ERASMUS UNIVERSITY ROTTERDAM

ERASMUS SCHOOL OF ECONOMICS

Master Thesis MSc Econometrics

Business Analytics and Quantitative Marketing

Cost of equity estimation in the case of an unlisted company

Name Student:

Justine HONORAT

Student Number:

622000

Academic Supervisor:

Dr. Erik KOLE

Second assessor:

Dr. Jeffrey DURIEUX

Internship supervisors:

Jelle DEN HOLLANDER

Rianne LURVINK

The content of this thesis is the sole responsibility of the author and does not reflect the view of the supervisor, second assessor, Erasmus School of Economics or Erasmus University.

September 29, 2022

Abstract

This paper studies and compares different methods aiming to estimate the cost of equity of an unlisted company. The purpose is to find the technique with the best complexity/ accuracy trade-off. Two different approaches are used : estimating the systematic risk of the Capital Asset Pricing Model and of the returns by finding a relationship with other variables. It is safe to use several methods and to compare the estimates in order to have a more robust estimation of the company's cost of equity. Different models including varying factors are used in this paper. It seems that estimating the cost of equity by a systematic risk estimate gives more accurate results than a return estimation. The *k*-means clustering algorithm and the similarity metric are the methods with the most accurate results. Finally, in the case of a small sample in our case, it is better to use the model with the smallest number of factors because the factor coefficients are very sensitive to extreme values.

Contents

1 Abbreviations overview

The following table describes the significance and meaning of different abbreviations used in the following report.

Abbreviation	Meaning
AG PLC	Admiral Group PLC
AG SpA	Assicurazioni Generali SpA
AIC	Akaike information criterion
CAPM	Capital Asset Pricing Model
CMA	Conservative Minus Aggressive
EBIT	Earnings before interest and taxes
FTE	Full-Time Equivalent
HG PLC	Hansard Global PLC
HML	High Minus Low
PeGH PLC	Personal Group Holdings PLC
PhGH PLC	Phoenix Group Holdings PLC
RMW	Robust Minus Weak
SCA SpA	Societa Cattolica di Assurazione SpA
SMB	Small Minus Big
TA	Total assets
TC	Target company
TL	Total liabilities
TWSS	Total Within Sum of Squares
USA SpA	UnipolSai Assurazioni SpA
UIG AG	UNIQA Insurance Group AG
WW AG	Wuestenrot & Wuerttembergische AG
ZI Group	Zurich Insurance Group

Table 1: Abbreviations overview

2 Introduction

In order to grow, companies have to raise funds either through debt or through equity. The debt corresponds to different loans provided by financial institutions such as banks. The equity is a funding method that relies on the sale of stocks to individuals. In order to attract potential investors, the return on investment has to be high enough to balance out the risk taken by investing on the company's capital. This return represents a cost for companies which is why it is crucial for them to determine it, as they have to keep track of each of their financial operations.

The Capital Asset Pricing Model (CAPM) is one of the most common methods used to determine the cost of equity of a company Pereiro (2002). It describes the relationship between systematic risk and expected return for assets. The systematic risk represents how a company's market value changes when there's a change in the overall market. In the CAPM, the systematic risk can be found with a linear regression over a period of time - for example a monthly regression over 5 years that can be found in Appendix B or a weekly regression over 2 years. The necessary data can easily be found in financial data bases in the case of listed companies.

Unlisted companies, representing a substantial part of the economy Abudy et al. (2015), have the particularity to be owned by private investors. This implies, among other things, that market data of these particular companies is not available, unlike public companies. Therefore, the exposure to systematic risk cannot be found by a regression and it makes the valuation of a private company more challenging.

This thesis aims to use and compare different methods in order to determine the cost of equity of an unlisted company. As no literature comparing different estimation methods was found, it seemed important to use different approaches. The first approach focuses on estimating the exposure to systematic risk through different techniques. The second one aims to find a relationship between certain variables and the returns themselves.

This paper intends to extend the literature on estimating the cost of equity of unlisted companies by answering the following research question :

How to estimate the cost of equity of an unlisted company ?

The data used to answer this question was gathered using Bloomberg, Eikon and Yahoo Finance data bases. It is bi annual data regarding the companies such as the book value or the headquarter country from years 2012 to 2021. The first method focuses on estimating the systematic risk by choosing similar companies by hand. It is the most common method as it requires no machine learning knowledge. However, it is time consuming and does not give the most accurate results. In a second step, the k-means clustering was used in order to determine groups of similar companies. This technique is the less time consuming but it only takes quantitative variables into account, which led to the third method. The similarity metric method aims to create an algorithm that imitates a choice that would have been done by hand. Compared to the k-means clustering this method was expected to perform better because it takes both qualitative and quantitative variables into account. Compared to the first method, it is expected to be more accurate and less time consuming. A linear regression was also used in order to find a relationship between variables and the returns. This method focuses on estimating the returns and not the systematic risk. To finish, the Fama-MacBeth regression was applied. This method is adapted to panel data and was expected to perform best. However, both those techniques require certain assumptions to be valid which can lead to a time-consuming preprocessing phase.

This study shows that estimating the systematic risk gives more accurate results than estimating the expected returns directly. The difference in accuracy between the two approaches is very high and cannot be neglected. The *k*-means clustering and the similarity method are the ones that perform best and the regression methods are the ones that perform worse by far. Also note that the number of companies chosen for the estimation has to be large enough in order to obtain proper estimations. It is also a plus to use several efficient estimation methods to compare the results and be more confident about the obtained results. Finally, using a lot of companies for our estimation does not significantly increase the estimations' accuracy but using more variables and/or time periods would definitely give better results.

The existing literature on the CAPM rarely applies to unlisted companies. Furthermore, when it come to private companies, the studies focus on one specific method. Also, the used methods are not diverse and they are not compared to one another. This paper contributes to the existing literature by using and comparing several methods that aim to apply the CAPM on unlisted companies.

To answer the research question, several methods are studied and compared. The remainder of the paper is structured as follows. Section 3 focuses on describing the already existing models and estimation methods. Then, the data description and preprocessing are described in Section 4. Section 5 provides information about the methodology and the results are discussed in Section 6. Finally, the conclusion and limits of the study are addressed in Section 7.

This research was conducted in the context of an internship at a Dutch insurance company : Achmea. The purpose of that internship was to derive the company's cost of capital through an estimation of its cost of equity. In this paper, I focus on the different methods used to estimate the company's cost of equity using listed companies' actual cost of equity as the actual outcome in order to compare the different methods' outcomes.

Note that all the data used for this study is publicly accessible and can be found online.

3 Related Work

A small number of articles can be found concerning the derivation of a private company's cost of equity. Most of them describe methods that could be used without applying them on concrete cases or without comparing them to other techniques.

3.1 Determining the yield rate of a company

Over the past 30 years, several models were developed in order to determine the cost of equity of a listed company. Two approaches can be used :

- Estimating the returns through a relationship with other variables
- Estimating the returns through the systematic risk

3.1.1 Fama-MacBeth regression

Devised in 1973 by Eugene Fama and James MacBeth, this method is used to estimate the parameters of the CAPM Pasquariello (1999). It highlights the linear relation between the returns and other factors Fama & MacBeth (1973). It is a two-step approach :

1. T Cross-sectional regressions using the explanatory variables :

$$\begin{pmatrix} r_{i=1,t} \\ \dots \\ r_{i=I,t} \end{pmatrix} = \begin{pmatrix} 1 & c_{i=1,1} \\ \dots & \dots \\ 1 & c_{i=I,1} \end{pmatrix} \begin{pmatrix} \alpha_t \\ \lambda_t \end{pmatrix} + \begin{pmatrix} \varepsilon_{i=1,t} \\ \dots \\ \varepsilon_{i=I,t} \end{pmatrix}$$

r_{it} being the returns of company i at time t, α_t and c_i parameters to be estimated.

2. Time-series average in order to determine each coefficients' value :

$$\frac{1}{T}\sum_{t=1}^{T}\lambda_t^k$$

This method is appreciated because it also takes the temporal aspect into account. The following assumptions need to be respected in order to conduct this model :

- Linearity assumption : linear relationship with the variables used in the model
- Random sampling of observations
- Conditional mean of error terms equal to 0

Checking for the validity of those assumptions can unfortunately lead to a time-consuming preprocessing phase.

