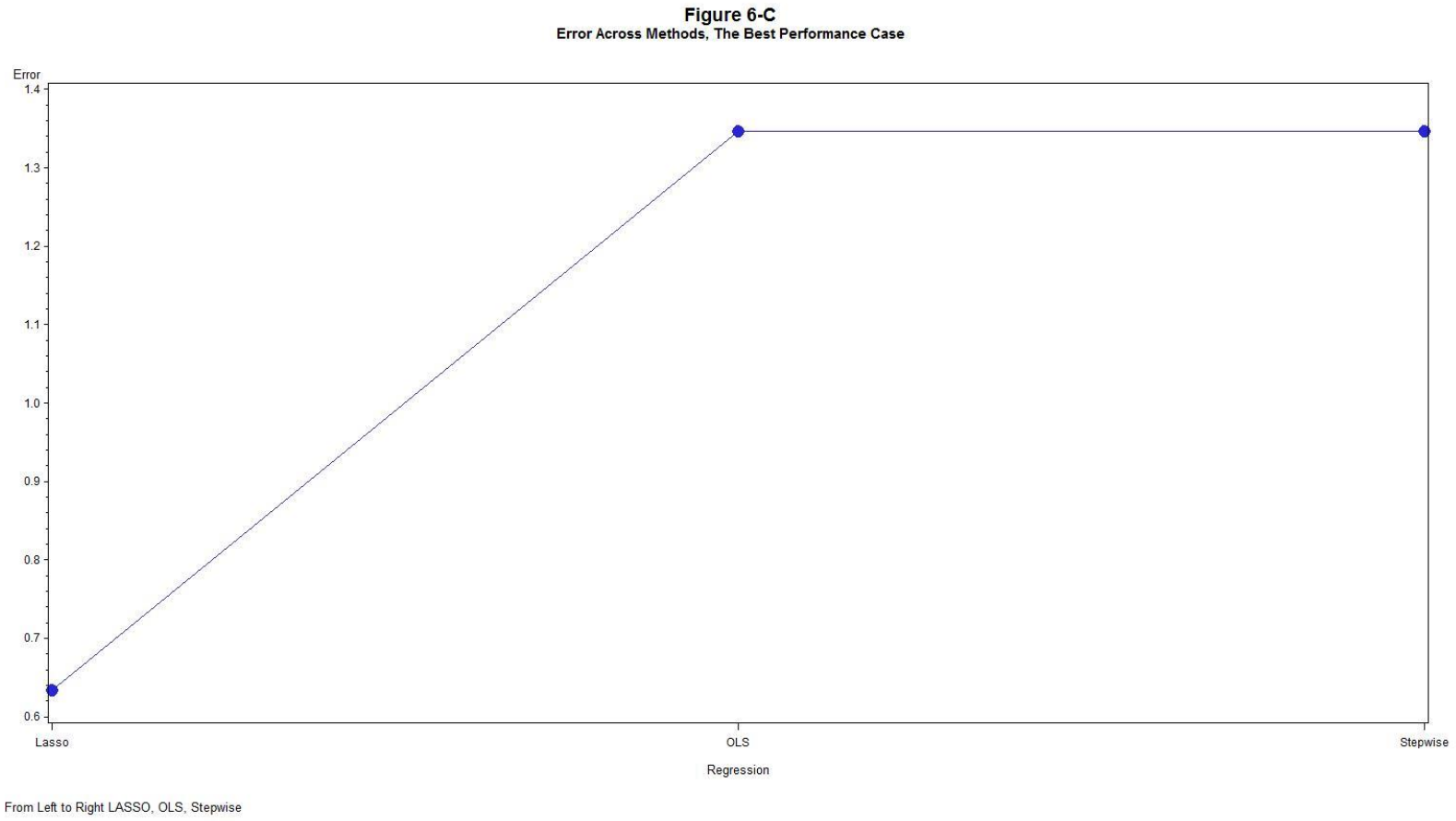
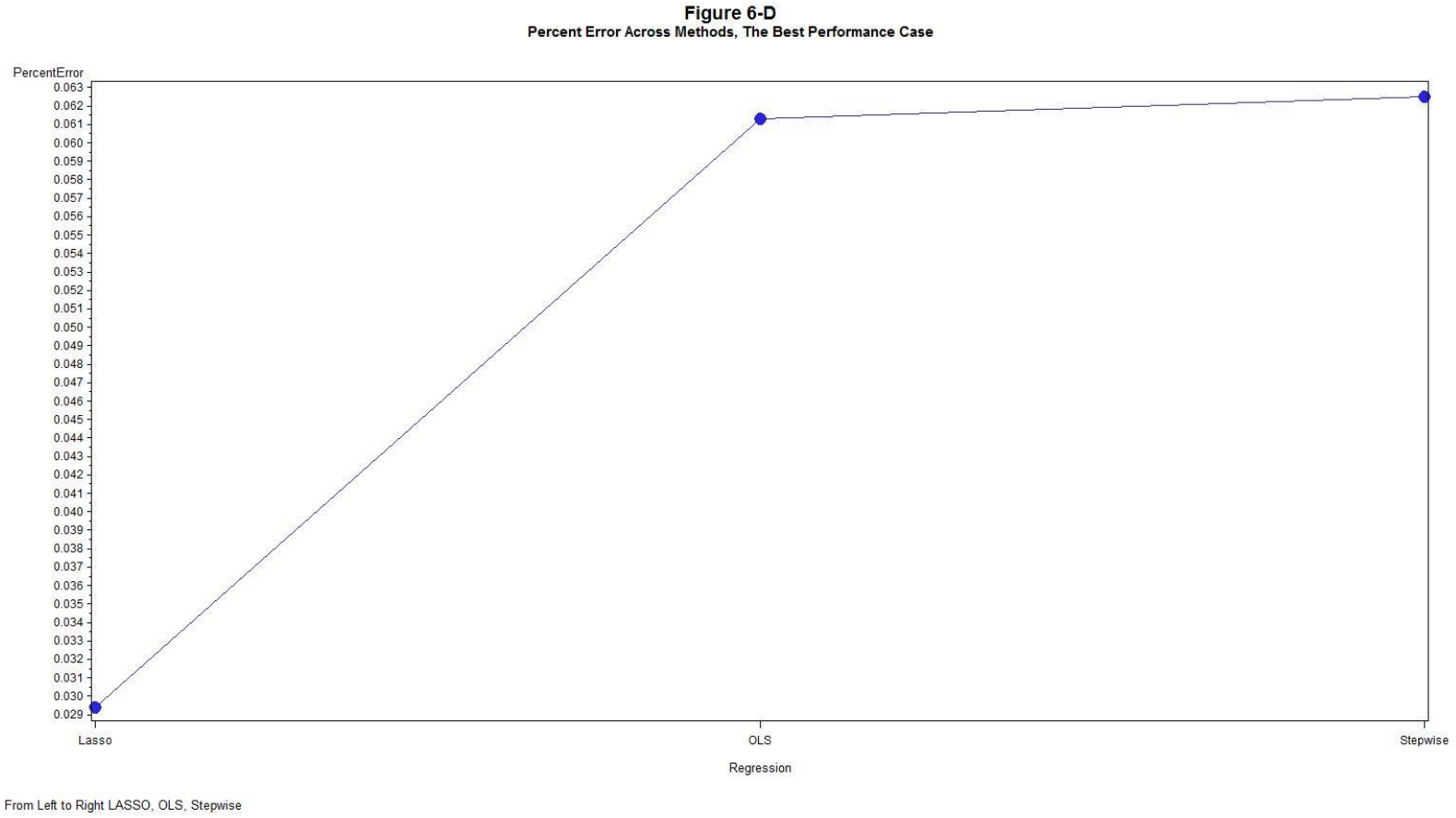


Figure 6-D: Percent Error Across Methods, The Best Performance Case

-



Chapter II: MEASURING DIVERSIFICATION LEVEL THROUGH STOCK AND INDUSTRY RETURN CORRELATION

Introduction

The question of whether diversified firms are more or less valuable than undiversified firms is of tremendous importance to corporate finance decision making. Indeed, throughout recent history a great number of firms have chosen to acquire other firms for various reasons, but presumably always to achieve greater shareholder value. Yet the literature stream on diversification's effect on firm value is contentious and generally inconclusive. In fact, a balance of the literature largely contends that diversification destroys value and that there exists a "diversification discount," calling in to question why any firm would choose to make diversifying acquisitions. For instance, early work in the area such as that of Lang and Stulz (1994) and Berger and Ofek (1995) find that greater diversification is associated with lower firm value, as does that of Rajan, Serveas and Zingales (2000). Meanwhile, using different data sources and methodologies, Villalonga (2004) finds evidence of a diversification premium rather than discount. Her findings are corroborated by the work of Hadlock et al. (2001) and Chang and Yu (1999). To say the least, the issue of diversification's effect on firm value is unsettled. The inability of the literature to reach a consensus encourages us to believe that perhaps the disagreement arises from non-comparable time series or data sources.

To avoid such hazards I develop an alternative method of identifying industry exposure that relies on investor perception of industry classification to re-evaluate the diversification discount. I do this by running regressions of firm level returns on industry level returns to determine a firm's industry exposures. The use of regression analysis to determine industry exposure allows me to develop a long time series and avoid the of drawbacks alternative industry identification such as self-reporting and regulation change. I find that overall an unconditional diversification premium exists but that it is cyclical in nature. Moreover, the results are robust to measuring premium a number of different ways including the use of Tobin's Q, market-to-book ratio and even across portfolio return alphas. Because of the long time series that return regressions provide, I am able to evaluate the premium at several of the time intervals used in previous literature and am able to provide some reconciliation across previously contrasting results. Finally, I provide evidence that both debt and macroeconomic variables play a part in explaining the cyclical nature of the premium or discount.

This remainder of this paper is organized as follows. Section II reviews the relevant literature on the value of diversification and elaborates on the shortcomings of the most commonly used data sources. Section III describes the framework under which I develop my methodology and Section IV describes the data. Section V provides my empirical results and explores the time trends in the value of diversification. Section VI examines reasons for the time trends I find. Section VII concludes.

Literature Review

A key contribution I make is to provide clarity to previously contradictory evidence regarding diversification's effect on firm value. The work of Lang and Stulz (1994), Berger and Ofek (1995), Rajan, Serveas and Zingales (2000) and many others find evidence of a diversification discount, yet these findings come with several caveats. For instance, Lang and Stulz (1994) examine diversification in the 1980s and, while not reaching any specific conclusions with respect to causality, account for the possibility that diversified firms are simply poorer performers than single segment firms and thus have lower valuations. They also suggest that perhaps diversified firms simply operate in lower valuation industries. Accordingly, Servaes (1996), in his study of the merger wave of the 1960s and 70s, adjusts his measure of valuation by industry to reach the conclusion that the diversification discount existed in the early part of his sample period but not the later. Meanwhile, Villalonga (2004) identifies multiple problems with prior studies, proposes new methods to circumvent them, and finds that there exists a strong diversification premium.

For the most part, however, the literature provides evidence that diversification destroys value, resulting in a diversification discount. Bradley et al (1988), Agarwal et al (1992) and Megginson et al (2000) all conduct event studies that show investors' reaction to merger announcements is negative. Similar results from corporate spinoffs are found by Desai and Jain (1999). The studies that conclude a diversification discount exists and seek to explain it can be broadly distinguished into one of two categories. The first is those studies that find diversified firms are discounted due to firm characteristics such as

capital allocation decisions or agency conflicts that arise from diversification. The second is studies that contend that diversification is either misinterpreted or measured incorrectly.

The literature on internal capital misallocation spans many origins such as inter-correlation of the multiple segments returns (Shin and Stulz (1998)), cross-subsidization hypothesis (Scharfstein (1998)), and diversity of individual investment opportunities across divisions (Rajan et al. (2000)). Hyland (1999), for example, argues firms with few growth opportunities use diversification through acquisition as a means to obtain growth that would be difficult to generate internally.

Managers may also deliberately engage in non-optimal decisions due to agency conflicts. Palia (1999) and Anderson et al. (1998) find negative correlations between the diversification discount and the governance of the firm, measured by higher pay-performance sensitivity, smaller boards and higher stock ownership. Also, May (1995) finds that managers tend to engage in diversification strategies of the firm when they have large amounts of their own wealth tied to the firm. The literature on manager to diversification discount level is robust, Maksimovic and Philips (2001) investigate the effects of asset sales and purchases to find that conglomerate firms are less efficient than single segment firms. They explain that one agency conflict that is exacerbated by diversification is empire building and managerial entrenchment. Morck, Shleifer, and Vishny (1990) provide evidence of this argument by documenting the stock market reaction to diversifying acquisitions, especially when the bidder's management had poor

past performance[3]. John and Ofek (1995) find that when firms sell off assets the remaining assets tend to perform better. Alternatively, Maksimovic and Philips (2002) suggest that firms do purchase assets efficiently across industries as they grow; they find that as prospects in their main industry significantly improve then managers will become more focused but not if the prospects are lower than other firms that choose to remain focused.

The relation between firm value and diversification may also be an endogenous one. Lamont and Polk (2002) address this issue by examining "diversity" – or the within-firm investment opportunities. They distinguish between changes due to the reporting of business segments (endogenous changes) from changes due to industry characteristics (exogenous changes) to conclude that diversification destroys value. This approach is noteworthy because it questions the validity of using self-reported Standardized Industrial Classification (SIC) codes as a measure of diversification, at least to the extent the researcher wishes to draw conclusions about the effect of diversification on firm value. Furthermore, Graham et al. (1999), Hyland (1999), Campa and Kedia (1999) and Villalonga (1999) all find support for the hypothesis that diversified firms trade at a discount prior to becoming diversified and so logically the discount may be endogenously determined. Chevalier (2000) explains that if firms exhibit the behaviors prior to merger that researchers are using as a basis, post-merger, to identify a discount then the results should be suspect. This possibility is consistent with Campa and Kedia (1999) who show that fundamentals of the firm (size, sales, growth rate, et cetera) are different across

[3] Lamont (1997) uses oil firms investment and shows that when the firm oil division revenues are facing negative pressure that they will under invest in the non-oil (positive NPV) segments.

diversified and non-diversified firms. Campa and Kedia (2002) reaffirm the empirical finding of a diversification discount, though they argue that diversification does not result in a discount per se, but rather that firm characteristics which may cause them to trade at a discount are also common to firms that choose to diversify.

Measuring diversification is critical, and Compustat Segment files are the most frequently used data for doing so. Indeed, these data are the primary source for segment classification such as that used in Lang and Stulz (1994) among many others. These data report each industry, identified by Standard Industrial Classification (SIC) code that is reported by firm management. There is at least one obvious and several more subtle shortcomings of these data. First, data in the Compustat Segment files begin in 1986, limiting the time series over which any analysis can be conducted. Second, the use of SIC industry classifications limits the ability of the firm to describe its involvement in, or more importantly its exposure to, multiple industry effects. Third, SIC industry classifications are self-reported.[4] As Botosan and Stanford (2005) find, management may have incentives to inaccurately report SIC classifications, potentially biasing the reporting to the extent management cannot or does not accurately determine the firm's industry exposure. Finally, segment reporting standards have changed significantly and frequently since the inception of segment reporting in 1977 under SFAS No. 14, most notably with the issuance of SFAS No. 131 in 1997.

In an effort to overcome some of these shortcomings, Villalonga (2004) uses the BITS dataset, which was constructed by the U.S. Census Bureau under contract to the

[4] For a more detailed explanation of SIC codes and their origins see Appendix 1

Office of Advocacy of the U.S. Small Business Administration. The use of the BITS data could overcome a central problem of many other studies. As Licthtenberg (1991) explains, the managerial authority to disaggregate the data revealed to investors may allow managers to underreport their activity in various business segments out of their own self-interest. Using the BITS data Villalonga (2004) reexamines the diversification discount and finds that diversification is in fact associated with higher firm value. She speculates that the discounts, previously found, are likely a result of noise in the segment data.

