CIMAT

Centro de Investigación en Matemáticas, A.C.

"Comparison between traditional risk measurement methods and machine learning methods in Mexican equity investment funds"

TESIS

Que para obtener el grado de

Maestro en Ingeniería de Software

P r e s e n t a

Grecia María Cortés Espinosa

Co Directores de Tesis

Dra. Natalia García Colin

MID José Guadalupe Hernández Reveles

Zacatecas, Zacatecas, 18 de enero de 2017

Table of Contents

Abstract

Among the investment instruments available to the general public, mutual funds offer the advantage of being a diversified and managed investment instrument that, on average, produces profits above those of bank notes, a favorite investment instrument among the Mexican population. These days, online brokers offer a low cost alternative to invest in Mexican mutual funds, but without the guidance of a financial advisor and at the investor's own risk. Any investment in a financial instrument (stocks, bonds, metals, etc.) relies on the investor's previous knowledge on the risks and liabilities of making such investments, or on paying for the services of a professional to advise them on the adequate investments instruments to accomplish their financial goals.

Different automated learning techniques and algorithms have been used in the analysis of data to discover unknown patterns of behavior in different fields, such as: engineering, medicine, biology, economy and finance. In finance, the daily price of the titles of investment instruments, such as mutual funds, are studied as time series of data. Applying known machine learning algorithms to the daily prices of Mexican mutual funds can uncover previously ignored patterns of behavior, which could help in the selection of mutual funds for new investors.

Keywords: finance, mutual funds, equity funds, Mexican mutual funds, Mexican equity funds, machine learning, clustering analysis, time series analysis, hierarchical clustering analysis.

Acknowledgments

To *Becas CONACyT*, for their scholarship grant.

To my teachers, Dr. Natalia García Colin, from ***INFOTEC***, and MID José Guadalupe Hernández Reveles, from ***CIMAT Unit Zacatecas***, for their support and patience.

To my family.

1. Introduction

Nowadays, the Mexican population has easier access to investment opportunities provided by online brokers and investment fund managers.

1.1. Motivation

For the general population, the evaluation and selection of mutual funds in which to invest in it's a complicated and daunting task:

- At this moment, public tools that allow the comparison and selection of the available mutual funds at the **Mexican Stock Exchange** (**BMV**, *Bolsa Mexicana de Valores*) for the general public do not exist.

- Most of the information that is required in order to find, evaluate and select an investment instrument it's scattered in three places:
  1. The list of available funds for *non-qualified* investors is only available at the **National Banking and Stock Commission** (**CNBV**, *Comisión Nacional Bancaria y de Valores*) website.

  2. The prospect and information sheets of each mutual fund is located at its respective investment manager website.

  3. The unit price series are available at financial news portals, such as Google Finance and Yahoo! Finance, and, sometimes, at each fund's investment manager website. The official source for the historical price series of all investment instruments that operate at the BMV can be purchased at the BMV website, but at a price of MXN 41,000 plus taxes per month of information (in August, 2015).

- Although, independent financial analysts, like Morningstar, offer public evaluation summaries and profiles on mutual funds and other investment instruments, this information is useful only to people with previous financial and statistical knowledge. Even some professionals with a background and experience in finance can find this task complex and confusing.

The effort and time required to collect, correlate and analyze the information of all available investment funds independently is enormous. Given this situation, most potential investors back away from the idea of placing their savings in mutual funds or any other investment instrument that is not endorsed by their own bank.

One of the consequences of a lack of investors diversity in Mexican funds, specifically in equity funds investing in the Mexican stock market, is a loss of income for investment management firms. For small and medium sized Mexican companies, this situation represents a lost source of capital.

1.2. Background

Besides the issues of information dispersion and gathering, and the need of a background in finance from potential investors, the financial information available about an investment instrument it's not easy to compare.

In order to measure the exposure risk of a fund, most professionals and some academics use two risk measures: the standard deviation and the Sharpe ratio. The most popular statistics to assess a fund's performance relative to a benchmark are the alpha, beta and r squared (coefficient of determination). Together, these five measurements are known as the **Modern Portfolio Theory** (MPT) statistics. However, as will be argued in detail, it's complicated to analyze the combination of these measures to provide an accurate evaluation of a fund's behavior and returns.

A common practice in the financial and investor sector, it's to include the traditional risk and performance measures from the MPT as part of the statistical analysis in their periodic mutual fund reports.

Investment and financial services companies have developed their own models to assess and measure the risk and performance of mutual funds. But such models are considered their intellectual property and are not available to the public.

Among professional investors, the Morningstar analysis is one of the most used references in the world. Its funds' ratings are a quantitative assessment of a fund's past

performance, in terms of return and risk, relative to mutual funds within its category. In this analysis ("*The Morningstar Approach to Mutual Fund Analysis—Part I*", 2010, p. 170-171), the overall rating for a fund is a weighted average of three period-specific ratings: the 3-year rating (based on the past 36 months), the 5-year rating (past 60 months), and the 10-year rating (past 120 months). Because the Morningstar Rating is evaluated every month, it's always changing. It's not recommended to invest in a fund, or any security, based only on its rating, as the fund's category, investment strategy, costs, and management style, must also be considered ("*The Morningstar Approach to Mutual Fund Analysis—Part I*", 2010, p. 172-173). This attests the enormous complexity of fund selection for a non-qualified investor.

1.3. Objectives

Although the lack of public tools for mutual funds financial advising can prompt the development of such tool or application, the effort required to solve the stated issues is unknown and not easy to measure. The construction of a single tool for automated financial advising can turn out to be a misleading solution. Also, the diversity and, sometimes, restricted nature of the information sources for Mexican mutual funds can be a problem on its own. Given this scenario, it could take a considerable time and resources, beyond those available for the elaboration of the current thesis, to propose a complete solution for these needs.

However, it possible to focus the present work in developing an exploratory thesis that can used as a guide, or basis, to build and automated mutual fund analyst and advisor with the help of machine learning methods.

1.3.1. Main objective

Build the basis for the development of an analysis tool of mutual funds investment in the Mexican market for the general public.

1.3.2. Specific objectives

- Define a method and procedure to extract mutual funds information from the CNBV website.
- Define and create a procedure to automatize the download of data of mutual funds registered and approved by the CNBV. Including their information sheets and historical price series.
- Calculate the traditional MPT measures for Mexican equity funds registered and approved by the CNBV.
- Select a machine learning method for time series clustering to analyze the funds' historic prices.
- Define patterns or profile groups based on the results of the clustering analysis performed in the historic funds' prices data and compare this analysis with the traditional MPT measures.
- Uncover behavioral patterns in equity funds that could help potential investors in the selection of funds, according to their investment objectives and risk aversion.
- Compare the clustering machine learning method results with the traditional MPT measures and evaluate if they can be substituted by this simpler and novel method.

1.3.3. Notes and clarifications

The first two objectives are focused on obtaining information on all the mutual funds available for investment in the BMV, but the objectives for the analysis and experimentation of this thesis are focused on the data of equity funds. This reasoning comes from the next reasons:

I. One of the goals of a financial advising tool for the general public should be protecting and providing investment suggestions that generate greater earnings than those of the debt certificates available at banks. Equity funds are preferable to debt

funds because their long term returns are much higher than those funds that focus only on government or corporate issued debt. Unfortunately, some debt funds do not have a much bigger return than debt certificates.

II. Because the focus of this exploratory work is on Mexican issued funds, the benchmark for their performance must be the index of the BMV. In order to do a fair and accurate comparison, only equity funds that invest in Mexican companies, whatever size and industry they belong to, can be used in the experiments. This excludes equity funds that invest, totally or partially, in foreign shares and international mutual funds.

At this point, it is unknown if the number of Mexican equity funds that invest in Mexican companies is enough to perform the expected analysis. Any unforeseen limitation and restriction about the discovered information and data is explained in the next chapters.

1.4. Research questions

The following research questions were stated as a guide to accomplish the previous objectives:

- Is it feasible to create a simple method or process to extract, clean and preprocess Mexican mutual funds' information from the CNBV website and other public financial information providers' websites for further analysis?

- Is there a correlation between the traditional Modern Portfolio Theory measures (alpha, beta, r squared, standard deviation and Sharpe ratio) and the clustering analysis results for equity funds?

- After applying a machine learning data analysis with a clustering method, do the resulting clustering of funds provides a meaningful grouping, or classification, of equity funds different from the classification provided by traditional financial entities (CNBV, Morningstar, etc.)?

- Do the resulting clustering of funds relates to an observable characteristic, or combination of funds' characteristics, that can be used to create investment profiles for guiding and advising novice investors in the choosing of an adequate investment?

- How does the clustering of mutual funds compares or relates to the Morningstar Rating of mutual funds?

1.5. Synthesis of the methodology

At the beginning, the expected work required to obtain and analyze the information for the experiments of this thesis was divided in three process:

1.5.1. Obtainment of information for analysis

For the data gathering and preprocessing phase of this thesis, it was expected to develop two applications or process:

1. An application to download the full list of mutual funds from the CNBV web site and their price series data from the Yahoo! Finance website.

2. An application or code process to preprocess the price series for the traditional and machine learning analysis.

1.5.2. Traditional financial data analysis

In order to calculate the traditional 5 MPT statistics for the funds' price series, two alternatives were considered:

1. To search for an existing and validated library with financial or econometric functions (from a trustable source, like the CRAN repository for the R language packages, the Python Package Index for Python language libraries, or from a

research department in economy and finance from a known university) that includes the code to calculate this measures. If possible, in the computer languages R or Python, due to those languages focus on statistical analysis or in the availability of open source libraries for statics and machine learning analysis.

2. In case there could not be found an existing library, or libraries, that could provide the functions for these measures, it would be required to develop and test them before performing the experiments.

1.5.3. Machine learning clustering analysis

For this step, it was planned to search and select a library that already provided functions for clustering analysis for data time series. If possible, it was expected to use a library in the same computer language, as the library.

1.6. Synthesis of the results

At the end of the experiments for this thesis, the planned processes were performed. However, the nature of the obtained data for analysis, and several errors and missteps with the chosen technologies to obtain them, compelled some changes in how these phases were achieved.

1.6.1. Results of the obtainment of information for analysis

The data gathering presented some interesting challenges (described in detail in chapter 3) that entailed the split of the data acquisition process in two applications:

1. An application to download the full list of mutual funds, with their classification information and benchmarks, from the CNBV website.

2. A script process to download the historical price series data of the chosen funds for study from the ***Yahoo! Finanzas*** web portal.
3. A group of SQL scripts process to perform each step of the historical price series' preprocessing for the traditional and machine learning analysis.

1.6.2. Results of traditional financial data analysis

At the CRAN repository for R language packages, it was possible to find a package for econometric and financial analysis which included the functions and their code for most of the MPT measures. The technical detail of this set of functions and the calculated statistics can be found in chapter 4.

1.6.3. Results of machine learning clustering analysis

The basis for the machine learning analysis comes from the algorithms explained in the article "*TSclust: An R Package for Time Series Clustering*" (Montero & Vilar, 2014) and the R package developed by its authors. The technical details of this library and the results of the clustering analysis can be found in chapter 5.

The results of the machine learning analysis and their comparison with the traditional risk and return measurements did not lead to a correlation between the MPT statistics and the discovered classifications. The complete comparisons and its detailed analysis are in chapter 5.2, while a detailed explanation of the conclusions of this work, its findings and future lines of research, are located in chapter 6.

2. Theoretical Background

This thesis proposes a time series clustering analysis for the classification of mutual funds based on their past performance. In order to assess its usefulness, it's necessary to perform a comparison with the parameters commonly used for mutual fund evaluation. This chapter begins by briefly explaining the financial theory behind the quantitative assessment currently used for mutual funds and investment instruments in the finance industry.

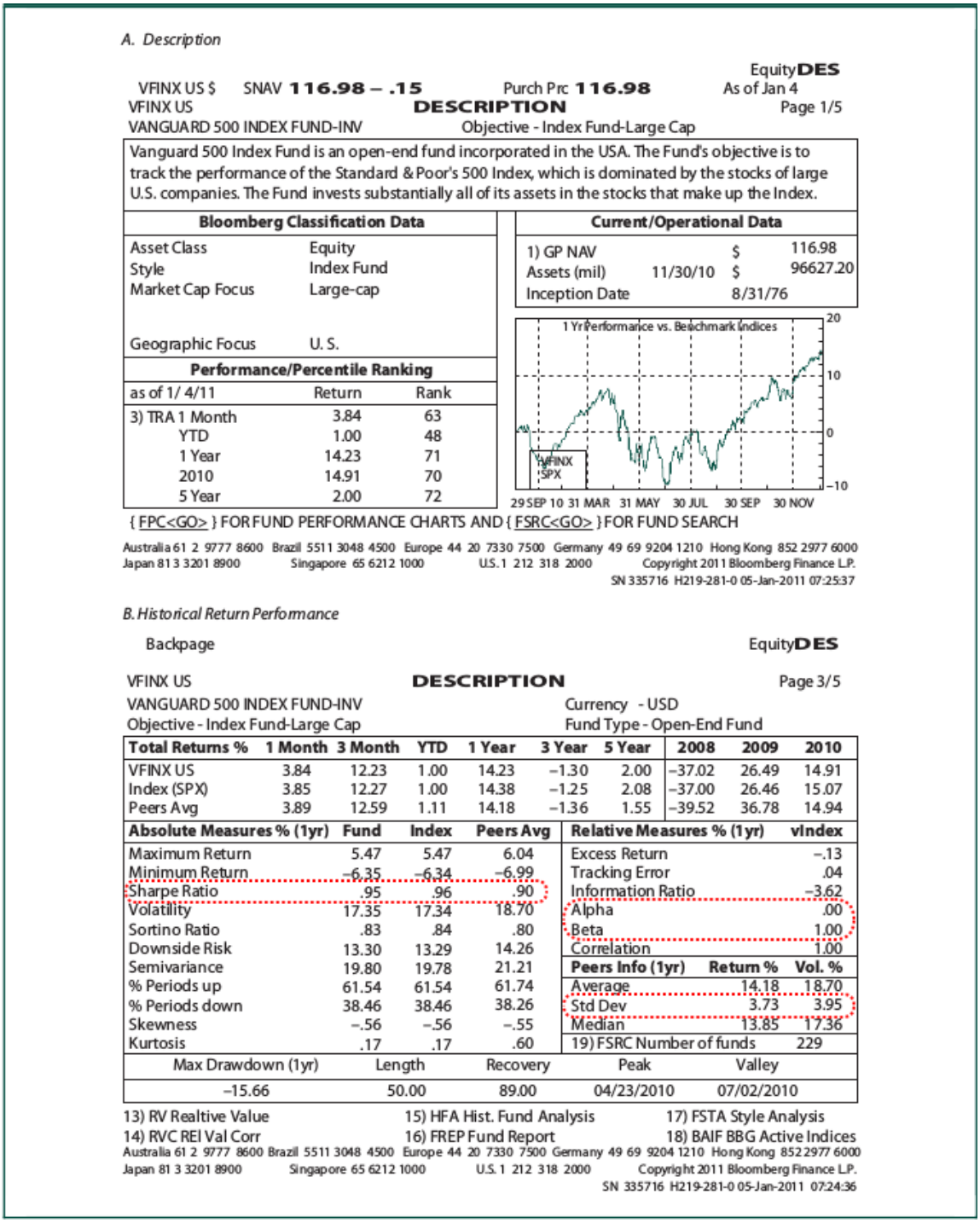
After the finance theory, an explanation of the algorithm chosen for the clustering analysis of the mutual funds' historical price series is presented. This description illustrates the advantages of the selected machine learning algorithm and the expected new knowledge that the resulting grouping of mutual funds can provide to seasoned and inexperienced investors.

2.1. Current financial assessment of investment instruments

When researching and evaluating which investment instrument, asset or security, better suits his or her investment goals, experienced investors want to know how risky is an investment and what is its expected profit. For mutual funds, brokers and fund management companies provide this information in a document called fund fact sheet, periodic reports of a fund's performance and risk over a period of time. Although, the content and name of this reports varies depending on local and national financial, commercial and fiscal laws, and the type of clients that the broker is marketing to, they are often published on a monthly, quarterly or yearly basis.

In the tenth edition of their book "*Investment Analysis & Portfolio Management*", Reilly and Brown (2012, p. 556) provide an example of a return performance report of the fund "Vanguard 500 Index Fund Investor Class" (ticker symbol "VFINX"), by independent financial analysis firm Bloomberg (see **Image 2.1.**). Among the many performance and

risk measures included in the report, the ***Sharpe ratio***, ***alpha***, ***beta*** and ***standard deviation*** statistics are part of the five MPT statistics, mentioned in the previous chapter.



A. Description

Equity **DES**

VFINX US $ SNAV **116.98 – .15** Purch Prc **116.98** As of Jan 4
VFINX US **DESCRIPTION** Page 1/5
VANGUARD 500 INDEX FUND-INV Objective - Index Fund-Large Cap

Vanguard 500 Index Fund is an open-end fund incorporated in the USA. The Fund's objective is to track the performance of the Standard & Poor's 500 Index, which is dominated by the stocks of large U.S. companies. The Fund invests substantially all of its assets in the stocks that make up the Index.

Bloomberg Classification Data	
Asset Class	Equity
Style	Index Fund
Market Cap Focus	Large-cap
Geographic Focus	U. S.

Current/Operational Data			
1) GP NAV		$	116.98
Assets (mil)	11/30/10	$	96627.20
Inception Date			8/31/76

Performance/Percentile Ranking		
as of 1/ 4/11	Return	Rank
3) TRA 1 Month	3.84	63
YTD	1.00	48
1 Year	14.23	71
2010	14.91	70
5 Year	2.00	72

{ FPC<GO> } FOR FUND PERFORMANCE CHARTS AND { FSRC<GO> } FOR FUND SEARCH

B. Historical Return Performance

Backpage Equity **DES**

VFINX US **DESCRIPTION** Page 3/5
VANGUARD 500 INDEX FUND-INV Currency - USD
Objective - Index Fund-Large Cap Fund Type - Open-End Fund

Total Returns %	1 Month	3 Month	YTD	1 Year	3 Year	5 Year	2008	2009	2010
VFINX US	3.84	12.23	1.00	14.23	–1.30	2.00	–37.02	26.49	14.91
Index (SPX)	3.85	12.27	1.00	14.38	–1.25	2.08	–37.00	26.46	15.07
Peers Avg	3.89	12.59	1.11	14.18	–1.36	1.55	–39.52	36.78	14.94

Absolute Measures % (1yr)	Fund	Index	Peers Avg
Maximum Return	5.47	5.47	6.04
Minimum Return	–6.35	–6.34	–6.99
Sharpe Ratio	.95	.96	.90
Volatility	17.35	17.34	18.70
Sortino Ratio	.83	.84	.80
Downside Risk	13.30	13.29	14.26
Semivariance	19.80	19.78	21.21
% Periods up	61.54	61.54	61.74
% Periods down	38.46	38.46	38.26
Skewness	–.56	–.56	–.55
Kurtosis	.17	.17	.60

Relative Measures % (1yr)	vIndex
Excess Return	–.13
Tracking Error	.04
Information Ratio	–3.62
Alpha	.00
Beta	1.00
Correlation	1.00

Peers Info (1yr)	Return %	Vol. %
Average	14.18	18.70
Std Dev	3.73	3.95
Median	13.85	17.36
19) FSRC Number of funds		229

Max Drawdown (1yr)	Length	Recovery	Peak	Valley
–15.66	50.00	89.00	04/23/2010	07/02/2010

13) RV Realtive Value 15) HFA Hist. Fund Analysis 17) FSTA Style Analysis
14) RVC REl Val Corr 16) FREP Fund Report 18) BAIF BBG Active Indices

Image 2.1. Example of a mutual fund performance and risk report, or fund fact sheet, from independent financial analysis firm Bloomberg (Reilly & Brown, 2012, p. 556).

Albeit detailed and focused on statistical measures, this type of report was designed for professional investors and finance researchers. Most mutual funds' performance reports and fact sheets focus on past return rates, total capitalization money and a few statistics.

Such is the case in the sample performance report in **Image 2.2.**, from fund manager Westchester Capital Funds' "The Merger Fund VL" (ticker MERVX). This report focuses on the funds' return rates in the short (from 3 months to 1 year), medium (3 years) and long (from 5 years to its creation year) terms, as well as its monthly performance. In the last section of the report, MPT statistics ***standard deviation***, ***Sharpe ratio*** and ***beta*** are included. This the type of report usually aimed to the general public.

Westchester Capital FUNDS

The Merger Fund VL

A Westchester Capital Fund

Fund Snapshot as of May 31, 2016

FUND FACTS

Inception Date:	5/26/2004
Total Fund Assets:	$29.2 million
Total Firm Assets:	$ 5.0 billion
Symbol:	MERVX
CUSIP:	589512102

As of the end of May, the Fund returned 1.34% for the month, 1.24% YTD, and 4.86% annualized since inception.

PERFORMANCE

as of Month-End: May 31, 2016

	3-month	YTD	1-year	3-year	5-year	Since Inception[1]
The Merger Fund VL	1.92%	1.24%	-1.58%	1.62%	1.01%	4.86%
Barclays Aggregate Bond Index	1.33%	3.46%	3.03%	2.92%	3.34%	4.66%
S&P 500 Index	9.12%	3.57%	1.72%	11.06%	11.67%	7.58%

as of Quarter-End: March 31, 2016

	3-month	YTD	1-year	3-year	5-year	Since Inception[1]
The Merger Fund VL	0.86%	0.86%	-1.41%	1.62%	1.11%	4.90%
Barclays Aggregate Bond Index	3.04%	3.04%	1.99%	2.50%	3.79%	4.69%
S&P 500 Index	1.35%	1.35%	1.78%	11.82%	11.58%	7.49%

[1]Performance is calculated for the period from June 1, 2004, the first full month of the life of the Fund. YTD and 3-month performance is not annualized. Performance data quoted represents past performance; past performance does not guarantee future results. The performance results portrayed herein reflect the reinvestment of all interest, dividends and distributions. The investment return and principal value of an investment will fluctuate so that an investor's shares, when redeemed, may be worth more or less than their original cost. Current performance of the Fund may be lower or higher than the performance quoted. As of the April 22, 2016 Prospectus, the Fund's total annual operating expense ratio was 2.60%. After applicable fee waivers and expense reimbursements (which may be terminated before April 30, 2017 only with approval of the Board of Trustees), total annual operating expenses were 1.82%, and after applicable fee waivers and before investment-related expenses, such as dividend and interest expense and acquired fund fees and expenses, total annual operating expenses were 1.40%. Performance data current to the most recent month-end may be obtained by contacting your financial advisor or the offering insurance company or by calling (800) 343-8959.

HISTORICAL PERFORMANCE SINCE INCEPTION

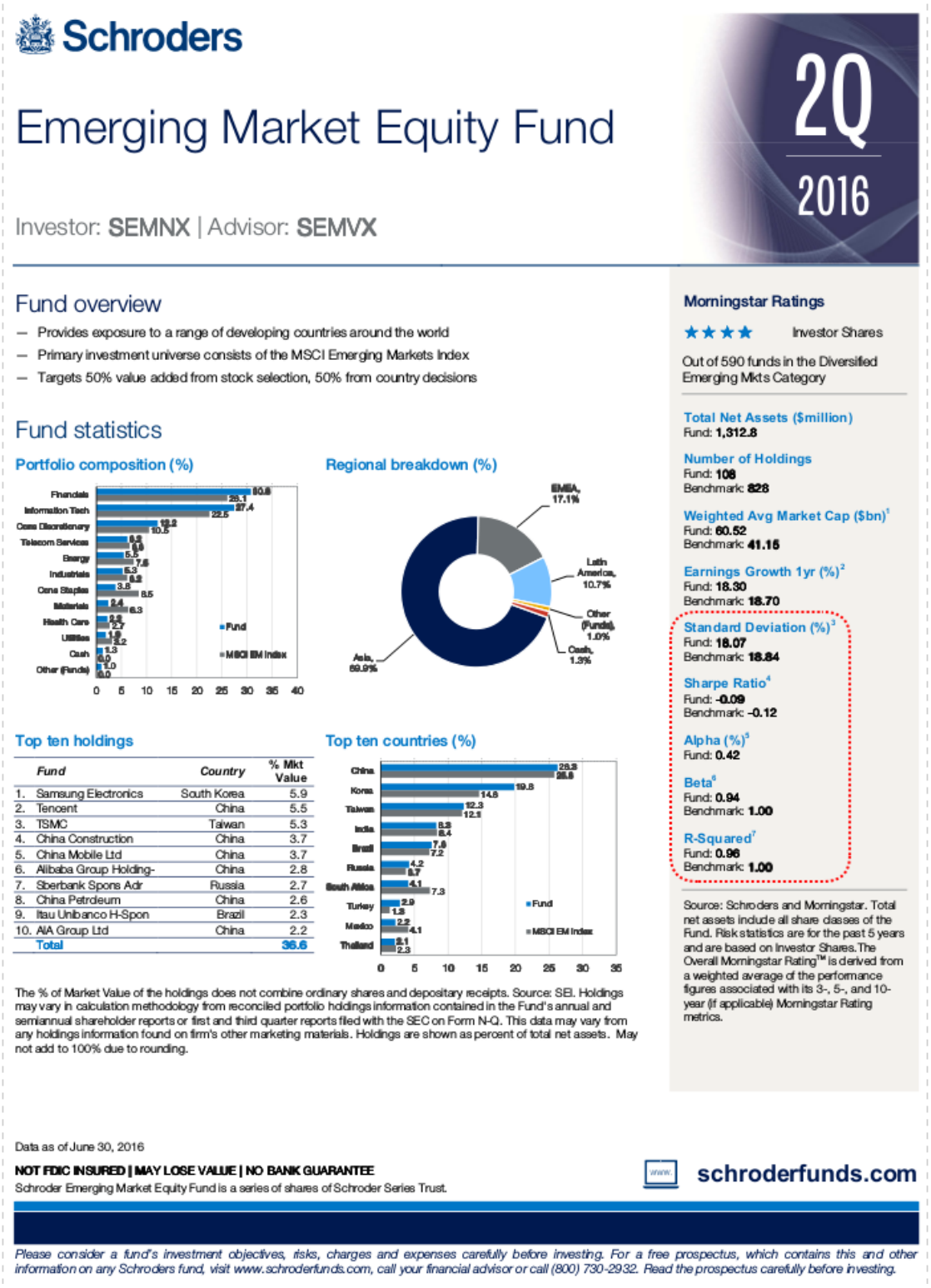
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Year
2016	-0.96%	0.29%	1.54%	-0.95%	1.34%								1.24%
2015	-0.18%	1.66%	-0.09%	0.09%	0.45%	-1.17%	-0.18%	-1.10%	-1.30%	1.31%	-0.46%	0.11%	-0.90%
2014	-0.55%	1.10%	0.09%	0.45%	1.00%	0.81%	-0.53%	0.89%	-1.33%	-1.17%	0.45%	0.18%	1.37%
2013	-0.57%	0.19%	0.67%	0.28%	0.09%	-0.28%	1.04%	0.09%	0.84%	0.74%	0.09%	0.63%	3.88%
2012	-0.10%	0.96%	0.28%	0.38%	-1.51%	0.19%	0.29%	0.95%	-0.19%	-1.04%	1.05%	1.26%	2.52%
2011	0.91%	0.45%	1.61%	0.88%	0.00%	-0.35%	-1.66%	-2.85%	-1.19%	2.13%	0.82%	0.24%	0.88%
2010	0.56%	0.74%	0.74%	-0.27%	-1.93%	0.84%	1.30%	0.83%	1.18%	0.45%	0.18%	0.59%	5.29%
2009	0.71%	-0.50%	4.04%	1.26%	0.77%	0.57%	0.57%	1.13%	0.74%	0.09%	0.55%	1.34%	11.80%
2008	-2.61%	2.37%	-1.51%	4.40%	3.62%	-3.59%	2.55%	2.68%	-3.17%	-2.50%	0.10%	1.85%	3.80%
2007	1.30%	1.45%	0.42%	0.92%	1.99%	-0.24%	-1.55%	1.16%	0.25%	1.31%	-4.36%	-0.39%	2.11%
2006	2.37%	1.96%	1.31%	0.52%	1.29%	3.05%	0.58%	1.55%	0.81%	0.80%	0.08%	1.14%	16.56%
2005	0.00%	0.09%	0.94%	0.00%	1.31%	0.46%	1.38%	0.72%	0.27%	-3.67%	1.86%	1.19%	4.53%
2004						0.30%	-1.40%	1.52%	0.80%	0.79%	1.97%	2.32%	6.42%

STATISTICAL ANALYSIS

	Standard Deviation	Sharpe Ratio	Beta	Correlation	Maximum Drawdown	Months to Recover
The Merger Fund VL	4.68%	1.04	0.19	0.34	-7.22%	4
Barclays Aggregate Bond Index	3.15%	1.46	0.00	0.00	-3.82%	2
S&P 500 Index	14.24%	0.59	1.00	1.00	-50.95%	37

Image 2.2. Performance report sample from fund management company Westchester Capital Funds (U.S. Securities and Exchange Commission, 2016, p. 1).

The group of statistical measures included in mutual funds reports varies from each fund manager company. For example, fund managers like Schroders include the five MPT statistics in the fund fact sheet for "Emerging Market Equity Fund" (SEMNX), **Image 2.3.**; still, none of the other metrics shown in previous performance report samples, like *Correlation*, were included.



Schroders

Emerging Market Equity Fund

2Q 2016

Investor: **SEMNX** | Advisor: **SEMVX**

Fund overview

- Provides exposure to a range of developing countries around the world
- Primary investment universe consists of the MSCI Emerging Markets Index
- Targets 50% value added from stock selection, 50% from country decisions

Fund statistics

Portfolio composition (%)

Regional breakdown (%)

Top ten holdings

	Fund	Country	% Mkt Value
1.	Samsung Electronics	South Korea	5.9
2.	Tencent	China	5.5
3.	TSMC	Taiwan	5.3
4.	China Construction	China	3.7
5.	China Mobile Ltd	China	3.7
6.	Alibaba Group Holding-	China	2.8
7.	Sberbank Spons Adr	Russia	2.7
8.	China Petroleum	China	2.6
9.	Itau Unibanco H-Spon	Brazil	2.3
10.	AIA Group Ltd	China	2.2
	Total		36.6

Top ten countries (%)

The % of Market Value of the holdings does not combine ordinary shares and depositary receipts. Source: SEI. Holdings may vary in calculation methodology from reconciled portfolio holdings information contained in the Fund's annual and semiannual shareholder reports or first and third quarter reports filed with the SEC on Form N-Q. This data may vary from any holdings information found on firm's other marketing materials. Holdings are shown as percent of total net assets. May not add to 100% due to rounding.

Morningstar Ratings

★★★★ Investor Shares

Out of 590 funds in the Diversified Emerging Mkts Category

Total Net Assets ($million)
Fund: **1,312.8**

Number of Holdings
Fund: **108**
Benchmark: **828**

Weighted Avg Market Cap ($bn)[1]
Fund: **60.52**
Benchmark: **41.15**

Earnings Growth 1yr (%)[2]
Fund: **18.30**
Benchmark: **18.70**

Standard Deviation (%)[3]
Fund: **18.07**
Benchmark: **18.84**

Sharpe Ratio[4]
Fund: **-0.09**
Benchmark: **-0.12**

Alpha (%)[5]
Fund: **0.42**

Beta[6]
Fund: **0.94**
Benchmark: **1.00**

R-Squared[7]
Fund: **0.96**
Benchmark: **1.00**

Source: Schroders and Morningstar. Total net assets include all share classes of the Fund. Risk statistics are for the past 5 years and are based on Investor Shares. The Overall Morningstar Rating™ is derived from a weighted average of the performance figures associated with its 3-, 5-, and 10-year (if applicable) Morningstar Rating metrics.

Data as of June 30, 2016

NOT FDIC INSURED | MAY LOSE VALUE | NO BANK GUARANTEE

Schroder Emerging Market Equity Fund is a series of shares of Schroder Series Trust.

www. schroderfunds.com

Please consider a fund's investment objectives, risks, charges and expenses carefully before investing. For a free prospectus, which contains this and other information on any Schroders fund, visit www.schroderfunds.com, call your financial advisor or call (800) 730-2932. Read the prospectus carefully before investing.

Image 2.3. Performance report sample from fund manager Schroder (Schroders, 2016, p. 1).

Hennessy Japan Fund — As of December 31, 2015

FUND FACTS

Ticker/CUSIP	
Investor Class	HJPNX/425 894 102
Institutional Class	HJPIX/425 894 201
Total Fund Assets	$128 million
Number of Holdings	21
Annual Total Expense Ratio	
Investor Class	1.49%
Institutional Class	1.08%
Inception Dates	
Investor Class	10/31/03
Institutional Class	10/31/03
Dividends Paid	Annually

PORTFOLIO CHARACTERISTICS

Portfolio Turnover	17%
Median Price/Earnings	27.32x
Median Price/Book	2.40x
Median Market Cap	$14.0 billion

MORNINGSTAR STYLE BOX

ABOUT HENNESSY

Hennessy Funds has a longstanding track record of proven performance and offers a broad range of mutual funds, with strategies that can play a role in nearly every investor's portfolio allocation.

Each of the Hennessy Funds employs a consistent and repeatable investment process, combining time-tested stock selection strategies with a highly disciplined, team-managed approach. Our goal is to provide products that investors can have confidence in, knowing their money is invested as promised, with their best interest in mind.

HENNESSY FUNDS

Investing, Uncompromised

HENNESSYFUNDS.COM | 800-966-4354

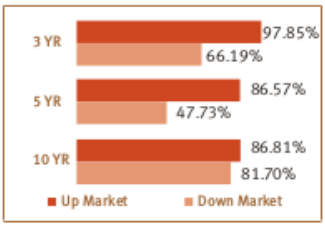
TOP TEN EQUITY HOLDINGS

Terumo Corp.	6.4%
Keyence Corp.	6.1%
Shimano, Inc.	5.9%
Rohto Pharmaceutical Co., Ltd.	5.7%
Ryohin Keikaku Co., Ltd.	5.4%
Misumi Group, Inc.	5.2%
Kao Corp.	5.0%
Unicharm Corp.	5.0%
Nidec Corp.	4.9%
Daikin Industries, Ltd.	4.9%
Total as % of Net Assets	**54.5%**

SECTOR WEIGHTING

Consumer Discretionary	21.4%
Consumer Staples	11.6%
Financials	3.0%
Health Care	12.1%
Industrials	24.6%
Information Technology	6.1%
Materials	3.4%
Telecommunication Services	4.8%
Cash & Other	13.1%

CAPTURE RATIOS RELATIVE TO TO RUSSELL/NOMURA TOTAL MARKET INDEX (INVESTOR CLASS)



RISK METRICS (INVESTOR CLASS)

	3 YR	5 YR	10 YR
Standard Deviation			
Japan Fund	13.51	12.34	15.34
Russell/Nomura TM Index	13.24	13.84	15.13
Risk Statistics (relative to Russell/Nomura Total Market Index)			
Beta	0.91	0.72	0.88
Alpha	4.59	7.06	1.09

Investors should consider the investment objectives, risks, charges and expenses carefully before investing. This and other important information can be found in the Fund's statutory and summary prospectuses. To obtain a free prospectus, please call 800-966-4354 or visit hennessyfunds.com. Please read the prospectus carefully before investing.