3.1.2 Capital Asset Pricing Model and β estimation

The Capital Asset Pricing Model, published in 1964 by William Sharpe (Sharpe, 1964) had a consequent impact in the area of financial management. It describes the relationship between expected return for assets and systematic risk. It is derived as follows :

$$E(R_i) = R_f + \beta_i E(R_m - R_f)$$

R_f corresponding to the risk-free rate, β_i to the systematic risk, R_m to the overall market return and $E(R_m - R_f)$ to the market risk premium.

The systematic risk β_i describes how much a stock moves compared to the market. A β_i equal to 1 means that the stock is expected to move exactly like the market. A higher β_i corresponds to a stock expected to be more volatile than the market which implies higher risk and returns. It is derived as follows :

$$\beta_i = \frac{Cov(R_i, R_m)}{Var(R_m)}$$

R_i being the return on the individual stock and R_m the overall return of the market. β_i is the result of a linear regression over time. It can be derived on a daily, weekly or monthly basis over different time periods such as six months, two years, five years, etc. Usually, the β_i is the result of a monthly regression over five years and corresponds to the slope coefficient of this regression. A company's β is levered by its debt which means that that we cannot compare the βs as they are and therefore (Drobetz et al., 2014) they should be unlevered with the following formula :

$$\text{Unlevered } \beta = \frac{\beta}{Leverage}$$

$$\text{Leverage } \beta = 1 + (1 - \text{Tax Rate}) * \frac{\text{Debt}}{\text{Equity}}$$

Although it is very known, this model is often criticised. Indeed, its assumptions are often described as unrealistic and its parameters are not always easy to derive (ACCA, 2020). However, it is still a very common technique that is not too complex, which is why this study focuses on it. In the case of an unlisted company, one has an additional issue as market data is not accessible. This implies that the β cannot be regressed over a period of time. The slope of the regression therefore needs to be estimated and the following part describes some of the methods previously used.

3.1.3 Valuation of an unlisted company

The Fama-Macbeth regression can be used to estimate the returns as it aims to show a relationship between independent variables and the returns. If the company is listed, the systematic risk is added to the regressors. If not, it simply just cannot be used as an explanatory variable.

In his article, Mirzayev (Mirzayev, 2021) proposes to estimate the slope of the β coefficient by finding public companies that are similar to the one one wants to estimate the β slope of. It is a simple method that is quite efficient if one chooses the companies carefully. However, this procedure takes a lot of time if one wants the estimation to be as accurate as possible. It can be tricky to find several companies that are comparable to another one in terms of size, activity, dynamic, etc. Indeed, listed companies are most of the time big companies that generate their revenue from several activities when unlisted companies are generally smaller and usually have a single operation (Favereau, 2015). This makes it hard to find a public company with a β that reflects the unlisted company's business. Also, this procedure has to be executed each time one wants to do the estimation as the similar companies might differ from one year to another. Other researchers use a different method such as using the company's previous earnings. However, not much information can be found concerning this method.

Although the CAPM is very famous and widely used to estimate the cost of equity of a company, some agree that the single factor β cannot capture all risk (Reinganum, 1981) and therefore cannot explain returns on its own. This is why some researchers made the statement that additional factors are needed.

3.2 CAPM expansion

As stated above, some studies have focused on expanding the CAPM in order to explain expected returns better. The following part describes some of the research conducted to add factors in order to the CAPM.

3.2.1 Fama-French Three Factor Model

This model, developed by Eugene Fama and Kenneth French, is an expansion of the CAPM. The researchers criticised the single factor model, stating that the returns could not be explained only by the systematic risk (Fama & French, 1992). This model takes two more parameters into account in order to make the model more flexible : Small Minus Big (*SMB*) and High Minus Low (*HML*). It is derived as follows :

$$E(R_i) = R_f + \beta_i E(R_m - R_f) + \beta_{SMB} E(R_{SMB}) + \beta_{HML} E(R_{HML})$$

SMB refers to the exposure to the size factor and compares the historic excess returns of small capital companies with the ones of big capital companies. Excess returns highlight how an investment performs compared to other investment alternatives (Alhassane Garba et al., 2019), being the risk free rate in our case. If the coefficient is positive, it implies that small companies have on average higher returns than big ones.

HML compares the returns of companies with a high book-to-market ratio with the ones with a low book-to-market ratio. In other words, it is the yield premium, at time t, related to the book-to-market ratio i.e. the returns of securities with a high book-to-market ratio minus the returns of securities with a low book-to-market ratio (Limaiem, 2009).

This model allegedly allows to explain more than 90% of a portfolio's return (*Fama-French Three-factor Model*, 2020) but can also be expanded which brings us to the Fama-French five-factor model.

3.3 Fama-French Five-Factor Model

This model adds 2 factors to the preceding one : Robust Minus Weak (*RMW*) and Conservative Minus Aggressive (*CMA*). It is derived as follows :

$$E(R_i) = R_f + \beta_i E(R_m - R_f) + \beta_{SMB} E(R_{SMB}) + \beta_{HML} E(R_{HML}) + \beta_{RMW} E(R_{RMW}) + \beta_{CMA} E(R_{CMA})$$

RMW returns the spread of the most profitable firms minus the least profitable ones. The profitability is derived as follows :

$$Profitability = \frac{EBIT}{BV}$$

EBIT corresponding to the earnings before interest and taxes and *BV* the book value. The book value is derived as follows :

$$BV = \text{Total assets} - \text{Total liabilities}$$

This coefficient highlights the return difference between robust and weak companies. If the company is considered neutral at time t, the β_{RMW} coefficient is equal to 0 for that time period.

CMA returns the spread of firms that invest conservatively versus aggressively. The investment is derived as follows :

$$Investment_t = \frac{TA_t}{TA_{t-1}}$$

TA corresponding to the Total Assets of the company at time t.

In his study (Jansen, 2019), Jansen states that the Five-Factor Model performs better than the two other models when it comes to predicting returns on the stock market in the Netherlands. In our paper, we will focus on which model performs better when it comes to estimating the yield rate of a company that is non-listed and on which estimation technique is the most accurate. However, it is important to keep in mind that his study was performed on the whole Dutch market whereas the present one is done on a smaller sample. However, even though the Three and Five factor models cannot be directly applied, they give motivation regarding the different characteristics that will be considered in the study.

4 Data

The data was gathered on Eikon and Bloomberg in order to have as much information from as many companies as possible. The data set contains 380 observations and 14 variables. Every company of the data set is a European insurance company specified in one or several of the following sectors : life, non-life or health insurance.

4.1 Data description

The data set has information about companies such as the total revenue, book value, full time equivalent, etc. all according to a specific period (bi annual data) from 2012 to 2021. They are very diversified in terms of size : the smallest full time equivalent value is 125 and the largest one is 125,411. They also differ in terms of book value, market capitalization, etc. This diversity is done on purpose, as different clustering methods will be used later in the study and it is preferable to have a large panel of companies. The data that was gathered is bi annual meaning that each company has 20 rows. The data set contains a total of 19 companies. The goal was to gather as many companies as possible with as much data as possible. This is why companies with too many missing values were not included in the study. This leaves us with only 19 companies but it was preferable to have less companies with accurate data than more companies with a lot of estimated values. The following table describes each variable used in the study and where it was retrieved from :

Variable	Description	Modalities	Source
Company	Name of the company	19	Eikon
Industry Name	Specification of the company	Life Nonlife Health	Eikon
Headquarter	Headquarter country	10 European countries	Eikon
Year	Half year period from 01/01/2012 to 31/12/2021	Half year period from 01/01/2012 to 31/12/2021	Eikon
FTE	Full-time Equivalent : unit of measurement equivalent to an individual worker (38h/w) (Beitone et al., 2012) Internal + external workers	From 125 to 125,411	Eikon Bloomberg
Book value	Total assets - Total Liabilities	From -1,926,749 to 84,596	Bloomberg
Total liabilities	Total legal obligations or debt owed to another person or company (Beitone et al., 2012)	From 8.21 to 2,857,444	Bloomberg
Total assets	Total resources with economic value that company owns or controls with the expectation that it will provide a future benefit (Beitone et al., 2012)	From 39.27 to 1,139,429	Bloomberg
Total debt per share		From 0 to 396.25	Bloomberg
Shares outstanding		From 30.06 to 3.083,95	Bloomberg
Total debt	Total debt per share * Shares outstanding	From 0 to 46,691.69	Bloomberg
Tax rate	Tax rate of the company	From 0 to 0.314	(KPMG, 2021)
Profit Before Taxes	Variable used instead of the Earnings Before Interest and Taxes for the *HML* factor.		Bloomberg
Returns	Returns of the company in %		Eikon

Table 2: Variable description : All the figures are in Millions except for FTEs, Debt per share, and Tax rate

The variables of the data set have different purposes.