Villalonga (2004) argues that there are three fundamental limitations to reported business segments. The first is that, because only those segments that make up more than 10% of sales, assets, or profits are required to be reported, firms can be more diversified than suggested by their reported segments. Lichtenberg (1991) argues this is a major issue, reporting that this is binding for 17% of all Compustat firms and Montgomery (1994) shows that it is even worse for the 500 largest firms with 56% being bound. Additionally, the definition of business segments is so flexible that firms can easily combined several industries that are in fact quite different[5]. The second limitation documented by Villalonga (2004) is bias in the measurement of Tobin's Q. First, the Q of an industry of diversified firms may not be representative of the Q for that industry for single segment firms. Additionally, as is noted by Villalonga, the segments of diversified

[5] Davis and Duhaime (1992) find that in 5% to 10% of businesses are grouped into industries that are not related. Denis et al (1997) and Hyland (1999) show that about one-fourth of the reported changes in segments are reporting changes only and have nothing to do with actual diversification level of the firm.

firms tend to be larger than the segments of standalone firms. Separately, Whited (2001) provides evidence that Tobin's Q is not an appropriate measure anyway.

Corporate diversification has been measured a variety of ways over the literature but the two most common approaches are business segment count or strategic approach. The two most common methods of counting the business segments are either SIC codes or FASB-mandated self-reported segments. SIC counting, as Sayrak (2003) explains, fails to consider the relative importance or distribution of the firm's involvement in each industry segment, and for this reason Berry (1971) and McVey (1972) suggest using the Herfindahl index to measure the industry concentration, or alternatively Jacquemin and Berry's (1979) Entropy measure to identify additionally the relatedness of the various indexes[6].

Across this vast literature no consensus has been reached, and in fact this has raised additional questions such as what is the appropriate definition of a diversification discount and what is the correct way to measure corporate diversification. Villalonga (2004) documents the explanations given during a round table discussion by diversification discount authors, answering many questions about the current state of the literature. They describe three types of diversification discount literature but that they are strong-form, semi-strong form, and weak form diversification discount hypothesis.

[6] Strategic measures of corporate diversification generally involve making judgment calls about what industry groups a firm falls into. Two measures commonly used within the strategic management literature are Wrigley (1970) and Rumelt (1974). Wrigley defines four categories; single business, dominant business, related business, and unrelated business. Rumelt expands the classification to 9 groups and develops procedures for how researchers should classify each one. With respect to Rumelt's method, he proposes assessing the firm's utilization of strengths, core skills, and purposes in the identification of segments. The largest criticism that these methods face in the finance literature is the subjectivity on the part of the researcher.

Strong-form asks whether diversified firms are worth less than what they would have been if their segments were operated as stand-alone businesses. Participants also explain that a strong-form discount can be explained by various agency arguments: risk reduction; empire building; managerial entrenchment; or by inefficient investment. The semi-strong form asks are diversified firms worth less than what they would be, were they to be split into pieces. Explanations of the semi-strong form include corporate refocusing: information asymmetries; analyst specialization; secular decrease in transaction costs of external funds; or market liquidity. Weak form diversification discount asks are diversified firms worth less than specialized firms in the same industry. This form is generally explained through value maximizing behavior, and is among the most common of the early literature on the subject.

Finally, Villalonga discusses multiple benefits that arise from diversification. Among those are efficient internal capital markets, debt coinsurance, use of non-tradable resources, economies of scope, or market power.

Another issue to be addressed when studying diversification's effect on firm value is simply how to measure firm value. Many researchers (Scharfstein (1998), Wernefelt and Monthgomery (1998), Lang and Stulz (1994)), have used Tobin's Q, defining a diversification discount as one in which the portfolio of diversified firms trades at a lower Tobin's Q as compared to a portfolio of comparable stand-alone firms.

Empirically Identifying Industries

Any discussion of a diversification discount or premium is hinges on the reliability of the data used to calculate the discount or premium. In this section I describe how my methodology is more reliable and can resolve many of the data shortcomings faced by other researchers. In Chapter 1 I discuss the theoretical motivation for the decomposition of the CAPM beta loading into industry level beta loadings. Additionally, I explain the econometric basis for regression analysis of the decomposition and provide simulations to determine the accuracy with which an empirical identification of industries can be performed. In what follows, I highlight the theory enabling us to empirically identify industry exposures from a firm's underlying returns[7].

CAPM defines the expected returns of any particular firm to be equal to the risk free rate plus the product of the beta loading and the excess return of the market portfolio

$$\tilde{F}_n = rf + \beta_{n,i}\left(\tilde{M} - r\right) + \varepsilon \tag{19}$$

We can rewrite the market portfolio excess returns as a summation of the weighted individual asset returns, where there are a total of J assets in the market, and those J assets can be assigned to I portfolios based on similarities of the underlying assets.

$$\tilde{M} - rf = \sum_{i=1}^{I} m_i\left(\tilde{C}_i - r\right) \tag{20}$$

If we assume that firms that hold a portfolio of the same assets (or similar assets) operate in the same industry and if the inter-correlation of all assets within portfolio_i is high enough, then we can consider the portfolio of these assets to be industries. The result is

[7] For a more thorough explanation of the theoretical motivation please see section 4 of French and Gibbs (2015).

that *J* assets can be categorized across *i* portfolios, where a larger i refers to more specific industries and a smaller *i* refers to more general industries[8].

Replacing the excess returns of the market portfolio with the excess returns on industries and moving the beta within the summation yields

$$\tilde{F}_n = rf + \sum_{i=1}^{I} \beta_n m_i \left(\tilde{C}_i - r\right) + \varepsilon \tag{21}$$

Where a) the summation of all weighted betas is equal to the CAPM beta, and b) the sum of the returns of the underlying assets of the firm must be equal to the systematic return of the firm.

$$\beta_n = \sum_{i=1}^{I} m_i \beta_{n,i} \tag{22}$$

Some substitution and re-writing leads to

$$F_n = r_f + \sum_{i=1}^{I} \sum_{j=1}^{J} m_i (\tilde{c}_i - r) m_j \beta_{n,j} + \varepsilon \tag{23}$$

This implies that both (a) and (b) are true. The industry beta for any industry in which the firm does not operate will be zero. Intuitively the derivation is easy to understand if we simply consider that all assets provide a return and that if we sum all the value weighted asset returns we will have a firm return. No matter how you separate the assets into value weighted groups, the summation of those value weighted group returns will still equal the firm's return or the return of the summation of value weighted individual assets.

Following Jensen (1969) we can re-write the expected returns from (5) into the realized returns equation

$$\left(F_n - rf\right) = \hat{\alpha} + \sum_{i=1}^{I} m_i \hat{\beta}_{n,i} \left(C_i - r\right) + \varepsilon_n \tag{24}$$

[8] Similar to SIC codes where a four digit SIC code is more specific than a two digit SIC code.

Because (6) only identifies the weighted beta we must rewrite (6) such that we can isolate estimated weighted betas, $\hat{\beta}_{n,i}\hat{m}_j$.

$$(F_n - rf) = \alpha + \left(\sum_{i=1}^{I} \sum_{j=1}^{J} \beta_{n,j} m_i m_j (\tilde{c}_i - r) \right) + \varepsilon_n \tag{25}$$

There remains a problem, however. Returns of industries are likely correlated. To mitigate this problem we define industry return as

$$\tilde{C}_i = rf + \beta_{C_i} \left(\tilde{M} - r \right) + \varepsilon \tag{26}$$

and to estimate that relationship

$$\left(C_i - rf \right) = \alpha + \hat{\beta}_{C_i} \left(M - r \right) + \varepsilon_{C_i} \tag{27}$$

Substituting (17) into (14) gives

$$(\mathrm{F}_n\text{-rf}) = \alpha + \sum_{i=1}^{I} \sum_{J=1}^{J} \underbrace{\beta_{n,j} m_i m_j}_{\beta_{I,C}} \underbrace{(\alpha + \beta_{c_i}(M - rf) + \varepsilon_{c_i})}_{C_i - r} + \varepsilon_n \tag{10}$$

As can be seen, (10) is equivalent to the CAPM regression but allows us to empirically identify the betas of each individual industry.

Data

In Chapter 1 I show that researchers can empirically identify industry level betas and that betas that are statistically greater than zero signify the firm's exposures to a particular industry. The identification of these industries allows me to determine the level of a firm's diversification in a less subjective manner than is commonly used in the literature and also enables me to evaluate a very long time series.

I collect firm returns from the CRSP daily stock file and CRSP monthly stock file for the period 1974-2013. I use daily returns data in the initial identification of industries, and monthly returns data to identify one month future returns to high and low diversification portfolios. Book-to-market and Tobin's Q are calculated using data from the quarterly fundamentals file from Compustat. My dataset has 4,815,683 firm month observations spanning December of 1969 through December of 2013, and contains 24,617 unique firms. Following French and Gibbs (2015) I use the industry returns provided in the Kenneth French online data library.

Because I wish to make time series comparisons with previous studies, I choose the time period of 1974-2013 in order to provide a dataset that encompasses the majority of time series previously used in the literature. Many methods have been used to identify diversification level, e.g. SIC codes, NAICS, 10-K filings, BITs, but the time series of each either fails to overlap with other studies or is simply not comparable. SIC codes for example are not assigned after 2004 and NAICS did not exist prior to 1997. Furthermore 10-K filings have undergone substantial overhauls multiple times that makes a time series unreliable.

It is worth noting that the returns used to generate the Fama French industry classifications are themselves determined by self-reported SIC codes. However, while there is still some room for subjectivity, it is unlikely to affect my results in a meaningful way. SIC codes used to determine the Fama French industry returns are still the primary SIC code reported by each firm and are categorized based on similarities across primary industries. The literature suggests that there are opportunities for the primary SIC codes

to be mis-represented (Lichtenberg (1991), Davis and Duhaime (1992) and others) but on average, because of Fama and French's categorization by commonality of firm within each assigned industry, the majority of primary SIC codes would still represent the largest class of assets, or cash-flows from those assets, and so deviations from optimal reporting would represent only noise. French and Gibbs (2015) provide simulation results to evaluate the accuracy of industry identification. Across specification of differing levels of noise, I find that there is little difference in the ability to empirically identify the true industries (using Fama and French (1997) categories), and this is especially true when a stepwise regression technique is applied. French and Gibbs (2015) also provide evidence that the primary self-reported industry is generally the most significant industry identified under m methodology, and it is rare for the primary industry to not appear in the measured model suggesting that this methodology successfully overcomes the noise introduced from mis-classification to identify both primary and secondary industry exposures.

Calculation of Diversification Level

To identify industry exposures I run stepwise regressions of market adjusted firm returns on market adjusted industry returns using the previous year's daily returns. This procedure involves introducing one industry's daily return series at a time into the model, descending according to its f-statistics with respect to the firm's returns, and measures the marginal explanatory power the industry returns have on the firm's returns. The output is a list of the industries that are able to improve the explained portion of the dependent

variable and the coefficients for each industry. In order to measure the diversification level, I count the number of industries that have positive statistically significant coefficients. I calculate this measure monthly using the prior 252 daily returns[9]. I use only positive coefficients because it is unclear how or why a firm that is diversified would engage in activities to create such an exposure or what the economic interpretation of negative exposures is.

The empirical measure is calculated in two stages. First, I run two regressions: a) a regression of excess industry returns on excess market returns and b) a regression of firm excess returns on market excess returns.[10] Second, I use the residuals from each of the prior regressions in a stepwise regression to determine the number of industries that have explanatory power over the firm's returns. A stepwise regression is more appropriate for this purpose because of the large number of independent variables and the inter-correlation between those variables. In an OLS setting I must be able to account for multi-collinearity. The standard stepwise parameters of entry and exit levels of .05 are applied. I calculate diversification level monthly using the previous year's daily observations of both industry and firm returns. I use one year of data for two reasons: a) 10-K filings are reported annually and there is ample evidence that over multi-year periods diversification levels can change (Denis et al (1997) and Hyland (1999)) and b) there seems to be little evidence that extending the time series improves a stepwise regression's accuracy[11].

[9] In unreported tests I check several other time lengths and find little difference in my results

[10] In this context "excess" returns are returns minus the risk free rate, and should be distinguished from the "residual" which is the output of the regression.

[11] See French and Gibbs (2015) on error across time horizon

The Sample

To determine whether a firm falls into a category of high or low levels of diversification I first calculate the monthly median values of statistically significant industries. I consider firms that have more statistically significant industries than the median to be diversified and those with fewer than the median to be un-diversified. I measure market-to-book as the market capitalization of the firm divided by the common equity as reported by COMPUSTAT, and Tobin's Q as the market value of assets divided by the book value of assets. The diversification premium or discount is the difference of the average market-to-book (or Tobin's Q) for the portfolio of diversified firms and un-diversified firms, for each month. My study evaluates the diversification premium or discount over 468 months using both measures. Following Villalonga (2004) I exclude firms in the financial sector (SIC 6000-6999), agriculture (SIC below 1000), government (SIC 9000), and other non-economic activities such as membership organization (SIC 8600), private households (SIC 8800), and unclassified services (SIC 8900).

My sample contains 22,724,452 firm month observations and 21,489 unique firms. Table 1 reports the summary of statistics for the sample. In Panel A I report that the median number of statistically significant industries is 2.00 and the mean is 2.09. The standard deviation is 1.46 and roughly 90 percent of firms have four or fewer industries. Panel B reports the mean and median number of industries across quintiles of market equity and debt to assets. Interestingly, firms with a higher market value of equity tend to have exposure to more industries which suggests that empire building may play a part in diversification. There does not appear to be a clear relationship between debt to assets

and industry exposures. Panel C reports the summary statistics for market-to-book and Tobin's Q. The median market-to-book level is 1.71 and the median Tobin's Q is 1.31. As is expected the standard deviation of market-to-book is substantially higher than that of Tobin's Q[12]. It is worth noting that there are firms that have zero industry exposure. In unreported results I verify that results are unchanged by exclusion of these firms.

Figure 1 plots the average diversification level through time. The diversification level appears to be fairly consistent on average, however there are time periods of increased average diversification. The late '80s, the late '90s and the late 2000s all appear to have short increases in average diversification level while the mid 90's has a small drop in average diversification level.

Diversification Discount or Premium

I now test the principal hypothesis of my study, that diversified firms are valued differently than non-diversified firms. I discuss implications for further research and provide evidence that the relationship is robust. I also consider time series trends in both diversification level and the diversification discount or premium.

In Table 2 I report results from a Fama-MacBeth (1973) regression. I report four models, Model 1 and 3 use the dependent variable market-to-book and Model 2 and 4 use Tobin's Q. I include the control variables of lagged asset growth and profitability. Models 1 and 2 use raw diversification level and Models 3 and 4 use log of diversification level. Newey-West auto-correlation corrections are applied to all four models to correct for the serial correlation that is likely to result from a measure as

[12] See Adam and Goyal (2008)

persistent as diversification level. The results in Table 2 show that, across all four models, there is a diversification premium as indicated by the statistically significant coefficients on either Diversification or Log (1+Diversification). The premium is always positive and is significant at the one percent level. The intercepts between 2.25 for diversification and 2.28 for log of diversification (1.63 and 1.65 for Tobin's Q) which are in line with the summary statistics reported in Table 2. Again lagged asset growth and profitability consistently appear negative, though not always significant.

The results from Table 2 imply an unconditional diversification premium. However the initial motivation of an empirical calculation of diversification level is to allow returns to tell the story of the effect of diversification. The motivation of empirical prediction relies on the assumption that investors are able to identify exposures of firms and will price them accordingly. If this is true, then the natural question is whether investors price the firms differently by adjusting their expected return. Tobin's Q and market-to-book are generally considered to be reliable valuation measures because they capture these investor expectations. Although it is not common across the diversification discount literature, I believe that a firm's ex-post returns may provide insight into whether a diversification discount or a premium exists. Table 3 shows the results of my analysis using this philosophy.

Because it is unclear that market-to-book can completely alleviate Whited's (and others) concerns, I additionally consider stock returns as a measure of discount or premium. While the presence of a diversification discount or premium is clearly important to corporate decision making concerns, it may also play a role in portfolio

management and so portfolio returns are an interesting way to determine the effects of diversification.