Mutual fund investing involves risk; Principal loss is possible. The Fund invests in small and medium capitalized companies, which may have more limited liquidity and greater price volatility than large capitalization companies. The Fund invests in the stock of companies operating in Japan; single country funds may be subject to a higher degree of market risk. The Fund may experience higher fees due to investments in pooled investment vehicles (including ETFs).

Each Morningstar category average represents a universe of funds with similar objectives. The Russell/Nomura Total Market Index is a market capitalization-weighted index of Japanese equities. The Tokyo Stock Price Index (TOPIX) is a market capitalization-weighted index of all companies listed on the First Section of the Tokyo Stock Exchange. The Russell/Nomura Total Market Index and TOPIX indices are presented in U.S. Dollar terms. One cannot invest directly in an index. Fund holdings and sector weightings are subject to change and should not be considered a recommendation to buy or sell any security.

Morningstar Proprietary Ratings reflect risk-adjusted performance as of 12/31/15. For each fund with at least a three year history, Morningstar calculates a Morningstar Rating™ based on a Morningstar risk-adjusted return measure that accounts for variation in a fund's monthly performance placing more emphasis on downward variations and rewarding consistent performance. The top 10% of funds in each category receive 5 stars, the next 22.5% receive 4 stars, the next 35% receive 3 stars, the next 22.5% receive 2 stars and the bottom 10% receive 1 star. Each share class is counted as a fraction of one fund within this scale and rated separately, which may cause slight variations in distribution percentage. Morningstar Percentile Ranking compares a fund's Morningstar risk and return scores with all the funds in the same Category, where 1% = Best and 100% = Worst. The Morningstar Style Box reveals the Fund's investment style as of 12/31/15. The vertical axis shows the market capitalization of the stocks owned and the horizontal axis shows investment style (value, blend, or growth).

Price/Earnings Ratio is the market price per share divided by earnings per share. Price/Book Ratio is the market price per share divided by book value. Standard deviation is a statistical measure of the historical volatility of a mutual fund or portfolio. Beta measures the volatility of the fund, as compared to that of the overall market. The Market's beta is set at 1.00; a beta higher than 1.00 is considered to be more volatile than the market, while a beta lower than 1.00 is considered to be less volatile. Alpha is an annualized return measure of how much better or worse a fund's performance is relative to an index of funds in the same category, after allowing for differences in risk.

The Hennessy Funds are distributed by Quasar Distributors, LLC.

Image 2.4. Performance report sample from fund manager Hennessy Funds Japan (Hennessy Funds, 2016, p. 2).

Note that not all fund performance reports include the MPT statistics; but the fund fact sheets reports consulted from literature and fund managers, including funds' reports from outside the U.S. stock exchange, imply that these are the most common measures of performance and risk for investors in the finance industry. Other statistical metrics are included in the performance reports, but not with enough frequency to be thought as standard measures of risk and performance. Therefore, the MPT statistics were considered as the main traditional financial measures for investment assessment in the present work.

2.1.1. Risk and risk-adjusted performance measures

To measure the risk of a mutual fund, the standard deviation and the Sharpe ratio of the funds' returns are some of the most frequently used metrics.

Standard deviation

A **risky security** or asset is an investment instrument whose future rates of return, or yield rates, are uncertain. The standard deviation of its expected returns measures this uncertainty. On the contrary, because the expected return on a **risk-free security** is entirely certain, the standard deviation of its expected return is zero (Reilly & Brown, 2012, p. 208-209).

A statistical measure of the volatility of a security's performance, the **standard deviation** measures the degree of variation of the series of returns of the security during a given period of time. The standard deviation of any given asset's total return is calculated as the square root of the average of the square of the difference between all daily returns over a period of time and its average total return over said period ("*The Morningstar Approach to Mutual Fund Analysis—Part I*", 2010, p. 165).

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (R_i - \overline{R})^2}{n}}$$

Formula 2.1. Formula to calculate a security's standard deviation.

Where:

- $\boldsymbol{\sigma}$**:** Standard deviation of the security, mutual fund or investment instrument.
- $\boldsymbol{R_i}$ **:** Series of rate returns, *i* to *n*, over the period of time
- $\boldsymbol{\overline{R}}$**:** Average rate of return over the period of time

In mutual funds, the standard deviation assumes that their returns follow a normal, or Gaussian, distribution. Based on this supposition, it is expected that a mutual fund's total returns differ no more than ±1 standard deviation from its average total return, approximately, 68% of the time, and be within a range of ±2 standard deviations from its average 95% of the time. For example, assuming a fund with an average total return of 10% and a standard deviation of 8%, hypothetically, 68% of the times this fund's total returns should be between 2% and 18% and, 95% of the times, between –6% and 26%.

The greater the standard deviation of a mutual fund's total returns over a time period, the greater is the fund's volatility. However, because is highly disputed that a fund's total returns follow a bell-shaped normal distribution most of the time, some researchers believe that standard deviation, or standard deviation–based statistics such as the Sharpe ratio, are of limited use ("*The Morningstar Approach to Mutual Fund Analysis—Part I*", 2010, p. 166).

Sharpe ratio

The purpose of investing in risky securities, like mutual funds, is to earn a rate of return above that which can be earned from investing in risk-free securities during the same period of time, such as: money market funds, government issued debt securities and

banknotes. The **excess return** on the riskier security, the difference between the total return on a risky security and the rate of return on a risk-free security, it's a measure that allows to know how well a security performed relative to a risk-free alternative ("*The Morningstar Approach to Mutual Fund Analysis—Part I*", 2010, p. 166). The excess return can be stated as:

$$ER = R - RFS$$

Formula 2.2. Adjustment to obtain an asset, or security, excess return against a risk-free security return.

Where:

ER **:** Excess return of an investment instrument, asset or security.

R **:** Return of a security over the same period of time than the risk-free security return.

RFS **:** Return of a risk-free security over the same period of time than the security return.

Developed by Nobel laureate William Sharpe, the Sharpe ratio is a risk-adjusted performance measure based on excess return and standard deviation. It is defined as the ratio of the average of a fund's excess returns of a security over the standard deviation of the security's returns during a period of time.

$$S = \frac{\overline{ER(R_i, RFS_i)}}{\sigma(R_i)}$$

Formula 2.3. Formula to calculate the Sharpe ratio of a security over a period of time.

Where:

S **:** Sharpe ratio of the analyzed security or investment instrument.

$\overline{ER(R_i,RFS_i)}$ **:** Average excess return of the security against a risk-free security over a period of time.

σ **:** Standard deviation of the security's returns over the analyzed period.

The ***Sharpe ratio represents the average excess return per unit of risk*** and allows comparisons of funds or securities that operate at different levels of volatility. The higher the Sharpe ratio of a fund, the better the fund's historical risk-adjusted performance. For example, suppose that Fund A and Fund B have these measures:

Fund A		**Fund B**	
Average annual excess return	6%	Average annual excess return	10%
standard deviation	10%	standard deviation	20%
Sharpe ratio	**0.6**	**Sharpe ratio**	**0.5**

Table 2.1. Example of Sharpe ratio calculation

On a risk-adjusted basis, Fund A outperformed Fund B because Fund A provided an excess return of 0.6 per unit of risk (or a Sharpe ratio of 0.6), while Fund B provided an excess return of 0.5 per unit of risk (Sharpe ratio of 0.5); although, Fund B had a higher average excess return, it was more volatile. In other words, Fund A was the better performer on a risk-adjusted basis ("*The Morningstar Approach to Mutual Fund Analysis—Part I*", 2010, p. 166).

Unfortunately, the Sharpe ratio has the disadvantage of been unable to identify the source of a security's associated risk; in other words, the *Sharpe ratio fails to distinguish between systematic and unsystematic risk.* As shown in the Sharpe ratio formula, a fund's risk is measured by the standard deviation of the returns, regardless of the sources of the volatility ("*The Morningstar Approach to Mutual Fund Analysis—Part I*", 2010, p. 166-167). Also called market risk or macro risk, the **systematic risk** comes from choosing to invest in a given securities market; it represents the generic risks of investing in a certain asset class or investment instrument type. On the opposite, the **unsystematic risk** is the risk that is specific to an individual asset or security, rather than to a whole class or type of securities. For example, an investor who owns shares in the company IBM takes on the *systematic risks of the stock market and the technology sector* as well as the *unsystematic risks associated with IBM itself: the company's specific competitive pressures, product line, stock valuation, and so forth* ("*The Morningstar Approach to Mutual Fund Analysis—Part I*", 2010, p. 154).

2.1.2. Modern Portfolio Theory statistics

As mentioned before, the **alpha**, **beta**, and **r squared** statistics are commonly used for the performance assessing of mutual funds relative to a benchmark, by investment practitioners and academics. These statistical measures are referred to as the **Modern Portfolio Theory** (MPT) statistics because they are derived from Harry Markowitz's theory of portfolio construction, developed in the 1950's, and considered modern relative to the approach to portfolio construction that preceded it ("*The Morningstar Approach to Mutual Fund Analysis—Part I*", 2010, p. 167).

The basis for these three measures is a linear regression analysis between the excess returns of an asset or security (as the *y* values) and an index or benchmark's excess returns (as the *x* values) over a period of time ("*The Morningstar Approach to Mutual Fund Analysis—Part I*", 2010, p. 167). The regression equation produced by the regression analysis is:

$$ER(R_a, RFS) = \alpha_{a,b} + \beta_{a,b}(ER(R_b, RFS)) + \varepsilon_{a,b}$$

Formula 2.4. In finance, this equation is known as the security's ***characteristic line*** with the index or benchmark (Reilly & Brown, 2012, p. 221).

Where:

- R_a : Security *a*'s returns over a period of time.
- R_b : The index, or benchmark, *b* returns to compare the security against to.
- RFS : Risk-free security's returns to calculate the excess returns of the security *a* and the index *b*, over a period of time.
- $ER(R_a,RFS)$: Excess returns of the security *a* against a risk-free security over the same period.
- $\alpha_{a,b}$: Intercept of the linear regression (alpha) of a security *a* compared to an index *b*.
- $\beta_{a,b}$: Slope of the linear regression (beta) of a security *a* compared to an index *b*.
- $ER(R_b,RFS)$: Excess returns of the index against a risk-free security over the same period.

$\varepsilon_{a,b}$ **:** Random error term accounting for the security a's unsystematic risk

In the United States of America, independent financial analysts, like Morningstar, calculate a mutual fund's alpha, beta, and r squared statistics by running a regression of the fund's excess returns over the 90-day U.S. Treasury bill (the risk-free security) compared with the excess returns of a selected index or benchmark (like the "S&P 500" index) as the standard index for the fund's category group.

Unfortunately, how the linear regression is calculated (the number and time interval of observations) can widely alter the values of the three statistics for a security. For example, Morningstar derives characteristic lines using monthly returns for the most recent five-year period (60 observations); while financial analyst Reuters Analytics calculates stock betas using daily returns over the prior two years (504 observations); and financial analyst Bloomberg uses two years of weekly returns (104 observations). Because *there is no theoretically correct time interval to perform the linear regression analysis* to calculate the MPT statistic measures (including standard deviation and Sharpe ratio), and estimate a security's risk, it's necessary to evaluate *how to make a balanced trade-off between enough observations* (to eliminate the impact of random rates of return) *and an excessive length of time* (such as 15 or 20 years), over which the subject security may have changed dramatically (Reilly & Brown, 2012, p. 221).

Beta

Beta is a measure of the sensitivity of a fund's, or security's, excess returns to movements in an index's excess returns; thus, beta provides a measure of the fund's systematic risk relative to the index. Also called the *beta coefficient*, the beta statistic can be calculated with the help of the Covariance formula.

$$\beta_{a,b} = \frac{CoV_{a,b}}{\sigma_b} = \frac{\sum_{1}^{i}((ER(R_a, RFS)_i - \overline{ER(R_a, RFS)})((ER(R_b, RFS)_i - \overline{ER(R_b, RFS)}))}{\frac{\sum_{1}^{i}(ER(R_b, RFS)_i - \overline{ER(R_b, RFS)})^2}{i}}$$

Formula 2.5. Covariance formula generally used to calculate the beta of a security (Reilly & Brown, 2012, p. 221).

Where:

$\boldsymbol{\beta_{a,b}}$ **:**	Beta of a security *a* compared to an index, or benchmark, *b*.
$\boldsymbol{CoV_{a,b}}$ **:**	Covariance of a security's excess returns compared to an index, or benchmark, *b*'s excess returns.
$\boldsymbol{\sigma_b}$ **:**	Standard deviation of the index *b*'s excess returns over the analyzed time period.
$\boldsymbol{R_a}$ **:**	Security *a*'s returns over a period of time.
$\boldsymbol{R_b}$ **:**	Index, or benchmark, *b* returns to compare the security against to.
RFS **:**	Risk-free security's returns to calculate the excess returns of the security *a* and the index *b*, over a period of time.
$\boldsymbol{ER(R_a,RFS)}$ **:**	Excess returns of the security *a* against a risk-free security over the same period.
$\boldsymbol{\overline{ER}(R_a,RFS)}$ **:**	Average of the excess returns of the security *a* against a risk-free security over the same period.
$\boldsymbol{ER(R_b,RFS)}$ **:**	Excess returns of the index *b* against a risk-free security over the same period.

Both, the linear regression equation and the above formula, produce the same estimate of beta for a given sample of security and index returns. However, the linear regression-based method, where ***beta is the slope of the linear regression***, is often preferred because it is a formal estimation process; this means that the statistical reliability of the estimate can be assessed (the $\beta_{a,b}$ estimate can be mathematically evaluated) (Reilly & Brown, 2012, p. 221).

The index always has a designed beta of 1. So, a fund with a beta of 1.10 indicates a tendency to generate an excess return 10% higher than that of the index when the market is up and 10% lower when the market is down, assuming all other factors remain constant. On the contrary, a fund with a beta of 0.85 would indicate that it has performed 15% worse than the index in up markets and 15% better in down markets.

It's important to note that ***a low value of beta*** does not mean that a fund has a low volatility; it only ***means that the fund's index-related risk is low***. For example, because its performance would be more closely tied to the price of gold and gold-mining stocks than to the overall stock market, a specialized fund that invest primarily in gold would

usually have a low beta (also, a low r squared). Thus, even if the return rates of this specialty fund fluctuate due to rapid changes in gold prices, its beta would continue to be low compared to the stock market ("*The Morningstar Approach to Mutual Fund Analysis—Part I*", 2010, p. 168).

Alpha

Alpha measures a fund's performance after adjusting for a fund's systematic risk, as measured by the fund's beta. It assumes that an investor could form a passive portfolio with the same beta as that of the fund by investing in the index and either borrowing or lending at the risk-free security rate of return to increase or decrease exposure to that index. alpha is calculated as the difference between the average excess return on the fund and the average excess return on the levered or unlevered index.

$$\alpha_{a,b} = \overline{ER(R_a, RFS)} - \beta_{a,b}\overline{ER(R_b, RFS)}$$

Formula 2.6. Basic formula to calculate the alpha of a security.

Where:

$\alpha_{a,b}$: Alpha of a security *a* compared to an index, or benchmark, *b*.

R_a : Security's returns over a period of time.

R_b : Index, or benchmark, returns to compare the security against to.

***RFS* :** Risk-free security's returns to calculate the excess returns of the security and the index, over a period of time.

$\overline{ER(R_a,RFS)}$: Average of the excess returns of the security against a risk-free security over the same period.

$\beta_{a,b}$: Beta of a security *a* compared to an index, or benchmark, *b*.

$\overline{ER(R_b,RFS)}$: Average of the excess returns of the index against a risk-free security over the same period.

For example, if a fund had an average excess return of 6% per year and its beta, with respect to the "S&P 500" index, was 0.8 over a period when the index's average excess return was 7%, the fund's alpha would be: 6% – 0.8 × 7% = 0.4%.

In a linear regression model, ***alpha is the intersection of the linear regression*** and ***aims to depict how a fund manager adds or subtracts value relative to a levered or unlevered index***, but its accuracy is limited. In some instances, *a negative alpha can result from the expenses present in the fund returns, but that are not present in the returns of the comparison index*. Also, *the usefulness of alpha is completely dependent on the accuracy of beta*. If beta is accepted as a conclusive definition of risk, a positive alpha would be a conclusive indicator of good fund performance ("*The Morningstar Approach to Mutual Fund Analysis—Part I*", 2010, p. 168).

R squared

Another statistic produced by the regression analysis, r squared is a number between 0 and 100 percent that measures the strength of the relationship between the excess returns of a fund and those of the index, or benchmark.

$$r^2_{a,b} = 1 - \frac{\sum_{1}^{i} (f(ER(R_a, RFS))_i - \overline{ER(R_b, RFS)})^2}{\sum_{1}^{i} (ER(R_b, RFS)_i - \overline{ER(R_b, RFS)})^2}$$

Formula 2.7. Formula to calculate the r squared of a linear regression model (Yau, n.d.).

Where:

$r^2_{a,b}$: R squared of a security *a* compared to an index, or benchmark, *b*.

R_a : Security *a*'s returns over a period of time.

R_b : Index *b* returns to compare the security against to.

***RFS* :** Risk-free security's returns to calculate the excess returns of the security *a* and the index *b*, over a period of time.

***ER(R_a,RFS)* :** Excess returns of the security *a* against a risk-free security over the same period.

***f(ER(R_a,RFS))* :** Calculated, or expected, excess returns of the index *b* with the formula generated from the linear regression of the security *a*'s excess returns against the index *b*'s excess returns.

ER(R$_b$,RFS) **:** Excess returns of the index *b* against a risk-free security over the same period.

ER(R$_b$,RFS) **:** Average of the excess returns of the index *b* against a risk-free security over the same period.

Also known as the *coefficient of determination*, ***the purpose of r squared is to judge the significance of the beta estimate***. Generally, the higher the value of r squared, the value of beta is more reliable.

When r squared has a value of 0%, it means that there is no relationship between the fund and the index, while an r squared of 100% means that the relationship is perfect; thus, U.S. stock index funds that track the "S&P 500" index will have an r squared very close to 100%. On the contrary, a low r squared indicates that the fund's movements are not well explained by movements in the index. For example, a r squared measure of 35% means that only 35% of a fund's return movements can be explained by movements in index returns ("*The Morningstar Approach to Mutual Fund Analysis—Part I*", 2010, p. 169).

2.2. Time series clustering

Machine learning encompasses any algorithm that learns from the behavior of a dataset. Although, any chosen algorithm to analyze a set of data still requires help from an expert in the knowledge domain to classify and provide a significance to the results of such analysis.

Because one the objectives of this work is to explore the possibility of a novel classification of investment funds for inexperienced investors, a time series clustering analysis is the best suited algorithm to analyze and learn from the previous performance of funds (their historical price series).

2.2.1. Introduction

In their paper "*TSclust: An R Package for Time Series Clustering*", Montero and Vilar (2014) explain that ***clustering is an unsupervised learning task aimed to partition a set***

of unlabeled data objects into homogeneous groups or clusters. The group partitioning is performed in a way that objects in the same cluster are more similar to each other than objects in different clusters, according to a defined criteria. Time series clustering problems arise in a wide variety of fields: economics, finance, medicine, ecology, environmental studies, engineering, and many others. Frequently, the grouping of series plays a central role in the studied problem, as in the objectives of the present work. In many real life problems, the cluster analysis must be performed on time series data; for example: finding stocks that behave in a similar way, determining products with similar selling patterns, identifying countries with similar population growth or regions with similar temperature (p. 1-2).

In cluster analysis, *a crucial question is* establishing what "similar" data objects means; in other words, *how to determine a suitable similarity/dissimilarity measure between two objects*. In the context of time series of data, the concept of dissimilarity is particularly complex due to the dynamic character of the series. Dissimilarities usually considered in conventional clustering could not work adequately with time dependent data because they ignore the interdependence relationship between values. Because of this, different approaches to define dissimilarity between time series have been proposed in the literature.

One of these approaches, measures the dissimilarity by comparing sequences of serial features extracted from the original time series, such as autocorrelations, cross-correlations, spectral features, wavelet coefficients, and so on. These feature-based approaches have the objective of represent the dynamic structure of each series by a feature vector of lower dimension, thus allowing a dimensionality reduction (time series are essentially high-dimensionality data) and a meaningful saving in computation time.

There exist a broad range of measures to compare time series and the choice of the proper dissimilarity measure depends largely on the nature of the clustering; this means that it depends on determining what the purpose of the grouping is for. Once the dissimilarity measure is determined, an initial pairwise dissimilarity matrix can be obtained and a conventional clustering algorithm be then used to form groups of objects.

In summary, ***a clustering algorithm is composed of two measures: a time series dissimilarity algorithm*** and ***the proper clustering algorithm for the resulting dissimilarity measures*** (Montero & Vilar, 2014, p. 1-3).

2.2.2. Time series dissimilarity algorithm CORT

The first step in the clustering analysis of mutual funds is to measure the dissimilarity between their historical price series. To measure the proximity between time series $X_T = (X_1, \ldots, X_T)$ and $Y_T = (Y_1, \ldots, Y_T)$, it is necessary to use a metric based on the closeness of their values at specific points of time, such as ***CORT***, ***an adaptive dissimilarity index covering both proximity on values and on behavior***.

Introduced by Douzal Chouakria and Nagabhushan (2007), CORT is a dissimilarity measure focused to ***cover both conventional measures for the proximity on observations and temporal correlation for the behavior proximity estimation***. The proximity between the dynamic behaviors of the series is evaluated by means of the first order temporal correlation coefficient, defined by:

$$CORT(X_T, Y_T) = \frac{\sum_{t=1}^{T-1}(X_{t+1} - X_t)(Y_{t+1} - Y_t)}{\sqrt{\sum_{t=1}^{T-1}(X_{t+1} - X_t)^2}\sqrt{\sum_{t=1}^{T-1}(Y_{t+1} - Y_t)^2}}$$

Formula 2.8. Formula of the correlation coefficient of the CORT dissimilarity index.

The result of $CORT(X_T, Y_T)$ is a value in the interval $[-1, 1]$, where:

- $\mathbf{CORT(X_T, Y_T) = 1}$, means that both series show a similar dynamic behavior, meaning, that their growths (positive or negative) at any instant of time are similar in direction and rate.

- **$CORT(X_T, Y_T) = -1$**, implies a similar growth in rate, but opposite in direction (opposite behavior).
- **$CORT(X_T, Y_T) = 0$**, expresses that there is no monotonicity between X_T and Y_T, and their growth rates are stochastically linearly independent (different behaviors).

The dissimilarity index proposed by Douzal Chouakria and Nagabhushan (2007) modulates the proximity between the raw-values of two series X_T and Y_T using the coefficient $CORT(X_T, Y_T)$, as defined in the following formula:

$$d_{CORT} = \varphi_k[CORT(X_T, Y_T)]d(X_T, Y_T)$$

Formula 2.9. Dissimilarity index formula, as proposed by Douzal Chouakria and Nagabhushan.

Where:

- **$\varphi_k(\cdot)$**, is an adaptive tuning function to automatically modulate a conventional raw-data distance, according to the series temporal correlation.
- **$d(X_T, Y_T)$**, is the conventional raw-data distance approach, which can be:
  - **d_{Lq}**, the Minkowski distance of order *q*:
    - **2**, for an Euclidean distance
    - **1**, for the Manhattan distance
  - **d_F**, the Fréchet distance
  - **d_{DTW}**, the dynamic time warping distance.

The modulating function $\varphi_k(\cdot)$ should increase (or decrease) the weight of the dissimilarity between observations as the temporal correlation decreases from 0 to −1 (or increases from 0 to +1). Additionally, $d_{CORT}(X_T, Y_T)$ should approach the raw-data discrepancy as the temporal correlation is zero. Instead of a linear tuning function, Douzal Chouakria and Nagabhushan propose that the modulating function $\varphi_k(\cdot)$ is an exponential adaptive function given by:

$$\varphi_k(u) = \frac{2}{1 + exp(ku)}, \; k \geq 0$$

Formula 2.10. Exponential adaptive tuning function to modulate a conventional raw-data distance, as proposed by Douzal Chouakria and Nagabhushan.

As consequence, both, $\varphi_k(\cdot)$ and k, modulate the weight that the correlation coefficient $CORT(X_T, Y_T)$ has on the calculated distance $d_{CORT}(X_T, Y_T)$ (Montero & Vilar, 2015, p. 17).

2.2.3. Clustering algorithm for dissimilarity measures

Once the dissimilarity measures are obtained with the dissimilarity index $d_{CORT}(X_T, Y_T)$, the clustering of the time series can be completed with the help of a ***hierarchical clustering algorithm based on p values***.

As Montero and Vilar (2014) explain "… introduced by Maharaj (2000), takes as starting point the $m \times m$ matrix $\mathcal{P} = (p_{i,j})$, whose (*i*, *j*)-th entry, $p_{i,j}$, for $i \neq j$, corresponds to the p value obtained by testing whether or not $X^{(i)}_T$ and $X^{(j)}_T$ come from the same generating model. Then, the algorithm proceeds in a similar way as an agglomerative hierarchical clustering based on $\mathcal{P}$, although in this case *will only group together those series whose associated p values are greater than a significance level α previously specified by the user*. In other words, the *i*-th series $X^{(i)}_T$ merge into a specific cluster C_k formed by the series $\{X^{(j1)}_T, \ldots, X^{(jmk)}_T\}$ if $p_{i,jl} \geq \alpha$, for all $l = 1, \ldots, m_k$. Analogously, two clusters will be joined together if the *p* values of all pairs of series across the two clusters are greater than *α*. This algorithm behaves similar to the single linkage procedure because the dissimilarity between two clusters is the smallest dissimilarity (the greatest *p* value) between series in the two groups. Unlike the single linkage, two clusters will not be joined with the Maharaj's algorithm when a significant difference between a pair of series of the candidate clusters is obtained. Note also that, unlike the conventional hierarchical methods, this algorithm presents the advantage of providing automatically

the number of clusters, which obviously depends on the prefixed significance level. Furthermore, the amount of compactness of each cluster can be evaluated by examining the *p* values within each cluster… ” (p. 19).

3. Data Extraction for Analysis

The first part of the experiment for this thesis focused on the acquisition and preprocessing of data of Mexican equity funds for its analysis. This process was performed in three phases:

1. Acquiring the names of the existing Mexican mutual funds registered (until October, 2015) at the National Banking and Stock Commission (**CNBV**, *Comisión Nacional Bancaria y de Valores*) official website.

2. Downloading historical prices for the selected equity funds.

3. Data preprocessing of the equity fund prices for the traditional risk fund analysis and for the proposed novelty statics analysis.

3.1. Mutual funds data extraction with Scrapy

As the first part of the data extraction process, it was indispensable to download the information of all the mutual funds registered for trading by the CNBV (fund type, CNBV classification, fund's start date, benchmark, and etcetera). While doing the initial research for Mexican mutual funds data sources, the CNBV did not offer an option or service to download information about the registered Mexican mutual funds for operation in Mexico at its website. Given this scenario, it was decided to develop a process or application to extract the needed information from the CNBV website. This type of data extraction process is called web scraping.

The technology selected to download the CNBV Mexican mutual funds information was **Scrapy 1.0.3**, an application framework for web scraping in **Python 2.7** used to write applications to extract structured information from webpages and websites.

3.1.1. Extracting the list of mutual funds

The Scrappy tutorials described in the blog entry *"The freedom to develop is priceless ...: Extracting data from web pages with scrapy" ("La libertad de desarrollar no tiene precio...: Extracción de datos de páginas web con scrapy")* and the official Scrapy documentation website "*Scrapy Tutorial — Scrappy 1.0.3 documentation*", were used as references for the development of the Scrapy project to search and download the mutual funds' information from the CNBV website.

The developed Scrapy project downloaded all the mutual funds' data in a JavaScript Object Notation (JSON) formatted text file, which was converted to a comma-separated values (CSV) formatted file and, then, to a Microsoft Office Open XML (XLSX) formatted file. This process made easier reviewing the funds' information and the definition of the selection criteria of the equity funds to analyze.

The reasons for choosing the Scrapy as the technology for extracting data were:

- Available official documentation, tutorials and existing user community forums where to search and ask for support.
- Reputation as one of the most complete set of libraries and utilities to create applications for data extraction from structured data sources.
- Opportunity to practice the development in the Python programming language and to learn to develop data mining process from structured data sources.

To develop the Scrapy framework project, the following Python libraries and components (**Table 3.1.**) were used:

Python 2.7 libraries and versions used in a Scrapy 1.0.3 project			
Component	**Version**	**Component**	**Version**
cffi	1.2.1	pycparser	2.14
characteristic	14.3.0	pyOpenSSL	0.15.1
cryptography	1.0.2	queuelib	1.4.2

cssselect	0.9.1	**Scrapy**	**1.0.3**
enum34	1.0.4	service-identity	14.0.0
idna	2.0	six	1.9.0
ipaddress	1.0.14	Twisted	15.4.0
lxml	3.4.4	w3lib	1.12.0
pyasn1	0.1.9	zope.interface	4.1.2
pyasn1-modules	0.0.8		

Table 3.1. List of libraries and components used in the mutual funds data extraction application.

Challenges

Several problems and challenges were discovered and solved during the data extraction application development and the information extraction tests from the CNBV website.

While analyzing the HTML code of the basic mutual fund search web page at the CNBV website, it was found that this search form is an embedded web page with an <IFRAME> HTML tag to the web address:

```
http://lt.morningstar.com/7ap7omrzjm/fundquickrank/default.aspx
```

This address implies that *the CNBV has leased the mutual fund search and information service to a private financial service provider:* ***Morningstar Mexico***. Due to this circumstance, this web address was used as the source for the mutual funds' information extraction. Also, the web address for each fund profile at the Morningstar Mexico web site was added to the list of data to extract, in case additional information was needed.

Technical problems

A. The first functional test of the Scrapy project, the console reading of the first page of the list of funds, uncover this error message:

```
"/usr/lib/python2.7/site-packages/boto/utils.py", line 210,
in retry_url r = opener.open(req, timeout=timeout)"
```

According with a proposed solution in the Stak Overflow portal, "*python - Scrapy gives URLError: <urlopen error timed out> - Stack Overflow*.", this instruction was added to the configuration file `settings.py` to fix this error:

```
DOWNLOAD_HANDLERS = {
     's3': None,
```

```
}
```

B. No other failures or errors were found during the first data extraction tests: downloading of the fund key name, its CNBV classification and fund's information webpage. A list with **2,837 registered funds, until October, 2015**, was gathered (see "Thesis Repository 00 - Thesis documentation, data and graphs" for the file with all the downloaded mutual funds information).

When the extraction process was modified to add the download and scraping of information for each fund information webpages, only 140 to 198 fund entries were downloaded, some funds were listed 2 or 3 times. The basic and advised design for data extraction from lists, such as online store catalogs, is by using recursive calls. By default, Scrapy executes web page requests in depth and then in width (see *"Frequently Asked Questions — Scrapy 1.0.3 documentation"*). The following changes (**Table 3.2.**) were applied and tested in the Scrapy project:

Problem or issue	Proposed or applied solution
At first, it was suspected that the CNBV server (the Morningstar Mexico subcontracted search form) was rejecting the sudden amount of webpage requests:	● In file `cnbvt01_spider.py`, the project's local configuration variable '**download_delay**' (waiting time in seconds between webpage requests) was increased. Tests with 1, 3 and 5 seconds were made; the requests were delayed, but the download information continued to be incomplete. ● In the `settings.py` file, other project global configuration variables that were changed from their default values were: ▪ '**CONCURRENT_REQUESTS**', to decrease and increase the concurrent number of petitions to the same website. ▪ '**REACTOR_THREADPOOL_MAXSIZE**', to increase the number of memory threads and ease the concurrent request processing. None of this configuration changes solved the incomplete data extraction problem.
By reviewing the request log, it was found that some requests for	Suspecting a request timeout error, the value of the configuration variable '**download_timeout**' was

funds' web pages were sent, but never returned.	increased from its default value of 180 seconds to 240 and 360. This change did not work either.
Another possibility was a memory management error.	• Following the instructions for debugging memory leaks in the official Scrapy documentation, *"Debugging memory leaks — Scrapy 1.0.3 documentation"*, the logic errors that were discovered were fixed. • Besides, the local configuration variable '**download_maxsize**' was set to 0, in order to eliminate the requested web page maximum size. The extraction process reduced its execution time, but the funds' information continued to be incomplete.
Additional exceptions to identify and exclude possible communication errors with the CNBV website were included in the request functions.	The most common HTTP error codes were included in the error code list to receive and process, including the Python libraries to catch them and a function to process every failed fund's web page request. The list of the HTTP error codes handled were: ▪ 500: Internal Server Error ▪ 501: Not Implemented ▪ 502: Bad Gateway ▪ 503: Service Unavailable ▪ 504: Gateway Timeout ▪ 505: HTTP Version Not Supported ▪ 400: Bad Request ▪ 401: Unauthorized (RFC 7235) ▪ 402: Payment Required ▪ 403: Forbidden ▪ 404: Not Found ▪ 405: Method Not Allowed ▪ 406: Not Acceptable ▪ 407: Proxy Authentication Required (RFC 7235) ▪ 408: Request Timeout ▪ 409: Conflict ▪ 410: Gone ▪ 411: Length Required ▪ 412: Precondition Failed (RFC 7232) ▪ 413: Payload Too Large (RFC 7231) ▪ 414: URI Too Long (RFC 7231) ▪ 415: Unsupported Media Type ▪ 416: Range Not Satisfiable (RFC 7233)

	■ 417: Expectation Failed However, because the failed webpage request never returned an error code that could be catch, the failed request process could never be executed. Also, the successful web page requests never returned an HTTP error code.
Insisting in a possible memory management failure, it was decided to change the project's web page request recursive design to an iterative design. Instead of ending in a recursive call to itself, until the link for the next list page is disabled, the function that loads every page with the funds list is called by an external iterative cycle inside the main parsing, which controls the amount of list pages (iterations) and the amount of fund's pages requests per page executed.	By reducing the number of iterations to 9 (meaning, by limiting the amount of requested list pages to only 9 at a time), the information of every fund listed in those 9 pages was downloaded without any error or any duplicated entry. In total: ■ 1 request was made by downloading the fund's search main page. ■ 9 requests for each list page. ■ 20 requests for each fund's web page with their data. This sum up to a total of 190 simultaneous requests each time the extraction process is executed. The default number of 20 funds per page could never be increased in the CNBV search page. Given the 2,837 funds registered at the time of the execution of this process, 142 pages with 20 listed funds, in sets of 9 pages, were downloaded.

Table 3.2. List of issues and errors found during the development and testing of the mutual funds data extraction application and their proposed solutions.

Problems with the official CNBV fund search website

A. All the search form elements call a JavaScript function that changes the value of the calling element and immediately executes a search with the new value. Due to lack of access to the library with the code of this JavaScript function, the number of listed funds per result page could not be changed. By default, *the basic fund searcher displays all registered funds in alphabetical order, in pages with 20 funds each*.

Web Scraping data from dynamically generated web pages is limited; this technology is focused to the gathering of information from massive static sources,

which do not include web applications or dynamically generated content, such as Macromedia Flash applications.

Several requests to the visible elements of the basic search page were tested, in order to narrow the downloaded list of funds to the preselected fund classifications and to reduce the total data extraction time:

- In the option to increase the number of listed funds per page, it was requested to list "500" funds per page. But the received web page always listed the default number of funds, **"20"** per page.

- Choosing a specific classification of the listed funds did not work. The default option was always selected: **"*All*"** ("*Todos*").