- Variables used to determine the similarity with the TC or to fins a relationship with the returns : Industry Name, Headquarter, FTE, Book value, Total Assets, Total Liabilities
- Variables used to unlever the companies' βs : Total debt, tax rate

4.2 Pre-processing

As explained above, the different companies were chosen according to the amount of available data so there were not a lot of missing values to deal with. Indeed, if a company had too many missing values, it was simply not added to the data set. Companies that did not have data for the whole study period were not included. The ones who had a variables missing values for more than 50% of the time period were not included either. Finally, the ones with not enough available market data were also not included.

Concerning the full time equivalent variable, each missing value was replaced by its preceding value, or the next one if the missing value was from the first year. The variable with the most missing values was the earnings before interest variable. This has no impact on the CAPM estimation as this variable is not used in its formula 3.1.2. It is not used either as a factor in order to determine the similarity between the TC and another company. However, it is used in the three and five factor models. If one missing value was in between two existing values, it was replaced by the mean of those values. It was decided not to use the mean of all the company's values to avoid sudden jumps throughout the time. However, if a company had several missing values in a row, the company's mean was computed and chosen as new value.

The variables with the most missing values are the ones used in the CAPM expansions. This implies that it can have an impact on the result of the models, especially as the data set size is small.

To finish, note that all the monetary information of the data set is in euros.

5 Methodology

This part aims to describe the different methods used in order to estimate a private company's cost of equity. The risk-free rate used in this study is the 31/12/2021 Euro short-term rate which is equal to -0.59% (ECB, 2021), as all the companies in the data set are in Europe. The risk premium that was used is the Q4 2021 risk premium equal to 5% (KPMG, 2022) and the tax rates were found in the Corporate Tax Rates table by KPMG (KPMG, 2021).

As no paper was found comparing different estimation methods on the CAPM, finding a comparison metric was challenging. The RMSE seemed to be the one that reflects the most the accuracy of an estimation method and it can be used for every one of them. It shows the average distance between the estimated value and the real one Hodson (2022) and is derived as follows :

$$RMSE = \sqrt{MSE}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)$$

MSE being the mean squared error, n the number of observations, Y_i the real value and $\hat{Y}_i$ the estimated value. From now on, we will call target company (TC) the company we want to do our estimations on. No literature was found in order to determine what a good RMSE value is. It was decided to aim for a RMSE equal to 0.5. The choice of this value does not rely on any previous study as none was found on this matter.

5.1 Comparable Companies Analysis

This method is most clearly the easiest one. It consists in finding a couple of companies that are similar to the TC in terms of size, revenue, etc. Five companies were chosen as they seemed similar to each other in terms of the different variables but also in terms of tendency. Not only the raw values were used but also their evolution throughout time. The next step is to make sure that the βs can be compared. A company's β is levered by its debt which means that that we cannot compare the βs as they are and therefore they should be unlevered with the following formula :

$$\text{Unlevered } \beta = \frac{\beta}{Leverage}$$

$$\text{Leverage } \beta = 1 + (1 - \text{Tax Rate}) * \frac{\text{Debt}}{\text{Equity}}$$

To estimate the TC's β, the median of the unlevered β of the five chosen companies is selected. It is then relevered with the median of the companies' leverage :

$$\hat{\beta} = \text{Median}(\beta \text{ similar companies}) * \text{Median (leverages similar companies)}$$

One of the cons of this method is that it is very subjective. Indeed, the similarity can be perceived differently from one individual to another. This is one of the reasons why it is better to have other estimation methods even only to compare the results and to make sure they make sense. The following part describes an other method to estimate a company's cost of equity.

5.2 Choosing similar companies with k-means clustering

This part consists in finding similar companies to the TC with the k-means clustering method.

5.2.1 The k-means algorithm

The k-means algorithm aims to divide a population into subsets based on the similarity between those subsets. It returns the population division that both maximise the intra subset similarity and minimise the inter subset similarity (Liu et al., 2018). This algorithm works as follows:

1. Pick k the number of clusters to divide the population in.

2. Select randomly k distinct data points in the population that will be our centroids.

3. Assign each point to the nearest cluster centroid (Euclidean Distance).

4. Recompute the centroids of the newly formed clusters.

5. Repeat step 3 and 4 until the algorithm converges.

5.2.2 Choosing the optimal k

The clusters have to be as much different as possible from one another. However, within the clusters, the population needs to be as homogeneous as possible. To have an idea of the heterogeneity of the population of each cluster, the within variation is derived. It reflects how much the points deviate from the centroid within each cluster so this value has to be as small as possible. The Total Within Sum of Squares is derived as follows :

$$\text{Total Within Sum of Squares} = \sum_{k=1}^{K} (x^{\bar{(k)}} - x_i^{(k)})$$

$x^{\bar{(k)}}$ being the centroid of cluster k and $x_i^{(k)}$ the value of the ith individual of cluster k.

The moment the reduction of the Total Within Sum of Squares (TWSS) is negligible corresponds to a good choice of k (Sinaga & Yang, 2020). Indeed, it is not efficient to add a cluster if it only makes the clustering slightly better.

After creating the clusters, the most similar one to the TC is used for the β estimation. The βs are levered with the median of the cluster's leverage

$$\hat{\beta} = \text{Median}(\beta \text{ similar companies}) * \text{Median (leverages similar companies)}$$

Unfortunately, the clustering algorithm only takes quantitative variables into account. However, it seems important to take into account other variables such as the activity of the company or the country it makes most of its sales revenue. The following part describes an algorithm that aims to take into account every variable regardless of its type.

5.3 Similarity metric

This method consists in coding a method that imitates a choice that would have been done by hand. It takes the same variables into account as usual and studies several aspects of it. For each variable, the idea is to compare one company's value to the target company's and assign a score to this company. The following parts will describe how each of those variables were taken into account for the similarity metric derivation. Several similarity metrics were tested and the one described below is the one that performs the best according to the chosen metric.

5.3.1 Building the similarity metric

The following part aims to describe the steps that lead to the derivation of the similarity metric for each variable and company of the data set. No similar work was found on such a method which is why the different figures were chosen randomly according to the importance given to each criteria. Changing the scores means changing the importance of a criteria i.e changing the similarity ranking. The results obtained with other scores gave less satisfying estimations which is why the following part describes the scores giving the best accuracy. The said results can be found in Appendix A.2 and A.4.

Specification

This variable refers to the specification of the insurance company. It can be specified in life, non life or health insurance. The companies of this data set have the following specifications :

- Non life
- Life and health
- Life, health and non life

This variable seems to be very important in order to determine the similarity between 2 companies. Companies that generate revenue from the same activities obviously are quite similar. The score were attributed as follows :

Company's specification TC's specification	Non life	Life and health	Life, health and non life
Non life	400	0	200
Life and health	0	400	200
Life, health and non life	200	200	400

Table 3: Score attribution for the specification variable

Headquarter

This variable refers to the country where the company's headquarters are located. This usually implies that the company makes most of its revenue in this country. This also accounts as a similarity between two companies. If the target company has the same headquarter country than another company, it gets the score of 400. Otherwise, it gets the score of 0.

Numeric variables

Three aspects of each numeric variable were evaluated.

Aspect 1 i.e the raw values : An interval was created centered on each TC's value in year y TC_y and depending on the standard deviation of the variable in year y σ_y :

$$[TC_y \pm 0.3\sigma_y]$$

For each year, if a company's value belongs to this interval, it gets the score of 5. Several values were tested for this interval and this one was chosen because it seemed to give efficient result. Meaning that the score of 10 was not attributed to too many or too few companies. Companies with their value outside the interval were given the score of 0 in year y. The maximum value a company can reach is 200 (10 points by period) which is equal to the headquarter score. Indeed, it seems that a company that falls in the interval for every year should not be penalised by the fact that it has a different headquarter country.