In the previous tables I have fairly strong evidence that a premium exists, under the assumption that valuation measures such as Tobin's Q or market-to-book are an appropriate method to measure a diversification premium or discount. In Table 3 I show the results of a test in which I am concerned with stock returns because it is widely accepted that returns are the derivative of value. That is, if a firm has a higher value, holding all else constant, I expect to see lower stock returns because investors will be willing to pay a higher price today for the same future cash flows. I evaluate the returns for two value weighted portfolios, high diversification (more than median industries) and low diversification (less than the median number of industries) across the CAPM, Fama and French 3 factors, and Carhart (1997) 4 factor models. Consistently across all three models, non-diversified firms have lower risk adjusted alphas than do diversified firms. The magnitude of the difference is quite substantial with annualized alphas ranging from 3.50% (Carhart), 3.55% (Three Factor), to 4.18 (CAPM) monotonically changing. While additional risk factors are not able to provide much additional explanatory power, there does appear to be a relationship across the HML factor loading and diversification level. Strangely, non-diversified firms have statistically higher loadings on the HML factor than do diversified firms. The results from Table 3 imply that there appears to be a strong diversification premium given that I hold cash flows constant then investors must be paying a lower price for diversified firms today.

It is worth noting that the adjusted R Squares are substantially different across the low and high portfolios. Under the CAPM specification they are 66.14% for the low portfolio, 99.04% for the high portfolio. Under the Carhart (1997) specification they are 66.90% for the low portfolio and 99.05% for the high portfolio. This may be of concern given that those firms with the highest number of industries seem like natural candidates, based on the method of identifying industries, for high diversification and therefore the relationship may be an endogenous one. However, given the method of regressing only market adjusted returns, this should not result in problems with the CAPM regression but could lead to some of the diversification premium (discount) being explainable by additional factors. This however does not appear to be the case because I see little difference in R Square between CAPM and the four factor model. The alternative is that idiosyncratic risk may have pricing power when I identify diversification level as non-systematic.

Although the Fama French analysis provides further evidence that diversification affects prices and returns, further investigation is needed to explain why researchers have arrived at such drastic difference in diversification premium or discount. Fortunately one of the benefits of an empirical evaluation of industries is a longer time series across which to compare results. In the next section I analyze the time series of diversification premium or discount and reconcile differences across the previous literature that was unable to reach consensus results.

Time Trends in Diversification Discount

Measures of diversification level vary in the time series of data. 10-K segment reporting, for example, began in 1977 under SFAS No. 14 and became available through Compustat in 1986, however the time series underwent substantial changes in 1997. Additionally, it is plausible to assume that there may be time series trends within segment reporting itself. That is, differing trends in external factors may affect the accuracy or method in which managers or agents report segments. Even if there are no trends in reporting accuracy, many of the explanations of the diversification discount previously discussed in the literature seem to be more relevant at different time periods. For example, the price of debt is time varying (Cross-Subsidization Hypothesis), as is the tax rate (Tax Shields), and the growth possibilities (Growth Opportunities).

Initially, I consider the question of whether there appears to be any time series trend in a diversification discount or premium. To evaluate this question I measure the difference five ways as the difference of Tobin's Q or market-to-book of the portfolio of diversified firms and the portfolio of non-diversified firms monthly, across the Fama MacBeth coefficients of Tobin's Q or market-to-book monthly, and across returns of a portfolio of diversified and undiversified firms. Figure 2 graphically represents the difference between each of the diversification premium measures of firm value[13]. Observations are plotted monthly from January of 1975 to December of 2013. While the magnitudes differ between across the four measures, for the most part the diversification discount (premium) is highly correlated across them. The more interesting observation is that the cyclical nature of the premium. There tends to be short periods of persistence in

[13] Appendix 2 reports the yearly average premiums for portfolios of Tobin's Q and market-to-book in table format, and Figure 3 reports only the statistically significant premiums.

one direction (premium vs. discount) followed by short periods of persistence in the other. I can see that in the late 70's and early 80's there is a clear premium to diversification, but then there is a short period in the mid 80's where a discount takes over, followed by another premium lasting from the late 80's to the mid 90's, and finally some pretty substantial discount in the mid 2000's. Figure 2, when compared to Figure 1, shows that there does not appear to be a clear relationship between the average level of diversification and the diversification discount magnitude, although a relationship appears to exist with respect to absolute value of magnitude.

To examine the time series in more detail, I consider only those times when the discount (premium) is statistically different from zero. Figure 3 plots the discount or premium only when it is significantly different from zero at the 10 percent level. I see that while the relationship with respect to time is approximately the same, the majority of the time there is neither a discount nor a premium, but rather appears for short intervals in time. It is interesting to note that, according to Figure 2B, the results do not appear to be drastically different from several previous studies, but that the time series used in the study may not be long enough to capture the entire relationship. For example, Berger and Ofek (1995) report discounts over the time period of 1986-1991 and I corroborate their findings. Rajan Servaes, and Zingales (2000) expand their time period to 1979-1993 to find a discount, and once again I also report similar findings[14]. Campa and Kedia (2002) include an even longer series of data (1978-1996) and have mixed findings. My results over this time period are periods of sustained discounts and sustained premiums and so I

[14] At least from the perspective of market-to-book, with Tobin's Q the premium becomes insignificant.

cannot clearly state that on average either should prevail. Finally, Villalonga (2004), using a different dataset, reports a significant premium over the time period of 1989-1996, which is consistent with my findings as well[15]. My results suggest that perhaps some of the problems identified by the diversification discount literature, explaining opposing results from their own, may be a relic of a dynamic relationship between diversification level and time.

In Table 4 I report results of diversification premium or discount, using my measure of diversification level, for the time periods used in the papers Berger and Ofek (1995), Rajan, Servaes, and Zingales (2000), Campa and Kedia (2002), and Villalonga (2004) across four method of identifying premium or discount, both Tobin's Q and market-to-book portfolio differences, and across Tobin's Q and market-to-book Fama MacBeth coefficients. The results show that using market-to-book portfolios that I find discounts and premiums at exactly the same time periods as the four papers show there should be. When considering the other three measures of premium or discount, I do not always find a discount when the authors do, but across all three methods the strength of the premium changes such that during times they find discount I find the lowest premium, during times they found premium I find the strongest premium. The results from Table 4 strongly imply that the results found by previous literature are likely an artifact of a time trend in diversification discount.

[15] Not only the direction but also a similar magnitude.

In the following section I will attempt to determine an explanation for the time series changes I observe in the diversification discount. I will consider multiple time series variables and evaluation whether the discount is related to macroeconomic states.

Causes of Time Trends

In this section I consider the time series relationships between a host of variables and the diversification discount or premium. The variables I include are debt, investor sentiment, and market volatility.

Diversified firms tend to have higher leverage levels than non-diversified firms. Explanations for the relationship between diversification and debt levels have been provided by the literature (such as Jensen (1986)), and typically involve the financing of diversifying mergers with debt. Additionally, conglomerates can sustain higher levels of debt because diversification reduces earnings variability (Lewellen (1971)). That is, investors are more willing to lend to diversified firms because the cash-flows tend to be more stable. Shleifer and Vishny (1992) argued that during bad states diversified firms can more easily sell assets and this leads to an increased debt capacity. The extent to which a firm is able to access capital may be an important determinant in the choice to diversify. Because of the previously documented explanations of financing constraints, I include the time series of debt levels and costs of debt as a proxies for difficulty of external financing. Jensen and Meckling (1976), and Jensen (1986) both suggest that agency conflicts could lead to diversifying events such as mergers and empire building. If

agency theory is correct then I should find that during times of high debt that those firms diversifying are sold at a discount.

The state of the economy may also affect investor's preferences toward risk. During times of high risk investors should require higher returns on risky firms and be willing to pay less for the series of cash flows that the firm is expected to produce. These risky time periods can be proxied by stock returns because investors may prefer diversified firms during periods of high market volatility (May (1995)). The risk may also show itself during periods of informational risk. When the informational environment is opaque between investors and firms, then firms may be willing to forego positive NPV projects, and diversified firms may cease to take on diversifying projects (Williamson (1970) and Myers and Majluf (1980)). Additionally, Maksimovic and Phillips (2002) document that firms diversify when the prospects in their primary industry are lower than those of their peers, but once the prospects increase they will return to more concentrated investments. This implies that diversification should not be considered as a negative choice per se, but rather making the best of a bad situation during periods of low performance.

The cross- subsidization hypothesis proposed by Scharfstein (1998) suggests that during periods when the cost of debt is high, diversified firms will be more able to perform internal financing through the movement of money from one segment to another. Rajan et al. (2000) however finds that, generally, managers of diversified firms will invest too evenly across segments and will underperform. In order to evaluate whether the cross-subsidization hypothesis is able to assist in the explanation of the time trend in

diversification discount, I include two cost of credit proxies; credit spread and Treasury yield. The credit spread (the difference in the yield of AAA and BBB bonds) provides insight into the riskiness of loaning money. When the cost and level of debt increases I expect that, if the cross-subsidization hypothesis is true, a diversified firm should be able to internally finance and will trade at a premium. Alternatively, Treasury rate also provides information towards general cost of borrowing. That is, when the cost of borrowing is high, we should find that investors will pay a premium for firms that do not need to go external to finance their NPV projects.

Baker and Wurgler (2006) find that when investor sentiment is low returns are high for firms of small stocks, young stocks, high volatility stocks, un-profitable stocks, non-dividend paying stocks, extreme growth stocks, and distressed stocks. In general they find that there is a relationship between stock returns and investor sentiment levels. Following Baker and Wurgler I hypothesize that diversified firms are less risky than non-diversified firms during periods of low sentiment, and therefore returns should be higher for non-diversified firms and a diversification premium should exist. Additionally, Akbulut and Matsusaka (2010) evaluate diversification discount using merger announcement returns and find that there is a premium, but that the premium is time varying. Their results suggest that investor sentiment is able to explain why some times investors are willing to pay more for a diversifying merger.

In Table 5 I present the results of an investigation to the empirical question of what affects the time series trends in diversification discount or premium. This table shows the results of regressions of the diversification discount or premium on each of the

variables that previous literature suggests should have explanatory power. I group the variables into three sets of independent variables; debt and cost of debt, investor sentiment, and market trends such as returns, institutional investment et cetera.

Data for Time Series Trend Explanation

Tnote is the yield of ten year U.S. Treasury note and ΔTnote is the change in the yield of the ten year Treasury note over the prior month. Long term debt is average firm level long term debt for each month. Short term debt is the average firm level short term debt for each month. These data are obtained from Compustat. DebtLT is the interaction between month average long term debt and the 10 year Treasury note yield. DebtST is the interaction between the month average short term debt and the 10 year Treasury note yield. Spread is the difference in the yield of the AAA level corporate bond and the BBB level corporate bond yield. Debt to Assets is the average firm level debt to asset level for the month, both variables come from Compustat. DAxSpread is the interaction between the average firm level debt to asset and the spread, calculated from above. Additional variables are reported in appendix 5. NIPO, Sentiment, SD, Institutional, and Turn Over all come from Jeffrey Wurgler's online database and are calculated following Baker and Wurgler (2006). NIPO is the number of IPOs that took place during the month. Sentiment is an investor's sentiment measure calculated as the first principal component of six sentiment measures. SD is the monthly level of debt issued and is reported from the Federal Reserve Bulletin. Institutional is the level of institutional investment during the period, and Turn Over is the average NYSE turnover from NYSE Factbook. Size is the

total market capitalization during the month. It is measured as the price multiplied by shares outstanding averaged across all firms reported by CRSP monthly stock file. Returns are the equally weighted average returns of all stock reported in CRSP monthly stock file for each month.

Results of Time Series Trend

Table 5 reports the results from an OLS regression of the time series of the diversification discount or premium. I report the results of the affect that debt has on diversification premium or discount and report additional explanatory variables within Appendix 6. The variables evaluated are explanatory power of debt, and in Appendix 6 I examine explanatory power of sentiment, explanatory power of market conditions, and I finally break down the explanatory by sub-periods. I find that the diversification premium described above can best be explained by varying debt levels and availability of credit such that between 21 and 34% of the premium is explainable by this one trend. There also appears to be weaker yet still significant relation between investor sentiment and the diversification premium as well as a relation with market effects.

Scharfstein (1998) supposes that cross- subsidization of different segments within the same firm may lead to differential pricing when the cost of external capital is high. I use a combination of variables to evaluate whether different time series characteristics of debt are able to explain whether the market will price up or down diversified firms. I include six variables, yield on ten year treasury, one month change in yield on ten-year treasury, average long term debt, average short term debt, corporate bond spread, average debt to assets, and three interaction terms, interaction between long term debt and

Treasury, interaction between short term debt and Treasury, and interaction between debt to assets and corporate bond spread, as independent variables. Diversification premium is the independent variable and is measured both as the premium between a diversified and non-diversified portfolios market-to-book and the premium between a diversified and non-diversified portfolios Tobin's Q. Panel A is separated into four models based on the type of debt variable that I am including in the model.

Model 1 describes the effect of debt on the diversification discount or premium. The cross- subsidization hypothesis suggests that the level of debt is important when the cost of debt is high, and so I include interaction with 10 year treasury yield to proxy for changing debt expenses. I include average long term debt, the Treasury yield, the one month change in Treasury yield, and the interaction between average long term debt and the Treasury Yield.

Models 1, 2, and 3 all support the notion that there is a strong relationship between the diversification premium/discount and the level and cost of debt. Across all three models the proxies for cost of debt are negative and statistically significant, proxies for level of debt are negative and statistically significant, and the interaction is positive and statistically significant. Although the cross-subsidization hypothesis would suggest that the premium should increase as external financing becomes more expensive, I find the opposite. My results suggest that investors are not willing to pay firms to diversify for them whenever the diversification becomes more expensive. This likely results from the fact that diversified firms have higher debt levels[16] and are financing their diversification

[16] See Table 1

with additional debt. Model 1, 2, and 3 are able to explain 20.6, 33.8, and 22.2 percent, respectively, of the market-to-book premium, and 24.3, 32.4, and 21.1 percent, respectively, of the Tobin's Q premium.

There is extensive literature on the effects of investor sentiment on firm and return characteristics and I wish to determine whether sentiment has any place in explanation of time series trend in diversification discount. Shleifer and Vishny (2003) have suggested that diversification may be the result of investor trends and that investors lack the resources to make prices fully reflect the information. Akbulut and Matsusaka (2010) find that there is time series variation in the diversification discount and that it is explainable by investor sentiment towards diversification. Their model however looks at the stock price reaction to diversifying mergers, which occurs infrequently, while mine considers general market perception of diversification in terms of monthly pricing of nearly all stocks.

In Panel A of Appendix 5 I evaluate the effect that investor sentiment measures have in explaining the time variation of the diversification discount or premium using four variables used by Baker and Wurgler (2006). The four variables are the number of IPOs during the month, the number of debt contracts issued in the month, the amount of institutional investment during the month, and their aggregate sentiment measure which is the first principle component of known sentiment measures.

I find that both the number of IPOs and the sentiment index are statistically significant in explaining the diversification premium, but are economically not very significant. While in general diversified firms tend to trade at a premium when sentiment

is high and a discount when the number of alternatives is high, the explanatory power is quite low. Only 3.7 percent of the time series trend for market-to-book, and 1.2 percent for Tobin's Q, are explainable by the two sentiment measures. While sentiment may play a role in explaining the trend, it does not appear to be a major source of explanatory power.

In Model 2 of Panel A I see, as with the previous measures of sentiment, premium measured by Tobin's Q is not explained very well using institutional investment. When measuring premium using market-to-book I do see a little explanatory power, specifically 10.8 percent of the variance can be explained.

In Model 3 of Panel A I include both sets of sentiment measures. Across both valuation measures, the signs do not change however the significance level does. Both the number of IPOs and the general sentiment index lose a great deal of significance although the changes are significantly different if I consider market-to-book or Tobin's Q. Finally, with the full model I am able to explain 11.8 percent of the market-to-book premium and only 3.3 percent of the Tobin's Q premium. The results from Panel B suggest that, contrary to the findings of Akbulut and Matsusaka (2010), sentiment does not appear to play a significant role in the time trends of the diversification premium. The results of Akbulut and Matsusaka may be specific to mergers that increase diversification.

A visual examination of Figure 2 brings to mind many potential explanations. Although many of these explanations have been examined throughout the literature, the second order consequences are likely to reveal themselves as market effects. In Panel B

of Appendix 6 I examine the explanatory power of various market effects. In addition to those reported I considered many alternative measures such as volatility and variance of industry returns. I report the results of four variables that appear to have the most explanatory power; market capitalization, growth in market capitalization, market turnover, and market returns.