3.1.2. Selection criteria for funds

After obtaining, in parts, the list with the complete information of the 2,837 mutual funds registered in the CNBV, the data was reviewed and the criterion for the selection of the funds to include in the experiment was set.

The criterion for the selection of the equity funds to analyze was defined as follows (**Table 3.3.**):

Fund category	Only "Equity Funds" (*"Renta Variable"*)
CNBV category for funds	Only those that invest (all or part) in Mexican companies: ▪ "Mexican Equity Funds" (*"RV México"*) ▪ "Mixed Aggressive Equity Funds" (*"Mixtos Agresivos"*) ▪ "Mixed Moderated Equity Funds" (*"Mixtos Moderados"*) ▪ "Mixed Conservative Equity Funds" (*"Mixtos Conservadores"*)
Acquirer	All, except only "Moral Entities" (*"Personas Morales"*)
Benchmark	All, except: ▪ *"S&P Valmer MEX CETES 28D"* (allowing percent mixtures with other benchmarks, 40%, 60%, etc.) ▪ *"FTSE-PiPG Cetes 24hrs"* (allowing percent mixtures) ▪ *"FTSE-PiP Cetes 364d 24hrs"* (allowing percent mixtures)

Start date	Before 2010-12-31 (to ensure that the funds would have, at least, five years of recorded prices)

Table 3.3. Defined criteria for the selection of equity funds to analyze.

3.2. Equity funds historical price data downloading

The funds that comply with the selection criteria were 197. Because neither the CNBV nor the Mexican Stock Exchange (**BMV**, *Bolsa Mexicana de Valores*) supply free historic data of the funds traded, their information was obtained from the financial news service *Yahoo! Finanzas* (the Latin American Spanish version of the news portal **Yahoo! Finance**).

3.2.1. Historical price data download with a Scrapy application and a MongoDB database

The objective of the funds historical price data extractor was to download the prices of the selected funds and to save them in a NoSQL database, **MongoDB 2.4.14**, to perform the next data preprocessing phase. For this purpose, the tutorial "*Web Scraping with Scrapy and MongoDB - Real Python*" was used as reference to develop a Scrapy application to download the information.

The decision to store the historical prices in a NoSQL database was taken due to:

- The data management flexibility and data reading speed offered by NoSQL databases.
- For the opportunity to learn about data management in non-relational databases.

To program this data extractor, the following Python libraries and component (**Table 3.4.**) were used:

Python 2.7 libraries used in this project			
Component	**Version**	**Component**	**Version**
cffi	1.2.1	pycparser	2.14
characteristic	14.3.0	**pymongo**	**3.0.3**
cryptography	1.0.2	pyOpenSSL	0.15.1
cssselect	0.9.1	queuelib	1.4.2
enum34	1.0.4	**Scrapy**	**1.0.3**
idna	2.0	service-identity	14.0.0
ipaddress	1.0.14	six	1.9.0
lxml	3.4.4	Twisted	15.4.0
pyasn1	0.1.9	w3lib	1.12.0
pyasn1-modules	0.0.8	zope.interface	4.1.2

Table 3.4. List of libraries and components used in the first equity funds historical price data extraction application with Scrapy and MongoDB.

Challenges

Although, it could not be possible to identify and solve the missing price registries errors encountered while downloading price data (as will the explained below), the main issue found during the execution of the historical price extraction process was not technical, but the lack of a ***ticker symbol***[1] to identify each investment instrument (stocks, mutual funds, etc.) in the downloadable text files available at the *Yahoo! Finanzas* website.

Technical problems

A. The process log in console reported a successful data registry in the MongoDB database, but it was not accurate. As a test, the dates 2015-10-30, 2015-10-29 and 2015-10-16 were randomly selected to verify that all equity funds had a price entry on those dates; for the funds that did not have an entry in the database on any of those dates, the data download process was repeated. Unfortunately, there were funds where the first downloading process only registered 600 price registries and the second download added 500 more price registries. Even when the missing test dates prices were added, the complete historical price data download could not be assured.

1 Also known as stock symbol is an abbreviation used to uniquely identify publicly traded shares or titles of a particular stock or investment instrument on a particular stock market.

Due to the lack of error codes or messages, it was difficult to pinpoint if the exact source of the failure is due to an error in the Scrapy extraction process, a data transmission error or to an error in the setup and configuration of the MongoDB database.

Problems with the *Yahoo! Finanzas* website

A. The text file's data structure with the historical prices for all the investment securities available at this website has these fields:

```
Date,Open,High,Low,Close,Volume,Adj Close
```

This data structure does not include the name or ticker symbol of the financial instrument to which the data belongs to, because the text file with this information it's intended to be manually downloaded by a user that knows or is reading the instrument's ticker at the instrument's information page provided by *Yahoo! Finanzas*.

This situation makes the automatic downloading of multiple funds' historical prices difficult. An application developed with the Scrapy framework can extract information from webpages or from structured data files, but not from two different types of data sources at the same time. As a result, each file was processed semi-manually; the web address and fund's ticker was edited in the Scrapy project code for each fund, in order to download its historical prices. This extraction process took a lot of time to complete for all of the 197 selected funds (approximately 4 days).

B. Another source of fund identification issues was when the historical price data do not exist for an equity fund or a fund is registered under a different ticker symbol.

- **Nonexistent Funds.** 10 funds have an information page in the *Yahoo! Finanzas* website, but do not have any registered historical prices (for example, fund "F-INDIC B"). In some cases, partial information was available, but after a download retry no data was found (example, "GOLD3+ B2-A").

- **Funds with different ticker symbols.** In some cases, a fund is registered in the *Yahoo! Finanzas* website under a different ticker than the name it uses in the

CNBV website (fund "SELECTC B1" is registered with the ticker "SELECT2B1.MX"). To obtain their historic data, a manual search was required to verify their existence at the *Yahoo! Finanzas* web portal.

- **Funds with historic data that starts years after their registered start date.** In these cases, historic price data was available, but not for all the years that the fund has been in operation (fund "SCOTIPC L" started in 1987, but only has historic data since 2000).

- **Funds with only a year of data (2015).** 5 funds were excluded due to incomplete historic price information (for example, fund "ST&ERBM F", which started in 2007, but only has historic price data since June, 2015).

Due to the issues presented, the use of Scrapy was considered inadequate for the retrieval of the funds prices from this source.

3.2.2. Historical price data download with Python yahoo-finance library

Instead of using a Scrapy framework project, it was decided to use the Python 2.7 library **yahoo-finance 1.2.1**. A Python language script was developed to download the selected funds' historical price data in a CSV text file, which was manually uploaded to a **MariaDB 10.0.25** database to continue with the data preprocessing phase.

The **database management system MariaDB** is a relational database management system (RDMS) forked from MySQL, which offers the advantage to allow the programming of functions and sets of instructions, called *stored procedures*, that execute complex sets of operations requiring frequent access to information in a database. Because MariaDB is developed from MySQL's same source code, both RDMS still share

many functionalities, have almost identical capabilities and use the same sequential query language (SQL) version to program functions, stored procedures and scripts.

The reasons to store the downloaded historical prices in a MariaDB database were:

- Previous experience with MySQL and other RDMS.
- Availability of official documentation and large user community forums where to ask for help and support to expected memory limitations and problems when using dynamic tables in MySQL and, by extension, in MariaDB.

The following Python libraries (**Table 3.5.**) were used to create the script to download the selected equity funds' historical price data:

Python 2.7 components used in this project			
Component	**Version**	**Component**	**Version**
simplejson	3.8.1	**yahoo-finance**	**1.2.1**

Table 3.5. List of libraries and components used in the second equity funds historical price data extraction application with the Python yahoo-finance library and MariaDB.

Challenges

The Python script code included exceptions to identify and report errors during the download and extraction data process, but there were not severe issues to solve.

Technical problems

A. In order to download the information of a financial instrument, a '*Share*' from the library yahoo-finance must be created with the ticker symbol of the instrument. This object's methods make the necessary calls to the Yahoo! Finance web services to allow the download of the requested information. The creation of a '*Share*' object and methods' calls are made inside blocks of `try: ... except JSONDecodeError: ...`, in order to catch and easily recover from errors. During testing, the most common error was 'JSONDecodeError', due to incomplete acquired data in JSON format. All funds that reported this error were registered in an

error log file to be reprocessed later. In all the logged cases, the historical price data was downloaded after the second or third retry.

Problems with the *Yahoo! Finanzas* website

A. The most recurring issues were network communication errors with the *Yahoo! Finanzas* website and its web services. '*Share*' objects and their methods' calls were not completed or the data was received incomplete.

Most of these issues were solved by setting pauses of 1 second after a '*Share*' object creation and a call to a method and a pause of 8 seconds between each fund information request.

B. As with the web scraping extraction process, several problems with the funds' ticker symbols were encountered: nonexistent funds, funds with different ticker names, funds with price data years after their start date and funds with only a year of price data (2015).

Also, **funds with an '&' in their ticker symbols** (like "ST&ER-D B1") **could not be downloaded automatically**. Due to an error in the Yahoo! API query language, **yql**, that does not allows it to correctly parse an '&' inside a query and to the lack of proper handling of tickers with this symbol in the **yahoo-finance** library code. Because there were only 4 funds in this situation, their historic prices were downloaded manually.

3.2.3. Additional data downloaded

Besides the described process it was necessary to manually download additional data:

- The **historic prices of the BMV index** (ticker symbol "BOLSA A") from 2008-01-01 to 2015-12-31 were downloaded. The BMV prices series was used as benchmark for most of the **Modern Portfolio Theory** (MPT) statistics in traditional risk analysis.

- In order to calculate some of the traditional fund risk analysis metrics, the yield rates percentages of the **364 Days Treasury Certificates**, or Bills 364-days (**CETES364D**, *Certificados de la Tesorería a 364 días*), were downloaded from the **Bank of Mexico** (**Banxico**, *Banco de México*) Economic Information System.

- The first historical prices data download was stopped on 2015-11-27. Because the most recent date with prices in all the selected funds is 2010-12-31, in January, 2016, it was decided to download the remaining December, 2015, data in order to complete 5 years of historical prices.

- During this complementary data download, the fund "BMERPAT F" was discarded because its price recording stopped in 2015-10-16 and there was not any new price registry until January, 2016. This reduced the list of **equity funds to analyze** to **182**.

3.3. Historical price data preprocessing

In order to execute the next experiment phase, the downloaded historical data had to be preprocessed. The preprocessing phase consisted of four steps:

1. **Data interpolation.** Before 2012-07-18, most of the funds report a final price per month. After this month most funds report a daily price, with the exception of Mexican holidays and weekends; some funds have breaks of several days or weeks after this date or only start to report on a daily basis on 2013. It is unknown if this price reporting irregularities are due to a change in the Mexican Banking and Investment laws or if this is because *Yahoo! Finanzas* limits the amount of free and public historical information. In order to have an homogeneous data set, the price series of the 182 funds, the index of the BMV, from 2010-01-01, and the yield rates of CETES364D were linearly interpolated.

2. **Data normalization.** The interpolated data series from each fund and the BMV index was normalized. Due to being the first date with price data in all the funds, the

date 2010-12-31 was selected to seed the normalization process for each data series. This allowed the experiments to be executed with 5 years of price data.

3. **Yearly return rates.** To calculate the MPT statistics, it is necessary to have each fund's yield rate, a benchmark yield rate (BMV index) and a free risk yield. The fund's and BMV annual yield rates were calculated using the instruments' normalized price data from 2010-12-31 to 2015-12-31, which produced 4 years of annual-rolling daily rates.

4. **Free risk rate adjustment.** As an alternative to building the Mexican yield curve, the CETES364D annual yield rate was chosen as the risk-free rate. Because the debt return rates are set at the start date of each debt instrument, the date of each rate was offset 364 days, to its maturity date, in order to be compared with the return rates on that date of the funds and benchmark.

As stated before, after uploading the gathered historical price data, the data preprocess was performed using a MariaDB 10.0.25 relational database. For the purposes and requirements of price data storage and price data preprocessing, MariaDB 10.0.25 and MySQL 5.5 can be considered as interchangeable platforms.

3.3.1. Data interpolation

In the finance industry, it is standard practice that all time series of data only include work or labor days (Mondays to Fridays, even holidays) of the calendar and exclude weekends (Saturdays and Sundays).

The linear interpolation process was divided in three parts:

1. Copying the downloaded dates with price data.

2. Generating sets of dates (weekdays) between the downloaded time periods of each equity fund.

3. Interpolating prices for the new dates between existing prices.

- Or to copy the price of the previous valid date, if it is only a day between two days with prices (a holiday).

On a XY coordinate plane, given the known points (x_1, y_1) and (x_3, y_3) and the value x_2, the formula to find the unknown value y_2 is:

$$y_2 = \frac{(x_2 - x_1)(y_3 - y_1)}{(x_3 - x_1)} + y_1$$

Formula 3.1. Linear interpolation formula used in the historical price data interpolation.

For example, the linear interpolation for the equity fund "IXE1 BI", from 2012-03-30 to 2012-04-30, consisted in:

1. Copying the original dates and prices from table prices to table '`prices`' to table '`prices_interpolated`'.
2. The weekdays between 2012-03-30 and 2012-04-30 were calculated and added to table '`prices_interpolated`'.
3. The interpolated prices between the dates were calculated and updated to their corresponding dates.

The results can be seen next, in **Figure 3.1.**:

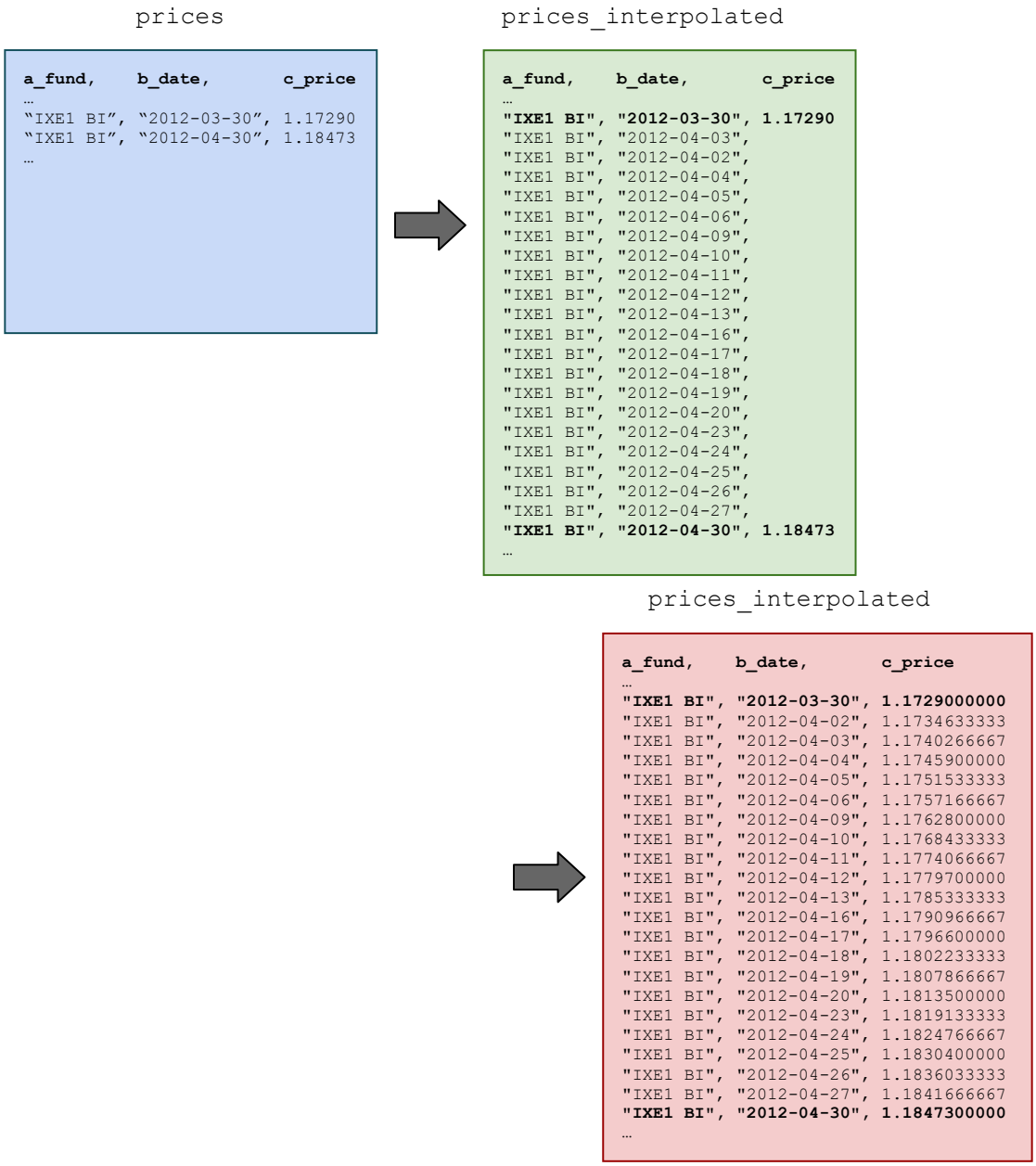


Figure 3.1. Example of the interpolation process for missing weekdays’ price data in equity fund “IXE1 BI”.

Challenges

Most of the problems consisted of time execution of dynamic queries and the memory intensive use of cursors while reading tables in MariaDB databases. All these problems were solved, but the total time required to test, find and fix logic errors was considerable (around six days).

Technical problems

1. In order to keep a backup of the original historical price data, the interpolated data is saved in a different table: ‘`prices_interpolated`’ for the funds, '`market_interpolated`' for the BMV index and ‘`market_rrates_interpolated`’ for the risk-free yield security CETED364D. *To reduce the time to generate valid dates to interpolate, the process is divided in two parts*:
   - To copy the original downloaded price data into the respective table ‘`<TABLE>_interpolated`’.
   - For each fund, the prices dates are read in ascending order and the valid dates (workdays) between each date range are registered.

   Note: In the stored procedures code, the automatic 'COMMIT' was disabled. With these instructions, the disk writing operations (INSERT and UPDATE sentences) are reduced to only one, at the end of the procedures' code, which reduced the number of disk access and the total execution time of the weekdays calculation subprocess.

2. Because the dates between the time periods to interpolate are not consecutive (due to the exclusion of the weekend dates), it was decided to use ***dynamic tables*** between the existing time periods to calculate the row number.

   Neither MariaDB nor MySQL include native functions to number or identify the row number in a query result set (unlike other RDMS like MSSQL Server Database and Oracle Database). Given this constraint, the position of a date in a search result (the value for the ‘x_2’ variable in the interpolation formula) must be calculated with the

support of a temporary variable. Queries that use dynamic tables take too much time to execute, but this was the most viable solution found during the development of this process.

Also, the automatic 'COMMIT' was disabled to help reduce the disk writing operations each time a price interpolation operation is performed.

The linear interpolation subprocess generated 261 entries of interpolated price data per year for every equity fund, index and debt instrument for further analysis. Overall, the price data interpolation took a little bit over 120,006 seconds (approximately, 1 day, 9 hours and 21 minutes) to be executed.

3.3.2. Data normalization

As stated before, the latest common date with price data in all the funds is 2010-12-31. The normalization process for each fund begun on that date by dividing each following price by the price in 2010-12-31.

All the normalized prices were saved in a table called '`prices_normalized`', for the funds, and '`market_normalized`' for the BMV index.

For example, in **Figure 3.2.**, the interpolated data normalization for the fund "MAYA B1", from 2010-12-31 to 2011-01-31, consisted in reading its price in 2010-12-31 (25.5628 MXN) from table '`prices_interpolated`', dividing the following dates' prices by 25.5628 MXN, and saving its dates and results in table '`prices_normalized`':

prices_interpolated

a_fund,	b_date,	c_price
…		
"MAYA B1",	"2010-12-31",	25.5628000000
"MAYA B1",	"2011-01-03",	25.5252380952
"MAYA B1",	"2011-01-04",	25.4876761905
"MAYA B1",	"2011-01-05",	25.4501142857
"MAYA B1",	"2011-01-06",	25.4125523810
"MAYA B1",	"2011-01-07",	25.3749904762
"MAYA B1",	"2011-01-10",	25.3374285714
"MAYA B1",	"2011-01-11",	25.2998666667
"MAYA B1",	"2011-01-12",	25.2623047619
"MAYA B1",	"2011-01-13",	25.2247428571
"MAYA B1",	"2011-01-14",	25.1871809524
"MAYA B1",	"2011-01-17",	25.1496190476
"MAYA B1",	"2011-01-18",	25.1120571429
"MAYA B1",	"2011-01-19",	25.0744952381
"MAYA B1",	"2011-01-20",	25.0369333333
"MAYA B1",	"2011-01-21",	24.9993714286
"MAYA B1",	"2011-01-24",	24.9618095238
"MAYA B1",	"2011-01-25",	24.9242476190
"MAYA B1",	"2011-01-26",	24.8866857143
"MAYA B1",	"2011-01-27",	24.8491238095
"MAYA B1",	"2011-01-28",	24.8115619048
"MAYA B1",	"2011-01-31",	24.7740000000
…		

prices_normalized

a_fund,	b_date,	c_price
"MAYA B1",	**"2010-12-31",**	**1.000000000000**
"MAYA B1",	"2011-01-03",	0.998530602876
"MAYA B1",	"2011-01-04",	0.997061205756
"MAYA B1",	"2011-01-05",	0.995591808632
"MAYA B1",	"2011-01-06",	0.994122411512
"MAYA B1",	"2011-01-07",	0.992653014388
"MAYA B1",	"2011-01-10",	0.991183617264
"MAYA B1",	"2011-01-11",	0.989714220144
"MAYA B1",	"2011-01-12",	0.988244823020
"MAYA B1",	"2011-01-13",	0.986775425896
"MAYA B1",	"2011-01-14",	0.985306028776
"MAYA B1",	"2011-01-17",	0.983836631652
"MAYA B1",	"2011-01-18",	0.982367234532
"MAYA B1",	"2011-01-19",	0.980897837408
"MAYA B1",	"2011-01-20",	0.979428440284
"MAYA B1",	"2011-01-21",	0.977959043164
"MAYA B1",	"2011-01-24",	0.976489646040
"MAYA B1",	"2011-01-25",	0.975020248916
"MAYA B1",	"2011-01-26",	0.973550851796
"MAYA B1",	"2011-01-27",	0.972081454672
"MAYA B1",	"2011-01-28",	0.970612057552
"MAYA B1",	"2011-01-31",	0.969142660428
…		

Figure 3.2. Example of the normalization process in equity fund "MAYA B1".

Challenges

There was not any difficult or specific challenge, problem or error during the development of this preprocessing step. The normalization of the prices series did not require the use of complex calculation, only arithmetic operations.

The automatic 'COMMIT' was disabled to help to reduce the disk writing operations time in the store procedure that executes the data normalization.

Beginning in 2010-12-31 and ending in 2015-12-31, each fund and index generated 1,305 entries of price data, 261 entries per workday per year for analysis, and took 6 seconds to be executed.

3.3.3. Yearly return rates

As explained, to calculate the MPT statistics **alpha**, **beta**, **r squared** and **Sharpe ratio**, an investment instrument, its benchmark and a risk-free security return rates are required. To obtain these yield rates, the annual yield rate for the chosen equity funds for analysis and the BMV index must be calculated.

The process to calculate the annual yield rates from the normalized price data is the following:

- A fund's or index's price is divided between its previous year price and subtracted 1.
    - If the previous year a date was a Saturday, then, the price of the closest workday's price available (the previous Friday) is used (a year and a day back).
    - Similarly, if the previous year a date was a Sunday, then the price of the immediately previous Friday is used (a year and two days back).

The calculated annual yield rates were saved in a table called '`prices_return_rates`', for the funds, and '`market_return_rates`' for the BMV index and benchmark.

For example, for fund "PRINRVA FA" the annual yield rate of 2012-01-02 is calculated by dividing its normalized price (0.963166873075 MXN) by the normalized price of 2010-12-31 (1 MXN) and subtracting 1. In **Figure 3.3.**, the calculated annual return rates for the period of 2012-01-02 to 2012-01-31 are displayed:

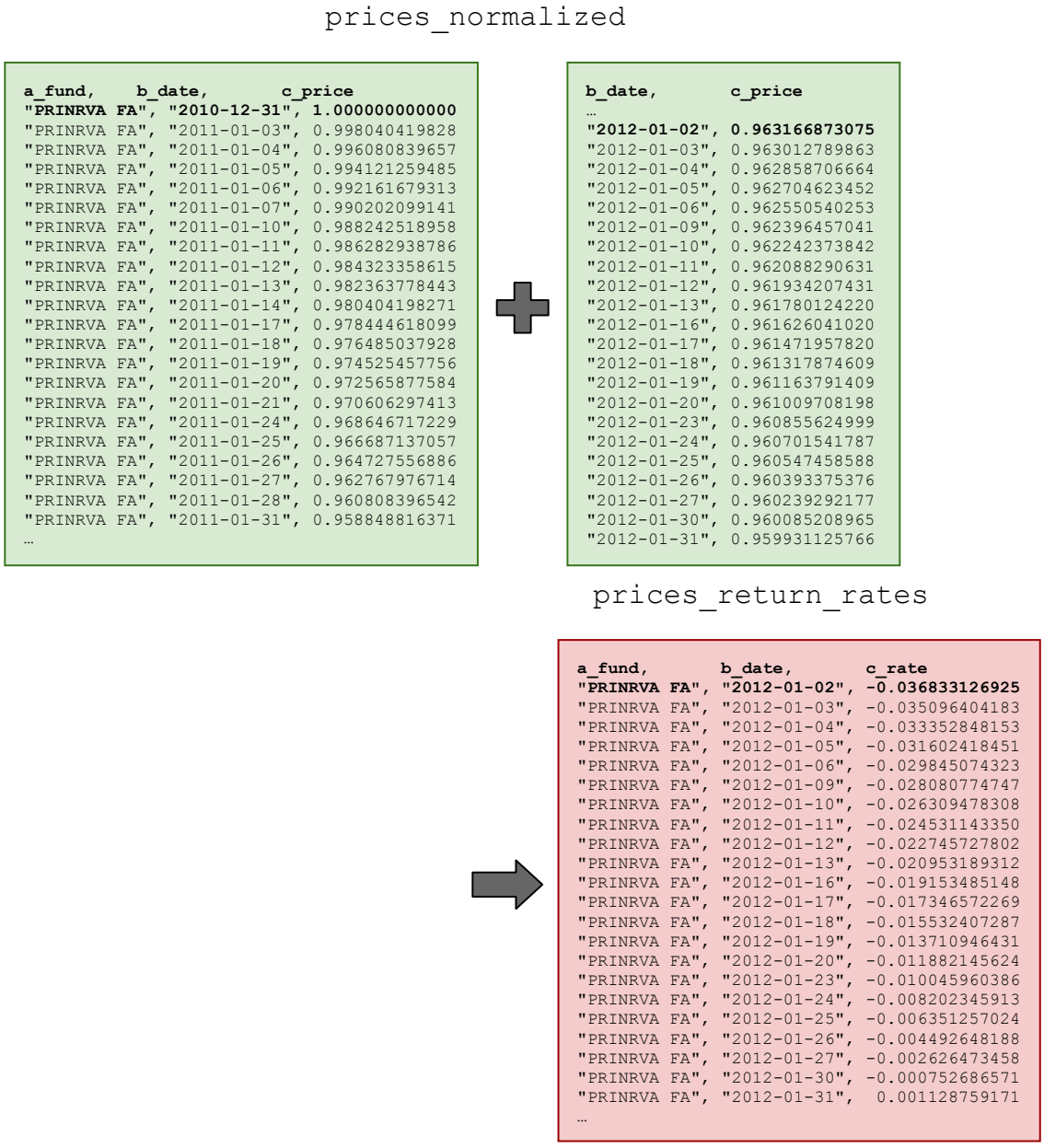


Figure 3.3. Example of the annual return rate calculation in equity fund "PRINRVA FA".

Challenges

This step did not incur in any technical challenges. The only important consideration was validating if the previous year normalized price does not exist, the price of the closest workday should be used.

Also, the automatic 'COMMIT' was disabled at the end of the store procedure code to help reduce the disk writing operations and reduce the overall execution time. The total execution time was 8 seconds.

The yearly yield rates were calculated from 2012-01-02 to 2015-12-31, which resulted in a total of 4 years of annual yield rates and internal rate of return rates to use in the analysis.

3.3.4. Free risk rate adjustment

In order to obtain the risk-free security yield or rate of return it was decided to use the CETES364D annual yield rate.

The rate adjustment consisted in subtracting 364 days to each annual yield rate in order to find the yield rate at the end of the instrument's term. Because Banxico reports all debt securities return rates in percentages, the CETES364D rates had to be divided by 100 in order to be compared them with the calculated equity funds annual rates.

As explained, the calculated annual return rates of the funds and BMV index are the return rates at the end of an investment in the instrument. The yield rates for government securities are compromised at the beginning of the investment term. To be compared, the maturity date for a CETES364D instrument must equal the date of the end rate for the calculation of the rate of return of a fund and the BMV index.

The adjusted CETES364D yield returns were saved in a table named '`market_rrates_return_rates`'.

Using this procedure, the yield returns from 2012-01-02 to 2012-01-31 are the dates of the previous 364 interpolated yield rates, the results are the next adjusted yield rates, as shown in **Figure 3.4.**:

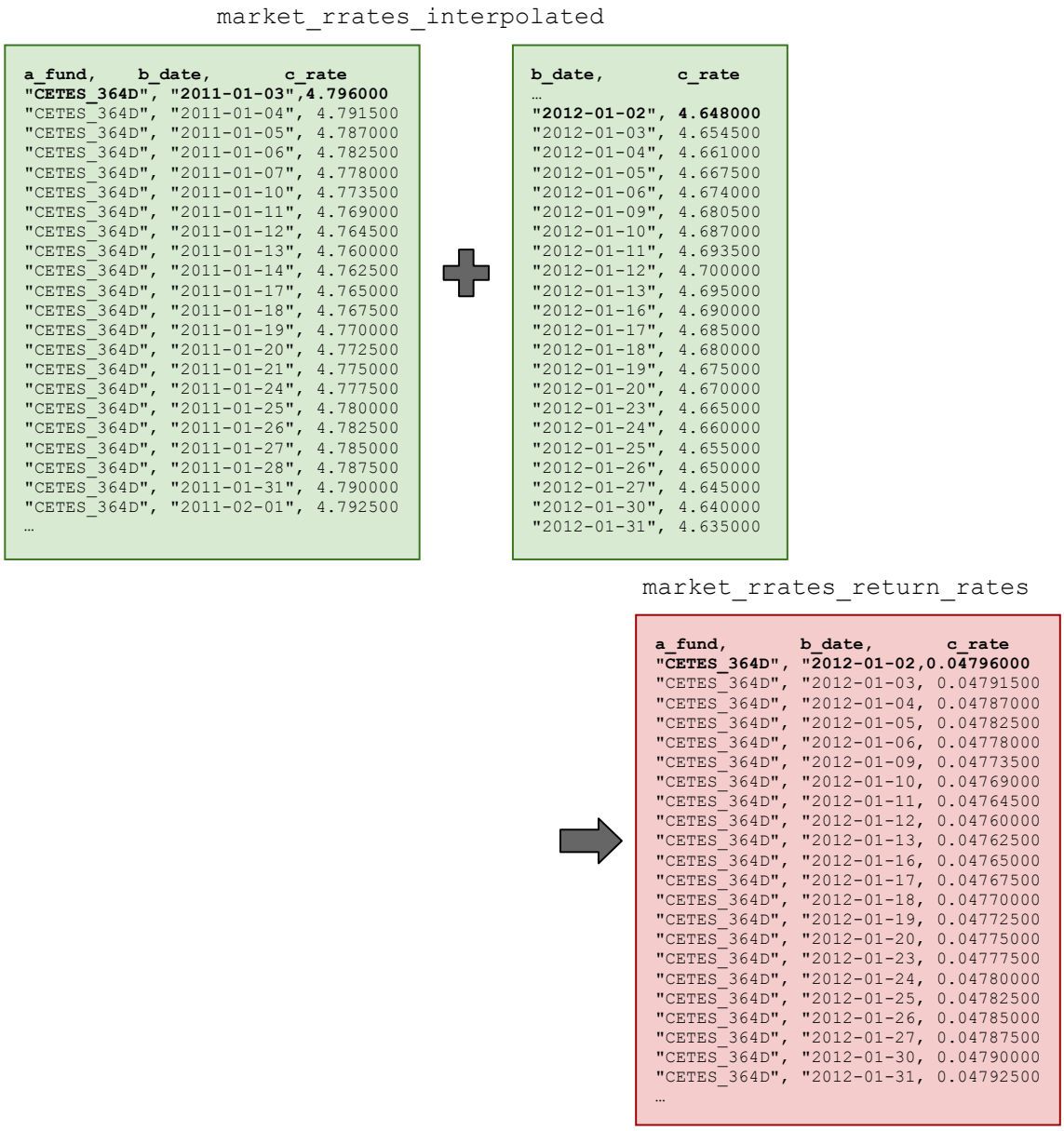


Figure 3.4. Example of the date adjustment for the risk-free annual yield rates.

Challenges

This step did not incur in any technical challenges. The most important consideration was to validate if the subtracted dates were valid workdays and to calculate the alternative dates if the subtracted dates were not valid.

As in the other subprocess, the automatic 'COMMIT' was disabled at the end of the store procedure code to help to reduce the disk writing operations and reduce the overall execution time (2 seconds).

4. Traditional Financial Analysis of Equity Funds

The objective of the second part of the experiment was to calculate the five **Modern Portfolio Theory** (MPT) statistical measures to compare it with the clustering analysis, which is described in the next chapter.

The MPT measures of the selected 182 equity funds are calculated with the yearly yield rates from the previous preprocessing phase. To calculate this financial measures, the yearly rates of the **Mexican Stock Exchange** (BMV, *Bolsa Mexicana de Valores*) index are used as the benchmark, and the yearly rates **364 Days Treasury Certificates** (**CETES364D**, *Certificados de la Tesorería a 364 días*) are used as the risk-free security or asset.

4.1. Calculation of the five Modern Portfolio Theory statistics

To continue with the next phases of the experiment, the computer language **R 3.2.3** (2015-12-10) *"Wooden Christmas-Tree"* was chosen. The reasons behind this decision were:

- Includes built-in functions to perform traditional statistical analysis (linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, etc.) and data graphical display.
- Provides an environment that allows the development of new functions and an archive of third-party developed packages with functions and data structures for field specific data analysis (such as finance and economics).
- Availability of official manuals from the R Project members and contributors; and statistics and data analysis tutorials with examples in the language.
- To learn the R programming language and its widely announced advantages among the data analysis community.

To calculate the MPT statistics for the 182 equity funds, the produced code includes the package **PerformanceAnalytics 1.4.3541**, a collection of econometric functions for performance and risk analysis aimed at the research analysis of non-normal return streams of investment instruments (Peterson & Carl, 2016), and the function `lm` from the **stats 3.2.3** base package, to calculate the linear integration of each fund and obtain the r squared measure.

4.2. Analysis of the results

The annual rates of return used to calculate the MPT measures covers the period from 2012-01-02 to 2015-12-31. Their results are discussed and explained below.