Aspect 2 i.e the overall evolution : This metric aims to take into account the overall evolution of the variable. Indeed, two companies that have doubled their FTEs between 2012 and 2021 most certainly have similarities. The TC's overall evolution is derived as follows :

$$ratio_{TC} = \frac{value_{2021}}{value_{2012}}$$

Companies whose ratio belong to that interval :

$$[ratio_{TC} \pm 0.3]$$

were given the score of 200 when others were given the score of 0. This value was chosen in order to take into account companies that have had a similar evolution but do not specifically have a similar specification or headquarter country than the TC. It is also to make a difference between companies that only share the same specification or headquarter country with the ones that have a similar dynamic.

Aspect 3 i.e the evolution over the years : This metric aims to take into account the overall evolution of the variable throughout the years. In order to do so, a linear regression was performed on the variable in order to get the slope coefficient. For each company, the slope was compared to the TC's slope. Then, an interval was created centered on the TC's slope, $slope_{TC}$, and depending on the standard deviation of all the companies' slopes, $slopes$:

$$[slope_{TC} \pm 0.5slopes]$$

Companies with a linear regression coefficient within the interval got the score of 200 when others got 0.

Similarity scores

After the attribution of the scores, they were all summed giving one similarity score per company according to the chosen target company. Two estimations methods were tested from that point :

1. Selecting the k companies with the highest score and use them to estimate the TC's yield rate. Different k values were tested : 3, 5 and 7.
2. Selecting all the companies of the data set and estimating the yield rate by weighing the companies according to their similarity value.

5.3.2 Estimation methods

This part describes and compares the 2 estimation methods that were tested.

k most similar companies

For this estimation method, the k companies with the highest scores are used to estimate the target company's yield rate. Similarly as previously, we take the median of those leverages and the median of the unlevered βs and derive the estimated yield rate with the single factor CAPM formula.

Weighted estimation

In this case, all the companies are used for the estimation. After all the similarity scores have been attributed we derive their sum. Then each company will have the following weight :

$$Weight_i = \frac{S_i}{Total}$$

S_i being the similarity score of company i and $Total$ the sum of all the similarity scores. Then, the weighted median of the unlevered βs and the weighted median of the leverages were used in order to estimate the company's yield rate.

5.4 Cross-sectional regression

A cross-sectional regression aims to describe the relationship between different variables. It can be simple or multiple. A single linear regression is a model that is expressed as follows :

$$y = \beta_0 + \beta_1 x + \varepsilon$$

y being the explained variable, x the independent variable/ regressor, β_0 the intercept and β_1 the estimated coefficient corresponding to the slope of the line and ε the error term (Supichaya, 2015). A multiple linear regression takes into account several dependant variables. It can be expressed as follows for a model with k dependent variables :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon$$

5.4.1 Assumptions

In order to conduct a linear regression, several assumptions need to be fulfilled Montgomery et al. (2021):

- Linearity : The independent variables used in the linear regression need to have a linear relationship with the dependant variable y. This implies that the result of the model is a result of multiple linear regressions modelling the relationship between the dependent variable and each of the independent variables.
- Homoscedasticity : The variance of the error terms ε has to be independent on the dependent variables' values. A plot of the residuals VS fitted values can be used to check the validity of this assumption.
- Lack of perfect multicolinearity : The independent variables need to be uncorrelated to one another. It can be tested with a correlation matrix.
- Independence of errors : The error terms ε need to be normally distributed.

5.4.2 Linear regression goodness of fit

The quality of a linear regression, whether it is simple or multiple, can be determined with several factors.

The R^2 indicates how close the data is to the regression line (Guyader, 2012). It ranges from 0 to 1, 1 usually meaning that the regression fits the data perfectly. It is derived as follows :

$$R^2 = \frac{\text{Explained variation}}{\text{Total variation}}$$

However, even if it is often the case, a high R^2 does not always mean a good fit (Corbière & Larivière, 2020). It does not reflect the bias between the predictions and the real values which is why the residuals plots have to be checked. They should be randomly distributed and not show any pattern else it indicates a bad fit of the regression. This is why it is not the only parameter that should be taken into account.

The Akaike information criterion is used to choose which model performs best. There is no good AIC value because the value cannot be interpreted as it is. It is used to compare models with one another and the one with the lowest AIC is the one offering the best fit. It is derived as follows :

$$\text{AIC} = 2k * 2ln(L)$$

k being the number of estimated parameter and L the maximum Likelihood. The AIC parameter is used in the *stepAIC* function from the *MASS* package in *RStudio* in order to return the regression coefficients that minimize the AIC value. The *R* function can also try both forward and backward selection in order to choose the one with the smallest AIC. This refers to stepwise selection. There are two types of stepwise selection methods Supichaya (2015) :

- Forward selection: The procedure begins with only the regressor in the model. At each step, the variable that improves the model the most is added.

- Backward selection: The procedure begins with all the variables in the model. At each step, the variable that is the less significant is removed from the model.

The p-value is important to know if the coefficients are significantly different from 0 and if the model is overall significant. The hypothesises are :

- H_0 : The coefficient is equal to 0

- H_a : The coefficient is not equal to 0

If the p-value is below the threshold $\alpha = 0.05$, we reject the null hypothesis.

5.5 Fama-Mac Beth regression

This method is used to estimate the parameters of the CAPM Pasquariello (1999). As a linear regression was performed as a previous method, the assumptions are already checked and will therefore not be repeated. This regression was performed on *RStudio* using the *fpmg* function of the *plm* package.

5.6 Fama-French Three and Five factor Models

Both those models are based on the excess returns of companies. The excess return corresponding to the returns over the risk-free rate. The returns correspond to the bi annual returns of the company i.e the price change and any relevant dividends during that period (Eikon, 2022). It is better to use average monthly returns, but the values are not available monthly which is why it was chosen to use bi annual returns instead of monthly. The following parts will describe the construction of the different factors used in the models.

5.6.1 Small Minus Big Factor

This factor refers to the size of the company and it is used in both 3 and 5 factor models. The variable used for this factor is the Market Capitalisation. Companies with a market capitalization higher than the median are considered as big companies and the other ones as small ones. Then, the average return of each category at time t is derived. Finally, the difference between average returns of small VS big market capitalisation companies is derived, giving the average return difference of small companies compared to big ones. The coefficient highlights how small companies returns are compared to big companies'.

5.6.2 High Minus Low Factor

This factor corresponds to the book-to-market value of the company. It is also used in the 3 and 5 factor models and is derived as follows :

$$\text{Book-to-market ratio } = \frac{\text{Book value}}{\text{Market capitalization}}$$

Companies with a book-to-value ratio higher than the 70^{th} percentile are considered high book-to-market ratio companies. The ones belonging to the 30^{th} percentile are considered low book-to-market ratio companies. The rest are considered neutral (Jansen, 2019). After building the different portfolios, the average return of the high book-to-market ratio and low book-to-market ratio companies is derived. Then the average difference returns is derived. This value corresponds to how the returns of high book-to-market ratio companies are on average compared to the ones of low book-to-market ratio companies.

5.6.3 Robust Minus Weak Factor

This coefficient takes into account the profitability of companies. It is only used in the 5 factor model. Different studies such as Jansen's (Jansen, 2019) use the Earnings before Interest and Taxes to derive the profitability. However, as interests are a very important part of financial companies it is better to include them in the reported earnings. To address this issue, it was decided to use either the Profit Before taxes or net income instead of EBIT. The variable used was the net income which corresponds to the amount of profit made by a company after the payment of all its expenses (Bloomberg, 2022).
The profitability is then derived as follows :

$$\text{Profitability } = \frac{\text{Net Income}}{\text{Book value}}$$

Companies with a profitability higher than the 70^{th} percentile are considered robust profitability companies. The companies below the 30^{th} percentile are considered weak and the rest neutral. At each time period, the companies are put into categories according to this rule. Similarly to the previous factors, the average return difference between high and low profitability companies is derived in order to obtain the *RMW* coefficient.