Model 1 of Panel A is concerned only with the market capitalization level of the market while Model 2 measures the explanatory ability of turnover and market returns. Model 3 includes both variables so that I can differentiate between the growth of the market due to new investment vs. growth in the market due to increased returns. Across both market-to-book and Tobin's Q, in all models, size is always positive and significant but growth of the market is positive in only Model 1, but becomes negative and significant (marginally for Tobin's Q) when I include additional control variables. Market size variables are able to explain 21.2 percent of the market-to-book premium but only 6.4 percent of Tobin's Q variable. If the relationship between market-to-book premium and market size is endogenous then that would imply that the market-to-book level of diversified firms changes at a different rate with respect to the market size than does the level of non-diversified firms. Returns are positive and significant across all models but turnover is less persistent. In Model 1 for market-to-book the relationship is positive and significant but in Model 3 of both measures the relationship becomes negative and significant. The R Squared are 3.9% and 1.2 percent respectively. Model 3 can explain quite a lot by including both sets of variables. 40.1 percent of the market-to-book premium variation can be explained and 30.1 percent of the Tobin's Q premium can be

explained. The results of Panel B suggest that during times of expansion I can expect a diversification premium and during times of contraction I can expect a discount on average.

Panel C regresses the time series of diversification discount or premium against the four variables from Model 2 of Panel A of Table 5 over five specifications. Model 1 models the relationship post 2005, Model 2 before 2005, Model 3 before 2004, Model 4 before 2003, and Model 5 before 2002. Perhaps the most interesting note is that the explanatory power is drastically different before and after 2005. Prior to 2005 the R Square ranges between 53.5-59.94 percent for market-to-book premium and between 47.9-57.0 percent for Tobin's Q premium. Additionally, the R Square is monotonically increasing as I cross from Model 2 to Model 5 for both value measures. Also of interest is the difference in the relationship between the debt measures and the premium. All three variables, Treasury, short term debt, and the interaction between the two, have opposite relationships between Model 1 and Models 2-5. In Model 1 the relationship is positive for Treasury and short term debt but is negative for the interaction whereas in Models 2-5 the relationship is negative for both Treasury and short term debt and positive for the interaction. As an additional side note the relationship seen in Figure 5 appears to have a lag between credit spread and diversification premium or discount and so in Appendix 3 I evaluate the predictive power of various lags on the premium and find that there appears to be something between a one and three year lag between credit spread and diversification discount or premium.

To further understand the relationship of diversification of firm value across time I perform a two stage regression that intends to disentangle endogeneity of time series characteristics is measures of firm value from time series changes in diversification valuation. In the first stage I regress the variables from Panel A on the valuation measures market to book and Tobin's Q. I separate out the explained and unexplained portions of the valuation measure and evaluation the effect diversification has on each. I report the results in Panel B of Table 5. The results suggest across both Tobin's Q and market to book that the unconditional diversification premium is likely due to time series characteristics of determinants of valuation other than diversification level, namely changing debt levels and costs of debt. After removing these characteristics from the valuation measure there appears to be an unconditional diversification discount.

The results from Table 5 (and Appendix 6) do not perfectly explain the time series trend that I observe in the diversification premium or discount, however they do provide insight into when I will expect to see a premium or a discount. In general I will see that a premium will exist when markets are doing well, when investors are enthusiastic, and when debt is cheap. I also see that this relationship is not persistent to all time periods, specifically the explanatory relationship becomes unstable around the middle of the 2000's, a time period when markets were very enthusiastic.

Conclusion

In this paper I introduce a new method to measure firm diversification that overcomes many of the shortcomings of prior work in this area. My measure is less

subject to the data constraints previous work has faced such as time series restrictions, and is less prone to biases in segment reporting. I determine industry involvement, and by deduction, diversification, as the extent to which firm returns are explained by industry returns. This implies that if firm returns are indeed explained by a given industry's returns, the firm is involved in that industry, regardless of its reported industry segments. Using this measure of diversification I find that there is an unconditional diversification premium, however the premium is time varying. For extended periods of time in the sample period I find a diversification premium and other times I find a discount. And finally, after conditioning on time series trends in additional firm level characteristics there appears to in fact be a discount. This finding helps to explain the inconsistency of empirical results within the diversification literature. I provide marginal explanation for the time varying trend. The most important explanatory variables appear to be related to debt and the cost of debt, however they are related inconsistently with the Cross-Subsidization Hypothesis but rather provide evidence that investors are willing to pay a premium for a diversified firm only when the cost of debt is cheap. My results suggest that a diversification discount (premium) explanation which does not allow for time variance is unlikely to be the true reason for differential pricing of diversified and non-diversified firms.

Table 1: Summary Statistics on Diversification Level, January 1975 to December 2013

This table reports the summary of statistics for my entire sample. Panel A reports the summary statistics for the measure of diversification level. Diversification level is measured as the number of industry excess returns included in the regression to explain firm level excess returns. We determine the number of industries to be included through a stepwise regression with entry and exit p-value parameters of .05. The time period of the sample begins in January 1, 1975 and ends on December 31, 2013. The diversification level is measured using 250 daily observations and is calculated monthly. The dependent variable is firm level excess returns and the independent variables are Fama and French 48 industry excess returns. The minimum diversification possible is 0 and the maximum diversification level possible is 48. Panel A reports the diversification level's mean, median, standard deviation, twenty-fifth percentile and seventy-fifth percentile. Panel B reports the same summary statistics variables but for quintiles formed on size, and book debt to book assets. Assets are obtained through Compustat's Fundamentals quarterly total assets, liabilities are obtained through Compustat's Fundamentals quarterly total liabilities, and size is obtained from CRSP's monthly stock file as the product of the absolute value of price and total shares outstanding. Quintile 1 is the lowest for each of these variables and quintile 5 is the highest. Panel C reports the summary statistics for the main dependent variables used through the paper, Market-to-book, log of Market-to-book, Tobin's Q, and log of Tobin's Q. Both variables are windsorized at the one and ninety-nine percentile.

		Mean	Median	Standard Deviation	10th percentile	90th percentile
Diversification Level	All	2.09	2	1.46	0	4
Market Equity	1	1.53	1	1.23	0	3
	2	1.72	2	1.29	0	3
	3	1.97	2	1.36	0	4
	4	2.38	2	1.44	1	4
	5	2.83	3	1.49	1	5
Debt to Assets	1	1.93	2	1.39	0	4
	2	2.08	2	1.44	0	4
	3	2.21	2	1.48	0	4
	4	2.19	2	1.47	0	4
	5	2.1	2	1.47	0	4
Market-to-book	All	2.55	1.71	2.7	0.63	5.31
Tobin's Q	All	1.78	1.31	1.34	0.83	3.26

Table 2: Fama-MacBeth Regressions of Valuation Proxies on Measures of Diversification

This table uses Fama-MacBeth regressions to check for diversification premiums or discounts. The dependent variable in Model's 1 and 3 is market-to-book ratio, the dependent variable in Models 2 and 4 is Tobin's Q. We account for autocorrelation by applying Newey-West with one year lags. Diversification level is measured as the number of industry returns included in the regression to explain firm level returns. T-statistics are reported in parenthesis.

Fama-MacBeth				
	Model 1	Model 2	Model 3	Model 4
	Market-to-book	Tobin's Q	Market-to-book	Tobin's Q
Intercept	2.28	1.65	2.25	1.63
	-19.8	-26.6	-19.6	-26.1
Diversification	0.039	0.019		
	-5.30	-4.80		
Log (1+Diversification)			0.106	0.052
			-5.09	-4.71
Lagged Asset Growth	-1.78	-0.38	-1.78	-0.38
	(-2.0)	(-0.7)	(-2.0)	(-0.7)
Profitability	0	-0.01	0	-0.01
	(-0.0)	(-0.5)	(-0.0)	(-0.5)

Table 3: Risk Factor Explanation of Diversification Discount

This table reports the results of a time series regression of portfolio returns for low and high levels of diversification. A firm has a low level of diversification if it has less than two industries (the median and it has high diversification when it has more than two industries. In high minus low the variable is the difference in excess returns of the portfolio of high diversification level firms minus the excess returns of the portfolio of low diversification level firms. The independent variables are the market return, small-minus-big, high-minus-low, and up-minus-down, from the datasets on the Kenneth French website. T-statistics are reported in parenthesis.

		Less than 2	More than 2	Low minus High	Annualized
Panel A	Intercept	0.0030	-0.0004	0.0034	0.0418
		(2.25)	(-1.39)	(2.42)	
	Market	0.6790	1.0162	-0.3372	
		(22.65)	(164.83)	(-10.74)	
Panel B	Intercept	0.0025	-0.0004	0.0029	0.0355
		(1.86)	(-1.32)	(2.03)	
	Market	0.7004	1.0103	-0.3099	
		(21.59)	(151.57)	(-9.14)	
	SMB	-0.0102	0.0194	-0.0296	
		(-0.20)	(1.88)	(-0.56)	
	HML	0.1086	-0.0135	0.1221	
		(2.19)	(-1.32)	(2.36)	
Panel C	Intercept	0.0025	-0.0004	0.0029	0.0350
		(1.78)	(-1.30)	(1.96)	
	Market	0.7015	1.0104	-0.3089	
		(21.20)	(148.59)	(-8.93)	
	SMB	-0.0100	0.0195	-0.0294	
		(-0.20)	(1.88)	(-0.56)	
	HML	0.1106	-0.0133	0.1240	
		(2.17)	(-1.27)	(2.32)	
	UMD	0.0054	0.0004	0.0050	
		(0.17)	(0.06)	(0.15)	

Table 4: Comparison of Results
This table compares the results of the previous literatures effects of diversification to the results found using an alternative measure of diversification level based on regression analysis. I measure discount or premium as the difference in Tobin's Q for a portfolio of high diversification and a portfolio of low diversification stocks, or the coefficients of a Fama MacBeth regression. The second column reports for each publication the results of the study in terms of a discount, premium, or mixed result. The fourth through eighth columns reports the difference in Tobin's Q between single and multi-segment firms for each of time series using my method of calculating premium or discount.

Author(s)	Sample	Premium or Discount	MB portfolio	TQ portfolio	MB Fama MacBeth	TQ Fama MacBeth
Berger and Ofek (1995)	1986-1991	Discount	-0.114	0.002	-0.015	0.009
			(-10.95)	(0.5)	(-3.47)	(3.9)
Rajan et al. (2000)	1979-1993	Discount	-0.069	0.011	0.015	0.019
			(-10.57)	(3.48)	(3.47)	(10.08)
Campa and Kedia (2002)	1978-1996	Mixed	-0.004	0.040	0.028	0.025
			(-0.74)	(14.63)	(7.19)	(14.19)
Villalonga (2004)	1989-1996	Premium	0.186	0.142	0.054	0.041
			(22.21)	(34.23)	(12.45)	(19.08)
Entire Sample	1975-2013	Premium	0.117	0.062	0.045	0.024
			(30.04)	(32.35)	(15.39)	(15.77)

Table 5: Explanations for the Time Varying Diversification Premium

Panel A regresses the time varying diversification premium, the difference between market to book (Tobin's Q) on the average level of long term debt, the yield on 10 year treasury bonds, the one month change in yield on the 10 year treasury bond, the average short term debt level, spread between AAA and BBB corporate bonds, the average debt to assets levels, the interaction between long term debt and the yield on a 10 year treasury, the interaction between short term debt and 10 year treasury bonds, and the interaction between corporate bond spread and debt to assets. Panel 2 separates the valuation measure, market-to-book or Tobin's Q, into the portion explained and unexplained by level and cost of debt then compares the relationship across explained/unexplained valuation to diversification level. In the first stage I regress market to book or Tobin's Q on the three month treasury rate, the ten year treasury rate, total assets, total liabilities, the interactions between long term debt and ten year treasury, and the interaction between long term debt and three month treasury rates. In the second stage I separate the predicted values of the valuation measures from the residuals and evaluate the explanatory power of the diversification level. T-stats are reported in parenthesis.

Panel A: Debt Effects- Market-to-book				Debt Effects – Tobin's Q			
	Model 1	Model 2	Model 3		Model 1	Model 2	Model 3
Intercept	0.569	0.734	3.982	Intercept	0.376	0.470	2.059
	(9.800)	(12.030)	(4.300)		(13.980)	(16.090)	(4.660)
DebtLTxTnote	0.025			DebtLTxTnote	0.017		
	(5.380)				(7.660)		
Tnote	-0.062	-0.111		Tnote	-0.037	-0.059	
	(-8.970)	(-12.810)			(-11.550)	(-14.290)	
ΔTnote	0.053	0.029		ΔTnote	0.019	0.009	
	(1.530)	(0.890)			(1.150)	(0.610)	
DebtLT	-0.107			DebtLT	-0.085		
	(-5.560)				(-9.560)		
DebtSTxTnote		0.394		DebtSTxTnote		0.202	
		(11.080)				(11.840)	
DebtST		-0.002		DebtST		-0.001	
		(-10.180)				(-13.070)	
R2	0.206	0.338	0.222	R2	0.243	0.324	0.211
Adjusted R2	0.199	0.332	0.216	Adjusted R2	0.236	0.318	0.206

Panel B	MB		Tobin's Q	
Explained	Intercept	2.546	Intercept	1.773
		(8584.7)		(11475.5)
	Diversification	0.007	Diversification	0.002
		(29.8)		(19.9)
Unexplained	Intercept	0.008	Intercept	0.006
		(2.6)		(4.1)
	Diversification	-0.008	Diversification	-0.006
		(-3.4)		(-5.5)