4.2.1. Standard deviation results

The **PerformanceAnalytics** package function `StdDev` was used to calculate the standard deviation (**σ**) statistic. The implemented instruction was:

```
StdDev(<xts_timeseries>[,1], portfolio_method="single")
```

where the object *<xts_timeseries>[,1]* has a fund's return rates to analyze and the parameter `portfolio_method="single"` means that the calculation is for an univariate set of data (only one column or list of data). A few of the funds' calculated **σ** are listed in **Table 4.1.**

Fund	Standard deviation (σ)
DIVER-M GB	0.015711
ELITE-C B1	0.008203
ELITE-M B1	0.017188
FONBNM B3-A	0.076475
FONBNM C0-A	0.079747

Table 4.1. A sample of calculated funds' standard deviations.

The calculated **σ** of the BMV index was ***0.191691***. This means that the returns' volatility of the Mexican Stock Exchange during that period of time was of 19.1691%.

The equity funds' standard deviation (volatility) is close or lower than that of the BMV index; however, other funds experienced volatilities higher than those of the BMV index. While a couple of funds, "BNMPAT B3-A" (**σ** = 0.219873) and "ACTINTK FF" (**σ** = 0.222942), had volatilities over the BMV index **σ**, their differences with the BMV index are low (0.028182 and 0.031251 percent points, respectively) and could be inferred to be produced by each funds' unsystematic risk or to their investment management companies goals and investment strategies. On the contrary, funds "**BNMPAT B2-A**" (**σ** = 0.281213), "**BNMPAT C0-C**" (**σ** = 0.393606) and "**BNMPAT M0-A**" (**σ** = 0.419677), reported very high volatilities. Fund "BNMPAT B2-A" has also a difference (0.089522 percent points) that could be explained by its unsystematic risk and investment style; but, the differences of funds "BNMPAT C0-C" (0.201915 percent points) and "BNMPAT M0-A" (0.227986 percent points), could be due to an error in their historical price data download.

The historical price series of fund "BNMPAT C0-C" report lower price values at, apparently, random days between 2012-10-18 and 2015-12-10, period during which the prices fluctuated between 4.49813 and 5.85756 Mexican Pesos (**MXN**). The dates in question are listed in **Table 4.2.**

date	price		date	price		date	price
2012-10-17	4.84367		2013-04-16	4.90223		2013-06-28	4.69895
2012-10-18	1.20218		2013-04-17	1.25006		2013-07-01	1.2001
2012-10-19	4.8532		2013-04-18	4.86243		2013-07-02	4.86264
...	...		...	...		...	...
2013-08-01	4.90503		2013-08-27	4.88138		2013-09-09	4.81027
2013-08-02	1.25621		2013-08-28	1.20088		2013-09-10	1.23842
2013-08-05	5.03517		2013-08-29	4.73574		2013-09-11	5.02531
...	...		...	...		...	...
2013-10-01	4.8602		2013-11-27	5.0184		2014-01-10	5.15673
2013-10-02	1.25324		2013-11-28	1.27584		2014-01-13	1.31363
2013-10-03	4.98469		2013-11-29	5.15855		2014-01-14	5.23033
...	...		...	...		...	...
2014-02-25	4.9555		2014-03-07	4.88109		2014-03-20	4.81701
2014-02-26	1.21791		2014-03-10	1.21295		2014-03-21	1.22611
2014-02-27	4.81455		2014-03-11	4.81177		2014-03-24	4.96306
...	...		...	...		...	...
2014-12-01	5.51826		2015-01-02	5.35904		2015-02-05	5.14511

2014-12-02	1.3628		2015-01-05	1.31054		2015-02-06	1.3088
2014-12-03	5.36327		2015-01-06	1.27552		2015-02-09	5.29552
			2015-01-07	5.12556			
...	...		...	...		...	...
2015-03-27	5.3956		2015-09-21	5.45062		2015-12-09	5.36837
2015-03-30	1.36102		2015-09-22	1.36046		2015-12-10	1.34195
2015-03-31	5.48578		2015-09-23	5.41634		2015-12-11	5.3948

Table 4.2. Example of a fund's prices that inexplicable change consecutive between dates and cause to calculate an unusual high standard deviation or volatility.

As can be read in **Table 4.2.**, prices for "BNMPAT C0-C" drop from one day to another in 3 to 4 MXN and raise back to a price closer to their previous value the following day, in 19 different instances. Upon reviewing the original raw text files with the downloaded historical prices from the **Yahoo! Finanzas** web portal, it was found that these were the exact prices for the conflicting dates; therefore, errors during the preprocessing phases before calculating the return rates were dismissed.

Unfortunately, this case implies that the source of the mistaken prices could come from:

- Yahoo! Finance web portal itself.
- Yahoo! Finance fund's information providers, like **Morningstar, Inc.**, **Capital IQ** and **Thomson Financial Network** for financial information outside the U.S.A. (Yahoo - News Network, n.d.).
- Funds' investment management companies themselves.
- Errors during the data transmission process between them.

Whatever the source of the faulty data, it's not practical to manually review all the historical prices data to search for more unusual changes in price series behavior to exclude a fund from the experiment. As explained by the definitions of the types of risk, exceptional high fluctuations in funds' prices do happen.

Out of all the funds with a volatility higher than the BMV index's, other funds that have the same type of data error source are "BNMPAT B3-A" (with 4 dates with a sudden drop in price, compared to the previous and following dates), "BNMPAT B2-A" (11 dates with a price drop) and "BNMPAT M0-A" (657 dates with a price drop, almost all

2013, 2014 and 2015 dates). It's important to clarify that a price drop that occurs during a long period of time, almost 3 years, could be considered a normal price behavior explained by usual changes in the market; however, during the period of time covered by those 657 dates, there were records of prices with similar behavior as the prices reported before these price drops (in 2014-09-17, a price of 5.89861 MXN; 2015-05-07 with 5.70744 MXN and 2015-08-28 with 5.53369 MXN), which indicates that the lower prices could be due to data corruption.

Meanwhile, equity fund "ACTINTK FF" has a different explanation for the source of its high volatility. It's last 2010 reported price was on 2010-12-31 (at 0.35297 MXN); it's next one, was reported on 2011-01-31 (0.33995 MXN); it has no records for 2012, until 2013-12-30 (0.051857 MXN), when it continued its daily price reports (Yahoo - News Network, "ACTINTKFF.MX Gráfico básico | Valores de ACTINTK FF - Yahoo Finanzas", n.d.). According to its webpage at the CNBV fund searcher, this fund suspended its operations, and price reporting, for a non-specified period of time (and for not specified reasons) ("Informe de fondos. ACTINTK FF. Fondo Técnico Actinver SA de CV S.I.R.V. FF.", n.d.). The prices linear interpolation preprocess filled in the prices for the fund's missing dates; but the sudden price drop between 2011-01-31 and 2013-12-30 is the sole source of the presumed fund's high volatility. Whatever the reasons behind its suspension, this fund's sharp price drop was not an error, but a result of its unsystematic risk; therefore, its historical price data and return rates are valid.

Given their questionable price data quality, the funds "BNMPAT C0-C" and "BNMPAT M0-A" are not mentioned again in the next statics results. Funds "BNMPAT B3-A" and "BNMPAT B2-A" are included (due to their low data error rate, compared to the other funds), but their results are not taken as being very reliable.

4.2.2. Sharpe ratio results

This static is implemented with the function `SharpeRatio` with the parameters:

```
SharpeRatio(<xts_timeseries_fund>[, 1],
Rf=<xts_timeseries_rfr>[, 1], FUN="StdDev")
```

where the object `<xts_timeseries_fund>[, 1]` has a fund's return rates to analyze, object `<xts_timeseries_rfr>[, 1]` has the risk-free security CETES364D returns over the same period of time and the parameter `FUN="StdDev"` indicates which value must be used as the denominator for the Sharpe ratio calculation, the standard deviation. Some of the funds' Sharpe ratios are listed in **Table 4.3.**

Fund	Sharpe ratio with risk-free returns CETES364D
GBMPMOD B	0.266534
GBMV1 BO	1.142726
GOLD3+ B1-C	0.160107
HSBC-F2 BFP	-0.447478
HSBC-F2 BFV	-0.398507

Table 4.3. A sample of funds' Sharpe ratios.

The BMV index has a Sharpe ratio of ***0.076438***, while the lowest ratio was fund "DIVER-C MB", with -1.760450, and the highest "PRINLS2 FA", with a ratio of 1.879434. Overall, 58 funds reported a negative ratio (32% of the funds) and 124 a positive ratio (68% of the funds). ***According to the Sharpe ratio definition, 112 funds proved to be a better investment than the BMB index itself during the analyzed period of time.***

4.2.3. Beta results

This beta static is implemented with the function `CAPM.beta` and parameters:

```
CAPM.beta(<xts_timeseries_fund>[,1],
<xts_timeseries_i_bmv>[,1], Rf=<xts_timeseries_rfr>[,1])
```

where object `<xts_timeseries_fund>[,1]` has a fund's return rates to analyze, object `<xts_timeseries_i_bmv>[,1]` has BMV index's returns to compare the fund's to, and object `<xts_timeseries_rfr>[,1]` has the risk-free security CETES364D returns over the same period of time as the fund and the index's. Some of the funds' beta values are listed in **Table 4.4.**

Fund	Beta against BMV index with risk-free returns CETES364D
PRINRVA XB	0.293286
PRINRVA XC	0.248870
PROF-1A B	0.035594
PROF-3A B	0.055599
SBMIX B	0.008346

Table 4.4. A sample of funds' beta statistic.

This static is produced from a linear regression model of each fund against the BMV index; in the case of the BMV index, it means that it was regressed, compared, against itself and produced a beta of 1. ***All the calculated beta values for the funds were lower than the BMV, meaning that none of the funds outperforms BMV index when the market is up and are better prepared to face downs in the market.***

The lowest beta is fund "SELECTD B1", with -0.1324; and the closest to the BMV index is fund "NUMC B0-B", with a beta of 0.573768. In total, 19 funds have negative beta values and 163 have values above zero. According to the definition of beta, a negative beta value could imply that the fund's goal is to report (positive) growth during periods of time when the market is down. Nevertheless, none of the 163 funds with positive betas completely behave as the BMV index; this might be due to investments made to protect them from the downturns in the value of the index.

4.2.4. Alpha results

The alpha calculation is implemented with the function `CAPM.alpha` and parameters:

```
CAPM.alpha(<xts_timeseries_fund>[,1],
<xts_timeseries_i_bmv>[,1], Rf=<xts_timeseries_rfr>[,1])
```

where object *<xts_timeseries_fund>*`[,1]` has a fund's return rates to be analyzed, object *<xts_timeseries_i_bmv>*`[,1]` has BMV index's returns to compare the fund's to, and object *<xts_timeseries_rfr>*`[,1]` has the risk-free security CETES364D returns. A few of the calculated alphas are in **Table 4.5.**

Fund	Alpha against BMV index with risk-free returns CETES364D
IXEESP BF1	0.088918
IXEESP BF2	0.102165
IXEESP BI	0.120973
MAYA B1	0.006954
MAYA B2	0.015233

Table 4.5. A sample of funds' alpha metric.

As explained in the results of the beta static, the linear regression of the BMV index against itself calculated an alpha of ***-3.43605400016196e^{-18}***, a very small value that can be interpreted as 0. The lowest alpha is -0.361274 for fund "ACTINTK FF", and the highest is 0.140463 for fund "VALUEV6 B".

In total, 66 funds report a negative value for their alphas, and 116 have positive values. This means, that the managers of ***those 116 funds included investments that produced returns above the index***.

4.2.5. R squared results

As mentioned, r squared is not implemented in the PerformanceAnalytics package. To calculate it, it is necessary to make a linear regression with the excess returns of the funds and the BMV index with the function `lm`, from the **stats** package.

```
<xts_timeseries_excess-fund> <- Return.excess(
     <xts_timeseries_fund>[,1],
     <xts_timeseries_rfr>[,1]
     )
<xts_timeseries_excess-i_bmv> <- Return.excess(
     <xts_timeseries_i_bmv>[,1],
     <xts_timeseries_rfr>[,1]
     )
<df_excess-data> <- as.data.frame(
     na.omit(
          cbind(
               <xts_timeseries_excess-i_fund>,
               <xts_timeseries_excess-i_bmv>
               )
          )
     )
names(<df_excess-data>) <- c("a_found", "b_benchmark")
```

```
<df_excess-data>.lm <- lm(
     a_found ~ b_benchmark,
     data=<df_excess-data>
     )
summary(<df_excess-data>.lm)$r.squared
```

Where:

`<xts_timeseries_fund>[,1]` has a fund's return rates.

`<xts_timeseries_i_bmv>[,1]` has the BMV index's returns.

`<xts_timeseries_rfr>[,1]` has the risk-free security CETES364D rates.

`<xts_timeseries_excess-fund>` has the fund's excess returns.

`<xts_timeseries_excess-i_bmv>` has the BMV index's excess returns.

`<df_excess-data>` has the joined excess returns of the fund and the BMV index in column format.

`<df_excess-data>.lm` has the results of the linear interpolation (slope, coefficients, error, interception, etc.) between the excess returns of the fund and the BMV index, including the r squared.

The BMV index, as expected, has an r squared of 1. The lowest r squared value is from fund "PRINLS3 FA", 0.000604 (or 0.0604%), and the highest is fund "NUMC B0-B", with 0.707715 (or 70.7715%). Some the calculated r squared can be seen in **Table 4.6.**

Fund	R squared against BMV index with risk-free returns CETES364D
GBMPCON B	0.046817
GBMPMOD B	0.166699
GBMV1 BO	0.371316
GOLD3+ B1-C	0.026314
HSBC-F2 BFP	0.320366

Table 4.6. A sample of calculated funds' r squared statistic.

Of all analyzed funds, 20 have a relationship with the BMV index of less than 1%, while only 5 funds have a 50% or more relationship (or dependency) with the BMV index. This situation indicates that the movements of the BMV index only partially explain the overall behavior of the funds' returns, according to the definition of r squared.

4.3. Description of the script for Modern Portfolio Theory statistics calculation

The basis for the code to calculate the MPT statistics is in the reference manual from the PerformanceAnalytics package webpage, "*CRAN - Package PerformanceAnalytics*". The name and version of the components used in the coded script are listed below, in **Table 4.7.**

R 3.2.3 packages and versions used in the Modern Portfolio Theory statistics calculation script			
Component	**Version**	**Component**	**Version**
PerformanceAnalytics	**1.4.3541**	**stats**	**3.2.3**
xts	0.9	zoo	1.7-10

Table 4.7. List of packages used in the calculation of the Modern Portfolio Theory statistical measures.

Prior to its calculation, each fund's return rate series was used to create a `xts` time series object used as the input format for the data to be analyzed by most of the Performance Analytics functions. Once executed, the script saves the calculated MPT statistics in a data matrix, with the funds' ticker symbols on the left column and each static's name at the top row, in a CSV file named in the format: "`RVMexico.capm_analysis_<YYYYMMDDHHmmSS>.csv`".

The MPT statistics for the benchmark, the BMV index, were also included in the code as the first row of MPT statistics in the results data matrix.

Challenges

Following the R package installation instructions at the "*R Installation and Administration*" webpage manual, the default R language documentation and the R language tutorials listed in chapter 7, provided the guide and assistance to solve all the found issues during the development of the script.

The computer where the MPT statistics script was developed and executed has the specifications listed in **Table 4.8.**

Hardware and operative system specifications	
Computer:	**DELL Inspiron 15**
Processor:	**Intel® Core™ i5-4200U CPU @ 1.60GHz × 4**
Memory:	**8GB**
Operative system:	**Gnu/Linux OpenSUSE Leap 42.1 (x86_64) 64-bit**

Table 4.8. Description of the equipment where the R script was developed and executed.

All the encountered issues and problems during the development of the MPT statics calculation script were solved.

Technical Problems

A. Installation of the PerformanceAnalytics package. The first attempt to install this package in the R environment, using the package installation command `install.packages("PerformanceAnalytics")`, failed because this package requires to be compiled. The compilation process generated binary files that are saved in the root user access directories, alongside other R packages. This decision was taken to avoid any problem accessing and installing the additional required libraries. Additional R packages required should be compiled and installed with root user privileges. The compilation of the package Performance Analytics was performed with the gcc 4.8 compiler for C/C++.

B. Statics result list preallocation. Due to lists objects in R must be created with a fixed length, it is required to pre assign the length and data type of the list on the results of each MPT statistical measure for each statistic fund. To ensure that these statistics are correctly calculated, the number of rates of return of each fund must be the same as the length of the benchmark. Thus, funds with unequal number of rates with the benchmark must be identified and removed from the results lists. This was achieved by sorting the yearly rates file by fund name and date before the execution of the MPT statics calculation script. This way, it was easy to save the position (number of fund in the presorted yearly rates file) of the faulty fund in another list and removing this position from the preallocated results lists by copying the result lists back minus the faulty positions.

5. Machine Learning Analysis of Equity Funds

The third and last part of the experiment was the application of the machine learning techniques for the analysis of the selected 182 equity funds. In particular, the clustering analysis tools for time series collected and programmed by Montero and Vilar (2014) were used.

To perform a clustering analysis, is required to select an adequate dissimilarity measure for the historical price data sets and a clustering algorithm. As explained in chapter 2, the daily price series of equity funds is a time series that requires of a specialized dissimilarity measure, d_{CORT}, which not only takes into account the frequency of the observations (prices) and their temporal relationship, but the behavior of the funds compared with each other. This particular metric was chosen because it was assumed that the notion of "similarity" of two time series, corresponding to two investment funds, should relate to how closely correlated they are and how similar their performance (profit) is.

For the clustering algorithm, a hierarchical clustering algorithm, also explained in chapter 2, was chosen.

5.1. Calculation of the dissimilarity measures and their classification

Like the calculation of the Modern Portfolio theory statics, the script to calculate the dissimilarity of the equity funds' normalized prices and their classification was programmed in the R language (see technical details in chapter 4). The package **TSclust 1.2.3's** data structures and functions, explained and implemented in Montero and Vilar's paper (2014), were extensively used to perform all calculations.

5.2. Analysis of the clustering results

As mentioned, the clustering analysis was performed for the time period from 2010-12-31 to 2015-12-31 in two parts:

1. Calculation of the funds' dissimilarity matrix index with the normalized time series of the selected 182 equity funds.

2. Clustering classification of funds based on their dissimilarity matrix.

5.2.1. CORT dissimilarity index measure

To calculate the dissimilarity measure d_{CORT}, the TSclust' general dissimilarity function `diss` was implemented with the following parameters:

```
diss(<m_tsclust_funds>, METHOD="CORT", k=2,
          deltamethod="Euclid")
```

Where:

<m_tsclust_funds> is a numeric matrix with the normalized price series, per row, of each fund.

`METHOD="CORT"` is the dissimilarity measure method to use, in this case CORT.

`k=2`, the weight of the dissimilarity between dynamic behaviors (2 is the default value).

`deltamethod="Euclid"` is the method to measure the raw data discrepancy ("Euclid", or the Euclidean distance, is the default method).

The resulting `diss` object is a data structure from the TSclust package that includes the lower triangle of a matrix with the dissimilarity measures of the BMV index and the examined funds.

Due to the explorative nature of this work, it was decided to use the default, and most simple, options to calculate the dissimilarities among the BMV index and the funds. The parameter `k` is the value to weight the modulating the correlation coefficient CORT, while the Euclidean distance is the basic measure of vector distance (see Formula 2.9. and Formula 2.10.).

Unlike the others dissimilarity measures collected by Montero and Vilar (2014), d_{CORT} has the capacity to include the behavior of the studied data series (the correlation coefficient CORT, with help of the modulating function $\varphi_k(\cdot)$), into the value of the chosen raw-data dissimilarity approaches $d(X_T, Y_T)$ (Euclidean, Fréchet or DTW). A positive correlation between two series of data will increase the value of the dissimilarity measure; a negative correlation, will reduce the dissimilarity value; and the neutral will not affect the calculated dissimilarity. This added measure, increases the possibility that funds exhibiting similar investment behaviors, during the studied time period, will be grouped together.

The formulas and details about how the correlation coefficient CORT and the modulating function $\varphi_k(\cdot)$ affect the d_{CORT} dissimilarity index are explained in chapter 2.

5.2.2. Clustering results based on dissimilarities values

With the `diss` object containing the BMV index and the fund's dissimilarity indexes, the clustering of the price series is performed with function `pvalues.clust`:

```
pvalues.clust(<od_cort>, significance=0.05)
```

Where:

<od_cort>, is a `diss` object containing the p-values (measures or values) from testing the equality of each studied time series.

`significance`, the algorithm groups together data series whose associated p-values, or measures, are greater than this prespecified significance level.

The first classification was performed with the default significance level of 0.05, which found 5 groups, or clusters, of funds. To explore the possibility for a more specific clustering, the clustering function was repeated eleven times to test different significance levels (see **Table 5.1.**).

Significance	Number of clusters	Significance	Number of clusters
0.05	**5**	0.65	**51**
0.10	**8**	0.75	**64**
0.15	**10**	0.85	**71**
0.25	**14**	0.95	**83**
0.35	**20**	1.00	**88**
0.50	**34**		

Table 5.1. List of number of clusters per significance level.

5.2.3. Graphical comparison between clustering groups and their Modern Portfolio Theory statistics

With help of the package **ggplot2 2.1**, scatterplots graphs were created to compare the discovered fund classification with their **Modern Portfolio Theory** (MPT) statistics.

5.2.3.1. Scatter plots with the alpha statistic

With the variables **alpha** and beta as the x and y axis, respectively, the found classification with the significance level 0.05 produces the following scatter plot.

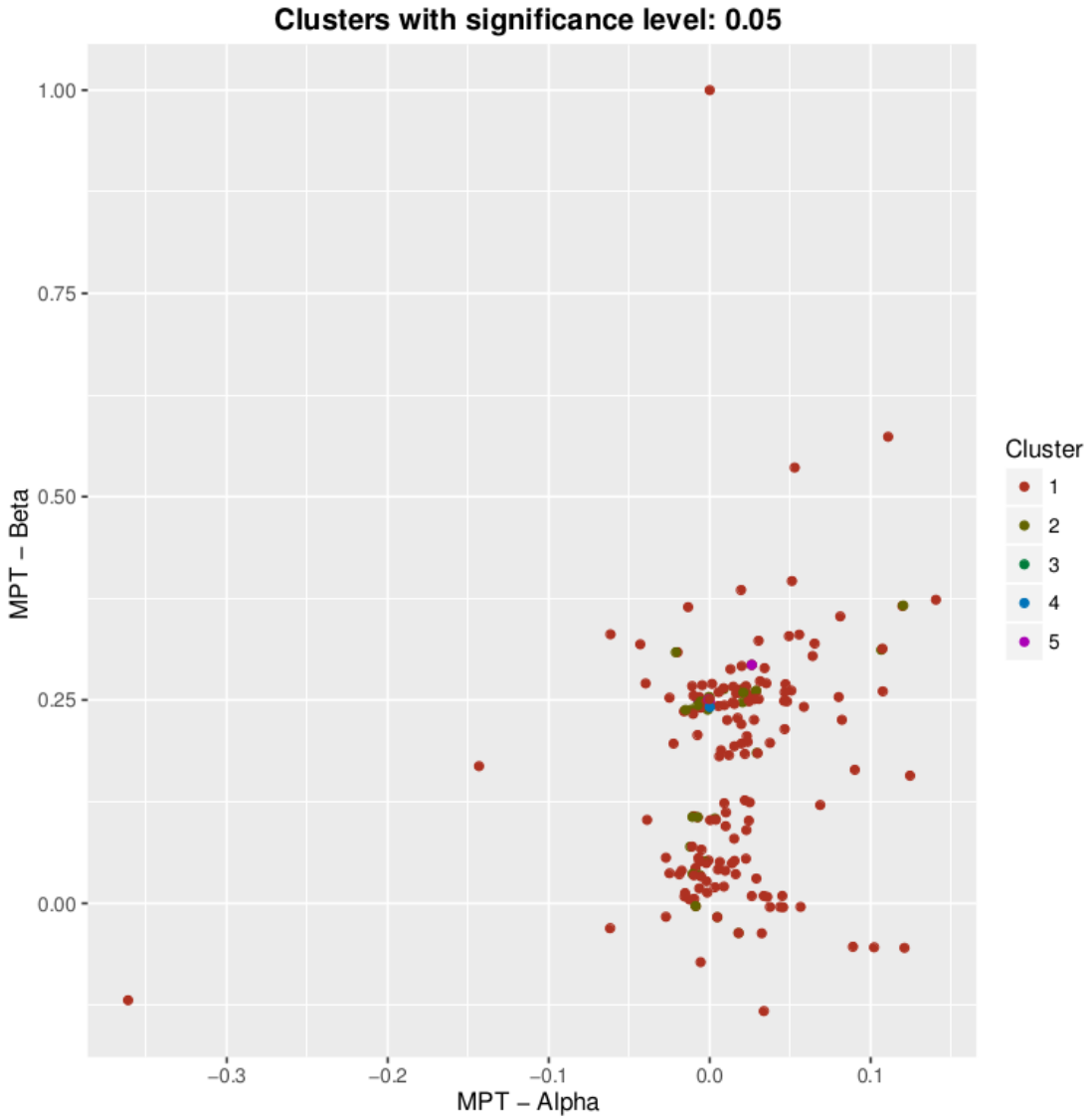


Image 5.1. Scatter plot of the relationship between the alpha and beta statistics with the color codes of the classification produced with a significance of 0.05.

As shown in the last graphic (**Image 5.1.**), most funds were concentrated in the inferior right side of the plot, between coordinates (-0.05, 0.0) and (0.05, 0.375). However, most of the funds were classified in group 1 and others groups' funds were also scattered among group 1's funds locations. For example, group 2's funds were dispersed all over the same areas were group 1's funds were distributed, but did not show any distinctive alpha and beta ranges of values to tell apart any of both groups' funds.

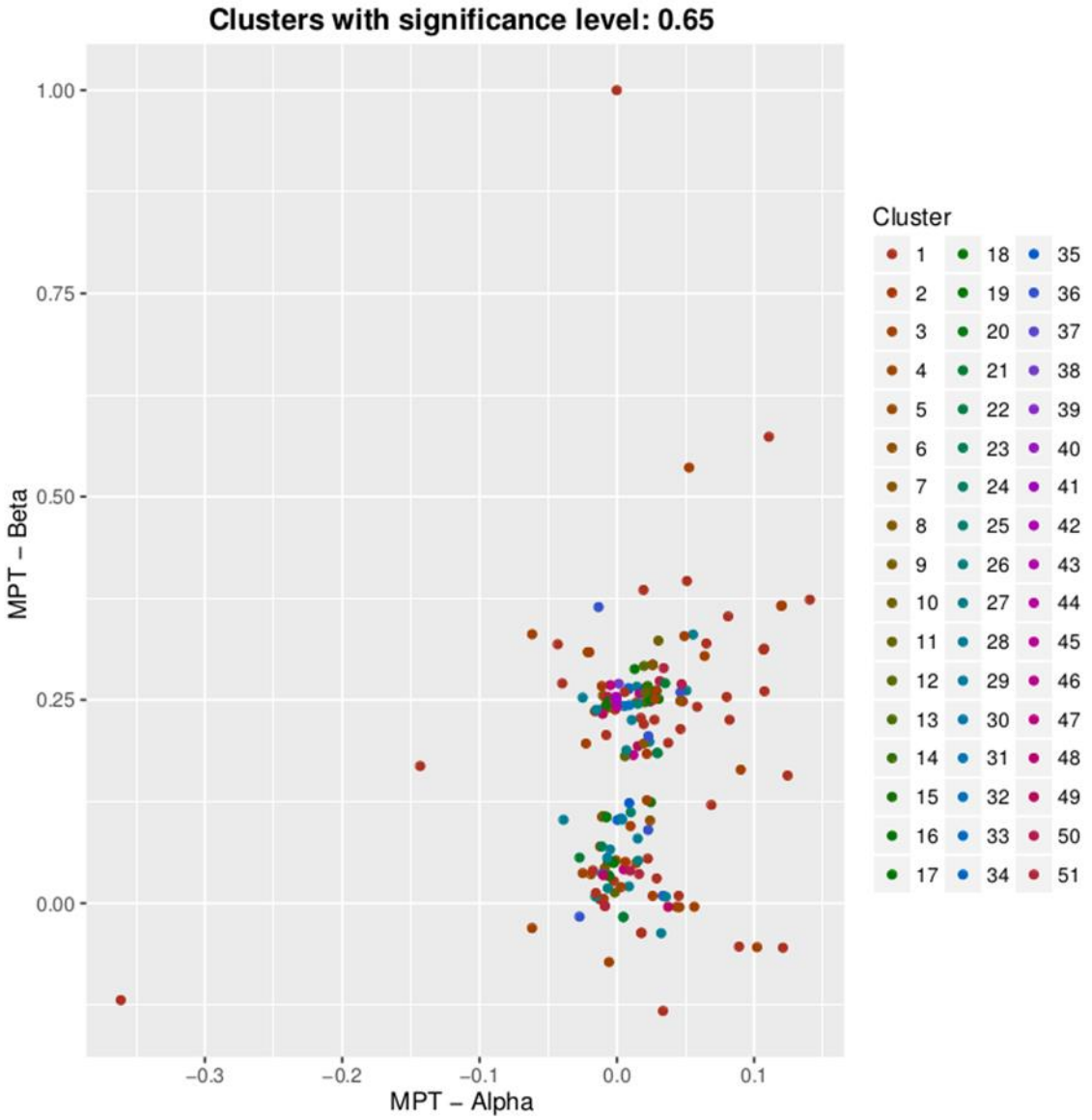


Image 5.2. Scatter plot of the relationship between the alpha and beta statistics with the color codes of the classification produced with a significance of 0.65.

With higher significance levels, the number of funds in group 1 decreased, as shown in **Image 5.2.** But the increased number of funds in the other groups did not mean that those funds were in closer quadrants in the scatter graphs. In fact, the dispersion among funds from groups 2 and above increased. The constant behavior, in the comparison between the alpha and beta statistics (see **Image 5.2.** and **Image 5.3.**), was that most of the extreme locations in the plot, the outliers, belonged to funds from groups 1 to 6.

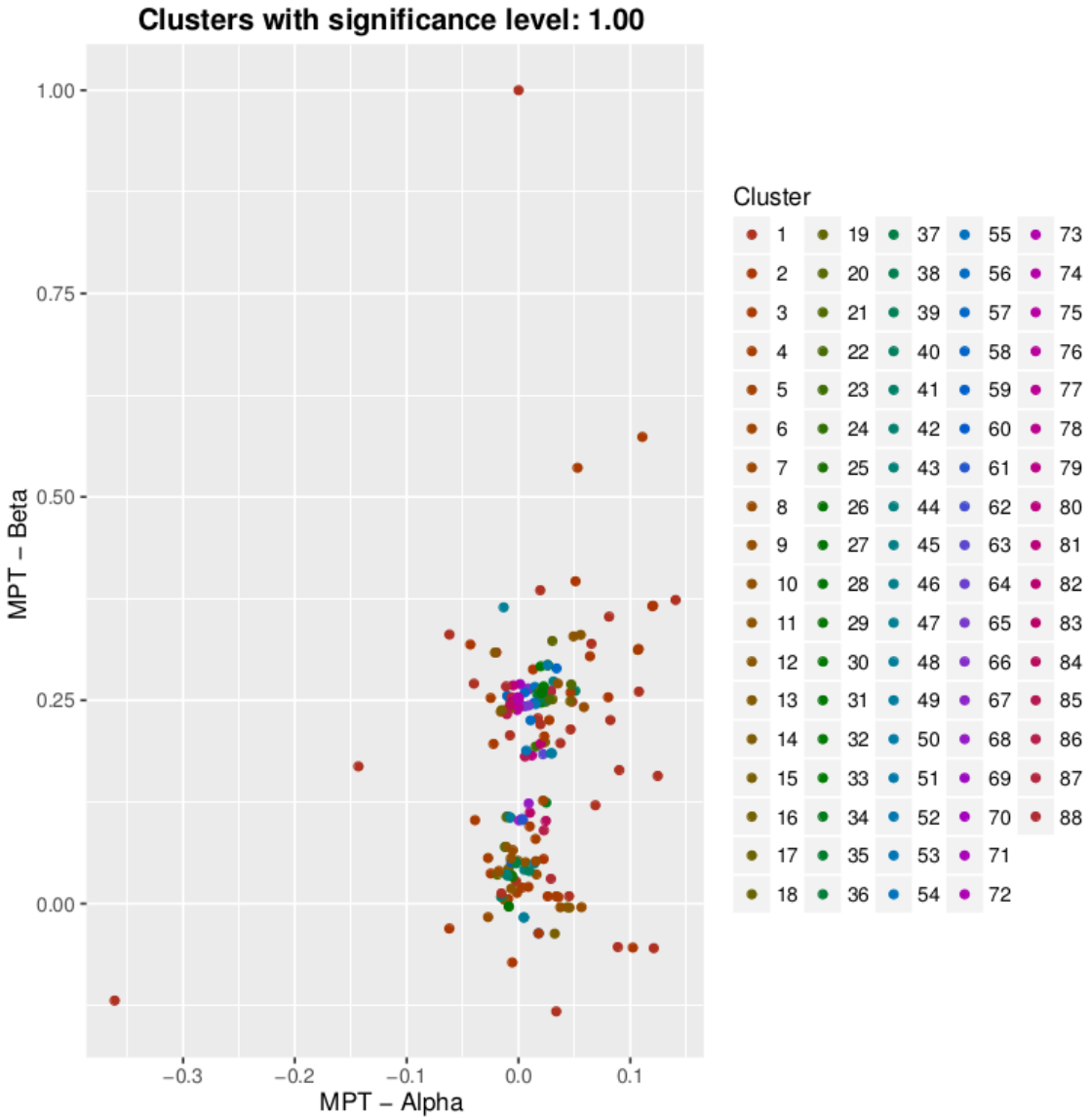


Image 5.3. Scatter plot of the relationship between the alpha and beta statistics with the color codes of the classification produced with a significance of 1.00.

As the level of significance increases, the funds' classifications did not show a clearer relationship, or behavior, in relation to a specific range of values of their alpha and beta statistics (**Image 5.3.**). Even some funds from groups 1 to 6 mingle with funds from other groups.

When compared with **r squared**, the alpha statistic produced a scatter plot (**Image 5.4.**), were most funds were, also, located in the inferior right side of the graph, between coordinates (-0.05, 0.0) and (0.05, 0.5).