5.6.4 Conservative Minus Aggressive Factor

This factor, used in the 5 factor model, corresponds to the difference in returns between firms with low and high investment policies (Amézola & Dolz, 2017). Companies with a value higher than the 70^{th} percentile are considered conservative investment companies. The companies below the 30^{th} percentile are considered aggressive and the rest neutral. Similarly to the previous factors, the average return difference between high and low profitability companies is derived in order to obtain the *CMA* coefficient.

6 Results

This part aims to describe the results obtained with the different methods. As all the companies in the data set are in Europe, the risk-free rate used in this study is the end of 2021 Euro short-term rate which is equal to -0.59% and the βs were regressed on the EURO STOXX 500 index. The risk premium that was used is the Q4 2021 risk premium equal to 5 (KPMG, 2022) and the tax rates were found in the Corporate Tax Rates table by KPMG (KPMG, 2021). The β coefficient correspond to the slope of the monthly βs regressed from 01/01/2017 to 31/12/2021.

In the remainder of the paper, the actual return corresponds to the following formula :

$$\mu_i = R_f + \text{Unlevered } \beta_i * \text{Median } (L) * E(R_m - R_f) \tag{1}$$

L corresponding to the list of the similar companies' leverages excluding the TC's leverage. The model implied return refers to :

$$\hat{\mu}_i = R_f + \text{Median } (B) * \text{Median } (L) * E(R_m - R_f) \tag{2}$$

B corresponding to the list of the similar companies' βs excluding the TC's β.

6.1 Comparable Companies Analysis

The results can be found in the following table :

Company	Unlevered β	Leverage	Levered β	R^2	Actual return	Model implied return
Aegon NV	1.20	1.34	1.85	0.48	8.65	6.10
Aviva PLC	0.85	1.64	1.04	0.63	4.62	6.4
Prudential PLC	1.15	1.33	1.81	0.52	8.48	6.15
Sampo PLC	0.93	1.27	1.47	0.64	6.78	7.00
WW AG	0.22	4.93	0.49	0.27	0.76	6.4

Table 4: Results of the CCA method

WW AG stands for Wuestenrot & Wuerttembergische AG

The results of the returns are expressed in %. Concerning Aegon NV, the CAPM using the β regressed monthly from 01/01/2017 to 31/12/2021 gave a yield rate of 8.65%. However, the R^2 associated to this company is the lowest compared to the other similar companies. The R^2 corresponds to a goodness-of-fit measure for linear regression models. If it is equal to 1, then the regression fits the data perfectly. If it is on the contrary equal to 0, the regression does not fit the data at all. In our case, it is equal to 0.476 which is a medium goodness-of-fit. The CAPM based on the β estimation gave a result of 6.10%. This is a result that could have been expected. Indeed, Aegon NV has the highest unlevered β and a medium leverage compared to the other companies. So, taking the median of the other companies' βs and leverage obviously leads to a lower value.

Concerning Sampo PLC, the CAPM result using the real company's regressed β is 6.78% versus 7.00% for the CAPM using the β estimation. Furthermore, this company's R^2 is the highest with a value of 0.64. It seems that the higher the R^2 is, the closest the estimated yield rate is to the real one but conclusions shouldn't be drawn already.

On the other hand, the estimation for Wuestenrot & Wuerttembergische AG is pretty bad. This is due to the fact that its leverage is very high compared to the other companies. As we take the median leverage of the four other companies in order to relever our estimated β, its estimation is necessarily much lower than its real value. If it had been chosen to take an average and not a median the problem would have remained the same, as the company's own leverage is not included in the estimation.

The estimations were relatively good especially compared to the low complexity of this method. However, the companies not only have to be very similar in terms of characteristics. If their leverages are too different from one another, the estimation is not accurate so the results have to be treated carefully. The RMSE is equal to 3.06 meaning that the estimations deviate on average from 3.06 percentage units from the real value which is a bit high. Predicting a yield rate of 3.06 when it is actually 0 could be very problematic which is why it is essential to try other estimation methods.

6.2 Choosing similar companies with k-means clustering

This part describes the estimation of a company's β slope coefficient using the k-means clustering method. As explained earlier, the TWSS was used in order to determine the numbers of clusters to create.

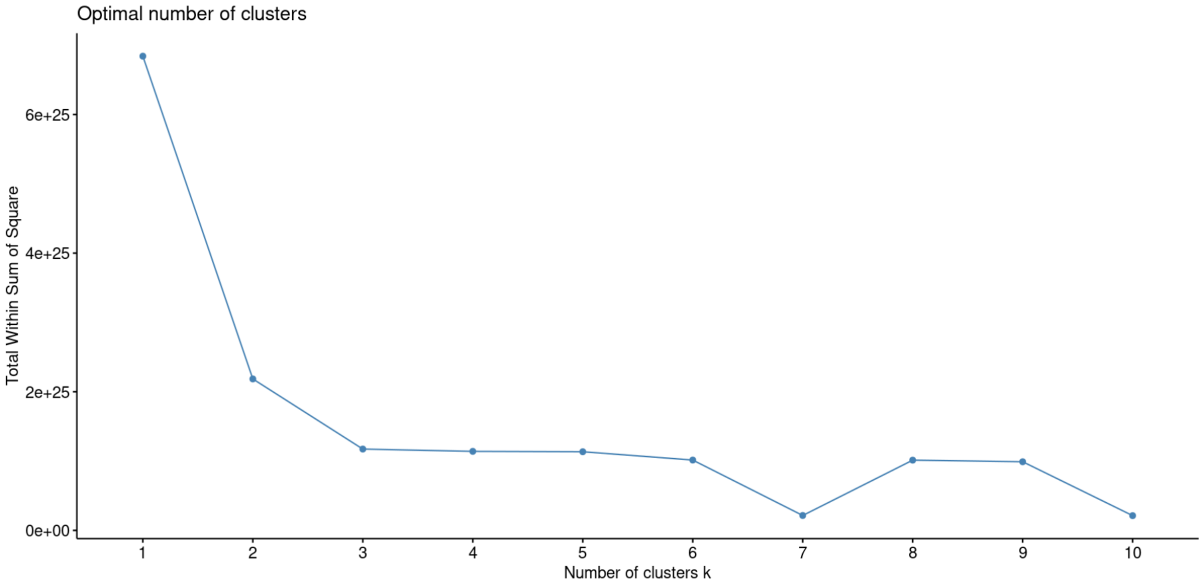


Figure 1: TWSS of the k-means clustering

From $k = 3$, the TWSS is not reduced significantly. However, it seems like the number of cluster that minimizes the TWSS is 7. This number of clusters was not chosen because it would create clusters with a small number of companies in each of them and the estimation cannot be based on a sample of 2 or 3 companies. This is why the population will be split in 3 clusters :

- *Cluster 1* : Smallest companies of the data set i.e. companies with the smallest values in FTEs, market capitalization, etc. such as Topdanmark.
- *Cluster 2* : Medium companies of the data set i.e. companies with the medium values in FTEs, market capitalization, etc. such as Aegon NV. The companies within this cluster are considered similar to Achmea.
- *Cluster 3* : Biggest companies of the data set i.e. companies with the highest values in FTEs, market capitalization, etc. such as AXA and Allianz SE.

The third cluster only contains 2 companies. The number of companies within each cluster is unfortunately unbalanced but that is due to the fact that the companies are very diverse. The companies within cluster 2 are not exactly the same ones as the one considered similar in Section 6.1. Indeed, the k-means algorithm takes the mean into account and not the evolution of the variables throughout the time, unlike what was done previously. Also, it only take quantitative variables into account i.e it does not take into account the headquarter country nor the specification of the company. A few companies belonged to several clusters over the 10 years. If a company was not considered medium in the year 2021, but had belonged to this cluster every year before, it was still considered medium. If the company was considered medium in the year 2021, but small or big in all the other years, it was not included in the cluster. Indeed, the values of the previous years will most likely disrupt the estimations. The results can be found in the following table. The results of the returns are expressed in %.