Figure 7: Diversification Level Through Time

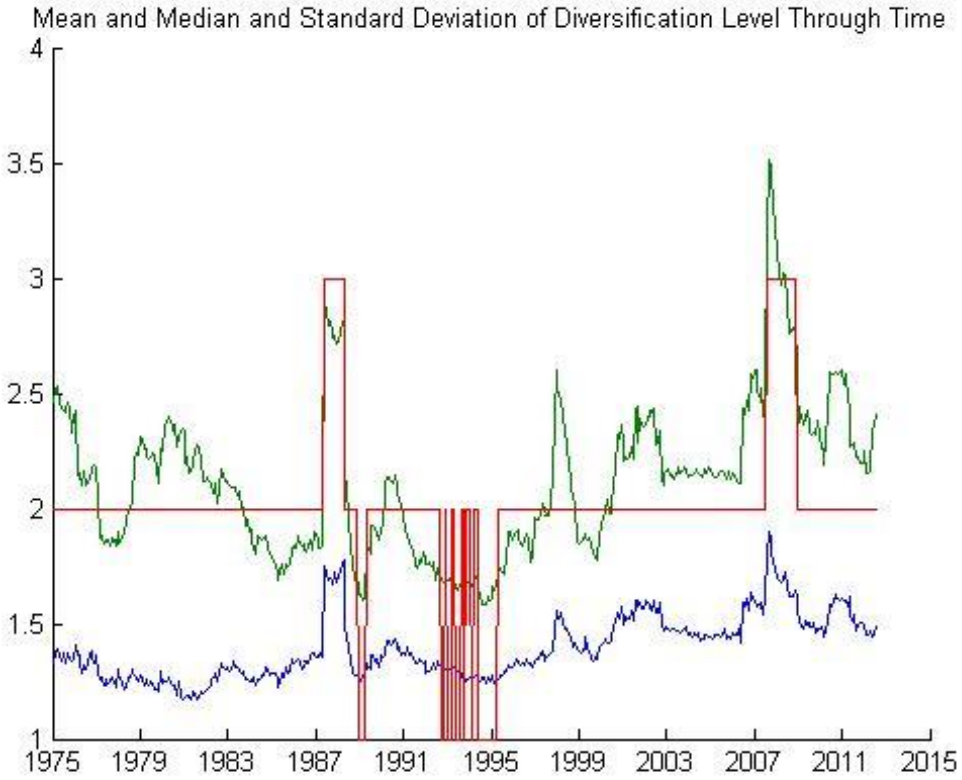


Figure 8A: Average Diversification Through Time-Market to Book

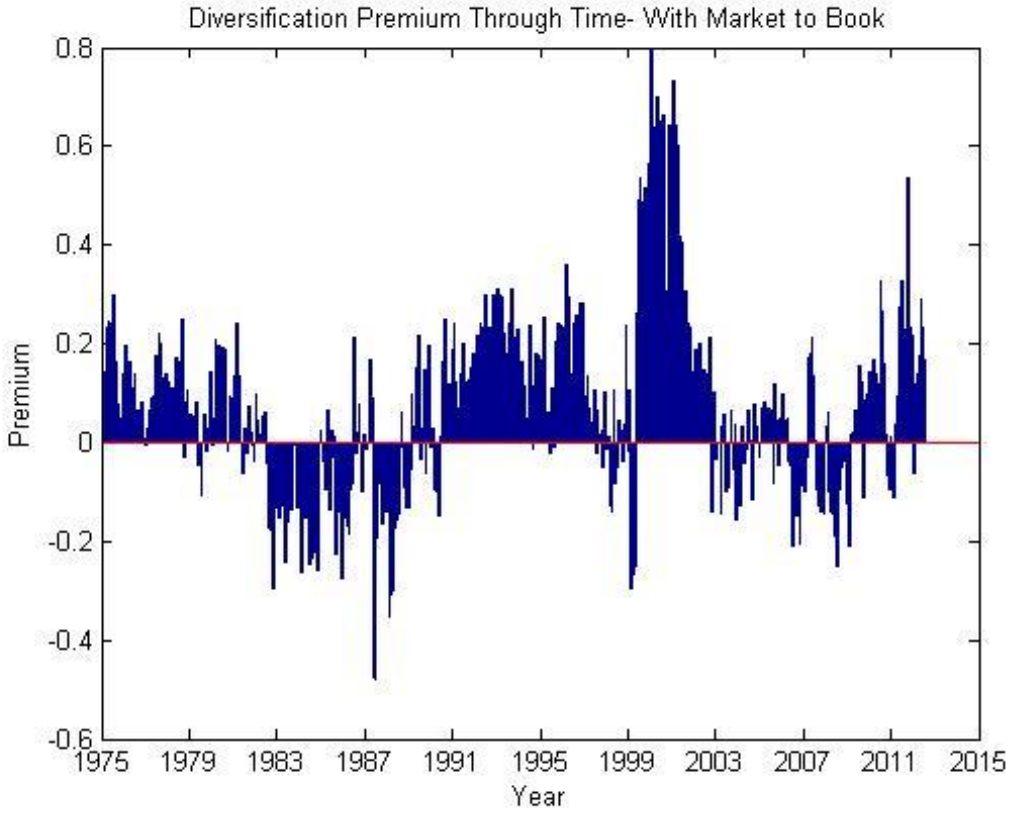


Figure 8B: Diversification Premium Through Time- Tobin's Q

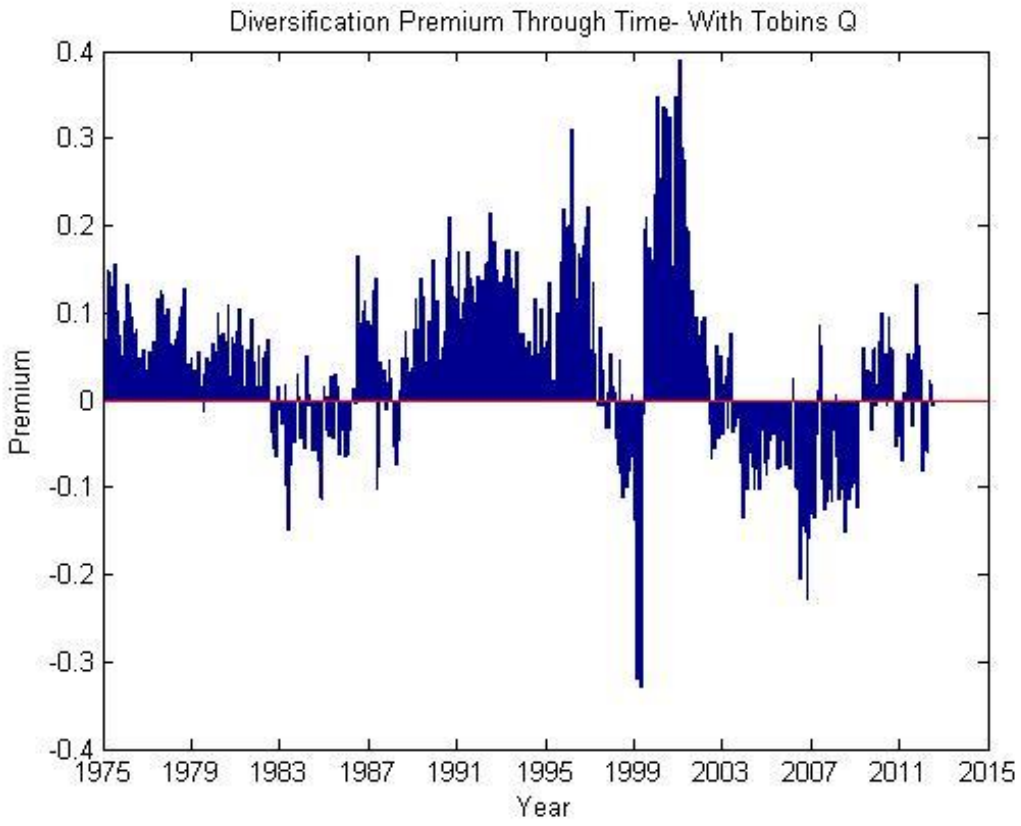


Figure 8C: Fama-MacBeth Through Time- Market to Book

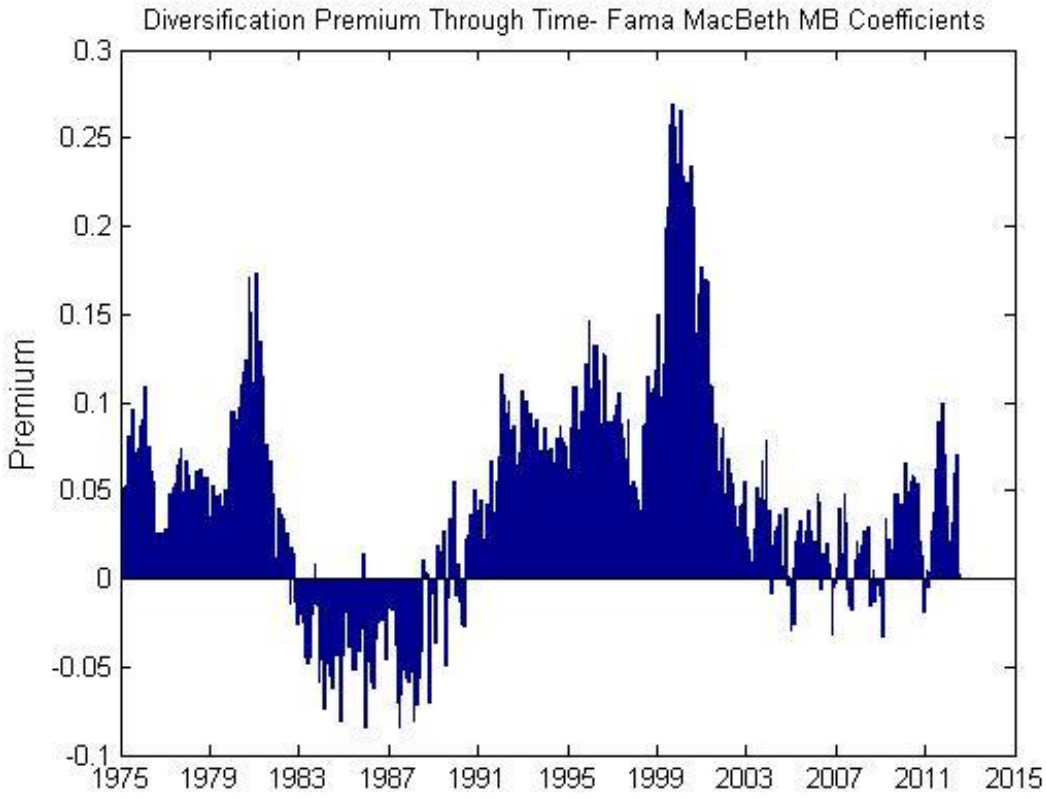


Figure 8D: Fama-MacBeth Through Time- Tobin's Q

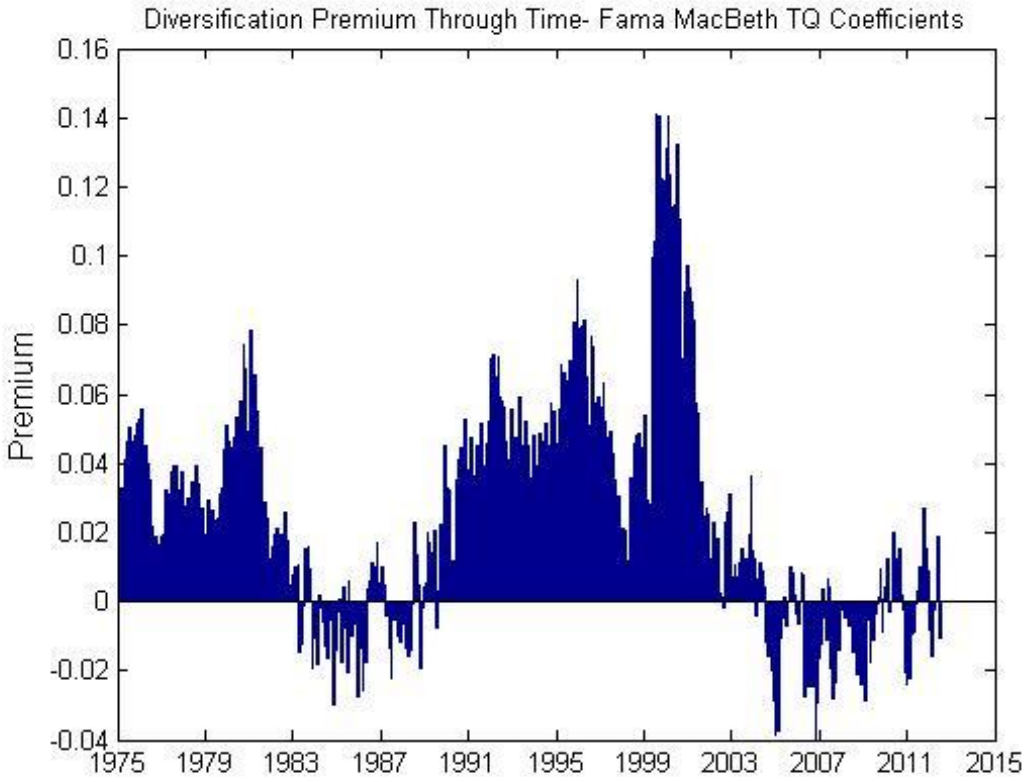


Figure 8E: Diversification Premium Through Time- Returns

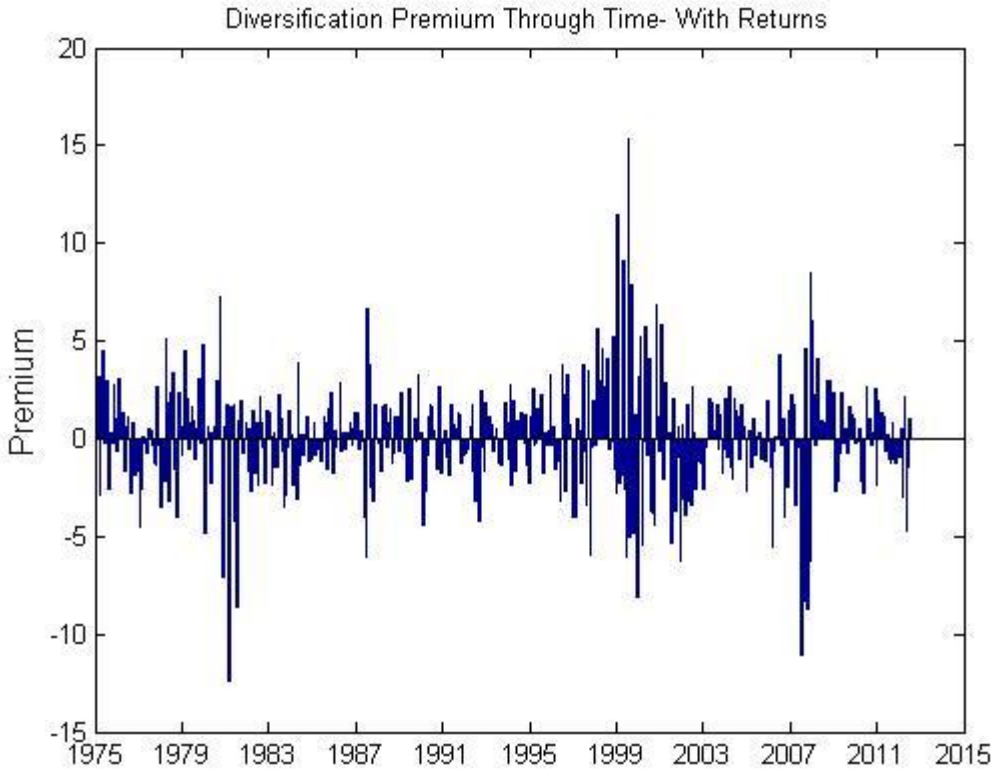


Figure 9A: Statistically Significant Premiums- Market to Book

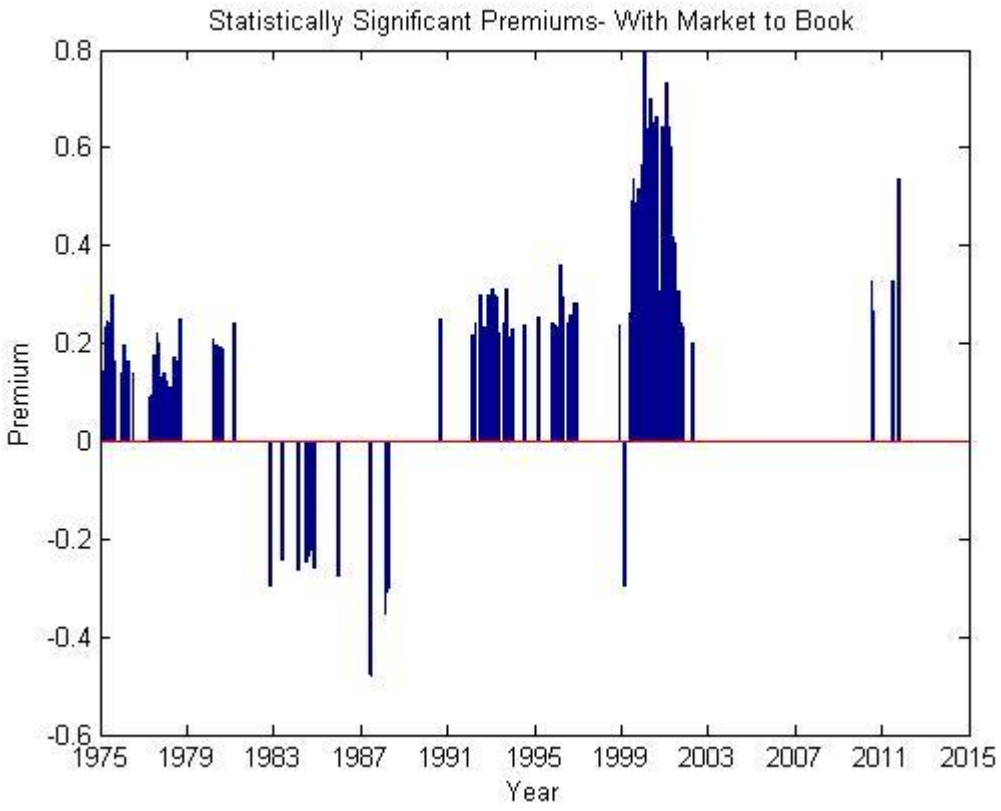
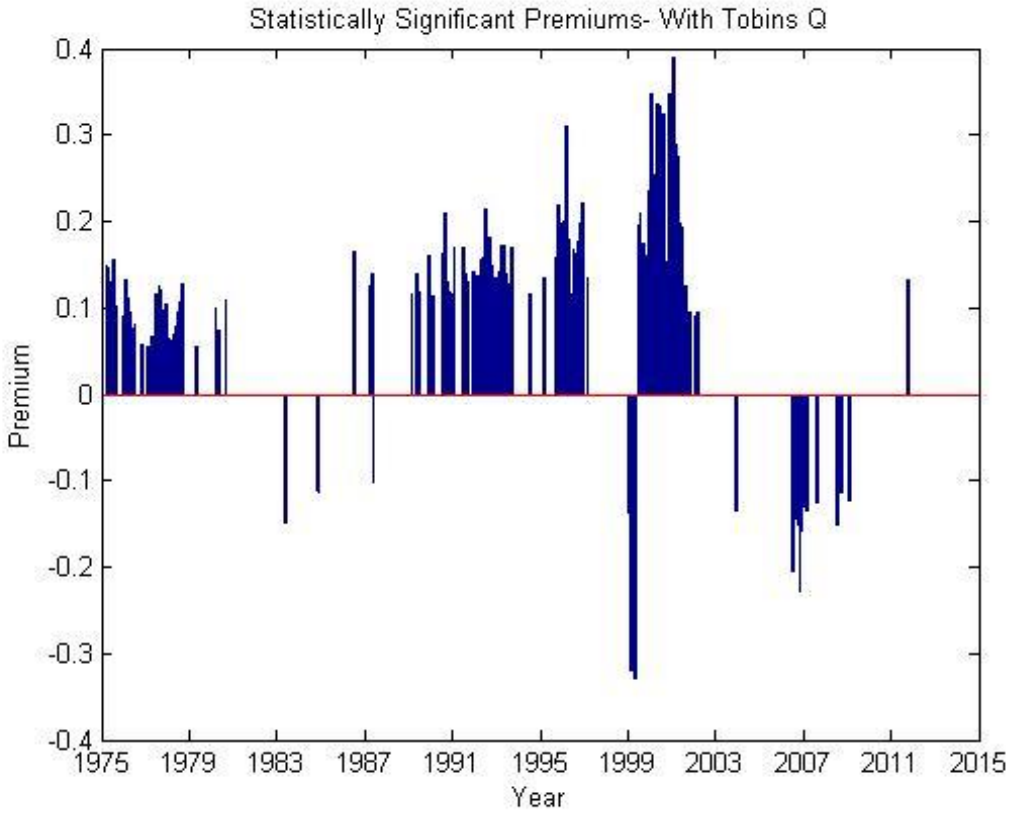