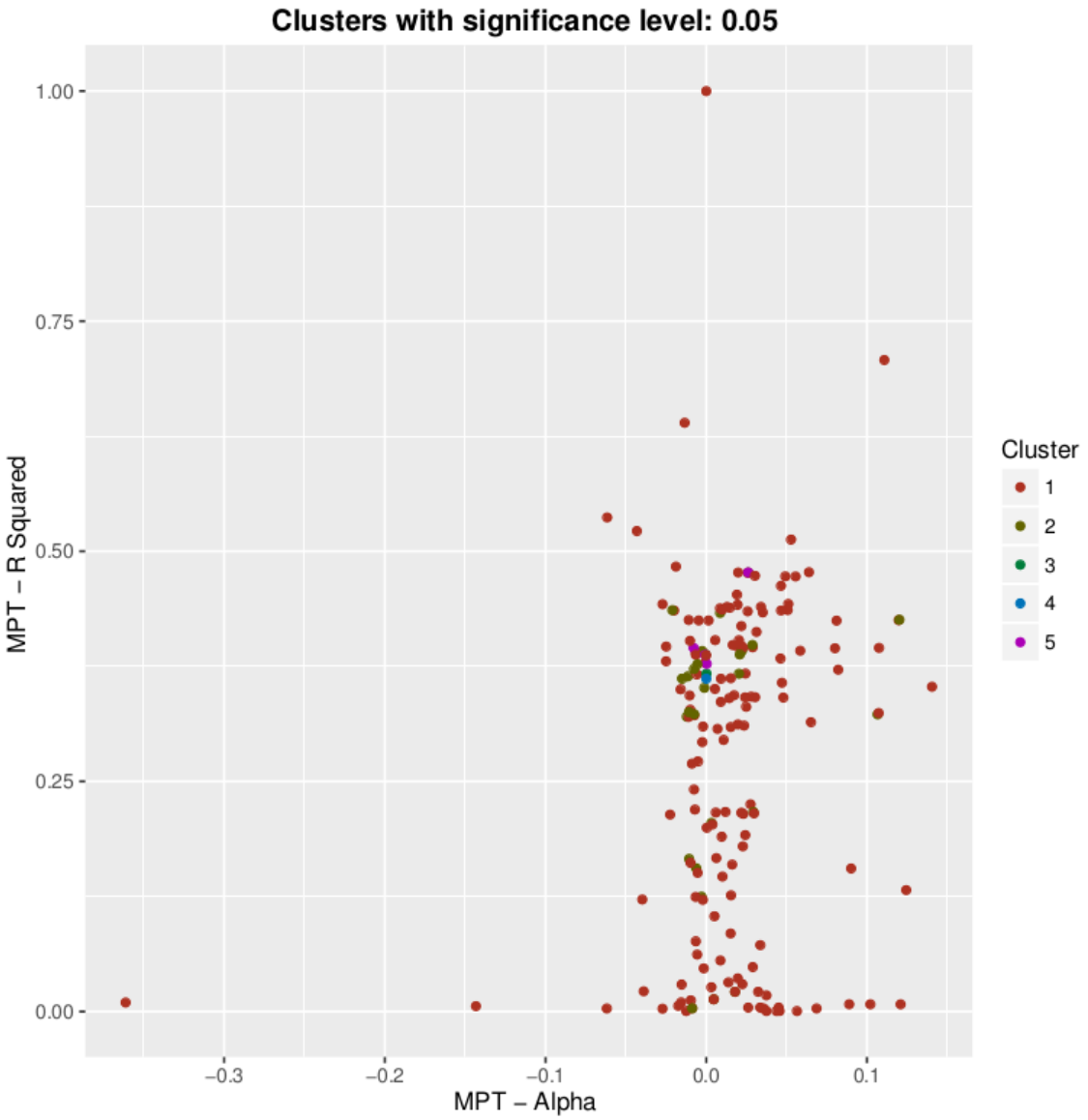


Image 5.4. Scatter plot of the relationship between the alpha and r squared statistics with the color codes of the classification produced with a significance of 0.05.

As in the comparison with the beta statistic, the past graph, with the classification generated with a significance of 0.05, shows that funds from groups 2 and above were scattered among group 1's funds.

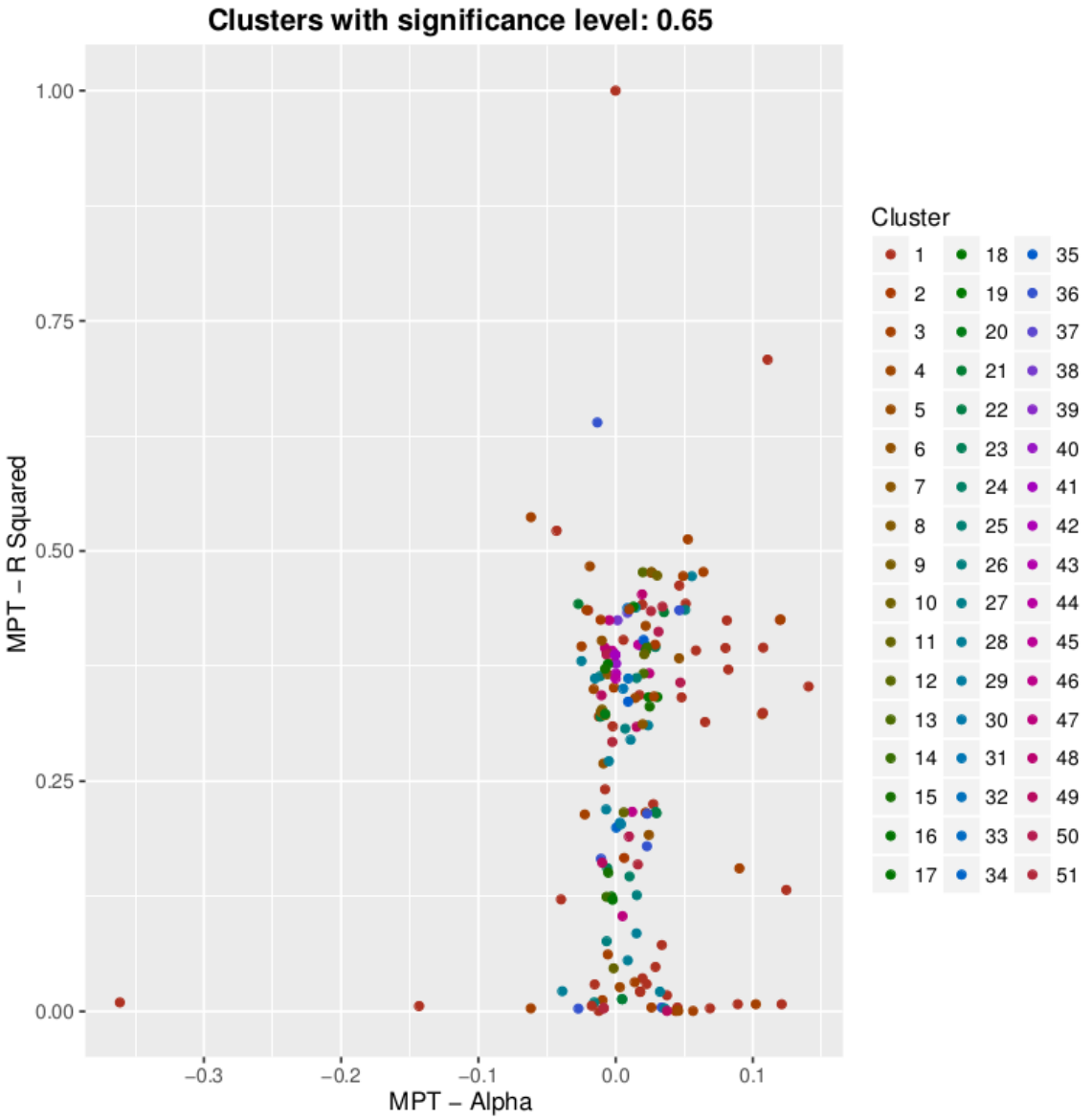


Image 5.5. Scatter plot of the relationship between the alpha and r squared statistics with the color codes of the classification produced with a significance of 0.65.

The previous **Image 5.5.** shows the significance 0.65's group classification. Unfortunately, as the comparison with beta, there was not a definitive relationship between the alpha and r squared range of values. Many group's funds continue to be plotted close to other groups' funds, but separated from their own group's other funds.

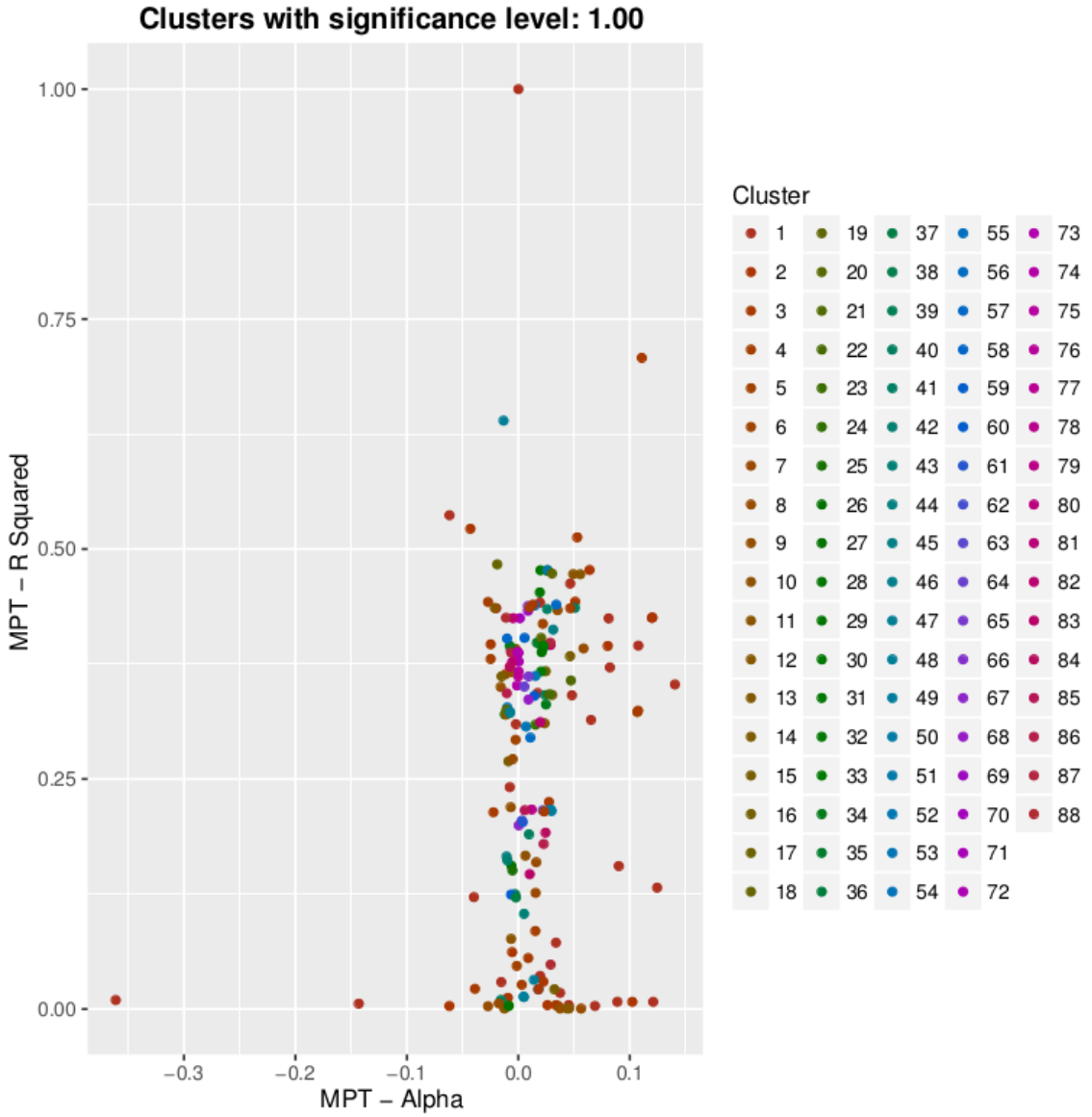


Image 5.6. Scatter plot of the relationship between the alpha and r squared statistics with the color codes of the classification produced with a significance of 1.00.

The increasing number of groups discovered, through testing higher significance levels (as in **Image 5.6.**), did not correspond with clearer or more defined ranges of values in the alpha and r squared statistics that could explain each group's behavior or their funds relationship with those MPT statistics. Although, funds from lower numbered groups (groups 1 to 18) tend to dominate the extreme locations (outliers) in **Image 5.5.** and **Image 5.6.**

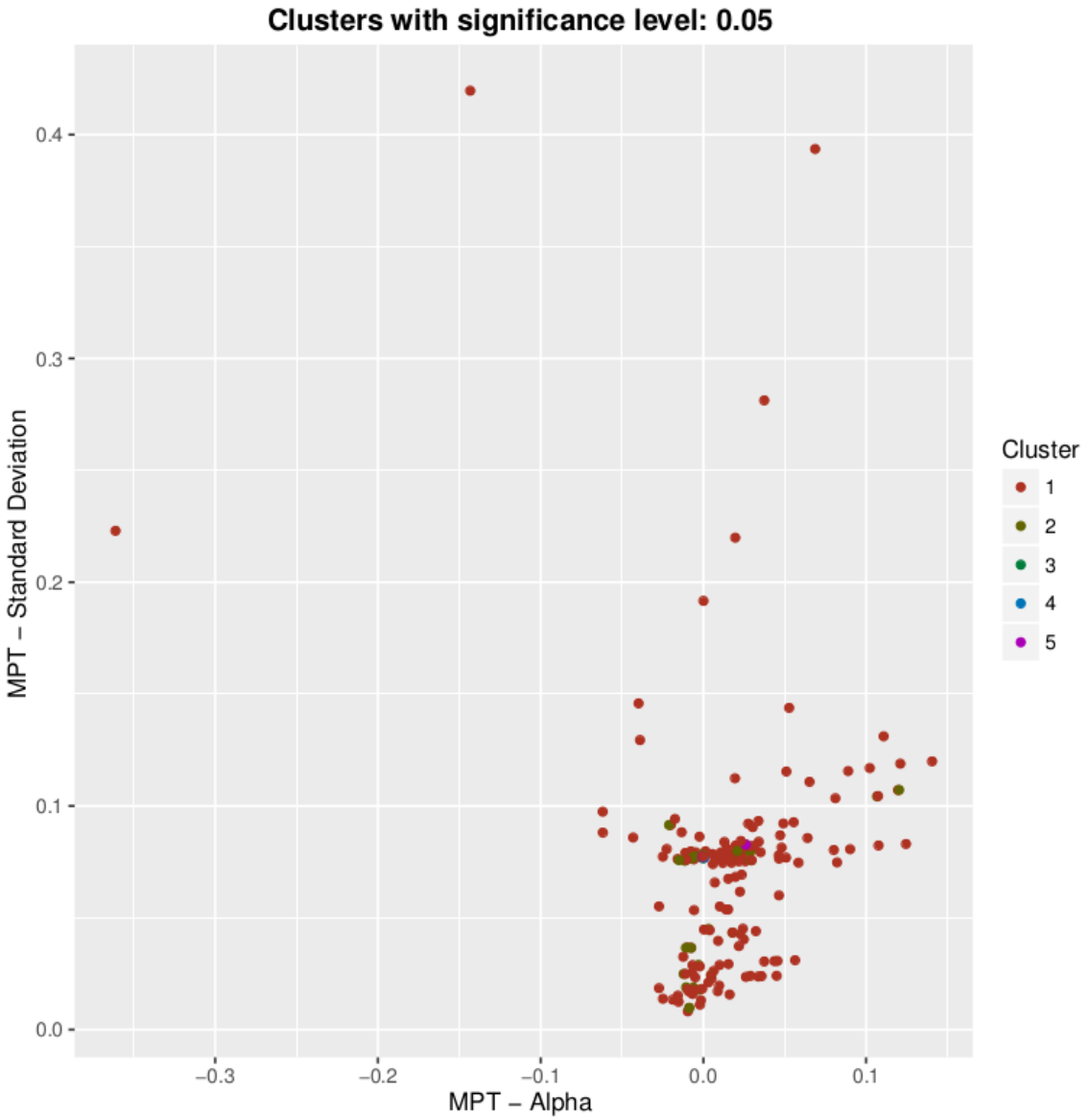


Image 5.7. Scatter plot of the relationship between the alpha and standard deviation statistics with the color codes of the classification produced with a significance of 0.05.

The scatter plot with the relationship between the alpha and the **standard deviation** statistics, as the y axis (**Image 5.7.**), showed the possibility of a closer grouping of funds. Most funds were clustered in the bottom right side of the graph, between coordinates (-0.05, 0.0) and (0.05, 0.1), and less scattered than in the previous relationship graphs, which usually hints to a closer relationship between the analyzed variables.

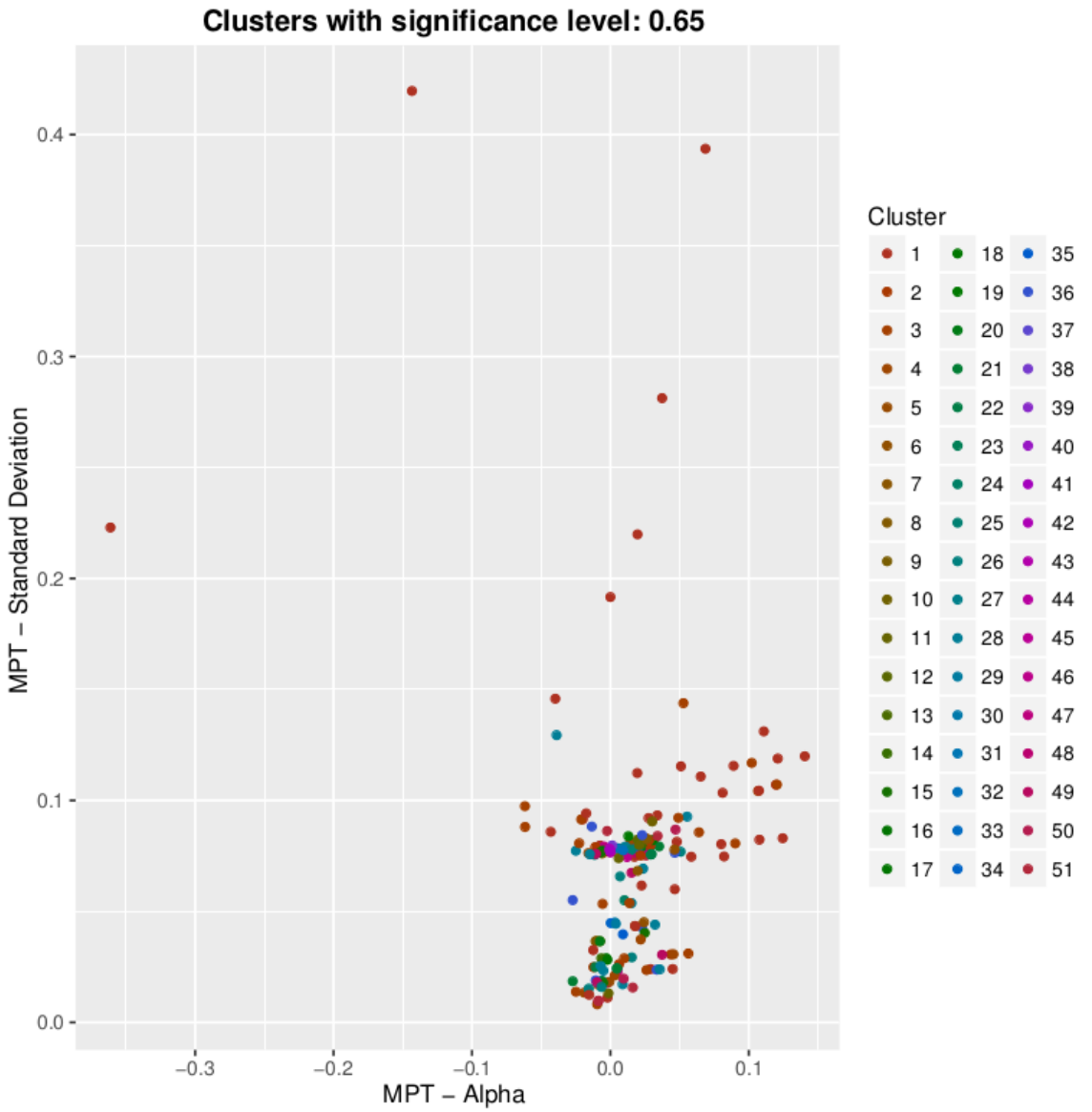


Image 5.8. Scatter plot of the relationship between the alpha and standard deviation statistics with the color codes of the classification produced with a significance of 0.65.

Like in previous graphs, significance 0.05's classification (**Image 5.7.**) grouped most funds in group 1. Unfortunately, under higher significances, such as 0.65, the cluster points in the graphs (**Image 5.8.**) were not plot together, o closer, as groups.

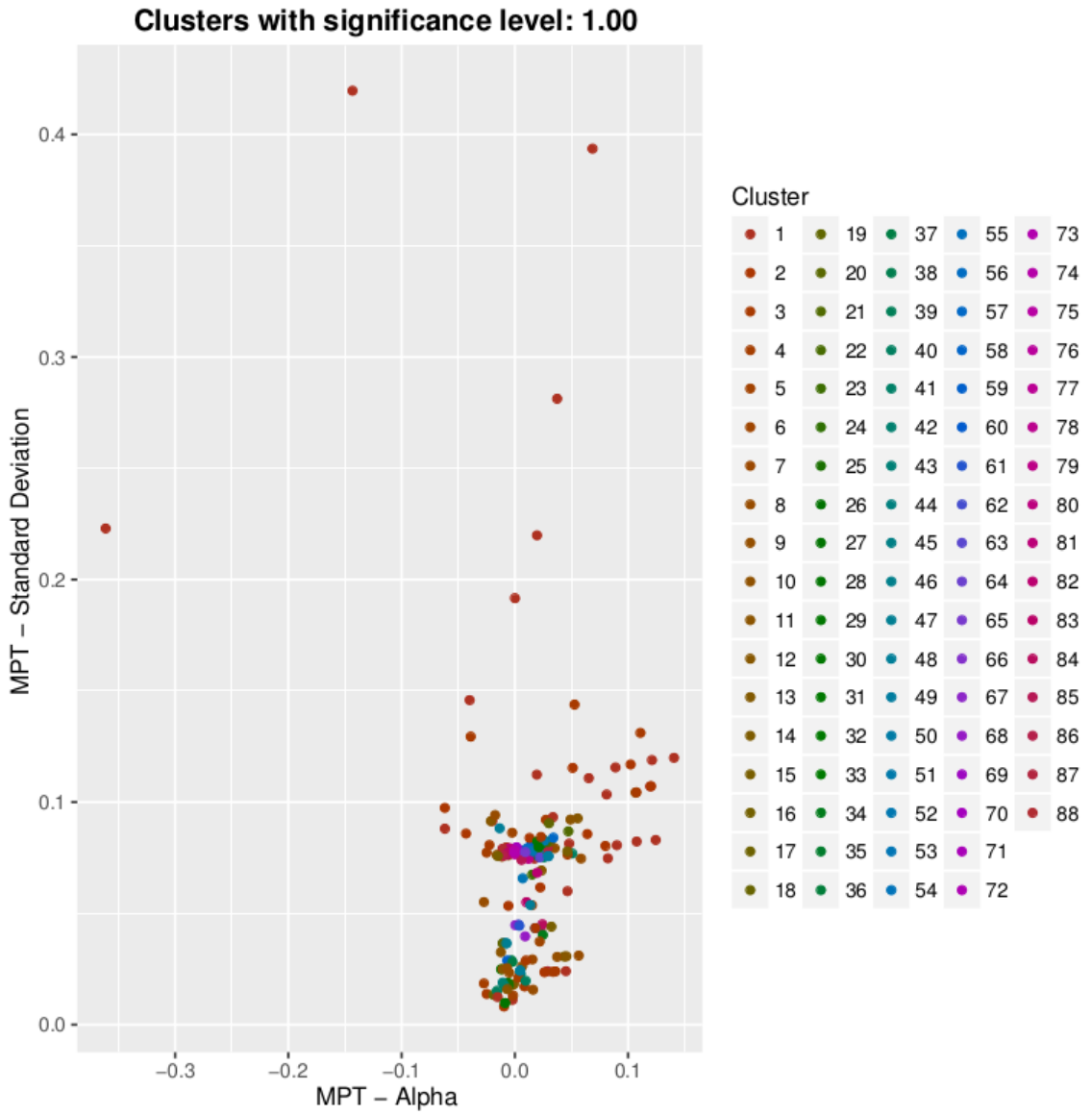


Image 5.9. Scatter plot of the relationship between the alpha and standard deviation statistics with the color codes of the classification produced with a significance of 1.00.

Some groups' funds were plotted together, according to their funds' standard deviation, while funds from other groups were also plotted with the same closeness (**Image 5.9.**). An observable pattern in the alpha and standard deviation graph, was that funds from groups 1 to 6 have a standard deviation higher than 0.1; but other funds from those same groups mingle with other groups' funds, below that standard deviation, as well.

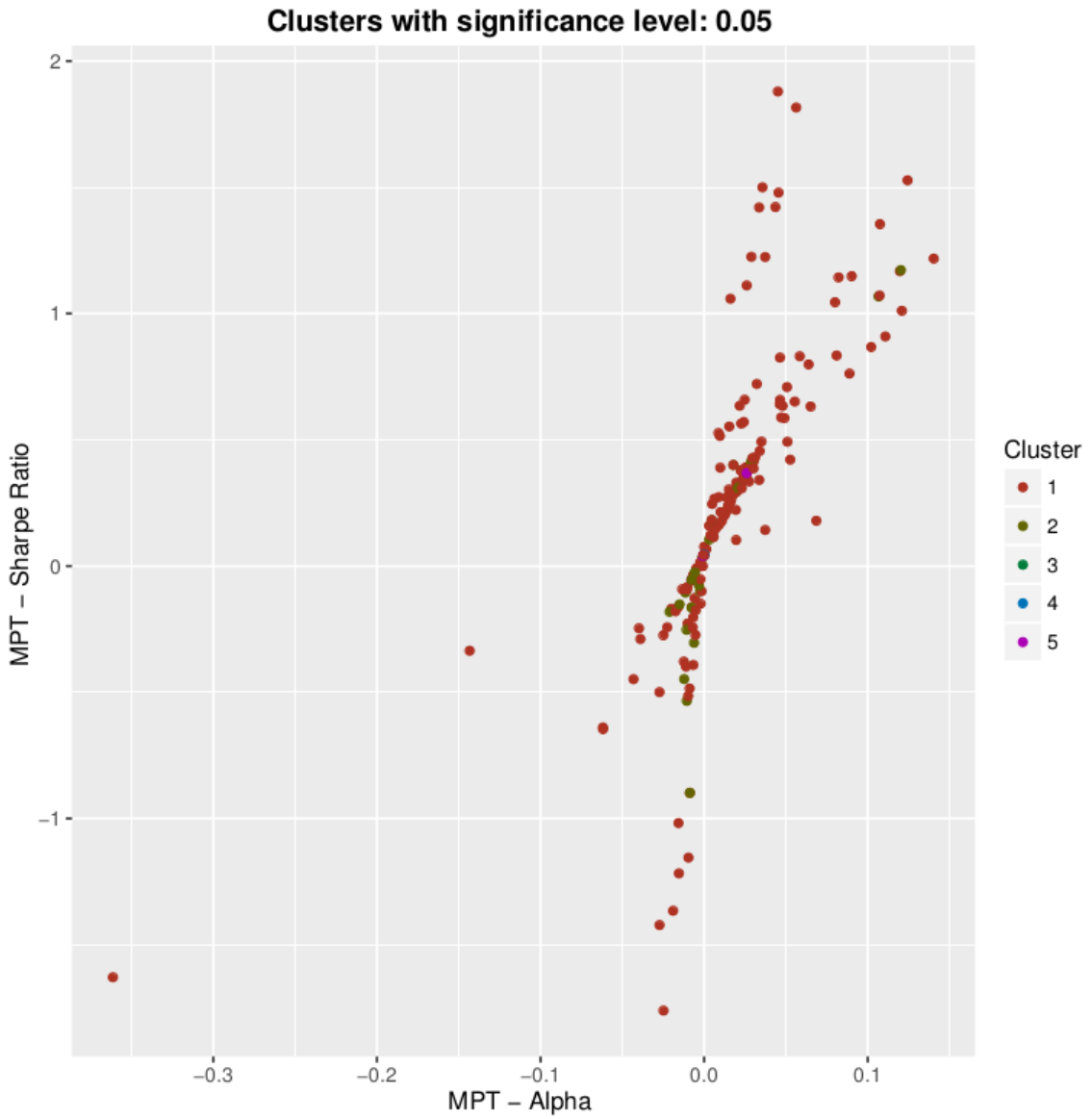


Image 5.10. Scatter plot of the relationship between the alpha and Sharpe ratio statistics with the color codes of the classification produced with a significance of 0.05.

When comparing the alpha measure with the **Sharpe ratio** (**Image 5.10.**), the generated scatter graph follows an 'X' shaped pattern in the middle right side, between coordinates (-0.05, -0.50) and (0.05, 0.50). Because most funds were classified in group 1, the significance level 0.05 classification was not very useful.

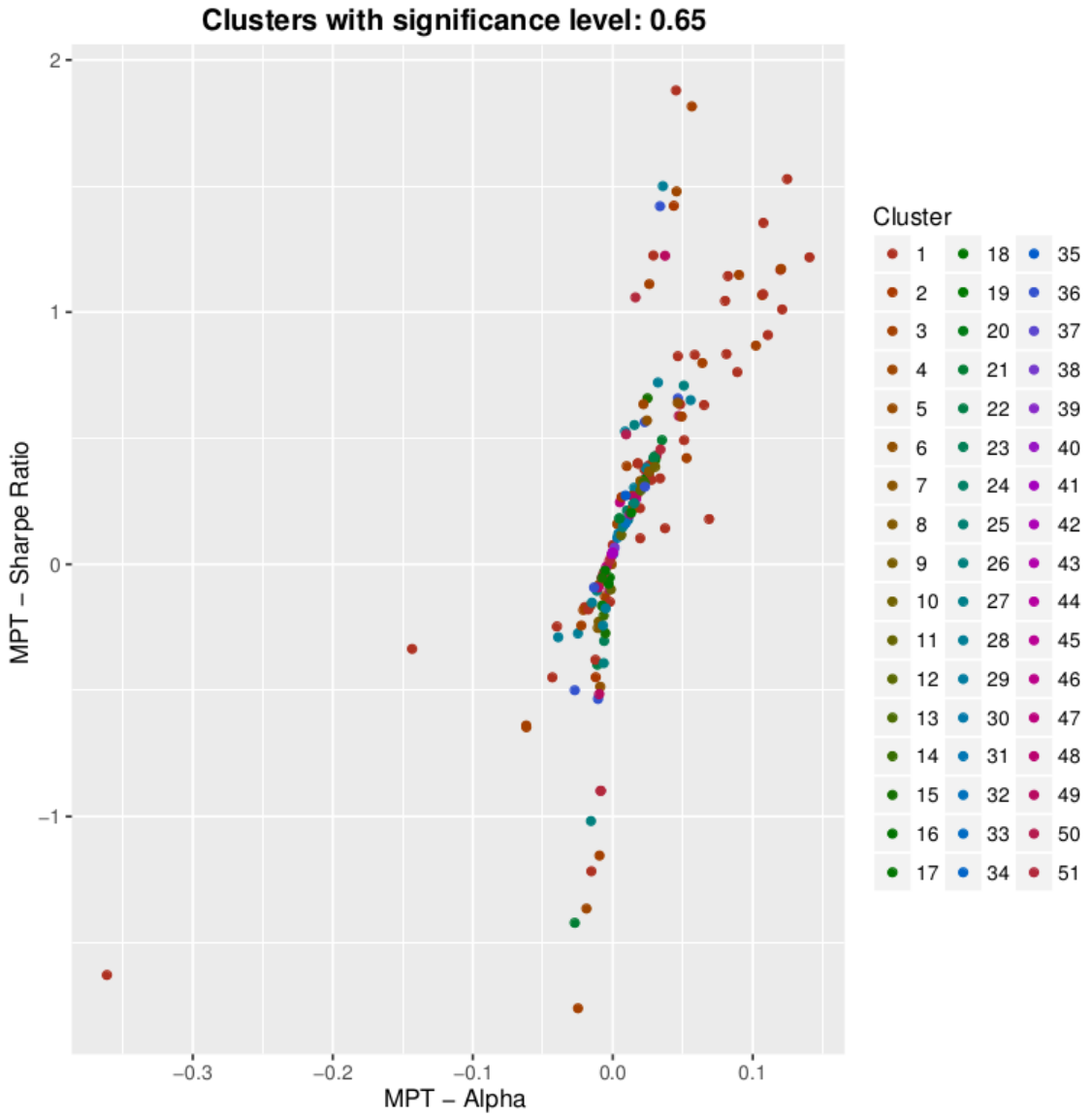


Image 5.11. Scatter plot of the relationship between the alpha and Sharpe ratio statistics with the color codes of the classification produced with a significance of 0.65.

Unfortunately, the plot of the clustering with significance level 0.65 (**Image 5.11.**) showed the same behavior: each group's funds were scattered along the main graph pattern.

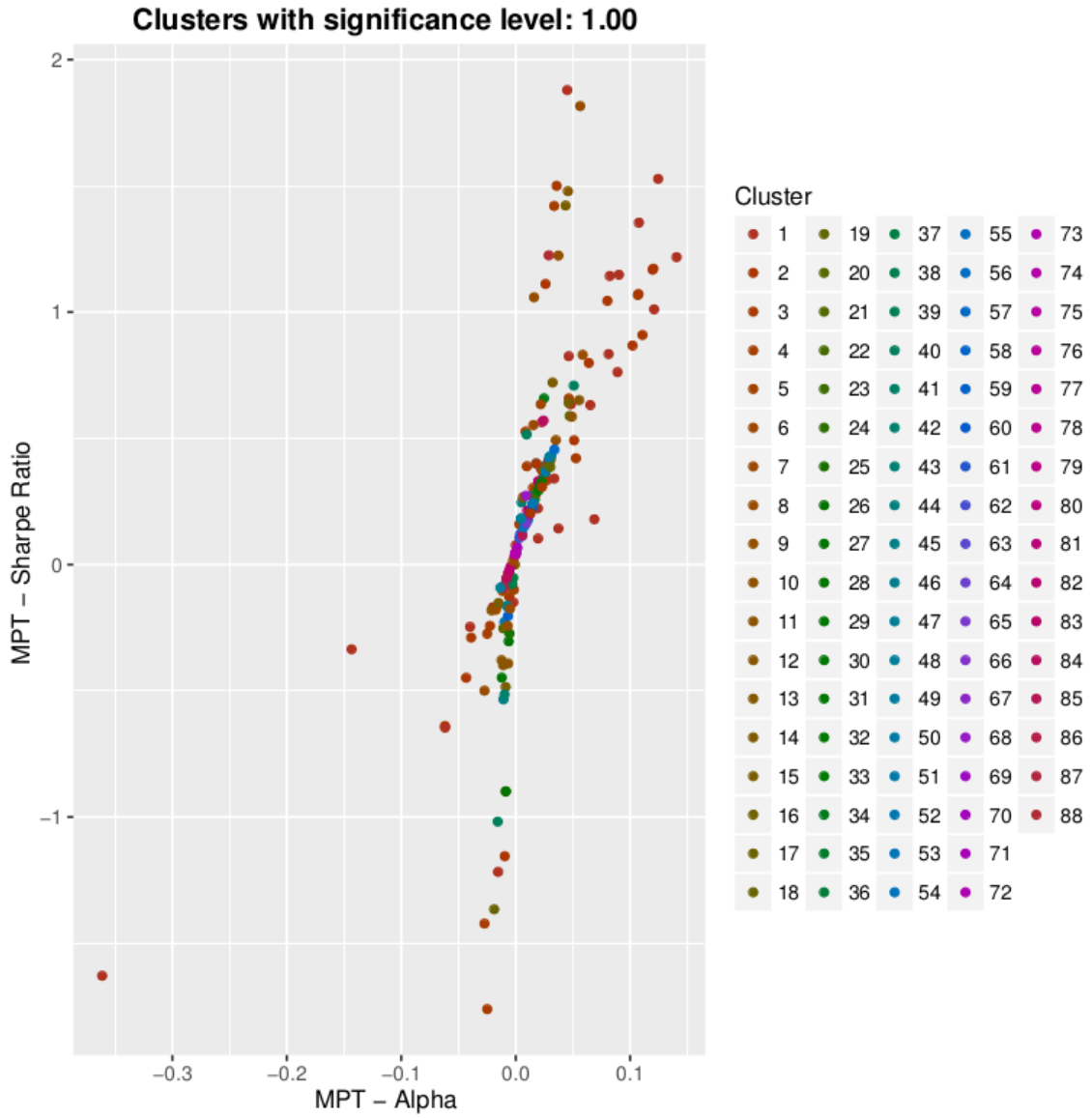


Image 5.12. Scatter plot of the relationship between the alpha and Sharpe ratio statistics with the color codes of the classification produced with a significance of 1.00.