Company	Unlevered β	Leverage	Levered β	Actual return	Model implied return	
AG PLC	0.21	1.41	0.24	0.62	2.81	Cluster 1
Chesnara PLC	0.45	1.08	0.55	2.17	2.93	
HG PLC	0.84	1.00	1.03	4.56	2.26	
PeGH PLC	0.46	1.00	0.56	2.23	2.93	
Sampo PLC	0.95	1.25	1.12	4.99	2.16	
SCA SpA	1.37	1.19	1.66	7.73	2.23	
Topdanmark A/S	0.36	1.26	0.43	1.55	2.81	
Tryg A/S	0.47	1.17	0.48	2.29	2.90	
USA SpA	0.68	1.16	0.83	3.57	2.26	
UIG AG	0.74	1.46	0.87	3.77	2.16	
WW AG	0.20	5.25	0.24	0.61	2.81	
Aegon NV	1.19	1.34	1.67	7.78	4.91	Cluster 2
AG SpA	0.90	1.40	1.2	5.41	4.67	
Aviva PLC	0.79	1.78	1.05	4.67	5.41	
PhGH PLC	0.49	1.49	0.66	2.72	5.41	
Prudential PLC	1.16	1.32	1.63	7.56	4.91	
ZI Group	0.65	1.33	0.91	3.97	5.70	
Allianz SE	1.04	1.38	1.35	5.8	8.53	Cluster 3
AXA SA	1.48	1.23	1.62	9.64	6.58	

Table 5: Results of the clustering method

The abbreviations' meaning can be found in the AO table 1

The R^2 value was not included as there did not appear to have any link with the quality of the estimations. None of the estimations are accurate according to the 0.5 threshold. The estimated yield rates range from 2.16 to 8.53. However, the yield rates using the regressed β values range from 0.62 and 9.64. For example, for the company *Societa Cattolica di Assurazione SpA* (SCA SpA), the estimated yield rate is approximately three times lower than the yield rate using the regressed β of the company. As the yield rate is estimated using the median of the companies' βs and the median of the leverages, it is not surprising that the results are very stable but they are unfortunately not always accurate.

For the companies *Allianz* and *AXA*, the estimations only depend on the other company as the cluster is only formed of the two companies. The RMSE for this method is equal to 2.34 meaning that the estimations are usually 2.34 percentage points away from the real value. This method is considerably faster than the previous one. However, the estimations are less accurate. Indeed, only the raw values of the quantitative variables are taken into account. The purpose of this method was to perform better than the previous one and it does. It is also less time consuming so this method is more efficient than choosing similar companies by hand. However, it does not take qualitative variables into account. The following part aims to describe a method that takes both quantitative and qualitative variables into account.

6.3 Similarity metric

6.3.1 *k* closest companies

As explained earlier, this method was tested with three different values of k : 3, 5 and 7. The following table reports the results of each company's estimated and real yield rate. The real yield rates differ according to the k as the leverages are not always the same.

		k = 3		**k = 5**		**k = 7**	
Company	R^2	**Act. return**	**MIR**	**Act. return**	**MIR**	**Act. return**	**MIR**
Allianz SE	0.74	5.90	3.51	5.73	4.38	5.90	4.36
AXA SA	0.69	9.42	5.12	9.47	5.15	9.47	5.33
AG SpA	0.66	5.33	7.19	5.33	7.19	5.12	6.92
Sampo PLC	0.64	4.71	1.97	5.14	1.88	4.92	2.07
Aviva PLC	0.63	4.62	7.23	4.64	5.45	4.59	4.34
Prudential PLC	0.52	17.51	4.61	6.63	2.25	7.13	2.44
ZI Group	0.50	4.06	5.27	4.06	5.43	3.72	5.18
WW AG	0.49	0.66	3.11	0.66	3.09	0.61	2.16
Aegon NV	0.48	4.06	5.27	4.06	5.43	3.52	5.18
PhGH PLC	0.44	1.98	1.79	2.21	2.01	2.43	3.71
UIG AG	0.37	3.96	1.31	4.06	2.94	4.18	2.10
USA SpA	0.34	11.27	4.82	3.69	2.03	3.53	2.17
SCA SpA	0.33	7.12	2.59	7.12	2.64	7.42	2.10
Tryg A/S	0.28	1.99	4.44	1.86	2.80	2.06	3.69
HG PLC	0.15	3.79	1.79	4.20	2.01	4.38	2.16
Topdanmark A/S	0.12	1.52	3.48	1.44	2.62	1.44	3.69
Chesnara PLC	0.11	2.18	4.73	2.10	3.10	2.17	3.43
AG PLC	0.07	0.61	2.07	0.61	2.07	0.65	2.23
PeGH PLC	0.04	2.24	4.73	2.24	1.96	2.24	3.43

Table 6: Results of the similarity-based method

MIR stands for Model implied return

The R^2 value was included in the tables because one expected that the estimations were better when the R^2 was high. However, it seems like this is not the case in this situation. The estimation is considered good if it is not further than half a percent to its true value. The numbers colored in green correspond to those estimations.

Some estimations are very close to their real value and others are very far from it. For example, the estimated yield rate for *Societa Cattolica di Assurazione SpA* $k = 5$ is 7.12% when the yield rate using the company's real β is 2.64%. However, some companies have very good estimations. The estimations of the company *Phoenix Group Holdings PLC* are very accurate for k equal to 3 and 5. In order to have an idea of the overall performance of each method, the RMSE was also derived and can be found in the table below :

Method	**RMSE**
$k = 3$	4.11
$k = 5$	2.24
$k = 7$	2.45

Table 7: RMSE of the third method according to the chosen k

The method with $k = 3$ clearly performs the worse and the accuracy between choosing 5 and 7 companies is almost the same. Compared, to the previous method 6.2, the similarity metric with k equal to 5 and 7 performs better. The downside of this method is that each company has the same impact on the final estimation and it might not give the most accurate results when the companies have very different similarity scores. This is why the following part aims to describe the method that assigns a weight to the companies so that the similarity level is taken into account in the estimation.

6.3.2 Weighted estimation method

This method consists in using all the companies of the data set into account. Each company takes part in the estimation according to its similarity score. The results of this method are described in the following table :

Company	R^2	Actual return	Model implied return
Allianz SE	0.74	5.89	3.68
AXA SA	0.69	9.20	4.61
AG SpA	0.66	5.30	4.57
Sampo PLC	0.64	5.30	4.57
Aviva PLC	0.63	4.61	5.00
Prudential PLC	0.52	6.92	3.81
ZI Group	0.50	3.76	5.06
WW AG	0.49	0.68	4.03
Aegon NV	0.48	3.76	5.06
PhGH PLC	0.44	2.43	3.69
UIG AG	0.37	3.69	2.52
USA SpA	0.34	4.06	3.49
SCA SpA	0.33	8.05	3.49
Tryg A/S	0.28	2.37	3.69
HG PLC	0.15	4.71	3.49
Topdanmark A/S	0.12	1.65	3.62
Chesnara PLC	0.11	3.43	3.69
Admiral Group PLC	0.07	0.69	3.68
PeGH PLC	0.04	2.31	3.69

Table 8: Results of the weighing method

The estimated CAPM result refers to :

$$\mu_i = R_f + \text{Median }(B_w) * \text{Median }(L_w) * E(R_m - R_f)$$

B_w corresponding to the weighted list of the similar companies' βs of the k most similar companies to the TC.

The estimations range from 2.52% to 5.06%. However, the real values range from 0.68% to 9.20%. The RMSE is equal to 2.31 which means that the estimations are on average 2.31 units of percentage below or above the real value. The goal, which was actually to reach a RMSE equal to 0.5% is unfortunately not reached but it is still below the RMSEs concerning the Comparable Companies Analysis.

6.3.3 Sensitivity analysis

This part aims to describe the impact of the different scores on the estimation results. The results can be found in the tables in Appendix A. Concerning the method where one chooses the $k = 3$ closest companies, it seems like it is preferable to give more importance to the qualitative variables or to give the same importance to each parameter.

For the $k = 5$ and 7 methods, the most efficient method seems to be the one where one gives the more importance to the specification and headquarter.

Concerning the weighing method, the results are quite similar to one another. Changing the scores does not seem to change the estimations a lot. It changes the ranking but still takes every company into account which obviously does not change the results by much. Giving the same level of importance to each criteria gives less accurate estimations. it seems important to make a distinction between the different criterion taken into account.

To sum up, changing the scores of the similarity metric slightly change the estimations. The difference between the RMSEs may not seem like much but even a 0.1 different is important especially when one is aiming for a low RMSE.