Figure 9A: Statistically Significant Premiums- Tobin's Q



Chapter 3: Classification of Firms into Industries Using Market Data
Classification of Firms into Industries Using Market Data

Introduction

I find that an iterative process reduces multi-collinearity creates a more even distribution of the number of firms in each industry. The iterative process I describe appears to lose accuracy as the process is expanded to more iterations, but increases accurately for firms in specific industries such as Coal, Agriculture, Oil, and Computers. This Results is likely due to more accuracy across these industries either because they have more unique returns or because they have less firms in their industry. The method appears to perform the most poorly for industries such as Steel, Finance, Aerospace, and Rubber. Additionally, I provide anecdotal evidence for the method from samples of firms and their empirically identified primary industries.

My First Essay develops a model that suggests that industry exposure can be empirically identified through regression techniques. The model shows that because market returns can be decomposed into weighted industry returns, a regression with the independent variables of industry returns should identify a firm's industry. The model relies on the concept that investors will alter the price appropriately when shocks affect an industry for every firm that is in that industry, conditional on the amount of weight that the industry has upon the performance of the firm.

If I accept the model proposed by my first essay then it is natural to question whether this technique can be take further to use those empirically identified industries to more accurately measure industry returns. I propose an iterative empirical method to create industry returns and suggest that when noise is small that industry returns can be refined by iterating returns over multiple processes and provide statistical analysis to understand whether this method is a reliable method.

In order to determine the validity of this method I evaluate it across several dimensions; how multi-collinear is each set of industry return returns, how evenly distributed are firms across industries, how accurately can this method identify primary industries, and how consistent are the industry classifications year over year?

Empirical Method

I identify a firm's industry or industries the industry that provides the highest R-squared from 48 univariate regressions of firm returns on each industry's returns individually. I require one year (252 trading days) of daily observations to be included in the sample.

Next, I re-create industry returns by equally weighting the firm returns of all firms that were identified in the first stage as that industry. I restrict a firm to only one industry and so we expect that there will be essentially the same number of firms included in the empirically identified industry as in the original. We equally weight returns to avoid large companies such as General Electric or Apple from dominating the industry returns that they are associated with.

The iterative process again regresses firm excess returns on the new industry returns and identifies significant industries. The new industry classification is again used to generate industry level returns. For the purposes of this paper we extend this forward to create five new sets of industry level returns, though I focus on comparisons of the initial returns and the first two iterations.

To empirically identifying a firm's primary industry we run univariate regressions of the firm's returns on each industries returns individually and label the industry that provides the highest R-square as the primary industry. We care about the primary industry for two reasons; it provides a clear empirical variable for testing, it logically should have the most economic meaning.

Data

I download Fama and French industry returns from the Kenneth French online data library. Daily returns come from the Center for Research in Security Prices (CRSP) and SIC codes come from the Compustat Names file as well as the CRSP Stock Names file. Our data spans January of 2005 to December of 2014. I include a random sample of 500 publicly traded stock on the NYSE, Nasdaq, and Amex that have at least one year of trading.

Methods and Results

At this point in the academic literature there appears to be no standard for identifying a firm's primary industry. The three most common methods are the Fama and French Industry Classification System (Fama and French (1997)) and Standard Industrial Codes (SIC). SIC were established in 1937 as a means to group together companies of

similar operations or creating similar products. In order to refine the SIC codes to more tailored industries Fama and French create industries by rearranging SIC codes that are often numerical distant but operationally close. In Table 6 I list the broad industry, or 2 digit, SIC Codes and which industries they represent. As can be seen they are quite broad and likely sub-optimal for industry adjustments. For a more detailed discussion on SIC codes see chapter 2. Compustat and CRSP are generally accepted alternative methods of identifying a firm's industry. It is worth noting that Kahle and Walking (1996) document that Compustat is superior to CRSP in general, and provide evidence through evaluation of firm operations and major SIC code. In this paper I concentrate mainly on Fama and French (1997) industry classifications but touch on Compustat and CRSP throughout the analysis.

Before analyzing whether an iterative process can assist with identifying more accurate industry returns I want to understand whether there are dramatic differences in applicability of the different industry classification systems. One applicable question when considering industry classification systems is which system allows us remove non-systematic risk most efficiently. Following researchers such as Campbell et al. (2001) and Xu and Malkiel (2003) I consider the amount of individual investments it takes to diversify a portfolio as information on which method most identifies industry.

To determine the optimal classification for diversification I evaluate the idiosyncratic volatility that remains after performing a market model on a portfolio diversified across industries identified by each system. The portfolios are made up of 48 stock selected from 48 different industries. Because Fama and French classify the

industries as 48 while CRSP and Compustat have closer to 100, not every industry will always be represented. This may bias Fama and French towards being more diversified given that I would assume that each individual industry will be more diversified holding the number of stock constant. Compustat and CRSP should be comparable though given their number of industries are similar. If an industry has less than 20 stocks with that industry listed as the major SIC code I do not select stock from it. I calculate implied idiosyncratic volatility as the summed absolute value of residuals. Because I have the same number of stocks in every portfolio all common methods of calculating idiosyncratic volatility should be functionally related and only the scalars should change. For each classification system I create 1000 portfolios and calculate the market model residuals and then implied volatility. The results below are the average of that volatility over all 1000 portfolios for each system.

The results show that the system with the highest idiosyncratic volatility is the CRSP industrial classification and the least is Compustat's. Not only is CRSP's higher but numerically it is over six times as high. Typically I cannot compare numbers of IV linearly, however because I do not square the residuals but rather take the absolute value a comparison is reasonable. More surprising is the dramatic difference across the standard deviations. The standard deviation of Compustat is less than one-third of that of CRSP. Perhaps the most interesting is that Fama and French are significantly higher than Compustat. Given that their portfolios include more stocks on average, taken together with the fact that Fama and French's system is simply re-organizing Compustat's, I might expect that it should have lower IV than Compustat. While this could be a relic of

randomness in the simulation it might be that I expect that they would have idiosyncratic volatility if each of their industries are accurate enough that market returns are orthogonal to the industry returns.

As I use industry returns to identify industry exposures, multi-collinearity is a concern. If an industry's returns are too closely correlated with another industry then empirical identification will be very difficult. For this reason I next evaluate the average correlation of industry returns across each iteration.

Table 8 reports the average of the correlation matrix of each of 48 industries. Initial is the average correlation of the 48 by 48 original Fama and French Industry returns. Iteration 1-5 are the average correlations of the re-created industry returns. As can be seen below, the mean correlation decreases increases dramatically from the initial data to each additional iteration but levels off afterwards at about 99%. The results strongly suggest that the process of iterating industry level returns produces results that change very little from additional iterations. It only appears to take a couple iterations to feel that the returns are reliable.

The next question I wish to address is the distribution of firms across industries. If most firms are located in a relatively low number of industries then it is likely that I will only be able to empirically identify those firms who are classified in uncommon industries because those firms' returns will play a larger role in the weight of the industry returns given that some industries would have more firms than others.

In Table 9 I report the percent of firms that are classified in each industry for the Initial Fama and French Data as well as the first and second iteration of re-classified

firms industries. Of interest is the dramatic differences in numbers for the Fama and French Data. Some industries such as Ships, Smokes, and Guns hover around .1% of all companies while others industries such as Banks, Business Services, and Finance range from 10-20% of all firms. Of particular interest is the smoothing of the cross-sectional distribution of percent. Industries that hold the largest portion of firms in the initial data decrease substantially their share on the subsequent iterations and those with small percent in the initial data increase their percent in the following iterations. As an example, 21.86% of firms are classified as finance firms in the Fama and French Industry Classification, but only 6.12% of firms are classified as finance by the second iteration.

Table 10 reports the percent of the time that the highest R-squared industry is also the industry reported based on the primary SIC code for CRSP, Compustat, and Fama and French Industries 48, or based on the previous iterations classification for the recreated industry returns. When comparing against established classification systems it is pretty clear cut, what percent of the time does the empirical result line up with the stated industry? When I use one iteration's results to derive the next iteration's industry level returns I have to be more careful with the interpretation because any error than may have occurred in a prior iteration will compound upon itself in the next iteration. I begin by identifying each firms industry by regressing firm returns on Fama and French 48 industry returns and placing a firm into each industry bin for which it is significant. Next I calculate equally weighted returns for each industry. Finally I regress the firm's returns on the new industry returns and determine whether they are significant.

In Table 10 I report the results from the test described above. I find that for both CRSP and Compustat the industry with the highest R-squared will be the industry reported by the data vendor approximately 40% of the time while Fama and French reported industry is the empirically identified industry only 22.7% of the time. I find that as I iterate the percent of the firms with an empirically identified industry being the same as their previous largest R-square industry shrinks. In the first iteration 26.3 percent of firms are classified as their primary SIC suggests they would be, however after five iterations only 19.5 percent of firms retain their primary SIC code. The result could result in more industries being significant and the highest R-square being an artifact of randomness.

The logical question to ask next is how persistent industry exposures are. Because I do not generally believe a firm is likely to move from one industry to another over a short time span I want a method that has a substantial amount of stability. In Table 11 Panel A I report the percent of the time that a firm remains in the same industry across all five iterations. Across all industries 73.9 percent of firms are consistently reclassified into the same industry. I break the percent down industry and find that the lowest percent is 50 percent and the highest is 100% of the time they maintain their classification. This is strong evidence that the iterative method is converging to a single set of industry returns and classifications. In Panel B evaluate whether the firms tend to remain in the same classification through time. For this test I restrict our sample to two years, 2006 and 2007, out of computation restrictions. I first identify the industry with the highest R-square on December 31, 2006 and on December 31, 2007 based on the first set of iterated returns. I

then calculate the percent of the time that they are the same industry the following year. Overall, the industry with the most unique explanatory power remains the same 100% of the time. The distribution is drastically different depending on which industry you are looking at though. Of those firms classified as the industry. The best explanation appears to be the number of firms in the industry. Because an industry with few firms will be affected more by any one set of returns this is not surprising. The lowest consistency is 50%. This is quite surprising given that we would expect many firms to change out of randomness. Additionally, the results I report could be unique to the time period tested across.

Table 12 reports how often the different iterations of identifying industries agree. I compare the classification initially reported by F&F, primary empirically identified industry, and first 5 iterations returns empirically identified industry.

The results are interesting, only about 26% of the time do the iterations line up with the Fama and French classifications, however across iterations the percent of the time they remain in the same category is very high, ranging from about 75% to about 99%. It is promising though that among the Fama and French industry comparisons, as the iterations progress I see that the results line up more frequently. Overall the results from Table 12 appear to provide evidence in favor of the empirical constructed returns.

Table 13 reports a sample of firms and their industry measured as the highest R-square industry in the regression of firm returns on each of the Fama and French 48 industry returns. I report those firms that are classified into the same industry across all five iterations.

The results are quite interesting. Out of the 95 companies reported it is difficult to find a firm classification that does not make sense. As we would expect Exxon Mobile is classified as oil and Gold Resource is classified as Gold. While many companies are unknown to the average researcher, cursory internet searches validate a great number of the firms. Moreover, it appears that certain industries find themselves with firms that have persistent classifications than others. Banks for example make up a large percent of the firms reported in Table 13 and we would expect them to have the least persistence because they also have many more firms in their industry than do other industries, which would encourage us to believe their returns would be less meaningful. Overall I believe that Table 13 provides evidence that an empirical identification of industries can provide benefits to researchers.

Conclusion

Overall I cannot conclude whether empirically constructed industry returns are superior to those constructed through SIC codes. Results from the tests of this study suggest that, across several spectrums, industry classifications work well. Firms tend to continue to be reclassified into the same industry through time, the returns of the industry appear to be consistent, and firms tend to be reclassified into the same industry across numerous iterations. Finally, from a visually inspection of real firms and their empirically identified industry, classifications make sense intuitively.

Table 6: SIC Code Definitions
This table reports the sic codes and their broad categorization. They can be commonly found from the internet at such locations as https://en.wikipedia.org/wiki/Standard_Industrial_Classification. Each range of codes refers to a different industry.

Range	Sector
0100-0999	Agriculture, Forestry and Fishing
1000-1499	Mining
1500-1799	Construction
1800-1999	not used
2000-3999	Manufacturing
4000-4999	Transportation, Communications, Electric, Gas and Sanitary service
5000-5199	Wholesale Trade
5200-5999	Retail Trade
6000-6799	Finance, Insurance and Real Estate
7000-8999	Services
9100-9729	Public Administration
9900-9999	Non Classifiable

Table 7: Implied Idiosyncratic Volatility By Classification
This Table reports the implied idiosyncratic volatility of three portfolios diversified across industries. I randomly select one stock from each of forty-eight separate industries and equally weight their returns to create an industry diversified portfolio return. Implied idiosyncratic volatility is calculated as the sum of the absolute value of the residuals of a market model regression. I generate one thousand portfolios and average the idiosyncratic volatility.

Segment Source	Mean Implied Idiosyncratic Volatility	Standard Deviation of Implied Idiosyncratic Volatility
Fama and French	0.91	1.05
Compustat	0.25	0.50
CRSP	1.68	1.53

Table 8: Correlations Across Iteration

This table reports the correlation matrix of industry returns across the original Fama and French 48 industry returns as well as each iteration of re-creation of industry returns through empirically determining constituents of industry portfolios. Iteration one regresses firm returns on industry returns identifies statistically significant industries at the 5% alpha, and creates a new portfolio of firms returns for those firms that appears significant then equally weights the firm returns of arrive at a new set of industry returns. Iteration two again regresses the firm returns on iterations ones industry returns and recreates industry level returns. Iterations 3, 4, and 5.