With classifications produced from higher significance levels, such as 1.00 (**Image 5.12.**), it was discovered that funds from groups 1 to 20 tended to dominate in the non-central locations. But, again, funds from those groups were also scattered among the central graph area.

From the analysis of the graphs where the alpha statistic was compared with the other MPT measures, it is concluded that the value of this variable has no direct correlation in the classification produced by the hierarchical clustering algorithm.

5.2.3.2. Scatter plots with the beta statistic

The comparisons of the beta statistic, as the x axis, with the other MPT variables, as the y axis, are explained in the next scatter plots. The comparison between beta and alpha variables was skipped to avoid redundancy.

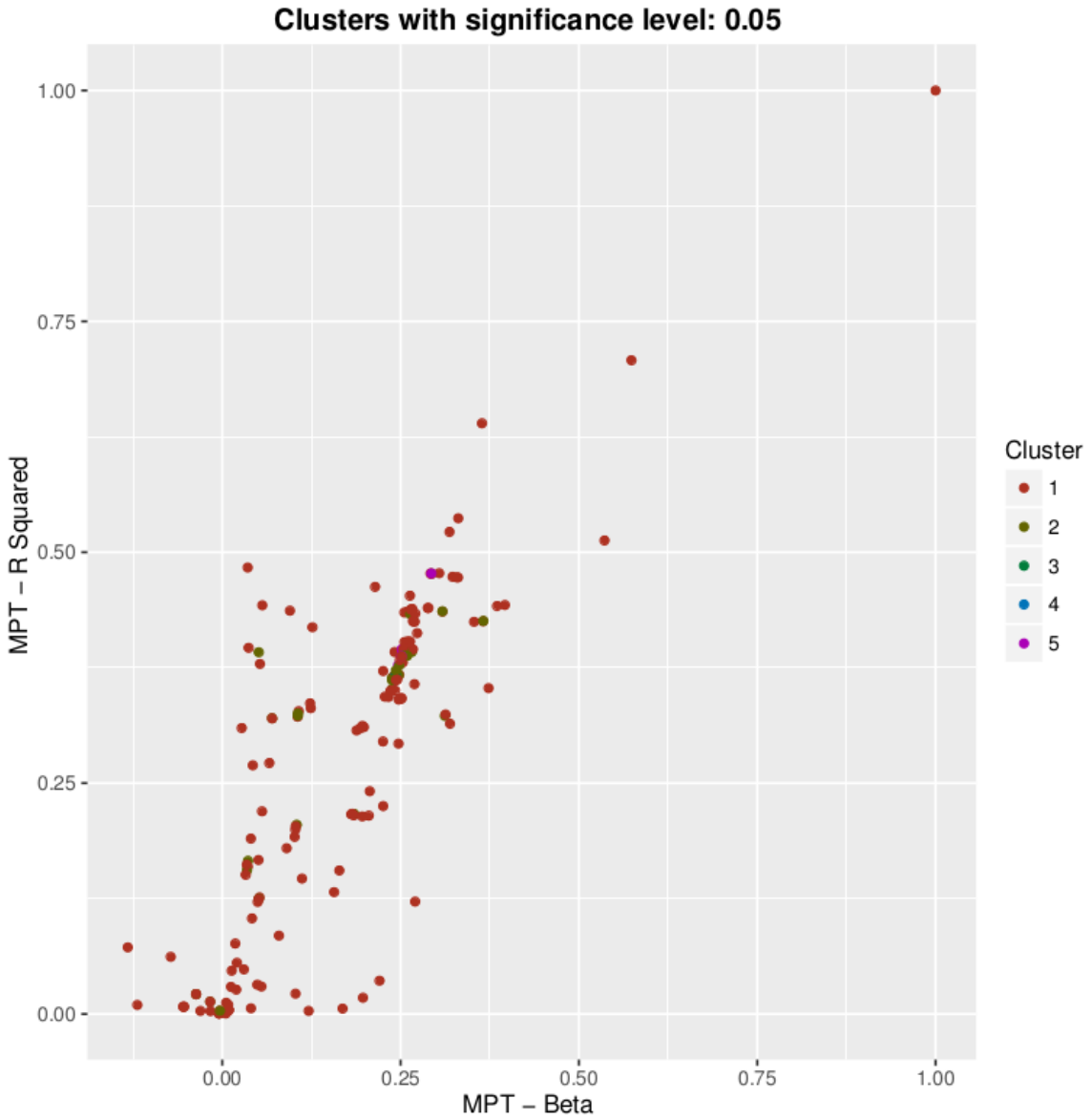


Image 5.13. Scatter plot of the relationship between the beta and r squared statistics with the color codes of the classification produced with a significance of 0.05.

Although the plot, with axes beta and **r squared**, showed a very dispersed pattern, most funds were concentrated in the inferior left hand side of the plot (from point (0.00, 0.00) to (0.375, 0.50)) and formed a few clusters of funds, as displayed in **Image 5.13.**

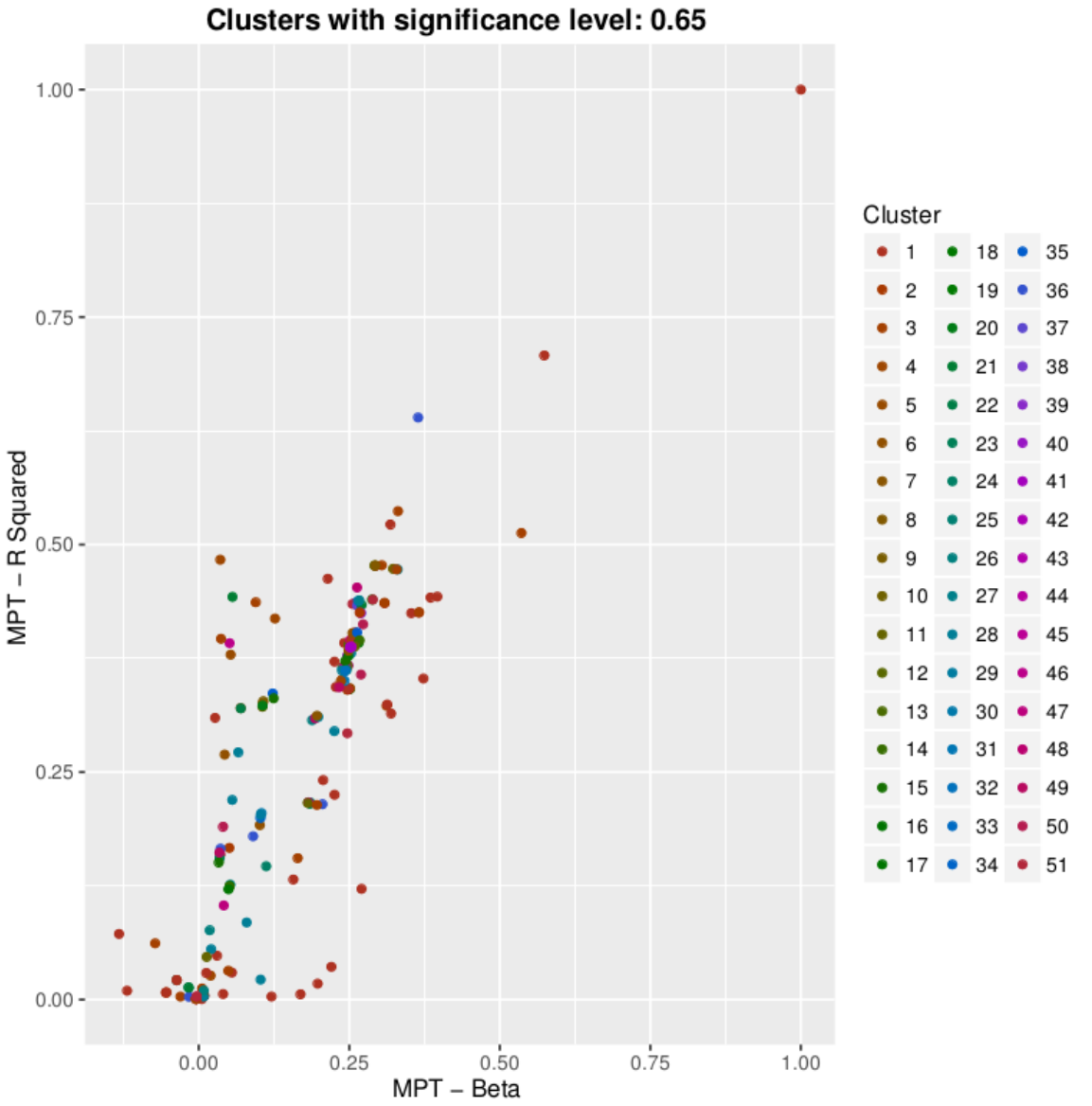


Image 5.14. Scatter plot of the relationship between the beta and r squared statistics with the color codes of the classification produced with a significance of 0.65.

Unfortunately, the classifications did not display their groups' funds plotted closer from other groups or forming a discernible pattern. (**Image 5.14.**).

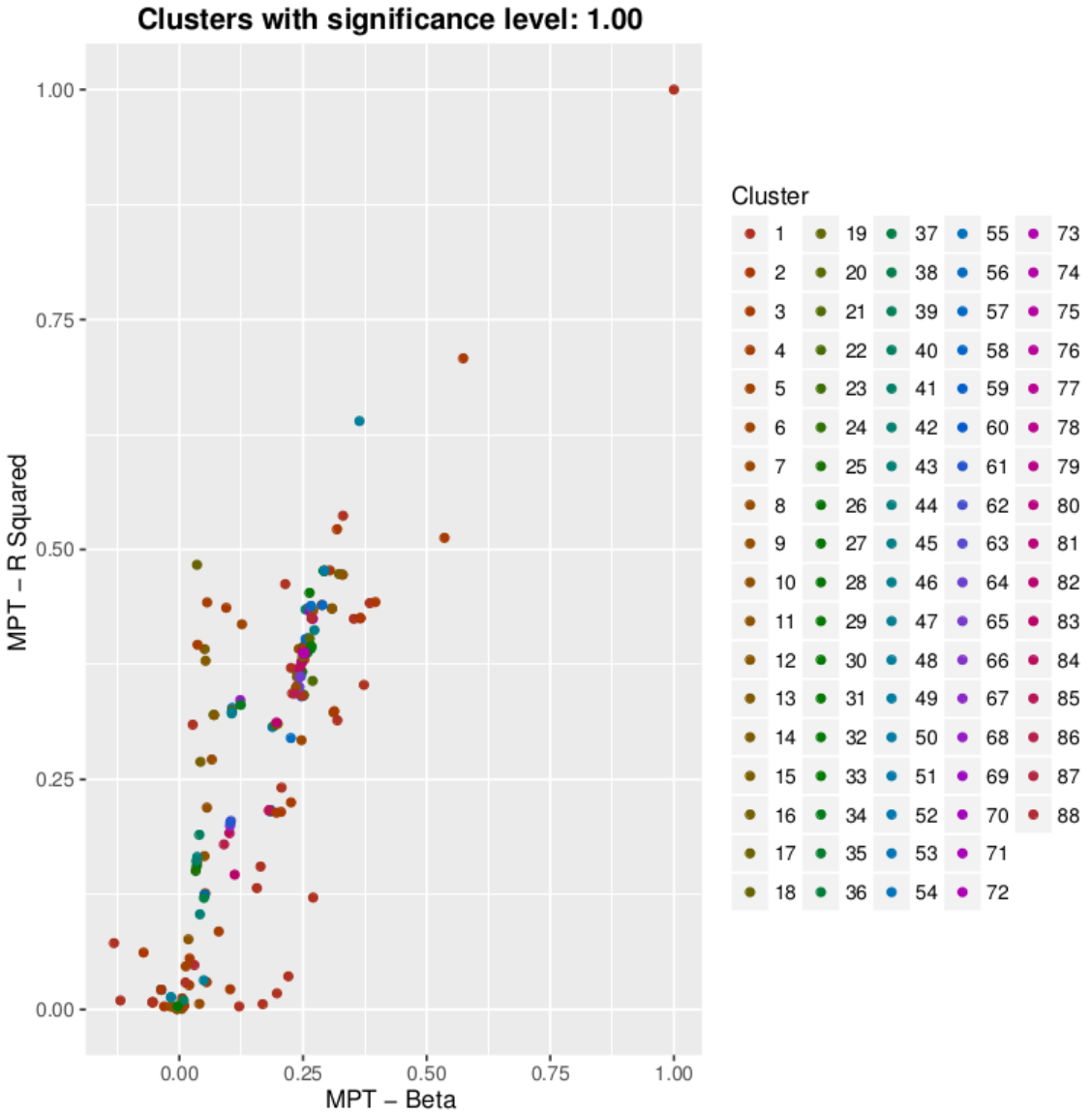


Image 5.15. Scatter plot of the relationship between the beta and r squared statistics with the color codes of the classification produced with a significance of 1.00.

For the plot with axes beta and r squared, the pattern where funds from lower numbered groups (1 to 18) showed a tendency to be located in the outside locations, far from the main clusters of funds, persists.

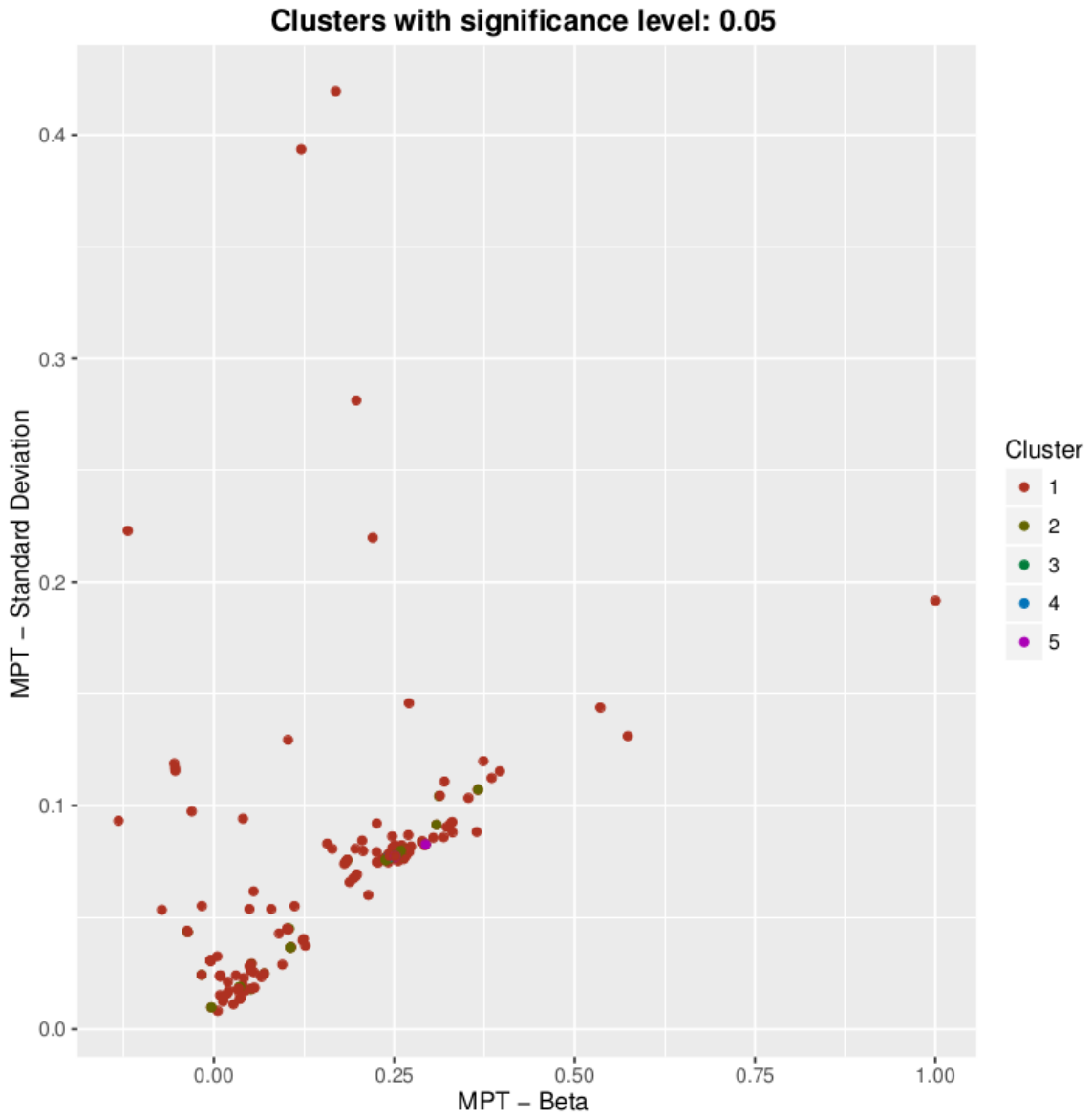


Image 5.16. Scatter plot of the relationship between the beta and standard deviation statistics with the color codes of the classification produced with a significance of 0.05.

In the previous graph, **Image 5.16.**, the funds plotted against the axes beta and **standard deviation** formed two main clusters, following a double 'V' shaped pattern in the bottom left side of the graph. The highest concentration of funds were located in the square from point (-0.125, 0.0) to point (0.125, 0.05), and the square between points (0.125, 0.05) and (0.375, 0.1).

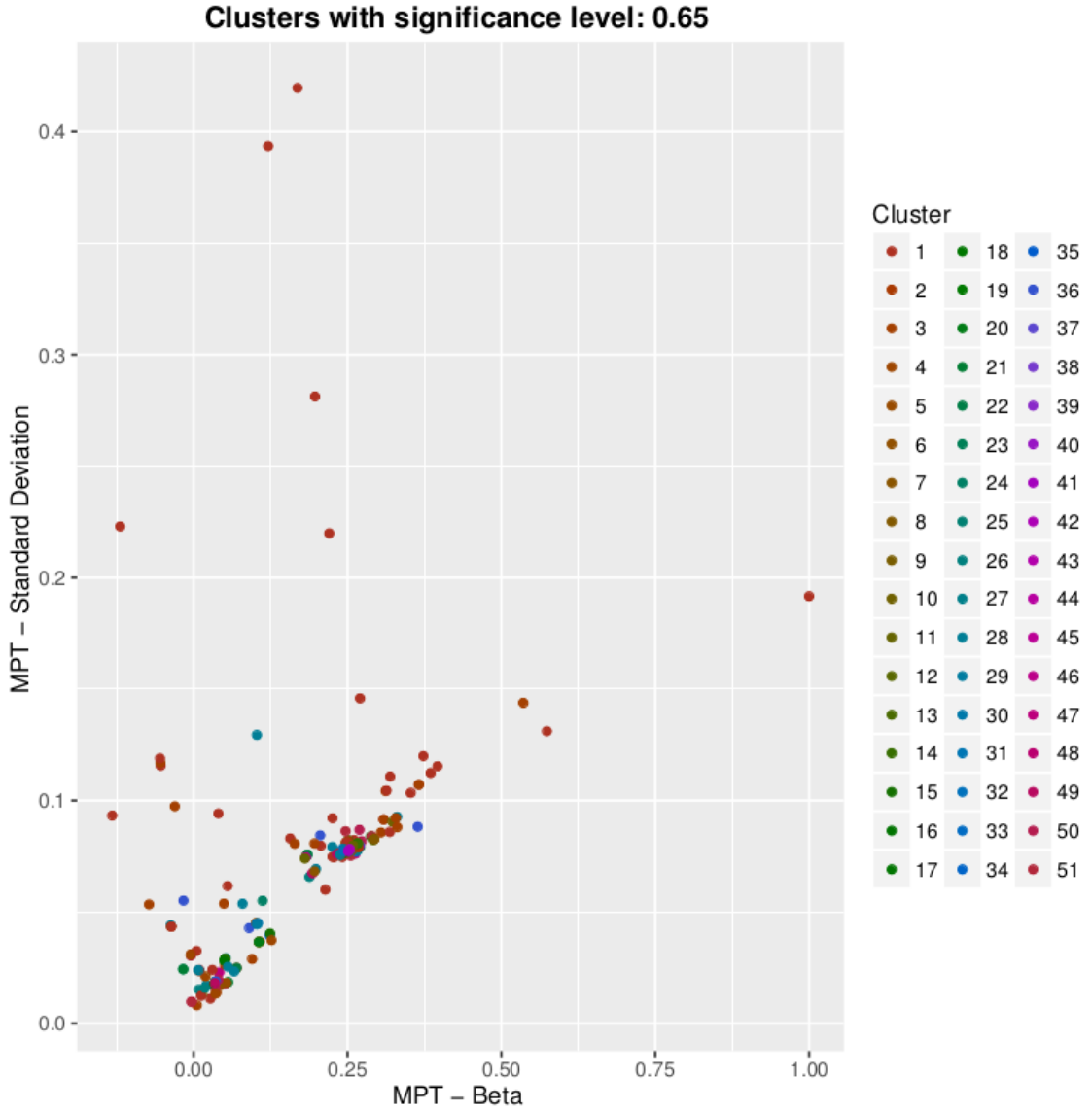


Image 5.17. Scatter plot of the relationship between the beta and standard deviation statistics with the color codes of the classification produced with a significance of 0.65.

The augmented significance level produced a classification with an increased number of groups. However, the classification did not display any particular pattern, as in previous graphs (**Image 5.17.**).

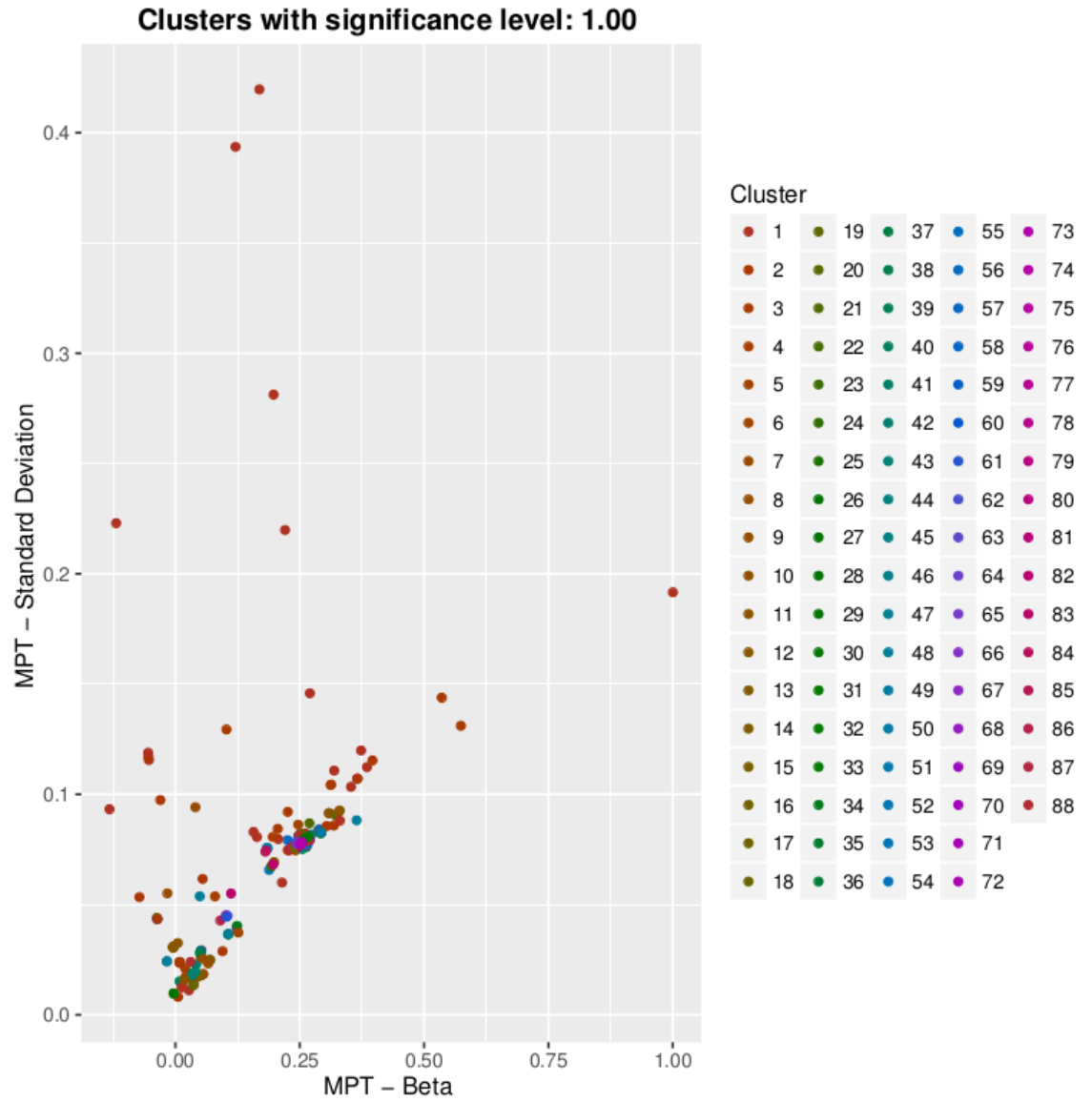


Image 5.18. Scatter plot of the relationship between the beta and standard deviation statistics with the color codes of the classification produced with a significance of 1.00.

Nonetheless, the funds pertaining to lowered numbered groups (1 to 18, in the classification with significance level 1.00), occupy most of the outlier positions (see **Image 5.18.**).

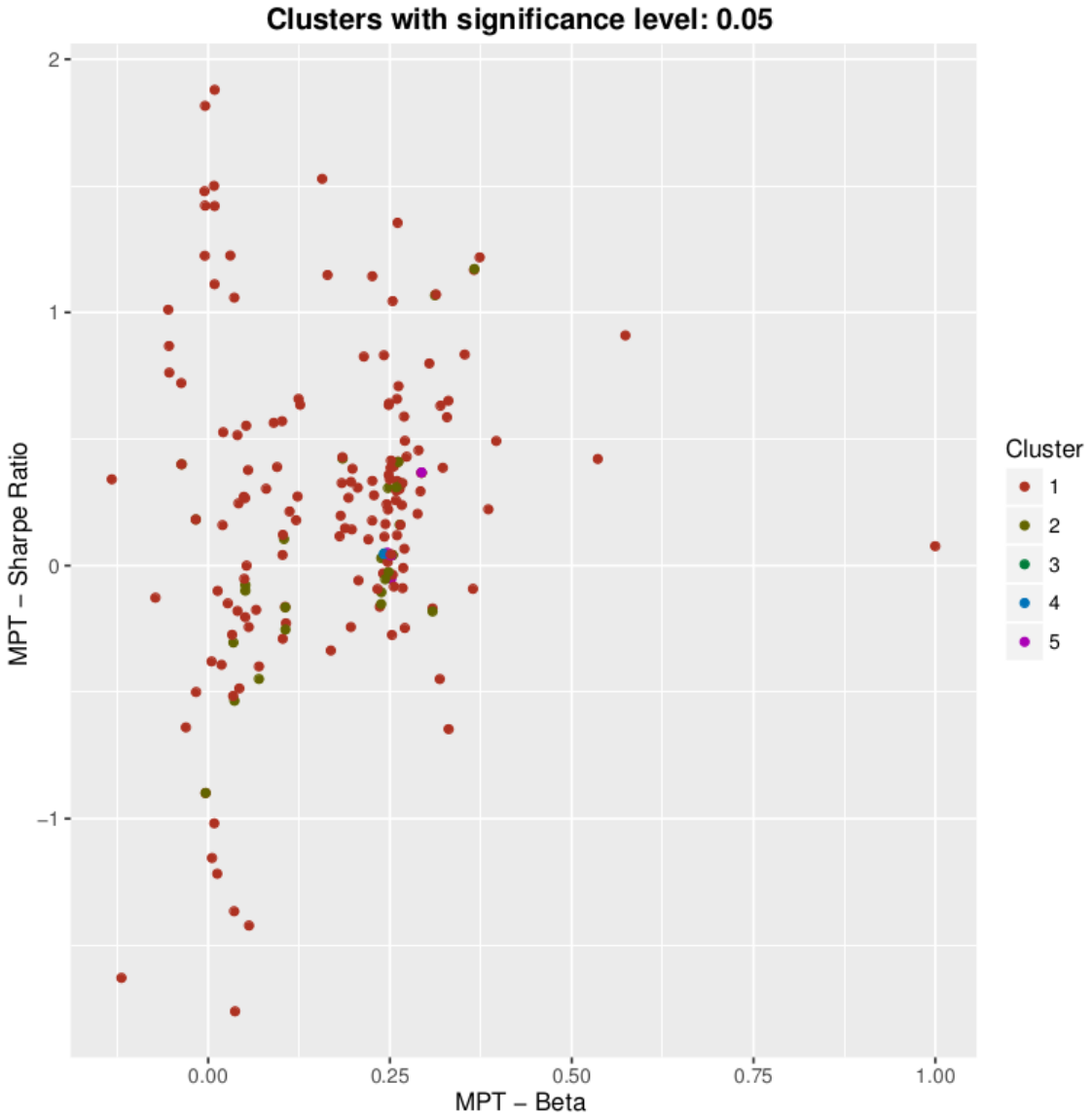


Image 5.19. Scatter plot of the relationship between the beta and Sharpe ratio statistics with the color codes of the classification produced with a significance of 0.05.

With a very scattered location of the funds, the graph with axes beta and **Sharpe ratio** displays a concentration of funds inside a wide square in the left side, between coordinates (0.0, -0.5) and (0.375, 1.0), as shown in **Image 5.19.**

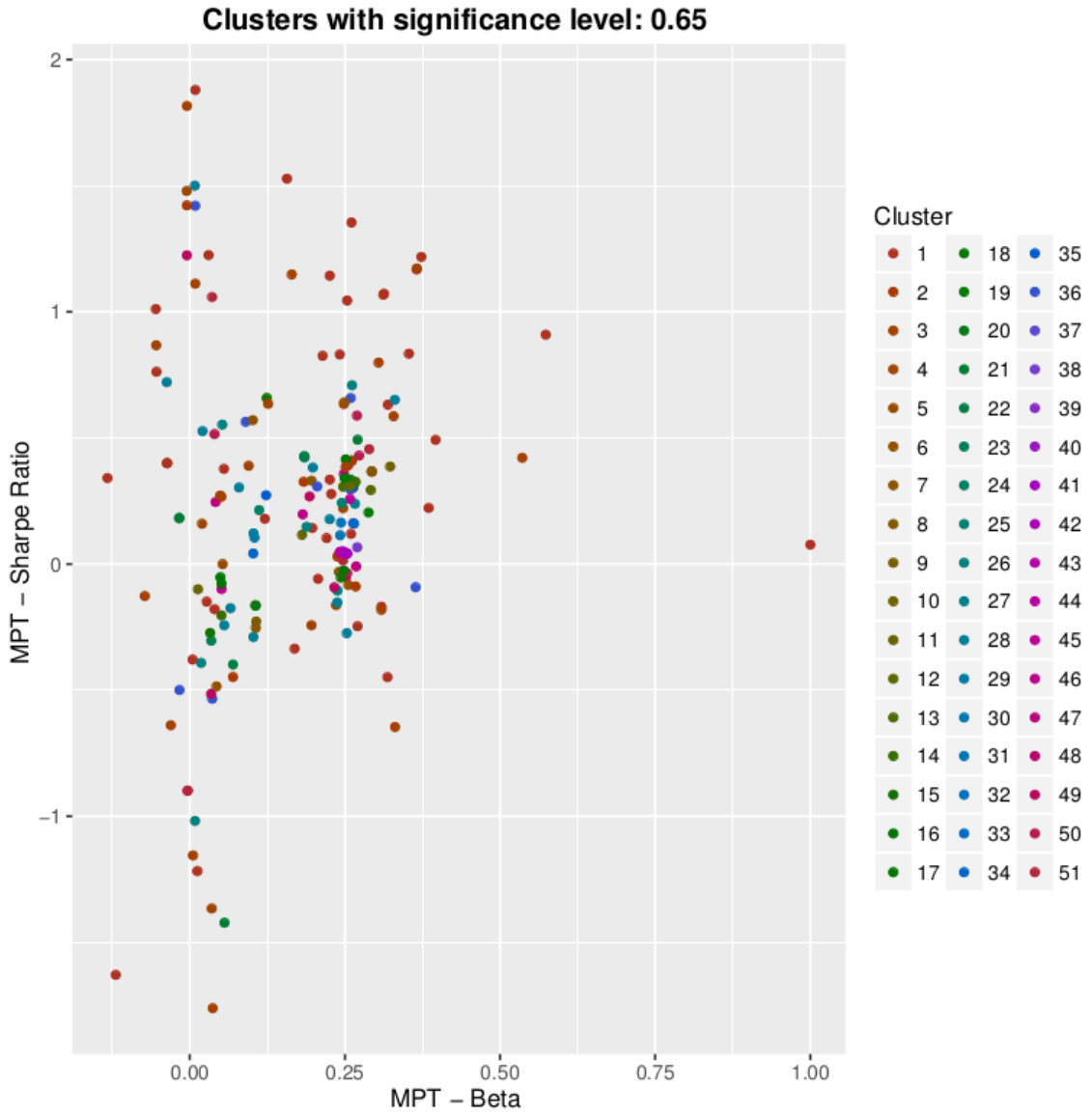


Image 5.20. Scatter plot of the relationship between the beta and Sharpe ratio statistics with the color codes of the classification produced with a significance of 0.65.

In **Image 5.20.**, funds from different groups mingle together without a clear division or limit among them.