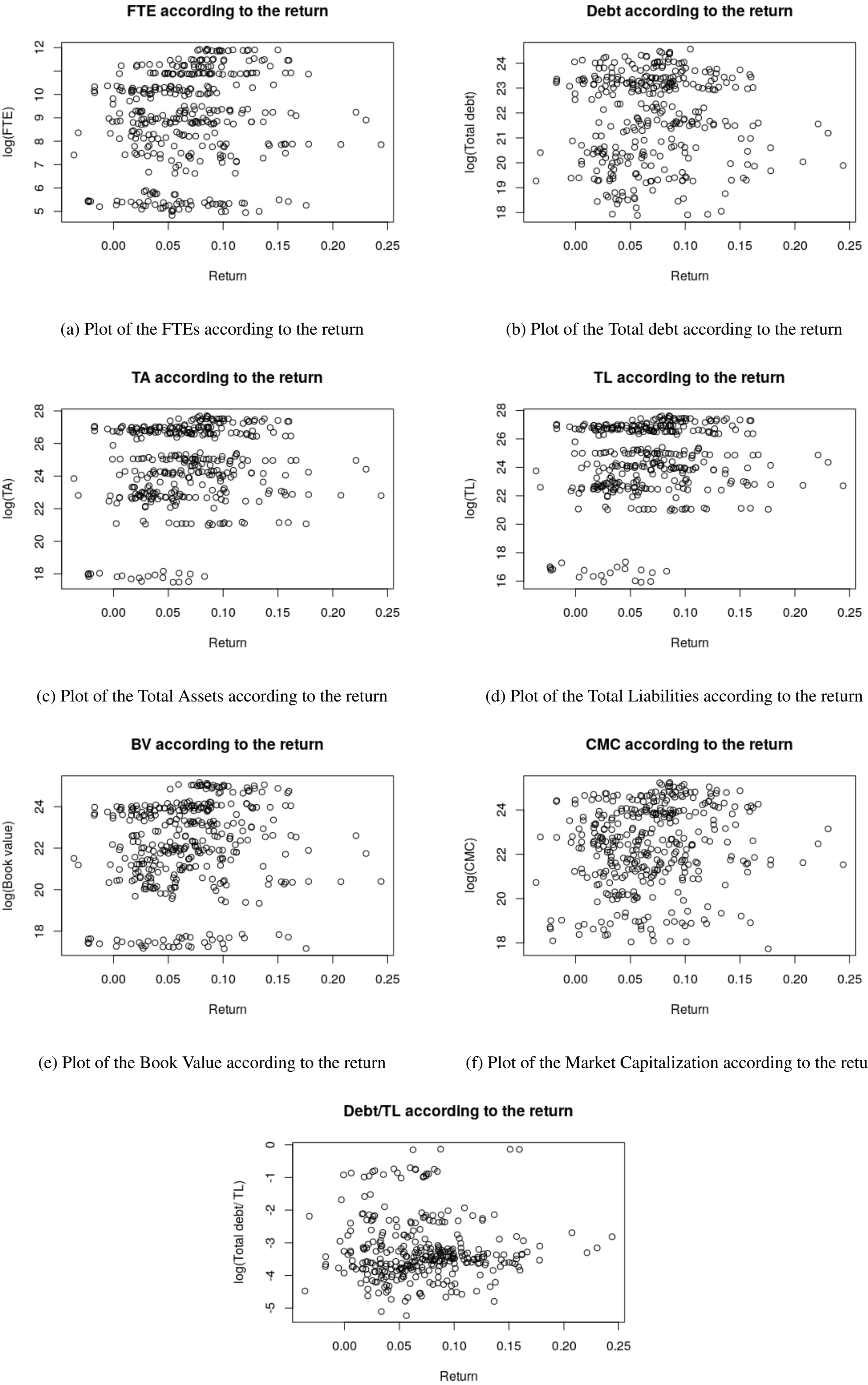
6.4 Cross sectional regression

This method is the only one focusing on estimating the returns directly without estimating the β coefficient beforehand. It aims to find a relationship between the returns and the variables of the data set. R proposes an option to do both forward and backward selection and choose the best regression out of it.

6.4.1 Assumptions

The variables used in the linear regression need to have a linear relationship with the returns. The following plots tend to give a visual idea of the validation of this assumption. In the Appendix C can be found the graphs with outliers and without the logarithm.

Linearity :



(a) Plot of the FTEs according to the return

(b) Plot of the Total debt according to the return

(c) Plot of the Total Assets according to the return

(d) Plot of the Total Liabilities according to the return

(e) Plot of the Book Value according to the return

(f) Plot of the Market Capitalization according to the return

(g) Plot of the Total debt/ TL ratio according to the return

Figure 2: Plots of the different variables

The linear relationship with certain variables such as Total Assets is not obvious but the relationship does not need to be perfectly linear either. In order to check the validity of the assumption, the following plot can be used :

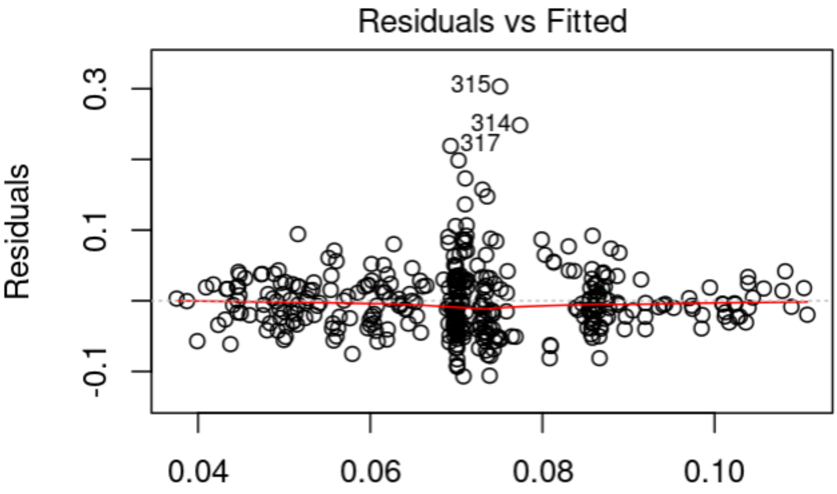


Figure 3: Residuals VS Fitted plot

The red line does not show any pattern so we can conclude that all the variables respect the linearity assumption.

Homoscedasticity :

The fitted VS residuals plot allows to determine whether there is homoscedasticity.

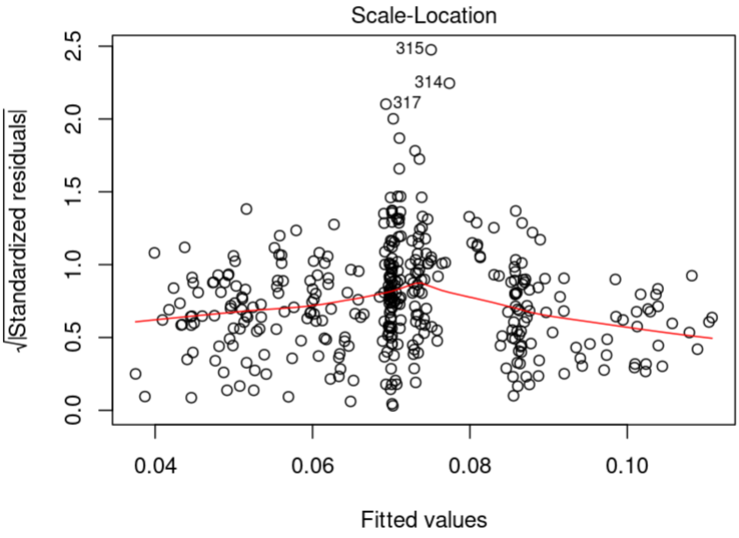


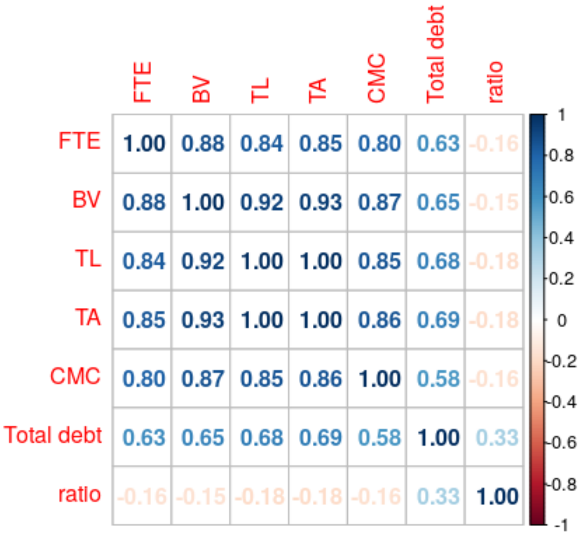
Figure 4: Scale location plot

The red line is not perfectly horizontal but the data points are equally spread from each side of the line so we can conclude that the assumption holds.