Industry Returns	Original	First	Second	Third	Fourth	Fifth
Original F&F	100.00%					
First Iteration	37.02%	100.00%				
Second Iteration	29.81%	94.06%	100.00%			
Third Iteration	29.48%	93.80%	99.80%	100.00%		
Fourth Iteration	29.40%	93.68%	99.70%	99.87%	100.00%	
Fifth Iteration	29.21%	93.61%	99.58%	99.81%	99.91%	100.00%

Table 9: Industry Summary Statistics
This table reports the number and percent of firms falling into each industry across the Fama and French Classification System, the first iteration of re-classification, and across the second iteration of reclassification. Classification is determined based on the highest R-Squared in a univariate regression of firm returns on industry returns.

Initial	Number	Percent	Iteration1	Number	Percent	Iteration2	Number	Percent
Aero	26	0.23%	Aero	88	0.79%	Aero	178	1.60%
Agric	24	0.22%	Agric	84	0.75%	Agric	127	1.14%
Autos	99	0.89%	Autos	123	1.10%	Autos	293	2.63%
Banks	1161	10.42%	Banks	418	3.75%	Banks	159	1.43%
Beer	31	0.28%	Beer	57	0.51%	Beer	145	1.30%
BldMt	105	0.94%	BldMt	150	1.35%	BldMt	144	1.29%
Books	45	0.40%	Books	54	0.48%	Books	107	0.96%
Boxes	19	0.17%	Boxes	222	1.99%	Boxes	111	1.00%
BusSv	1233	11.07%	BusSv	226	2.03%	BusSv	137	1.23%
Chems	138	1.24%	Chems	171	1.54%	Chems	239	2.15%
Chips	499	4.48%	Chips	826	7.41%	Chips	251	2.25%
Clths	90	0.81%	Clths	113	1.01%	Clths	242	2.17%
Cnstr	76	0.68%	Cnstr	279	2.50%	Cnstr	211	1.89%
Coal	25	0.22%	Coal	461	4.14%	Coal	503	4.52%
Comps	321	2.88%	Comps	317	2.85%	Comps	316	2.84%
Drugs	623	5.59%	Drugs	151	1.36%	Drugs	188	1.69%
ElcEq	109	0.98%	ElcEq	152	1.36%	ElcEq	150	1.35%
FabPr	21	0.19%	FabPr	597	5.36%	FabPr	234	2.10%
Fin	2435	21.86%	Fin	202	1.81%	Fin	682	6.12%
Food	101	0.91%	Food	54	0.48%	Food	127	1.14%
Fun	99	0.89%	Fun	132	1.18%	Fun	239	2.15%
Gold	73	0.66%	Gold	344	3.09%	Gold	415	3.73%
Guns	13	0.12%	Guns	185	1.66%	Guns	343	3.08%
Hlth	123	1.10%	Hlth	121	1.09%	Hlth	135	1.21%
Hshld	99	0.89%	Hshld	164	1.47%	Hshld	317	2.85%
Insur	301	2.70%	Insur	510	4.58%	Insur	738	6.62%
LabEq	139	1.25%	LabEq	66	0.59%	LabEq	87	0.78%
Mach	204	1.83%	Mach	167	1.50%	Mach	103	0.92%
Meals	122	1.10%	Meals	507	4.55%	Meals	578	5.19%
MedEq	287	2.58%	MedEq	108	0.97%	MedEq	164	1.47%
Mines	76	0.68%	Mines	111	1.00%	Mines	282	2.53%
Oil	410	3.68%	Oil	362	3.25%	Oil	547	4.91%
Other	48	0.43%	Paper	76	0.68%	Paper	161	1.45%
Paper	83	0.75%	PerSv	819	7.35%	PerSv	382	3.43%
PerSv	89	0.80%	RlEst	394	3.54%	RlEst	62	0.56%
RlEst	89	0.80%	Rtail	71	0.64%	Rtail	151	1.36%
Rtail	330	2.96%	Rubbr	128	1.15%	Rubbr	130	1.17%
Rubbr	55	0.49%	Ships	62	0.56%	Ships	144	1.29%
Ships	12	0.11%	Smoke	88	0.79%	Smoke	153	1.37%
Smoke	10	0.09%	Soda	52	0.47%	Soda	118	1.06%
Soda	20	0.18%	Steel	142	1.27%	Steel	143	1.28%
Steel	107	0.96%	Telcm	121	1.09%	Telcm	210	1.89%
Telcm	349	3.13%	Toys	226	2.03%	Toys	118	1.06%
Toys	56	0.50%	Trans	446	4.00%	Trans	105	0.94%
Trans	226	2.03%	Txtls	507	4.55%	Txtls	184	1.65%

Txtls	68	0.61%	Util	361	3.24%	Util	427	3.83%
Util	228	2.05%	Whlsl	125	1.12%	Whlsl	160	1.44%
Whlsl	243	2.18%						

Table 10: Accuracy of Classification System Using Highest R-Squared

This table reports the percent of the time that the highest R-squared industry is the same industry reported as its major industry according to each classification system. CRSP and Compustat classifications are two digit SIC codes as reported by each industry respectively. Iterated major SIC codes are from the highest R-squared industry from the previous calculation of industry level returns.

Source	Percent
CRSP	40.6%
Compustat	41.8%
F&F 48	22.7%
1 Iteration	26.3 %
2	22.3%
3	21.1%
4	20.7%
5	19.9%
6	19.5%

Table 11: Consistency of Industries
Using the years I calculate each firm's highest R-squared industry and compute the percent of the time that, in Panel A, the industry is consistent across all five iterations. In Panel B the percent of the time the industry remains constant over two consecutive years (2006-2007). Maximum is the industry consecutively appears as the highest R-squared the most frequently. Minimum is the industry that is the least consecutive.

Panel A- Across		
Industry	Frequency	Mean
All	502	73.9%
Aero	6	66.7%
Agric	5	100.0%
Autos	4	100.0%
Banks	40	90.0%
Beer	7	71.4%
BldMt	22	95.5%
Books	6	100.0%
Boxes	7	85.7%
BusSv	17	82.4%
Chems	14	35.7%
Chips	13	100.0%
Clths	6	66.7%
Cnstr	3	66.7%
Coal	5	100.0%
Comps	8	62.5%
Drugs	8	100.0%
ElcEq	7	71.4%
FabPr	5	80.0%
Fin	18	50.0%
Food	6	83.3%
Fun	11	90.9%
Gold	7	85.7%
Guns	3	100.0%
Hlth	8	100.0%
Hshld	2	100.0%
Insur	18	44.4%
LabEq	25	56.0%
Mach	30	80.0%
Meals	1	100.0%
MedEq	10	90.0%
Mines	9	44.4%
Oil	16	50.0%
Other	7	85.7%
Paper	4	50.0%
PerSv	7	85.7%
RlEst	15	73.3%
Rtail	4	75.0%
Rubbr	12	83.3%
Ships	4	75.0%
Smoke	4	100.0%
Soda	10	100.0%
Steel	13	61.5%
Telcm	9	33.3%
Toys	5	100.0%
Trans	5	100.0%
Txtls	2	100.0%
Util	12	33.3%
Whlsl	42	61.9%

Panel B- Across Time	
Total	78.9%
Max	100.0%
Min	50.0%
Aero	100.0%
Agric	75.0%
Autos	100.0%
Banks	81.0%
Beer	100.0%
BldMt	63.6%
Books	80.0%
Boxes	100.0%
BusSv	77.8%
Chems	100.0%
Chips	88.9%
Clths	100.0%
Cnstr	100.0%
Coal	100.0%
Comps	66.7%
Drugs	75.0%
ElcEq	50.0%
FabPr	100.0%
Fin	66.7%
Food	50.0%
Fun	62.5%
Gold	100.0%
Hlth	100.0%
Hshld	100.0%
Insur	80.0%
LabEq	80.0%
Mach	70.0%
MedEq	50.0%
Mines	50.0%
Oil	85.7%
Other	100.0%
Paper	100.0%
PerSv	50.0%
RlEst	71.4%
Rtail	100.0%
Rubbr	100.0%
Ships	50.0%
Smoke	66.7%
Soda	100.0%
Steel	100.0%
Telcm	100.0%
Toys	100.0%
Trans	66.7%
Txtls	100.0%
Util	75.0%
Whlsl	80.0%

Table 12: Comparability of Identification Method
I use all stock listed in the Compustat Names file from 2000-2010 and compare the percent of the time that each method agrees on the primary SIC code. Primary SIC codes Initial come from those defined by the Names file in Compustat or identified by Fama and French Industry classification. Regressed 1 identifies the industry that provides the highest R-square of each industry on each firm's returns. Iterated 1 is recreates industry returns using the regressed 1 identification of industry and again calculates the industry with the highest R-Square.

System 1	System 2	Percent of Time Agree
FF 48	First Iteration	26.30%
First Iteration	Second Iteration	84.30%
First Iteration	Third Iteration	78.30%
First Iteration	Fourth Iteration	76.10%
First Iteration	Fifth Iteration	75.10%
First Iteration	Sixth iteration	73.90%
Second Iteration	Third Iteration	91.40%
Second Iteration	Fourth Iteration	89.80%
Second Iteration	Fifth Iteration	87.10%
Second Iteration	Sixth iteration	86.50%
Third Iteration	Fourth Iteration	96.80%
Third Iteration	Fifth Iteration	95.20%
Third Iteration	Sixth iteration	93.20%
Fourth Iteration	Fifth Iteration	97.00%
Fourth Iteration	Sixth iteration	96.20%
Fifth Iteration	Sixth iteration	97.80%

Table 13: Sample of Companies From Each Computed Industry
This table reports firms that have the same industry based on all iterations. To be in an industry you that industry must provide the highest R-squared of any of the 48 industries in univariate of firm returns on initial industry returns.

Company	Industry	Company	Industry
ASM INTERNATIONAL NV	Mach	LABORATORY CP OF AMER HLDGS	Hlth
AIR T INC	Trans	CHEMICAL FINANCIAL CORP	Banks
CATO CORP -CL A	Rtail	TCF FINANCIAL CORP	Banks
DANA HOLDING CORP	Autos	PEOPLE'S UNITED FINL INC	Banks
EXXON MOBIL CORP	Oil	HANGER INC	Hlth
FIRST CITIZENS BANCSH -CL A	Banks	BAY VIEW CAPITAL CORP	Banks
GENERAL MOTORS CO	Autos	CENTRAL PACIFIC FINANCIAL CP	Banks
MOLYCORP INC	Steel	FIRST BANCORP/NC	Banks
AEP INDUSTRIES INC	Rubbr	FIRST CHARTER CORP	Banks
SUN MICROSYSTEMS INC	Comps	CITY HOLDING CO	Banks
CANYON RESOURCES CORP	Gold	COMMUNITY BANKS INC	Banks
MICROSOFT CORP	BusSv	FIRST BANCORP P R	Banks
ORACLE CORP	BusSv	INDEPENDENT BANK CORP/MA	Banks
OSI PHARMACEUTICALS INC	Drugs	HARLEYSVILLE NATL CORP/PA	Banks
CYPRESS SEMICONDUCTOR CORP	Chips	OFG BANCORP	Banks
LINEAR TECHNOLOGY CORP	Chips	PROVIDENT BANKSHARES CORP	Banks
GENZYME CORP	Drugs	OLD NATIONAL BANCORP	Banks
SIGMA DESIGNS INC	Chips	SVB FINANCIAL GROUP	Banks
WERNER ENTERPRISES INC	Trans	SANTANDER HOLDINGS USA INC	Banks
CYTOGEN CORP	Drugs	WSFS FINANCIAL CORP	Banks
AMERICAN WOODMARK CORP	BldMt	REPUBLIC BANCORP INC	Banks
SKYWEST INC	Trans	WEBSTER FINANCIAL CORP	Banks
WATTS WATER TECHNOLOGIES INC	BldMt	WESBANCO INC	Banks
BIO REFERENCE LABS	Hlth	UNITED BANKSHARES INC/WV	Banks
HEALTHSOUTH CORP	Hlth	SOUTH FINANCIAL GROUP INC	Banks
P.A.M. TRANSPORTATION SVCS	Trans	REPUBLIC FIRST BANCORP INC	Banks
FISERV INC	BusSv	F N B CORP/FL	Banks
HEARTLAND EXPRESS INC	Trans	COMMUNITY TRUST BANCORP INC	Banks
PLEXUS CORP	Chips	STERLING FINANCIAL CORP/WA	Banks
PHOTRONICS INC	Chips	WASHINGTON TR BANCORP INC	Banks
SANDERSON FARMS INC	Food	COMSTOCK RESOURCES INC	Oil
BARR PHARMACEUTICALS INC	Drugs	SUN HEALTHCARE GROUP INC	Hlth
CALGON CARBON CORP	Chems	GREAT LAKES BANCORP INC	Banks
BERRY PETROLEUM -CL A	Oil	RHINO RESOURCE PARTNERS LP	Coal
CELGENE CORP	Drugs	GOLD RESOURCE CORP	Gold
CESCA THERAPEUTICS INC	LabEq	PARK STERLING CORP	Banks
CAM COMM SOLUTIONS INC	Comps	IMRIS INC	MedEq
CYBEROPTICS CORP	LabEq	ACCESS MIDSTREAM PARTNERS LP	Oil
FASTENAL CO	Rtail	WESTMORELAND RES PARTNERS LP	Coal
MAXIM INTEGRATED PRODUCTS	Chips	QLIK TECHNOLOGIES INC	BusSv
RESPIRONICS INC	MedEq	AEROFLEX HOLDING CORP	Chips
MGM RESORTS INTERNATIONAL	Fun	INTRALINKS HOLDINGS INC	BusSv
WPP PLC	BusSv	NXP SEMICONDUCTORS NV	Chips
KCS ENERGY INC	Oil	AUTONAVI HLDG LTD	BusSv
BMC SOFTWARE INC	BusSv	INPHI CORP	Chips
GENLYTE GROUP INC	ElcEq	IAO KUN GROUP HOLDING CO LTD	Fun
PRIDE INTERNATIONAL INC	Oil	ELSTER GROUP SE -ADR	LabEq
LINDSAY CORP	Mach		

References

Agrawal, A., Jaffe, J.F., Mandelker, G.N., 1992. The post-merger performance of acquiring firms: a re-examination of an anomaly. Journal of Finance 47, 1605 – 1621.

Akbulut, Mehmet E., and John G. Matsusaka, 2010, 50+ years of diversification announcements. Financial review 45.2, 231-262.

Anderson, R.C., Bates, T.W., Bizjak, J.M., Lemmon, M.L., 2000, Corporate Governance and Firm Diversification, *Financial Management*, v29, 5-22.

Ahn, Dong-Hyun, Jennifer Conrad, and Robert F. Dittmar. "Basis assets."Review of Financial Studies 22.12 (2009), 5133-5174.

Baker, M. and J. Wurgler, 2006, Investor sentiment and the cross-section of stock returns, Journal of Finance,61, 1645-1680.

Berger, P.G., Ofek, E., 1995. Diversification's effect on firm value, Journal of Financial Economics 37, 39-65.

Berry, C.H., 1971. Corporate growth and diversification, Journal of Law and Economics, 14, 371 – 383.

Botosan, Christine A., and Mary Stanford, 2005, Managers' motives to withhold segment disclosures and the effect of SFAS No. 131 on analysts' information environment, The Accounting Review 80.3, 751-772.

Bradley, M., Desai, A., Kim, E.H., 1988, Synergistic gains from corporate acquisitions and their division between the stockholders of target and acquiring firms, Journal of Financial Economics 21 (1), 3-40.

Campa, J.M. and S. Kedia, 2002. Explaining the diversification discount, Journal of Finance 57,
1731–1762.

Campbell, J, Lettau, M Malkiel, B, Xu, Y, 2001, Have Stocks Become More Volatile? An Empirical Exploration of Idiosyncratic Risk. The Journal of Finance 56, 1540.

Carhart, M., 1997, On Persistence in Mutual Fund Performance, Journal of Finance, 52, 57–82.

Chang, C., Yu, X., 2004, The Informational Benefits and Costs in Conglomerate Mergers, The Journal of Business, 77, 45-74.

Chevalier, J., 2000, Why Do Firms Undertake Diversifying Mergers? An Investigation of the Investment Policies of Merging Firms. Unpublished manuscript, University of Chicago.

Clark, Richard N., 1989, "SICs as Delineators of Economic Markets," *Journal of Business* 62, 17-31.