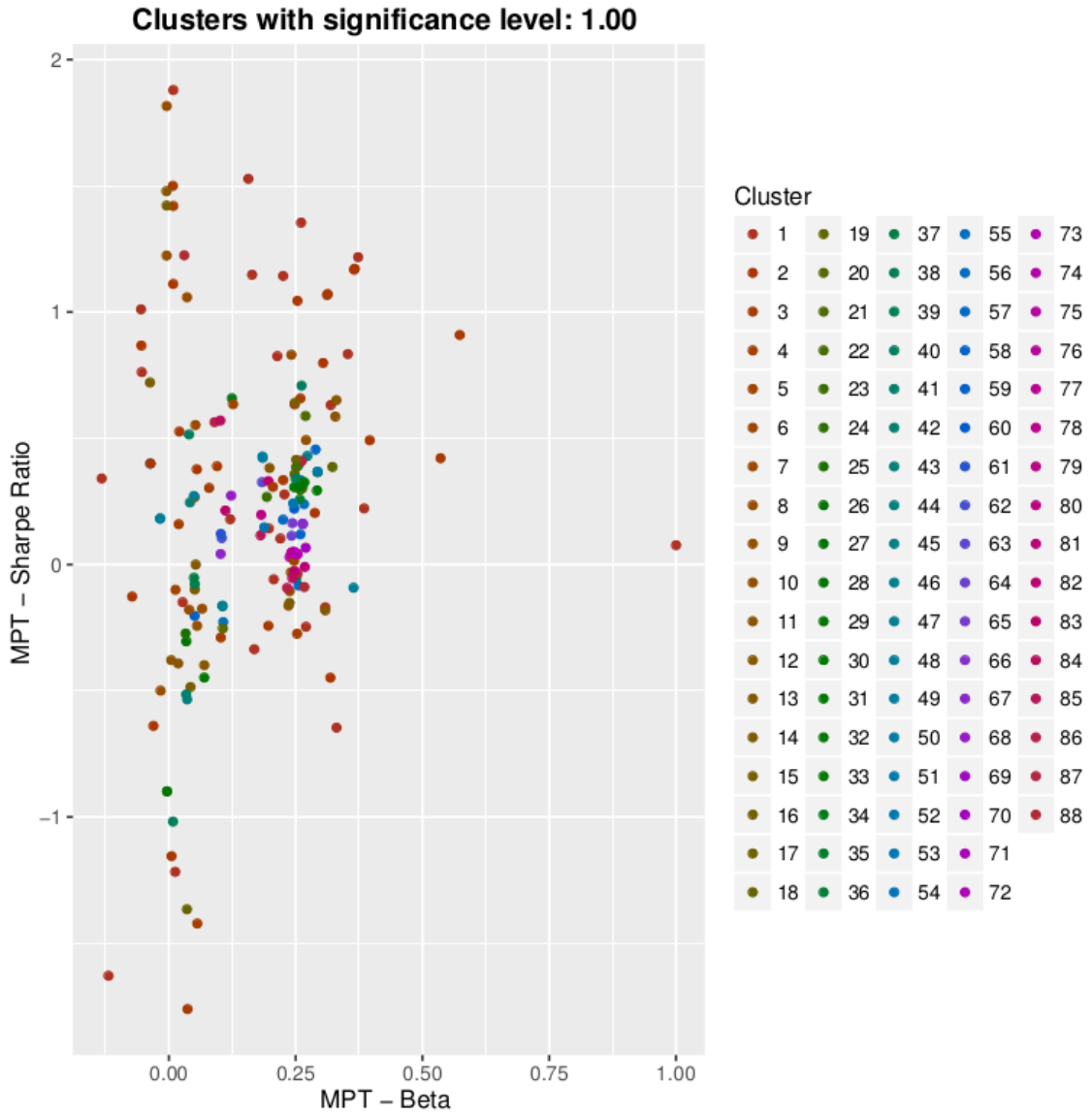


Image 5.21. Scatter plot of the relationship between the beta and Sharpe ratio statistics with the color codes of the classification produced with a significance of 1.00.

Also, as had happened before, as the number of classification groups increased, most funds in the outlier locations usually belonged to low numbered groups, 1 to 20, as seen in **Image 5.21.** For example, the outliers with Sharpe ratios higher than 0.75 and lower than -1.0, belong to groups 1 to 20; as well as funds with beta values higher than 0.375.

Because the comparison with beta, as the x axis, with the other MPT statistics did not yield any clear relationship in respect to the classifications, regardless of the significance level that produced the classifications, it is concluded that there is no direct correlation between this parameter and the fund classifications.

5.2.3.3. Scatter plots with the r squared statistic

The plots with the comparisons between the r squared, as the x axis, and the remaining MPT measures, as the y axis, are described below.

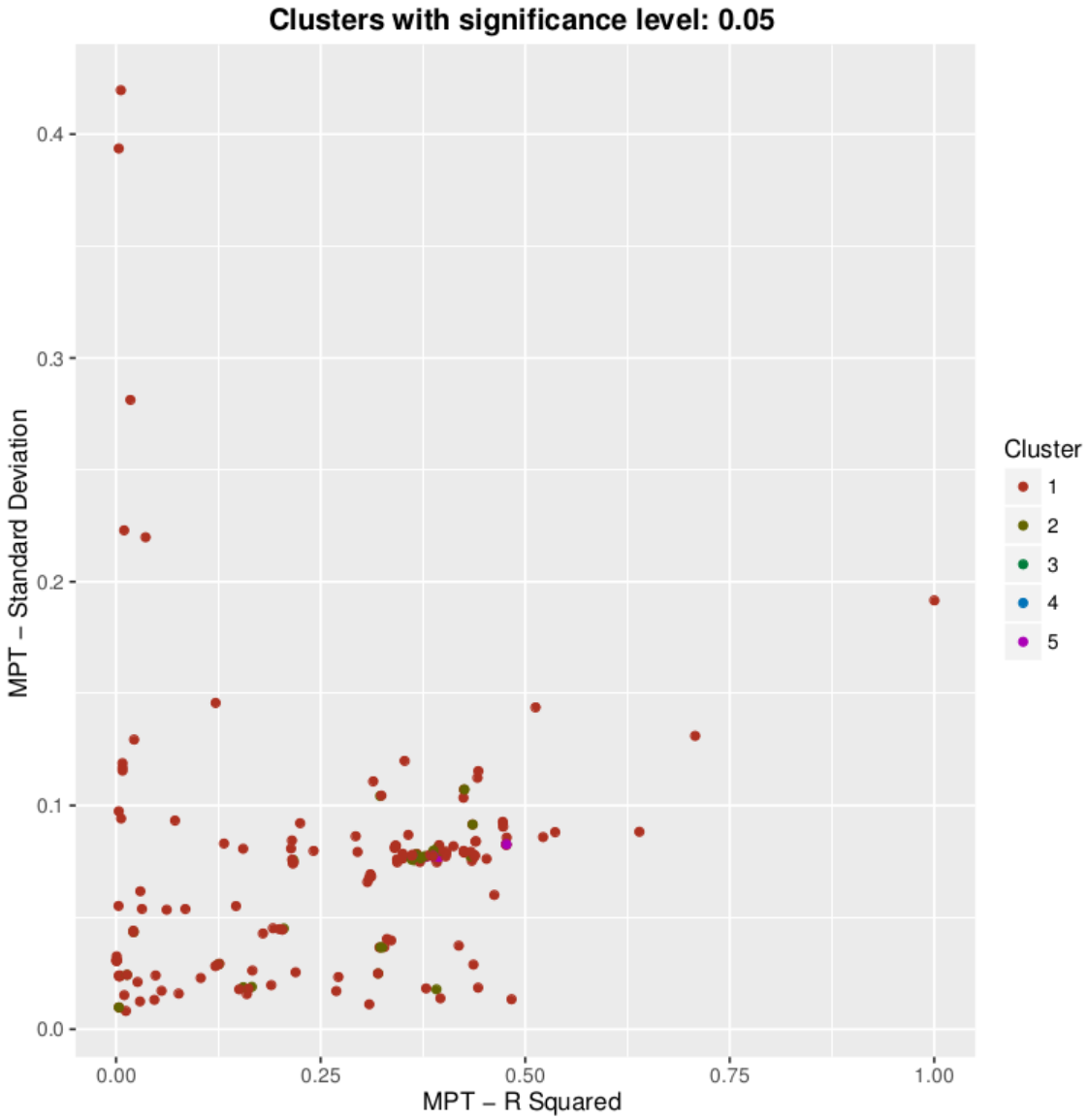


Image 5.22. Scatter plot of the relationship between the r squared and standard deviation statistics with the color codes of the classification produced with a significance of 0.05.

The comparison between the r squared axis and the **standard deviation** axis did not produce a very scattered plot, as displayed in **Image 5.22.** Most funds were concentrated in the part of the plot bounded by a square in the bottom left side of the graph, with lower left corner coordinate (0.0, 0.0) and upper right corner coordinate (0.5, 0.1). This area had a few clusters of funds that hinted to a defined clustering.

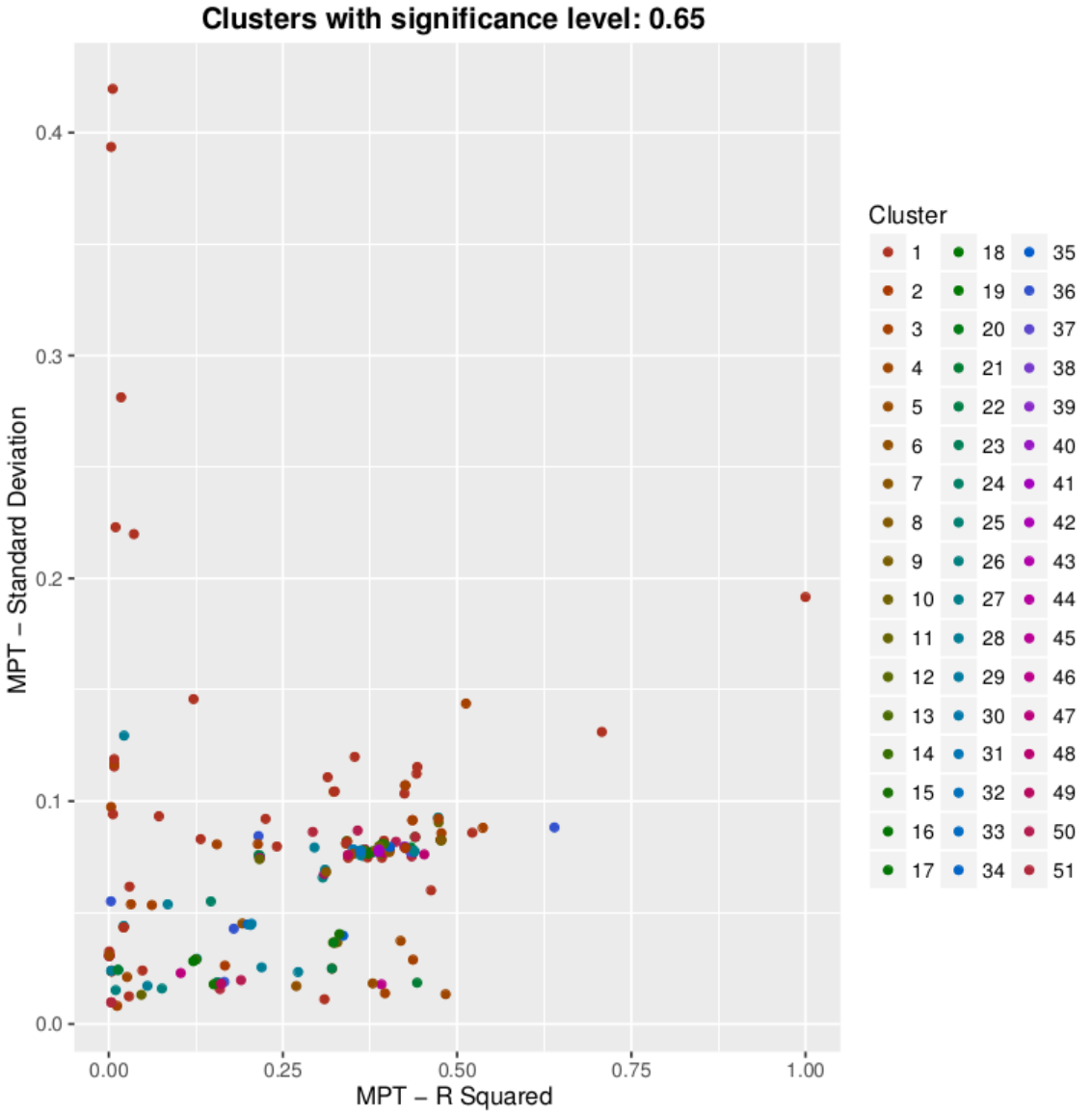


Image 5.23. Scatter plot of the relationship between the r squared and standard deviation statistics with the color codes of the classification produced with a significance of 0.65.

As the significance level increased, the resulting classifications grouped together funds with similar r squared and standard deviations values, belonging to different groups (see **Image 5.23.**).

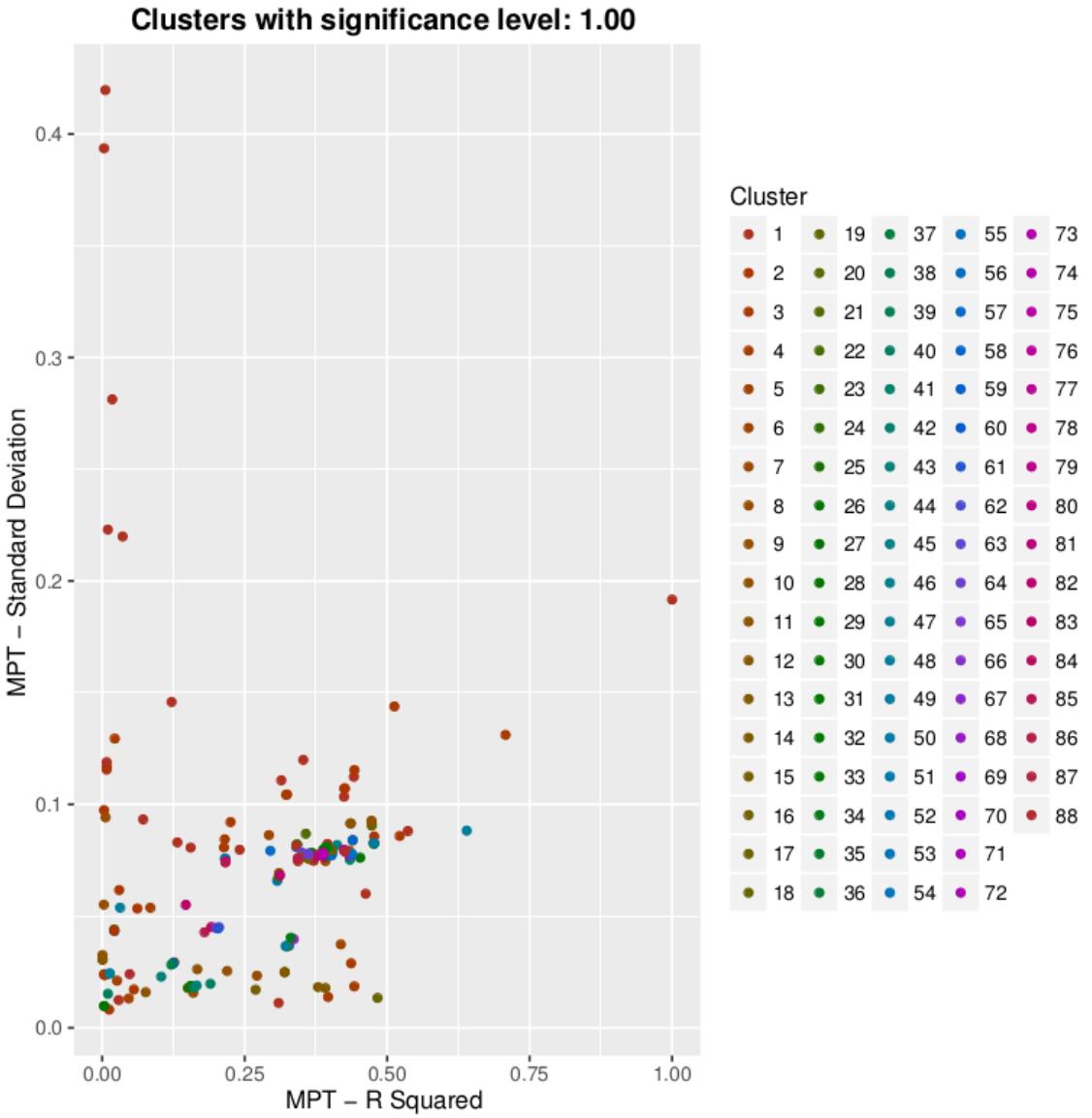


Image 5.24. Scatter plot of the relationship between the r squared and standard deviation statistics with the color codes of the classification produced with a significance of 1.00.

Nevertheless, the behavior where, as the significance level increased (as seen in **Image 5.24.**), the lower numbered groups of the classification included most funds in the outlier locations of the graphs, was seen again.

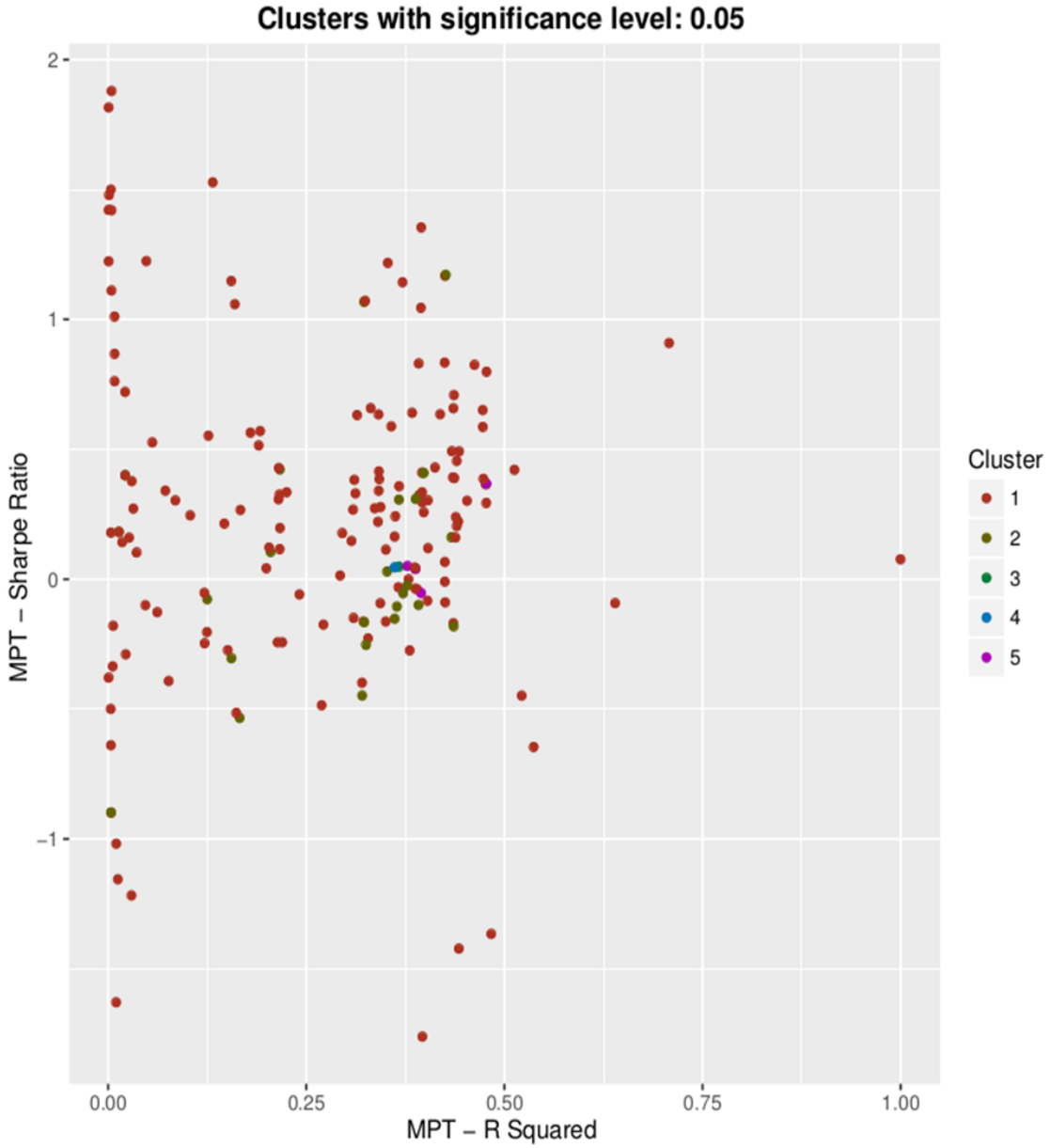


Image 5.25. Scatter plot of the relationship between the r squared and Sharpe ratio statistics with the color codes of the classification produced with a significance of 0.05.

The graph with the comparison between the r squared axis and the **Sharpe ratio** axis (**Image 5.25.**), shows a very scattered distribution of funds. However, most funds were concentrated in the left middle area of the plot, specifically in the square with lower left corner coordinate (0.0, -0.5) and upper right coordinate (0.5, 1.0).

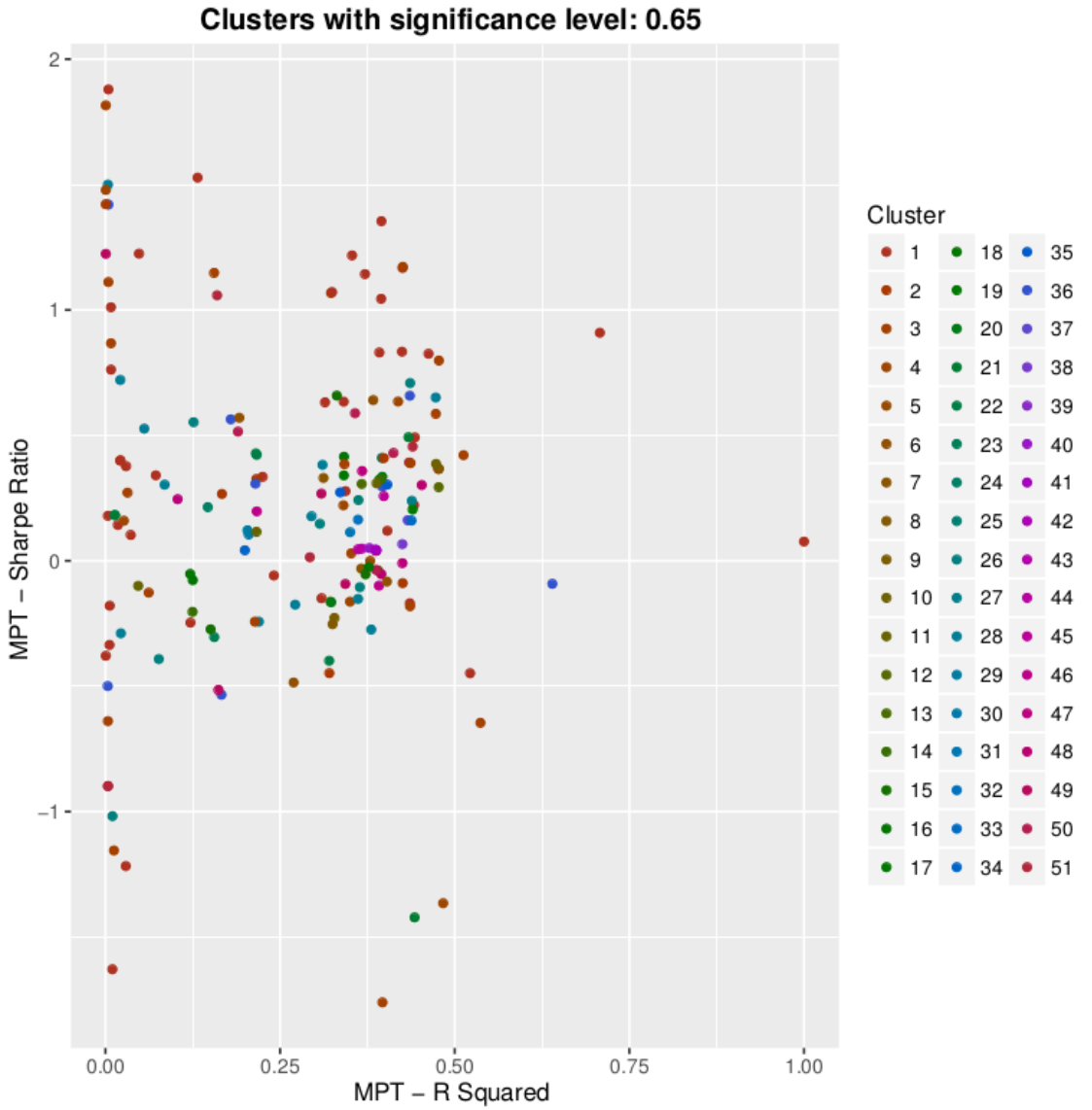


Image 5.26. Scatter plot of the relationship between the r squared and Sharpe ratio statistics with the color codes of the classification produced with a significance of 0.65.

Apparently, the funds' classifications did not bear a strong direct relationship with this MPT statistics comparison. As funds from the same groups were plotted near funds from other groups and separated from funds that belong to their own groups (see **Image 5.26.**).

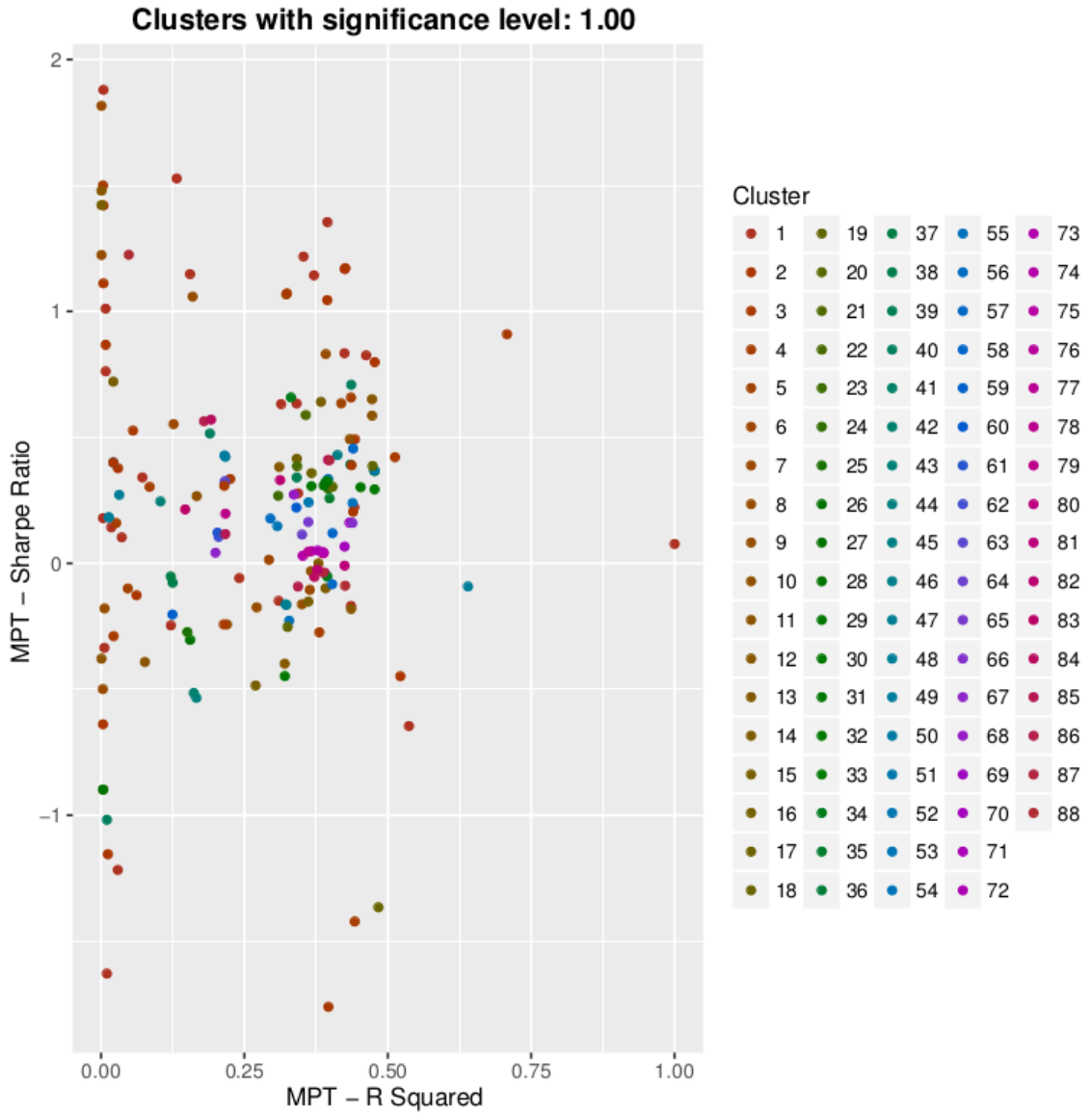


Image 5.27. Scatter plot of the relationship between the r squared and Sharpe ratio statistics with the color codes of the classification produced with a significance of 1.00.

Just as in previous cases, in the scatter graphs of the classifications produced with higher levels of significance, the behavior where most of the funds in the outlier locations of the graphs belonged to the lower numbered groups of the classification (groups 1 to 20), showed up again (**Image 5.27.**). For example, most funds with Sharpe ratios higher than 0.75 and lower than -1.0, regardless of their groups, were classified in those groups.

The high level of scattering of the classified funds in the graphs, provides an inkling that the r squared measure has no direct relation in the hierarchical classification process.

5.2.3.4. Scatter plots with the standard deviation statistic

Given the existing relationship between the standard deviation and the **Sharpe ratio**, there was an opportunity that these variables could uncover some unseen relationship between the price time series classification and their traditional risk measures.

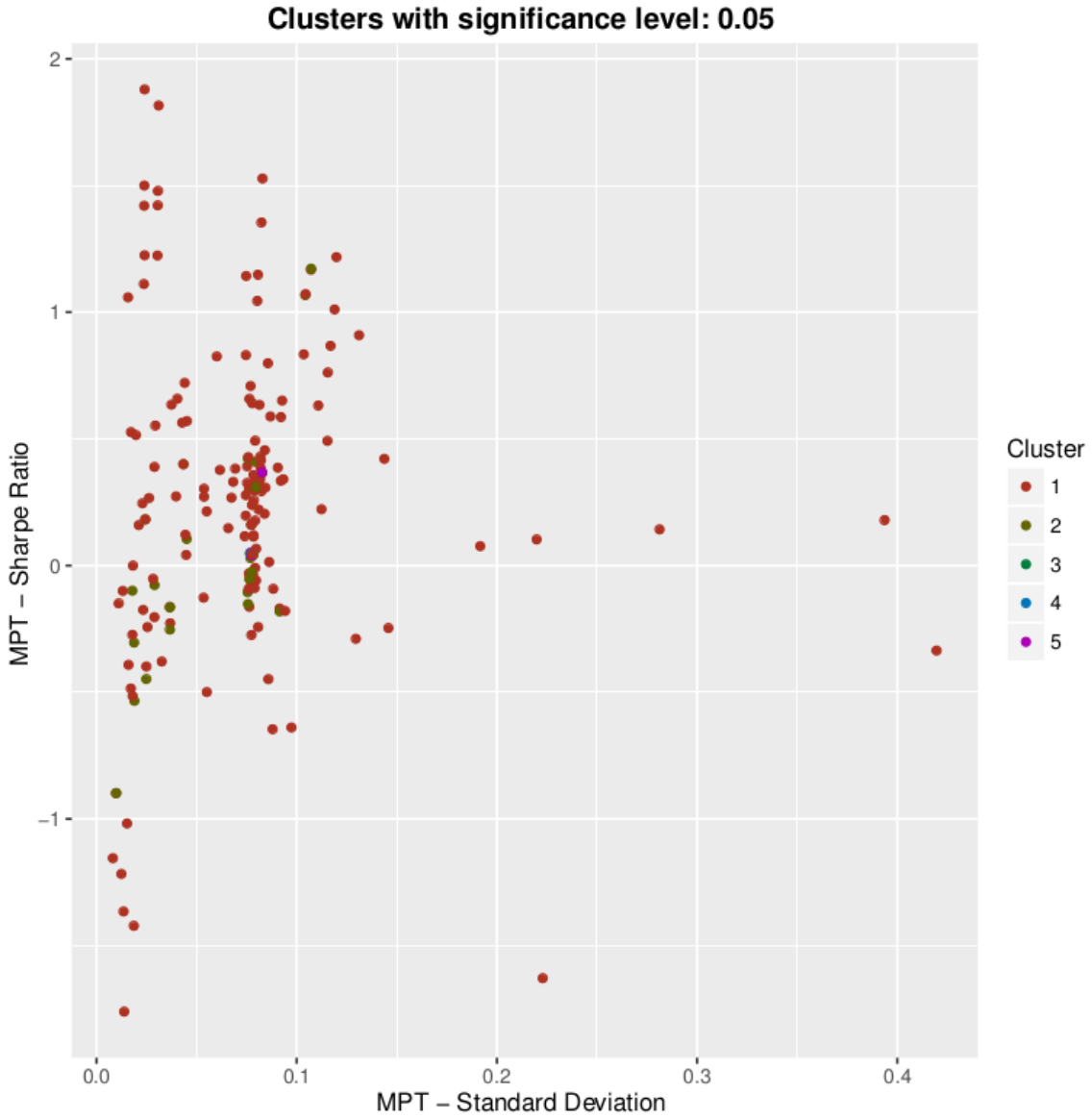


Image 5.28. Scatter plot of the relationship between the standard deviation and Sharpe ratio statistics with the color codes of the classification produced with a significance of 0.05.

The linear pattern, aligned with the Sharpe ratio axis, was repeated again, but with funds with standard deviations values close to 0.0 (**Image 5.28.**), instead of the r squared values (as seen in **Image 5.25.**). Unlike the comparison with the r squared metric, this graph shows a less scattered pattern of funds, with most of them concentrated in the furthermost left middle side of the graph, inside the square with left lower corner (0.0, -0.5) and upper right corner (0.1, 1.0).

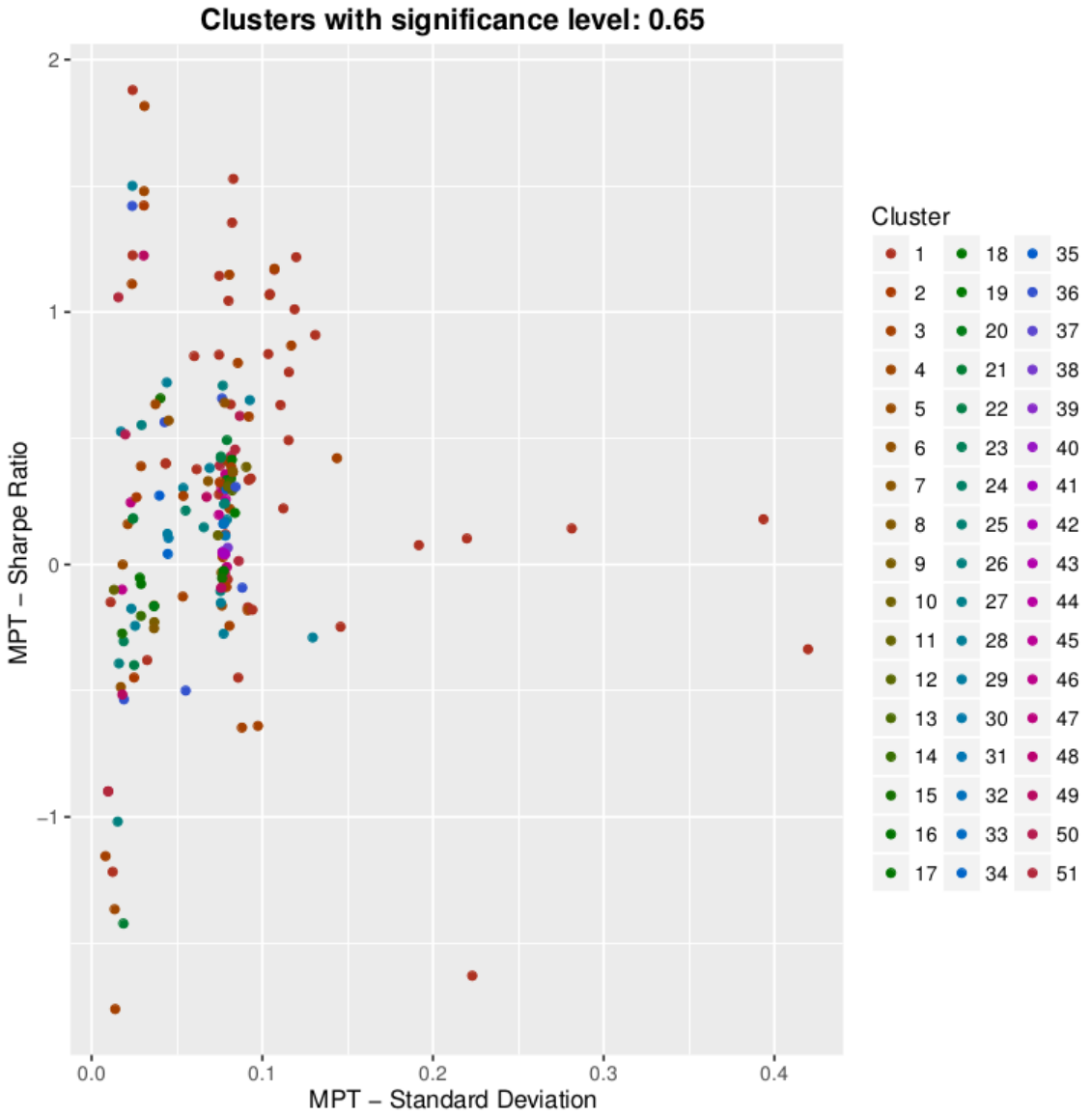


Image 5.29. Scatter plot of the relationship between the standard deviation and Sharpe ratio statistics with the color codes of the classification produced with a significance of 0.65.

Sadly, it seemed that neither of these statistics contributed strongly to the funds classification process. The funds from the same groups had a variety of standard

deviation and Sharpe ratio values that did not match or were close to the values of other funds in their own groups (**Image 5.29.**).

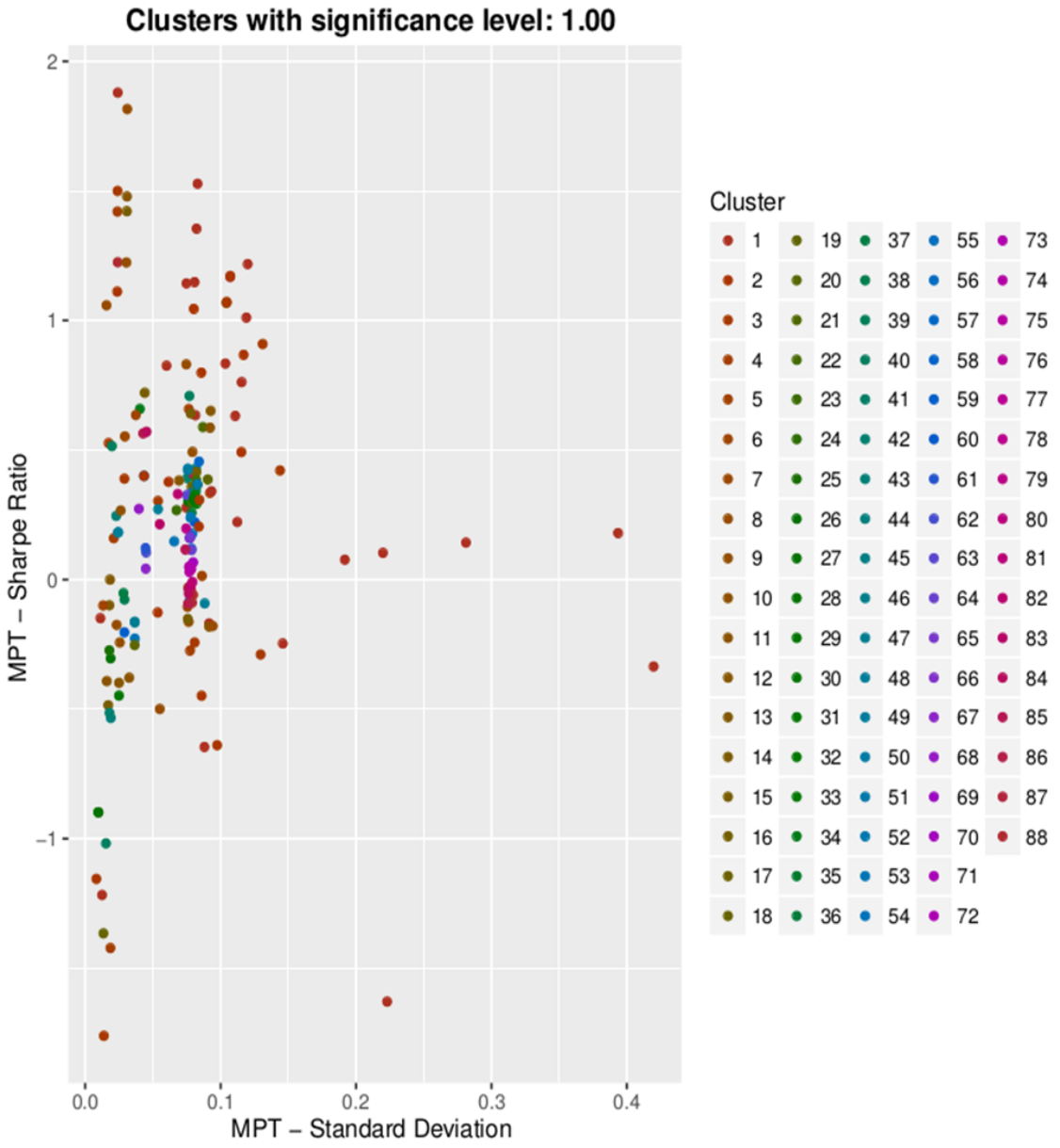


Image 5.30. Scatter plot of the relationship between the standard deviation and Sharpe ratio statistics with the color codes of the classification produced with a significance of 1.00.

As displayed in **Image 5.30.**, the constant observable behavior in the comparison graphs with classifications from higher levels of significance, was that most funds in the outlier locations belonged to the lower numbered groups of the classification (groups 1 to 20).

Most funds with Sharpe ratios higher than 0.75 and lower than -1.0, and funds with standard deviations higher than 0.1, were classified in those groups.

This very simple evidence leads to the conclusion that no particular, single MPT statistic, or pair of traditional financial metrics, can successfully show when two funds are similar. Clearly, if the MPT variables describe correctly the similarities between funds, the dependencies are much more elaborate than a simple eye comparison can elucidate.

This is, the MPT variables presented as such, are of little help to investors who are comparing among different funds.

5.3. Description of the scripts to perform the clustering analysis

To perform the analysis described in this chapter, two scripts were developed: the first one, executed the clustering analysis, and the second, created the scatter plots used in the MPT variables comparisons displayed in the previous section.

5.3.1. Description of the script to perform the clustering analysis

The examples from the TSclust package reference manual, from webpage "*CRAN - Package TSclust*", were used as reference to code the script for performing the clustering analysis. However, TSclust requires of other packages' data structures and functions support. They are listed below.

R 3.2.3 packages and versions used in the clustering analysis script			
Component	**Version**	**Component**	**Version**
TSclust	**1.2.3**	ifultools	2.0-1
Xts	0.9-7	MASS	7.3-45
Zoo	1.7-11	Pdc	1.0.3
Wmtsa	2.0-0	cluster	2.0.3
splus2R	1.2-0		

Table 5.2. List of R packages and versions needed in the clustering analysis of equity funds.

The script, named "`ananov_tsclust.r`", requires an input text file in CSV format, which contains the normalized historical price series for each fund, ordered by fund's ticker symbol and date (in format `YYYY-MM-DD`), with the data structure: fund, date, (normalized) price.

Once executed, the script produces two text files in CSV format:

- "`RVMexico.tsclust_analysis_CORT_<YYYYMMDDHHmmSS>.csv`", which includes a matrix with the d_{CORT} measures of the funds compared with each other.

- "`RVMexico.tsclust_analysis_CORT_pclust_<YYYYMMDDHHmmSS>.csv`", which lists the clustering results, per fund and significance level, based on the d_{CORT} measures of the funds.

The **Mexican Stock Exchange** (**BMV**, *Bolsa Mexicana de Valores*) index price time series was included in the script code as the first row of the time series data matrix used to calculate the dissimilarity measures. This decision was made with the purpose of forcing the clustering algorithm to always assign the BMV index in the cluster (group) 1 in the resulting classification and to ease its identification.

Challenges

TSclust was installed, and compiled, after the installation of the PerformanceAnalytics package. No other technical issues or problems occurred during the development and testing of the script.

5.3.2. Description of the script to create the scatter plots

The basis for code of the graphs comes from the examples in the website "*Cookbook for R*", a companion website to Winston Chang's book of the same name. The script required of the package **ggplot2 2.1** to create the scatter graphs used in section 5.2.

The script was named "ananov_tsclust_scatterplots.r", and its input is a CSV file named "RVMexico.capm_pclust_<*YYYYMMDDHHmmSS*>.csv", which contains: the ticker symbols of the funds and the BMV index, five columns with the values of the MPT statistics, and eleven columns for each of the significance levels' classifications.

The script chooses a combination of two the MPT variables to compare and, for each significance level classification, constructs the instructions to graph a scatter plot. The instructions are saved in a string variable that is used to create another R language file, named "ananov_tsclust_scatterplots_instructions.r", which is manually executed to create all the comparison graphs.

Besides image files (in PNG format) for each scatter plot, the ggplot package creates a default PDF file (named "Rplots.pdf", by default) with all the graphs generated during script execution, but with the default scale and size. Those graphs were scaled to size "4", 295 centimeters per side, to facilitate the analysis of the funds distribution. To ease their visualization, the scatter plots included in this chapter were the unscaled versions from the "Rplots.pdf" file.

The image files were named with the structure: "RVMexico.capm_pclust_sig<significance_level>_capm_<MPT_var1>X<MPT_var2>.png".

Challenges

The ggplot library was installed in the user session of the described equipment in chapter 4. During the development and testing of this script, no technical issues or problems occurred.

6. Conclusions, Findings and Future Work

6.1. Conclusions

To provide a clearer report of the results, the conclusions are divided in two sections, between the results of the objectives and the conclusions of the research questions set in chapter 1.

6.1.1. Were objectives achieved?

Overall, the main objective: "*Build the basis for the development of an analysis tool of investment mutual funds in the Mexican market for the general public*", was partially achieved. To explain why, is important to understand the conclusions for each of the objectives.

The specific objectives that were planned to aid in the achievement of the main objective were:

- Define a method and procedure to extract mutual funds information from the CNBV website.
- Define and create a procedure to automatize the download of data of mutual funds registered and approved by the CNBV. Including their information sheets and historical price series.
- Calculate the traditional MPT measures for Mexican equity funds registered and approved by the CNBV.
- Select a machine learning method for time series clustering analysis to be applied to the funds' historic prices.

- Define patterns or profile groups based on the results of the clustering analysis performed in the historic funds' prices data and compare this analysis with the traditional MPT measures.

- Uncover behavioral patterns in equity funds that could help potential investors in the selection of funds, according to their investment objectives and risk aversion.

- Explore and evaluate if the traditional MPT measures can be substituted by a simple and novel clustering machine learning method.

The achievement level of each objective is described below.

Define a method and procedure to extract mutual funds information from the CNBV website.

The developed web scraping project, built with the *Scrapy* framework, did achieve the objective of creating a procedure to extract the information of the mutual funds registered in the National Banking and Stock Commission (**CNBV**, *Comisión Nacional Bancaria y de Valores*) website. However, as detailed in chapter 3, the information extraction process is not completely automatized and the downloaded funds' information had to be converted from the JSON data structure format to the CSV text file format before the information could be used.

Define and create a procedure to automatize the download of data of mutual funds registered and approved by the CNBV. Including their information sheets and historical price series.

The process to acquire the data of the funds had to be divided in two because, as mentioned for the previous objective:

1. The source of the funds' information is located at the CNBV website and their historical price series data was obtained from the financial news portal **Yahoo! Finanzas**. At the moment when the web scraping and data downloading research was made, the Scrapy framework 1.0.3 did not have the functionality to extract information from two different sources, join the gathered data and save the unified information in a single file or database.

2. To extract the information an additional process was required. The web scraping process of the CNBV website implied the development of a custom process to extract information from an information source that did not provide a "friendly" data download service (like an online API). The price series download was facilitated by the availability of infrastructure and technology, namely, the **Yahoo Query Language** (***Yql***), which allows third party developers to program any set of code to aid in the gathering of data from any of the Yahoo! information services (such as the yahoo-finance Python library). The download of the funds' information sheets ***was excluded*** from the data targets gathering due to time constraints. These sheets are scattered (in PDF format) at the funds' respective management and brokerage companies; sometimes they were not available to the public at all. The extraction process would require to either manually search and download the information sheets or to develop an extraction process for each broker's website, with the help from the *Scrapy* framework. Because the information from the CNBV website and the historical price series from Yahoo! Finanzas constituted sufficient data to execute the experiments, the gathering of the information sheets was regarded as avoidable.

Calculate the traditional MPT measures for Mexican equity funds registered and approved by the CNBV.

As was described in chapter 4, the **Modern Portfolio Theory** (MPT) statistics were calculated for the 182 equity funds selected for analysis, with the aid of the R language package PerformanceAnalytics.

Select a machine learning method for time series clustering to analyze the funds' historic prices.

The clustering method chosen for the analysis of the selected equity funds was a hierarchical clustering process with the dissimilarity measure d_{CORT}, both described in chapter 2, and calculated with the help of the functions from the R package TSclust.

Define patterns or profile groups based on the results of the clustering analysis performed in the historic funds' prices data and compare this analysis with the traditional MPT measures.

As detailed in chapter 5, **no discernible pattern was discovered among the comparisons of the funds' classifications and their traditional MPT measures**. In some cases, regardless of the significance level of the classification, the funds within the same group were not plotted in clusters. In other cases, a group's funds were plotted together, but mixed with funds from other groups. However, a common behavior in many of the scatter plots with classifications from high significance levels, was that funds in the outlier locations usually belonged to the classifications with the low numbered groups (most of the times, the groups 1 to 18), which in hierarchical clustering is not surprising.

Uncover behavioral patterns in equity funds that could help potential investors in the selection of funds, according to their investment objectives and risk aversion.

The lack of a discernible relationship between the studied equity funds' classifications and their traditional MPT statistics in this experiment, suggests that the traditional risk and profit evaluation measurements of mutual funds do not provide enough information for their proper comparison. This implies that it is other characteristics of the funds that have an effect in the classification of funds.

Explore and evaluate if the traditional MPT measures can be substituted by a simple and novel clustering machine learning method.

With the gathered information of the funds and the machine learning analysis performed, it is difficult to determine if the traditional financial risk analysis can be substituted by a clustering algorithm. More experiments and analysis are required.

6.1.2. Where the research questions answered?

The research questions guiding the present work were:

- Is it feasible to create a simple method or process to extract, clean and preprocess Mexican mutual funds' information from the CNBV website and other public financial information providers' websites for further analysis?

- Is there a correlation between the traditional Modern Portfolio Theory measures (Alpha, Beta, R-Squared, Standard Deviation and Sharpe Ratio) and the clustering analysis results for equity funds?

- After applying a machine learning data analysis with a clustering method, do the resulting clustering of funds provides a meaningful grouping, or classification, of equity funds different from the classification provided by traditional financial entities (CNBV, Morningstar, etc.)?

- Do the resulting clustering of funds relates to an observable characteristic, or combination of funds' characteristics, that can be used to create investment profiles for guiding and advising novice investors in the choosing of an adequate investment?

- How does the clustering of mutual funds compares or relates to the Morningstar Rating of mutual funds?

Is it feasible to create a simple method or process to extract, clean and preprocess Mexican mutual funds' information from the CNBV website and other public financial information providers' websites for further analysis?

It was not feasible to create a single extraction process that could download information from, both, the CNBV website and from the Yahoo! Finanzas web portal. Instead of a single unified program, it was found that a process divided in three phases, each supported by a different technology, was the best solution.

Is there a correlation between the traditional Modern Portfolio Theory measures (Alpha, Beta, R-Squared,

Standard Deviation and Sharpe Ratio) and the clustering analysis results for equity funds?

It was not possible to reveal a visible relationship between the clustering of all the funds, produced with the chosen machine learning algorithms, and their traditional MPT statistics.

After applying a machine learning data analysis with a clustering method, do the resulting clustering of funds provides a meaningful grouping, or classification, of equity funds different from the classification provided by traditional financial entities (CNBV, Morningstar, etc.)?

The comparison of the funds' classifications was done against their MPT metrics; which, as mentioned in chapter 2, are some of most common variables used in the financial industry to assess and select an investment instrument. The search for a straight relationship between this statistics and the discovered classifications did not lead to a direct link between them.

Do the resulting clustering of funds relates to an observable characteristic, or combination of funds' characteristics, that can be used to create investment profiles for guiding and advising novice

investors in the choosing of an adequate investment?

It is possible that the classifications of funds are due to, or affected by, other characteristics of the funds (i.e. their investment composition, their main industry of investment, main type of investment instrument, investment style, etc.). But, given that the objective of this research was to look for a better, and easier, form to classify equity funds, the comparisons were focused on the existing methods of fund classification employed by the financial industry.

How does the clustering of mutual funds compares or relates to the Morningstar Rating of mutual funds?

As was detailed in chapter 3, the extraction of data from two (or more) different sources it's a feature unavailable in the *Scrapy* framework 1.0.3.

The mutual funds' pages, accessible by the CNBV mutual funds' searcher, were the main source of information; but those pages did not include the Morningstar rating, or the rating of a third party financial investments evaluator. Therefore, it was considered that the time needed for the gathering of this information did not correspond to the time frame of this research. Thus, this comparison was not performed.

6.2. Other findings

When the research for this thesis began, it was expected that the most difficult part would be the cleaning and preprocessing the price series, the calculation of the traditional MPT metrics or the execution of the clustering algorithm. Although, those parts of the research did take time to complete, the most surprising and time consuming tasks came from the data gathering and acquisition phases.

6.2.1. Information source for registered Mexican mutual funds

The only mutual funds covered by Mexican investment laws are the mutual funds registered at the **National Banking and Stock Commission** (CNBV). But, as mentioned in chapter 3, it was a surprise to find out that the mutual fund search engine is provided by the investment firm **Morningstar Inc.**, at this web address:

`http://lt.morningstar.com/7ap7omrzjm/fundquickrank/default.aspx`

Because the CNBV is a government division whose primary responsibility and goals are to oversee financial and banking law and its enforcement in banks and financial institutions in Mexico, it's understandable that they may outsource their financial information services to a specialized company.

Oddly, given that one of their primary services is financial and investment information, Morningstar did not include functionality or tools to download the information of the listed funds in the basic mutual fund searcher. The advanced searcher did provide a fund comparison tool, but it only worked online and did not allow the download of the information of the compared funds.

The CNBV should make it easier for the public to acquire (free of charge) the data of all the investment funds, in order to allow for the making of informed decisions when choosing Mexican investment instruments.

6.2.2. Historical price data of Mexican mutual funds

Some information services companies, like Yahoo!, are aware of the importance of allowing their users to have access to all the information they may require. They have built a platform that allows any third party developer to construct specialized tools for this purpose.

Given that the information required to make the current research had a high purchasing cost at the official source, the CNBV, it was extremely useful that Yahoo!, through its information service *Yahoo! Finanzas* and the third-party developed yahoo-finance Python library, allows this information to be downloaded free of charge to the public. However, as explained in chapters 3 and 4, the missing data from some funds and the suspected data errors in some others, make it difficult to perform its automated analysis. Thus, it is necessary to develop procedures to review and audit the data.

While some brokers and investment companies do provide the information and the historical price series of the mutual funds they manage, there is not a standard government protocol about how to offer this information to potential investors (periodicity, structure of the information, file format, temporality of the information, statistics, etc.), which makes data gathering a tiresome and daunting task.

6.3. Future work

Using the procedures developed for performance of the experiments within this thesis, other variables could be tested, such as: extending the time period of the historical price series, use different dissimilarity measures, etc.

1. As mentioned in chapter 5, the presented classification was executed with the d_{CORT} default values: dissimilarity weight of "2" and discrepancy method "Euclid".
   a. The weight, or effect, that the correlation CORT has on the raw dissimilarity measured with the "Euclid" method changes the dissimilarity values of the funds and could find different classifications for the funds.
   b. The raw data discrepancy can also be measured with the methods: "Fréchet" distance and "Dynamic Time Warping" distance (DTW). Both distance methods calculate different distance values for the funds and can help to find different classifications.

2. Other dissimilarity algorithms, compiled by Montero and Vilar (2014) in the TSclust package, could also provide different measures between the funds' price time series and lead to different classifications. Such as:

a. ***Dissimilarity based on the symbolic representation SAX***: its approach includes the normalization of data series and the division of the data in segments (just as the traditional financial assessment of investment instruments, when is performed in a periodic basis).

b. ***Dissimilarity measure based on the discrete wavelet transform***: the main use of this algorithm is as an aid in the choosing of an appropriate scale to obtain an accurate clustering for the analyzed time series. Although, the scale is for the wavelet approximation coefficients of the time series, this algorithm might allow to find a more accurate and meaningful clustering of funds without losing information from the original data series.

3. The preprocessed data could be used to perform other types of machine learning analysis, besides clustering.

4. This analysis could also be replicated with debt funds or with international equity funds.

6.4. Discussion

6.4.1. What was done right during the research work?

Limiting the selection of the dissimilarity measures and the clustering algorithms to a set of already implemented dissimilarity measures was a good decision. It could be argued that the compilation of dissimilarity measures made by Montero and Vilar (2014) restricted the type of analysis performed. But, given that the main objective of this work was to find new information about the funds' historical price time series with already available, tested and proved, machine learning algorithms, the selection provided by Montero and Vilar (2014) was a good starting point.

Also, the choosing of well known technologies (like MariaDB) to perform and store data for this research work was a good decision. While it would have been ideal to develop all

the experiment's processes using new and novel technologies oriented to the execution of machine learning algorithms and "big data" analysis (such as NoSQL databases, like mongoDB), it is important to keep in mind that many novel technologies simply offer a specialized functionality that can also be replicated by previous and generic technologies.

6.4.2. What must be done different to obtain better results?

Whenever possible, it would be ideal to stop using text files to load and store intermediate analysis results. Reading and writing CSV text files, with the results of the traditional financial measures and the clustering analysis, had the intent of eliminating possible communication errors while accessing the MariaDB database where the preprocessed data was stored. Also, because it was quicker to code the scripts and review the analysis results before continuing with the next analysis phases.

Unfortunately, this form of running the Python and R language analysis scripts requires the careful creation of the CSV text files with the price time series information and the constant reviewing of the created text files to avoid data corruption. Otherwise, human errors could propagate and void the results of the next analysis.

In retrospective, it could had been a better approach to automatically generate an extra file with the input data for the next script in the analysis process. The file with the MPT statistics had to be manually joined with the clustering results. The combination of the results into a single script process could have been more efficient, reducing the risk of a human induced error.

6.5. Overall conclusion

The goal of building the basis for an *automated mutual funds investor advisor* was not completely fulfilled, but it cannot be considered a failed effort.

First, the equity fund clustering suggests that there is not a direct relationship between the behavior of the chosen funds' price series, between the years 2011 and 2015, and the MPT statistics calculated for the same period of time.

The results hint that more analysis of the behavior of the price series of funds is required to properly provide a simpler classification of the available investment options to the general public. Although, the causes of the produced classifications may not be evident, it is still important to find the underlying causes for the classification and to analyze if those causes can be translated into meaningful information for the investors.

Second, due to the technical challenges and restrictions of the technologies used in the present work, joining the developed process in a single process to build the basis for an automated investor advisor was relegated in favor of acquiring and analyzing the data. As detailed in section 6.4.2. of this chapter, unless all the required process and programs can access the funds information without the intervention of a person, this goal is partially accomplished.

While the traditional financial risk and profit variables were not completely discarded as valid evaluation instruments, it is still important to remain open to new methods and metrics that facilitate the evaluation of funds and any other investment, in order to aid people in the achievement of their financial goals.

7. References

1. **Banasiak, L.** (2015, May 6th). *yahoo-finance*. Python package version 1.2.1. Retrieved on December 2, 2015 from Python Package Index: https://pypi.python.org/pypi/yahoo-finance

2. **Banco de México** (n.d.). *Sistema de Información Económica. Securities prices and interest rates. (CF300) - Prices of sovereign securities (on the run)*. Retrieved on March 18th, 2015 from: http://www.banxico.org.mx/SieInternet/consultarDirectorioInternetAction.do?sector=18&idCuadro=CF300&accion=consultarCuadro&locale=en

3. **Bolsa Mexicana de Valores, S.A.B. de C.V.** (2015). *::. Grupo BMV .::. Productos de Información.* Retrieved on August 4th, 2015 from: http://www.bmv.com.mx/es/productos-de-informacion

4. **Buscador Rápido de Fondos** (n.d.). *Morningstar, Inc*. Retrieved on September 21th, 2015 from: http://lt.morningstar.com/7ap7omrzjm/fundquickrank/default.aspx

5. **Chang, Winston** (n.d.). *Cookbook for R*. Retrieved on May 24th, 2016 from: http://www.cookbook-r.com

6. **Comisión Nacional Bancaria y de Valores** (n.d.). *Buscador de Sociedades de Inversión.* Retrieved on August 17th, 2015 from: http://www.cnbv.gob.mx/SECTORES-SUPERVISADOS/SOCIEDADES-DE-INVERSION/Buscador-de-Sociedades-de-Inversi%C3%B3n/Paginas/Buscador-de-Sociedades-de-Inversion.aspx

7. **Comisión Nacional Bancaria y de Valores** (n.d.). *Buscador de Sociedades de Inversión. Básico.* Retrieved on August 17th, 2015 from: http://www.cnbv.gob.mx/SECTORES-SUPERVISADOS/SOCIEDADES-DE-INVERSION/Buscador-de-Sociedades-de-Inversi%C3%B3n/Paginas/B%C3%A1sico.aspx

8. **Crespo, E.** (2015). *Ernesto Crespo presentations.* [Web log post] SlideShare. Retrieved on September 9th, 2015 from SlideShare: http://www.slideshare.net/ecrespo

9. **Crespo, E.** (2015, January 1st). *Extracción de datos de páginas web con scrapy*. [Web log post] La libertad de desarrollar no tiene precio... Retrieved on September 12th, 2015 from: http://blog.crespo.org.ve/2015/01/extraccion-de-datos-de-paginas-web-con.html

10. **Crespo, E.** (2015, February 3rd). *#MéridaTechMeetup.* [Web log post] La libertad de desarrollar no tiene precio… Retrieved on September 10th, 2015 from: http://blog.crespo.org.ve/2015/02/meridatechmeetup.html

11. **Dorneles, E.** (2014, July 14th). *XPath Tips from: the Web Scraping Trenches* [Web log post] The Scrapinghub Blog. Retrieved on October 23th, 2015 from: http://blog.scrapinghub.com/2014/07/17/xpath-tips-from-the-web-scraping-trenches/

12. **Goldman Sachs** (2016). *[Goldman Sachs India - FactSheet Monthly Fund Update December 2015]*. Retrieved on September 14th, 2016 from Goldman Sachs: http://www.benchmarkfunds.com/gs/Documents/Archives/FactsheetDec2015.pdf

13. **Hennessy Funds** (2016). [Hennessy Funds - Hennessy Japan Fund Fact Sheet - As of December 31, 2015]. Retrieved on September 17th, 2016 from: https://hennessyfunds.com/resources/docs/literature/factsheets/2015-12-31/Japan_Fund.pdf

14. **Informe de fondos. ACTINTK FF. Fondo Técnico Actinver SA de CV S.I.R.V. FF.** (n.d.). *Morningstar, Inc*. Retrieved on April 20th, 2016 from: http://lt.morningstar.com/7ap7omrzjm/snapshot/snapshot.aspx?id=F0000004TN&SecurityToken=F0000004TN%5d2%5d1%5dFOMEX%24%24ALL_1414&clearcache=true&ClientFund=1&LanguageId=es-MX&CurrencyId=MXN&UniverseId=FOMEX%24%24ALL_1414&BaseCurrencyId=MXN

15. **José Ricardo** (2015, June 25th). *python - Scrapy gives URLError: <urlopen error timed out>* . Retrieved on September 23th, 2015 from Stack Overflow: http://stackoverflow.com/questions/31048130/scrapy-gives-urlerror-urlopen-error-timed-out/31055000#31055000

16. **JPMorgan Funds (Asia) Limited** (2016). [J.P. Morgan India - Fund Facts December 2015]. Retrieved on September 14th, 2016 from: https://www.jpmorganmf.com/inec/en/Factsheet/Factsheet%20-%20Dec%2015.pdf.Bromium

17. **MariaDB Foundation** (2015). *Welcome to MariaDB! - MariaDB.* Retrieved on November 9th, 2015 from: https://mariadb.org/

18. **Miloslav, N.** (n.d.). *Attributes [XPath 1.0 Tutorial @ Zvon.org]*. Retrieved on October 22th, 2015 from Zvon.org: http://www.zvon.org/comp/r/tut-XPath_1.html

19. **MongoDB, Inc.** (2015). *MongoDB for GIANT Ideas.* Retrieved on September 25th, 2015 from: https://www.mongodb.org

20. **Montero, P. & Vilar, J. A.** (2014, November). *TSclust: An R Package for Time Series Clustering*. Journal of Statistical Software, 62(1), 1-43. http://www.jstatsoft.org/v62/i01/

21. **Montero, P. & Vilar, J. A.** (2014, November 18th). *TSclust: Time Series Clustering Utilities*. R package version 1.2.3. Retrieved on February 10th, 2016 from The Comprehensive R Archive Network: https://CRAN.R-project.org/package=TSclust

22. **Montero, P. & Vilar, J. A.** (2015, February 19th). Package '*TSclust'.* Retrieved on September 20th, 2016 from: https://cran.r-project.org/web/packages/TSclust/TSclust.pdf

23. **Morningstar** (2015). *Morningstar|Fondos de Inversion|Analisis de Fondos|ETFs|Precio|Rendimiento Fondos|Mejores Fondos.* Retrieved on September 28th, 2015 from: http://www.morningstar.com.mx/mx/

24. **ncalculators.com** (2015). *Linear Interpolation Calculator, Definition & Formula*. Retrieved on November 13th, 2015 from: http://ncalculators.com/geometry/linear-interpolation-calculator.htm

25. **Oracle Corporation** (2015). *MySQL*. Retrieved on November 9th, 2015 from: http://www.mysql.com/

26. **Peterson, B. G. & Carl, P.** (2014, September 16th). *PerformanceAnalytics: Econometric tools for performance and risk analysis*. R package version 1.4.3541. Retrieved on February 9th, 2016 from The Comprehensive R Archive Network: https://CRAN.R-project.org/package=PerformanceAnalytics

27. **Phillips, D. & Kaplan, P. D.** (2010). The Morningstar Approach to Mutual Fund Analysis—Part I. In John A. Haslem (Ed.), *Mutual Funds: Portfolio Structures, Analysis, Management, and Stewardship* (pp. 153-174). Hoboken, NJ: John Wiley & Sons, Inc.

28. **R Core Team** (2015). "*R Installation and Administration*". Retrieved on February 18th, 2016: https://cran.r-project.org/doc/manuals/r-release/R-admin.html

29. **R Core Team** (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/

30. **Real Python** (2014, December 31th). *Web Scraping with Scrapy and MongoDB*. Retrieved on September 25th, 2015 from Real Python: https://realpython.com/blog/python/web-scraping-with-scrapy-and-mongodb/

31. **Reilly, F. K. & Brown, K. C.** (2012). *Investment Analysis & Portfolio Management* (10th ed.). Mason, OH: South-Western Cengage Learning.

32. **Schroders** (2016, June 30th). *[Schroders - Emerging Market Equity Fund 2Q 2016 fund fact sheet]*. Retrieved on August 26th, 2016 from: http://www.schroders.com/getfunddocument?oid=1.9.68646

33. **Scrapy developers** (2015). *Scrapy Tutorial*. Retrieved on September 17th, 2015 from Scrapy 1.0.3 documentation: http://doc.scrapy.org/en/1.0/intro/tutorial.html

34. **Scrapy developers** (2015). *Debugging memory leaks*. Retrieved on October 8th, 2015 from Scrapy 1.0.3 documentation: https://doc.scrapy.org/en/1.0/topics/leaks.html#topics-leaks

35. **Scrapy developers** (2015). *Frequently Asked Questions*. Retrieved on September 25th, 2015 from Scrapy 1.0.3 documentation: https://doc.scrapy.org/en/1.0/faq.html

36. **U.S. Securities and Exchange Commission** (2016). *[Westchester Capital Funds - The Merger Fund VL - Fund Snapshot as of May 31, 2016]*. Retrieved on August 24th, 2016 from: https://www.sec.gov/Archives/edgar/data/1208133/000089418916010020/new4024b.pdf

37. **Wikipedia Foundation, Inc.** (2015). *Linear interpolation*. Retrieved on November 13th, 2015 from: https://en.wikipedia.org/wiki/Linear_interpolation

38. **H. Wickham.** (2009). *ggplot2: Elegant Graphics for Data Analysis*. R package version 2.1. Retrieved on March 8th, 2016 from: http://ggplot2.org

39. **Yahoo - News Network** (n.d.). ACTINTKFF.MX Gráfico básico | Valores de ACTINTK FF. *Yahoo Finanzas*. Retrieved on April 20th, 2016 from: https://es-us.finanzas.yahoo.com/q/bc?s=ACTINTKFF.MX&t=my&l=on&z=l&q=l&c=

40. **Yahoo - News Network** (n.d.). *Mercados, Cotizaciones, Divisas - Yahoo Finanzas.* Retrieved on August 4th, 2015 from: https://es-us.finanzas.yahoo.com

41. **Yau, C.** (n.d.). Coefficient of Determination. *R Tutorial*. Retrieved on March 12th, 2016 from: http://www.r-tutor.com/elementary-statistics/simple-linear-regression/coefficient-determination

42. **Yau, C.** (n.d.). Simple Linear Regression. *R Tutorial*. Retrieved on March 13th, 2016 from: http://www.r-tutor.com/elementary-statistics/simple-linear-regression

7.1. Code repositories

1. **Cortés, G.M.** (2016). Thesis Repository 00 - Thesis documentation, data and graphs. *GitHub*. Retrieved from: https://github.com/satsuki-chan/thesis_docs00

2. **Cortés, G.M.** (2015). Thesis Repository 01 - Registered mutual funds at the CNBV website information extractor. *GitHub*. Retrieved from: https://github.com/satsuki-chan/cnbvt01

3. **Cortés, G.M.** (2015). Thesis Repository 02 - Historical price series from Mexican equity funds downloader. *GitHub*. Retrieved from: https://github.com/satsuki-chan/yahoo-finance-t01

4. **Cortés, G.M.** (2016). Thesis Repository 03 - Database and SQL stored procedures, functions and scripts for historical price series preprocessing. *GitHub*. Retrieved from: https://github.com/satsuki-chan/preprocess_sql03

5. **Cortés, G.M.** (2016). Thesis Repository 04 - Traditional Modern Portfolio Theory (MPT) statistics of Mexican equity funds. *GitHub*. Retrieved from: https://github.com/satsuki-chan/anatrad_capm04

6. **Cortés, G.M.** (2016). Thesis Repository 05 - Novel machine learning analysis of Mexican equity funds with hierarchical clustering. *GitHub*. Retrieved from: https://github.com/satsuki-chan/ananov_tsclust05

7. **Cortés, G.M.** (2016). Thesis Repository 06 - Scatter plots of Mexican equity funds' Modern Portfolio Theory variables comparison. *GitHub*. Retrieved from: https://github.com/satsuki-chan/ananov_tsclust_graph06