Davis, R., Duhaime, I.M., 1992, Diversification, industry analysis and vertical integration: new perspectives and measurement, Strategic Management Journal, 511 – 524.

Denis, D.J., Denis, D.K., Sarin, A., 1997, Agency problems, equity ownership and corporate diversification, Journal of Finance 52, 135 – 160.

Desai, H., Jain, P.C., 1999, Firm performance and focus: long-run stock market performance following spinoffs, Journal of Financial Economics 54, 75 – 101.
Fama, E, French, K, 1997, Industry costs of equity. Journal of Financial Economics 43 (2), 153—194.
Fama. Eugene F. and James MacBeth, 1973, Risk, return and equilibrium: Empirical tests, Journal of Political Economy 81, 607-636.
Graham, John, Michael Lemmon, and Jack Wolf, 2002, "Does Corporate Diversification Destroy Value?", Journal of Finance, 57, 695-720.
Hadlock, C., Ryngaert, M., Thomas, S., 2001, Corporate structure and equity offerings: are there benefits
to diversification?, Journal of Business, 54, 613–635.
Hyland, David, and David Diltz, 2002, "Why firms diversify? : An empirical examination",
Financial Management 31, 51-82.
Jacquemin, A., Berry, C., 1979, Entropy measure of diversification and corporate growth. Journal of Industrial Economics 27, 359 – 369.
Black, Fischer, Michael C. Jensen, and Myron Scholes, 1972, The Capital Ssset Pricing Model: Some empirical tests, in Studies in the Theory of Capital Markets, pp. 79-121.
Jensen, Michael C., 1986, Agency costs of free cash flow. Corporate finance and takeovers. American Economic Review 76, 323-329.
Jensen, Michael C. and William H. Meckling, 1976, Theory of the firm: Managerial behavior, agency costs and ownership structure, Journal of Financial Economics 3, 305-360.
John, K. and E. Ofek, 1995, Asset sales and increase in focus, Journal of Financial Economics 37, 105-126.
Kahle, Kathleen M., and Ralph A. Walkling. "The impact of industry classifications on financial research." Journal of financial and quantitative analysis 31.03 (1996): 309-335.
Kraus, Alan, and Robert H. Litzenberger, 1973, "A state-preference model of optimal financial leverage", The journal of finance 28.4, 911-922.
Lamont, O., 1997, Cash flow and investment: evidence from internal capital markets, Journal of Finance 52, 83 -109
Lamont, Owen A., and Christopher Polk, 2002, "Does diversification destroy value? Evidence from the industry shocks", Journal of Financial Economics 63.1, 51-77.
Lang, L.H.P. and R.M. Stulz, 1994, Tobin's q, corporate diversification, and firm performance, Journal of
Political Economy 102, 1248–1280.
Lang, Larry, Eli Ofek, and RenéM Stulz. "Leverage, investment, and firm growth." Journal of financial Economics 40.1 (1996): 3-29.
Lewellen, W.G., 1971. A pure financial rationale for the conglomerate merger, Journal of Finance 26,
521–537.

Lichtenberg, F.R., 1991, The managerial response to regulation of financial reporting for segments of a business enterprise, Journal of Regulatory Economics 3, 241 – 249.

Maksimovic V. and G. Phillips, 2002, "Do Conglomerate Firms Allocate Resources Inefficiently? Evidence from Plant-Level Data, Journal of Finance, 721-767.

May, D.O., 1995, Do managerial motives influence firm risk reduction strategies. Journal of Finance 50, 1291 -1308

McVey, J.S., 1972, The industrial diversification of multi-establishment manufacturing firms: a developmental study. Canadian Statistical Review 47, 112 – 117.

Megginson, W.L., Morgan, A., Nail, L., 2004, Changes in Corporate Focus, Ownership Structure, and Long-Run Merger Returns, Journal of Banking and Finance, 28, 523-552.

Modigliani, F. and M. Miller, 1958, The Cost of Capital, Corporation Finance, and the Theory of Investment, American Economic Review, 48, 261-297.

Montgomery, C.A., 1994, Corporate diversification, Journal of Economic Perspectives 8 (3), 163 – 178.

Wernerfelt, B., Montgomery, C.A., 1988, Tobin's q and the importance of focus in firm performance. American Economic Review 78, 246 – 250.

Morck, R., A. Shleifer, and R.W. Vishny, 1990, Do managerial objectives drive bad acquisitions? Journal of Finance 45, 31–48.

Myers, S.C. and N.S. Majluf, 1984, Corporate financing and investment decisions when firms have information that investors do not have, Journal of Financial Economics 13, 187–221.

Himmelberg, C.P., G.R. Hubbard, and D. Palia, 1999, Understanding the determinants of managerial ownership and the link between ownership and performance, Journal of Financial Economics, 53, 353-85.

Rajan, R., Servaes, H., Zingales, L., 2000.,The cost of diversity: the diversification discount and inefficient Investment, Journal of Finance, 55 (1), 35 – 80.

Rumelt, R.P., 1974, Strategy, Structure and Economic Performance. Division of Research, Harvard University Press, Cambridge, MA.

Martin, John D., and Akin Sayrak, 2003, Corporate diversification and shareholder value, Journal of Corporate Finance 9, 37–57.

Scharfstein, David S., 1998, The dark side of internal capital markets II: Evidence from diversified conglomerates, Working paper no. 6352, National Bureau of Economic Research. Servaes, H., 1996. The value of diversification during the conglomerate merger wave. Journal of Finance 51, 1201 – 1226.

Shin, H.-H., Stulz, R.M., 1998, Are internal capital markets efficient. The Quarterly Journal of Economics 113, 531 – 552.

Shleifer, Andrei, and Robert W. Vishny, 1992, Liquidation values and debt capacity: A market equilibrium approach, The Journal of Finance, 47.4, 1343-1366.

Shleifer, A. and R.W. Vishny, 2003, Stock market driven acquisitions, Journal of Financial Economics 30, 295–311.

Tibshirani, Robert. "Regression shrinkage and selection via the lasso." Journal of the Royal Statistical Society. Series B (Methodological) (1996): 267-288.

Villalonga B., 2004,Diversification Discount or Premium? New Evidence from BITS Establishment Level Data, Journal of Finance, 59, 479-502.

Wernerfelt, B., Montgomery, C.A., 1988, Tobin's q and the importance of focus in firm performance. American Economic Review, 78, 246 – 250.

Whited, T., 2001, Is it inefficient investment that causes the diversification discount. Journal of Finance, 56, 1667-1691.

Williamson, O. E., 1970, Corporate Control and Business Behavior: An Inquiry into the Effects of Organizational Form on Enterprise Behavior, Englewood Cliffs, NJ: Prentice-Hall.

Wrigley, 1970, Divisional Autonomy and Diversification. Unpublished doctoral dissertation, Graduate School of Business Administration, Harvard Business School, Boston, MA.

Xu, Y., & Malkiel, B. G.. (2003). Investigating the Behavior of Idiosyncratic Volatility. The Journal of Business, 76(4), 613–645.

Appendix 1

Figure A1: Diversification Premium Mapped Against Credit Spread MB

Plots the market to book premium against the spread between AAA and BBB corporate bonds yields for premium measured as the difference across Market to Book Portfolios.

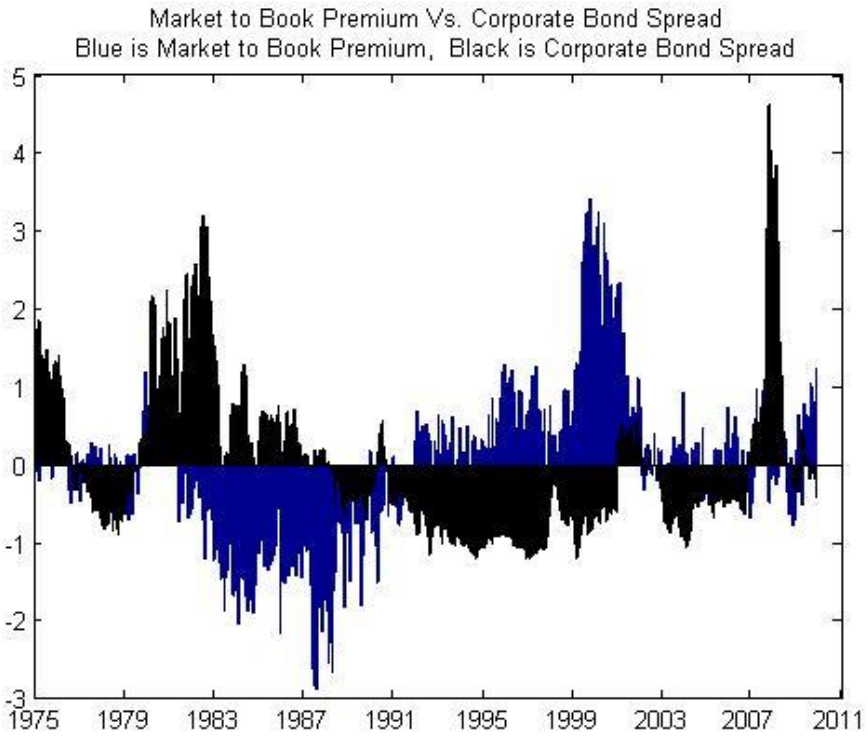


Figure A2: Diversification Premium Mapped Against Credit Spread MB

Plots the market to book premium against the spread between AAA and BBB corporate bonds yields for premium measured as the difference across Market to Book Portfolios.

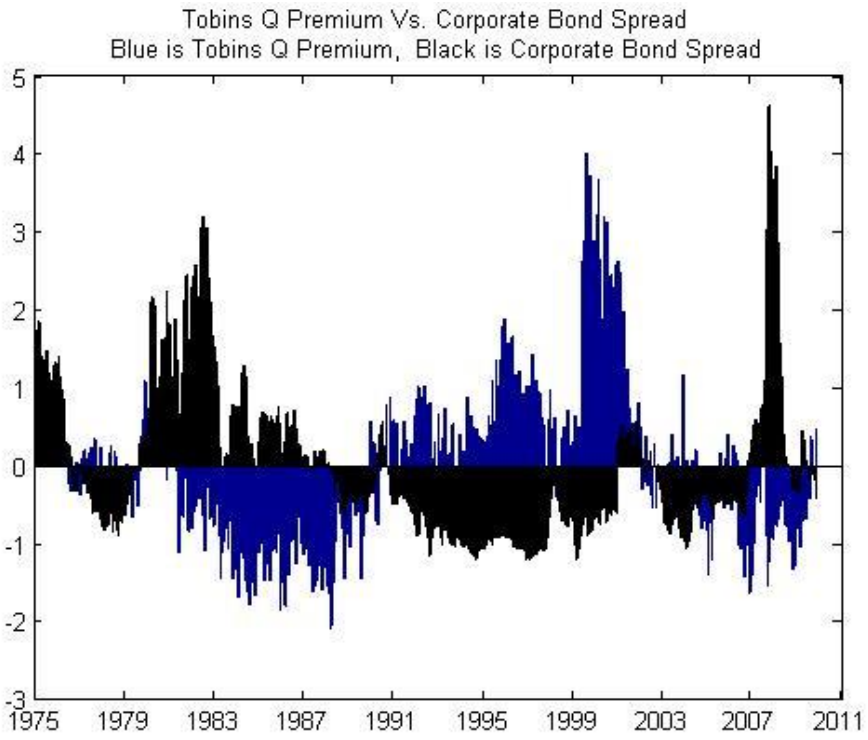


Table A1: The effect of Macro Variables on Diversification Through Time

Panel A includes the independent variables number of IPOs issued during the month, the first principle component of investor sentiment measures, the amount of debt issuances, the amount of and institutional investment. Panel 3 includes the independent variables market capitalization, one month change in market capitalization, turn over, and the equally weighted return for the month. Panel 4 includes the independent variables from all previous panels. Panel E uses the independent variables average short term debt, 10 year treasury bond yield, the one month change in the 10 year treasury bond yield, and the interaction between the 10 year treasury bond yield and the average short term debt Model 1 of Panel E evaluates the relationship post 2005, Model 2 includes only data before 2005, Model 3 includes only data before 2004, Model 4 includes data only before 2003, and Model 5 includes only data before 2002. All Panels on the left have the dependent variable of average premium or discount measured by market to book, all panels on the right have the dependent variable of average premium or discount measured with Tobin's Q. T-statistics are reported in parenthesis

Panel A: Sentiment Effects – Market-to-book				Sentiment Effects –Tobin's Q			
	Model	Model 2	Model 3		Model 1	Model 2	Model 3
Intercept	0.252	0.113	0.150	Intercept	0.131	0.100	0.116
	(12.860)	(6.780)	(5.960)		(13.840)	(12.110)	(9.240)
NIPO	-0.002		-0.001	NIPO	-0.001		0.000
	(-4.070)		(-1.990)		(-1.980)		(-1.610)
Sentiment	0.029		0.004	Sentiment	0.013		0.010
	(1.910)		(0.290)		(1.770)		(1.380)
SD		0.215	0.241	SD		0.493	0.495
		(0.620)	(0.690)			(2.850)	(2.860)
Institutional		0.009	0.008	Institutional		-0.001	-0.002
		(3.960)	(3.440)			(-1.140)	(-1.510)
R2	0.037	0.108	0.118	R2	0.012	0.026	0.033
Adjusted	0.033	0.104	0.109	Adjusted R2	0.007	0.021	0.024

Panel B				Market Effects - Tobin's Q			
	Model 1	Model 2	Model 3		Model 1	Model 2	Model 3
Intercept	0.049	0.095	0.264	Intercept	0.080	0.122	0.193
	(2.790)	(3.420)	(10.810)		(8.850)	(9.090)	(15.350)
Size	0.018		0.047	Size	0.045		0.020
	(10.640)		(16.200)		(5.170)		(13.280)
ΔSize	0.501		-0.841	ΔSize	0.234		-0.368
	(2.030)		(-2.040)		(1.840)		(-1.730)
Turn Over		0.132	-0.660	Turn Over		-0.012	-0.347
		(3.610)	(-11.610)			(-0.700)	(-11.840)
Returns		0.551	0.905	Returns		0.224	0.380
		(2.500)	(2.720)			(2.110)	(2.220)
R2	0.212	0.039	0.406	R2	0.064	0.012	0.301
Adjusted R2	0.209	0.035	0.401	Adjusted R2	0.060	0.008	0.295

Panel C: Full Model – Market-to-book					Full Model –Tobin's Q				
	Model 1	Model 2	Model 3	Model 4		Model 1	Model 2	Model 3	Model 4
Intercept	-2.402	0.904	0.78	0.354		-1.474	0.516	0.457	0.198
	(-2.320)	(10.390)	(8.320)	(3.000)		(-2.600)	(12.23))	(10.05))	(3.58))
ΔTnote	0.14	0.006	0.012	0.024		0.075	-0.002	0.000	0.007
	(-2.370)	(0.190)	(0.380)	(0.800)		(2.32))	(-0.140)	(-0.010)	(0.520)
Tnote	0.48	-0.167	-0.152	-0.098		0.324	-0.08	-0.073	-0.04
	(1.930)	(-13.340)	(-11.550)	(-6.120)		(-2.370)	(-13.250)	(-11.470)	(-5.330)
DebtSTxTnote	-0.001	0.001	0.001	0.000		-0.000	0.000	0.000	0.000
	(-1.870)	(13.010)	(11.230)	(5.390)		(-2.200)	(12.350)	(10.600)	(4.070)
DebtST	0.284	-.273	-0.219	-0.023		0.002	-0.137	-0.111	0.009
	(2.50)	(-9.030)	(-6.520)	(-0.480)		(2.560)	(-9.340)	(-6.820)	(0.390)
R2	0.298	0.538	0.559	0.594	R2	0.234	0.485	0.508	0.576
Adjusted R2	0.249	0.535	0.554	0.589	Adjusted R2	0.18	0.479	0.502	0.57

Vita

Michael Jarrod Gibbs was born in Morehead, Kentucky in 1984 and grew up in Mount Sterling, Kentucky. He completed his undergraduate degree in 2008, majoring in finance. Prior to joining the Ph.D. program at the University of Missouri, Michael split his time between taking graduate classes in mathematics/statistics and working as an electrician.