As the red line is close to the dashed one, the homoscedasticity assumption seems to be valid.

Lack of perfect multicolinearity :

One considers that the variables are not perfectly colinear if the correlation is between 0.9 and -0.9.



	FTE	BV	TL	TA	CMC	Total debt	ratio
FTE	1.00	0.88	0.84	0.85	0.80	0.63	-0.16
BV	0.88	1.00	0.92	0.93	0.87	0.65	-0.15
TL	0.84	0.92	1.00	1.00	0.85	0.68	-0.18
TA	0.85	0.93	1.00	1.00	0.86	0.69	-0.18
CMC	0.80	0.87	0.85	0.86	1.00	0.58	-0.16
Total debt	0.63	0.65	0.68	0.69	0.58	1.00	0.33
ratio	-0.16	-0.15	-0.18	-0.18	-0.16	0.33	1.00

Figure 5: Correlation plot

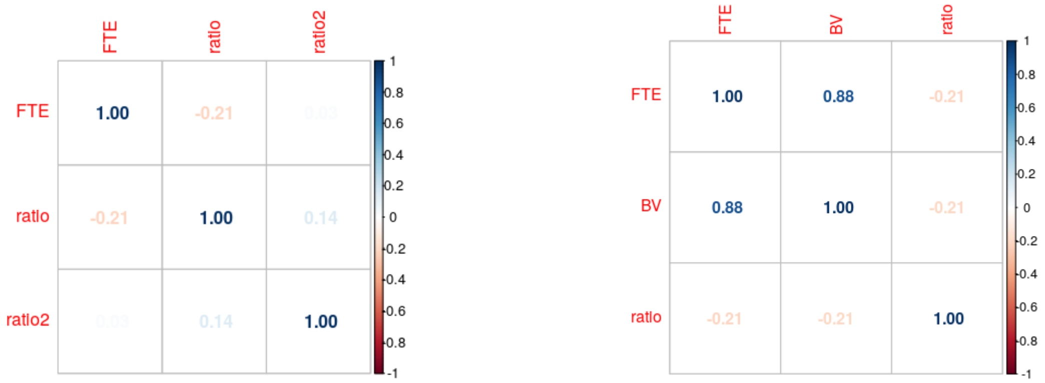
There seems to be a strong correlation between the Book Value and the Total Liabilities (0.92) and between the Book Value and Total Assets (0.93). So the TA and TL variables were removed from the model. The *ratio* variable corresponds to :

$$\frac{\text{Total debt}}{\text{Total Liabilities}}$$

The *ratio2* variable corresponds to :

$$\frac{\text{BV}}{\text{CMC}}$$

The graph on the left shows the correlation plot between the following variables : FTE, $\frac{Debt}{TL}$ and $\frac{BV}{CMC}$. The correlation is almost null between each variable which is a good point. The variables *debt* and *TL* had much higher correlation values before this transformation. In order to minimize the correlation effect of the *BV* variable, the *ratio2* variable is added. However, in the case of an unlisted company, the *CMC* cannot be used. This is why the *BV* was used as a proxy, hence the correlation plot on the right.



(a) Corrplot of other variables

(b) Corrplot of other variables

Figure 6: Correlation plots of different variables

The right correlation plot 6b shows a strong correlation between the variables *FTE* and *BV*. It will probably be preferable to leave the *BV* variable out in the regression.

Independence of errors :

Finally, the residuals need to be normally distributed.

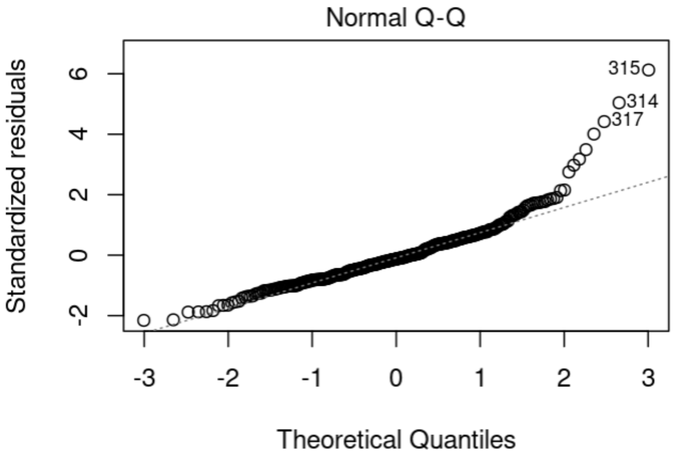


Figure 7: QQ plot of the residuals

The points seem to follow the dashed lines : the assumption is respected except at the extremes but the assumption is still considered valid.

6.4.2 Coefficients and estimations

The variable total debt was not used in the regression. The variable *ratio* was used as a proxy of the book-to-market ratio variable. In the case of an unlisted company, the Market Capitalization cannot be used which is why another variable was used in this study. In order not to count the debt aspect twice, the variable was only taken into account in the *ratio* variable.

The regression using all the variables gave the following results :

Variables	**Coefficients**	**p-value**
Intercept	0.05	$< 0.01**$
log(FTE)	< 0.01	0.71
log(ratio)	≈ 0.00	0.75
log(CMC)	< 0.01	0.58
log(BV)	≈ 0.00	0.84

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 9: Results of the linear regression

Note that the variables *BV* and *CMC* are expressed in millions. Only the intercept is significant at the level α = 5% in this case which is surprising. This implies that none of the variables of the model are significant.

The actual return corresponds to the total index return from 01/07/2021 to 31/12/2021 (Eikon, 2022). The model implied returns are derived as follows :

$$\text{Return} = \text{Intercept} + \beta_{log_{FTE}} * log(FTE) + \beta_{log_{ratio}} * log(ratio) + \beta_{log_{CMC}} * log(CMC) + \beta_{log_{BV}} * log(BV)$$

The results can be found in the following table.

Company	R^2	Actual return	Model implied return
Allianz SE	0.74	6.80	9.23
AXA SA	0.69	10.04	9.14
AG SpA	0.66	6.47	9.05
Sampo PLC	0.64	6.46	8.79
Aviva PLC	0.63	3.32	8.82
Prudential PLC	0.52	-0.05	8.97
ZI Group	0.50	6.92	9.10
WW AG	0.49	2.25	8.15
Aegon NV	0.48	6.86	8.73
PhGH PLC	0.44	4.06	8.73
UIG AG	0.37	4.66	8.64
USA SpA	0.34	8.71	8.68
SCA SpA	0.33	5.54	8.26
Tryg A/S	0.28	1.82	8.56
HG PLC	0.15	2.71	7.54
Topdanmark A/S	0.12	5.54	8.61
Chesnara PLC	0.11	5.64	8.15
Admiral Group PLC	0.07	2.61	8.75
PeGH PLC	0.04	5.98	7.57

Table 10: Results of the linear regression

The results are in %

The results obtained using the linear regression range from 7.57% to 9.23% when the actual returns range from -0.05% to 10.04%. This method does not seem very accurate and the RMSE value confirms that this method is less accurate than the preceding ones.

RMSE	4.40
R^2	< 0.01
p-value	0.72

The RMSE is not very low, especially compared to the complexity of this method. it requires a lot of preprocessing and performs even worse than choosing companies by hand.

The best regression in terms of AIC is regression using only the intercept meaning that the returns are all estimated as the same value :

This model is constant but the only coefficient is very significant. The RMSE is equal to 3.31, which is lower than the regression using all the variables. To finish, a regression without intercept was tested giving the following results :

Variables	Coefficients	p-value
Intercept	0.07	$< 2.00 * 10^{-16} ***$

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 11: Coefficients of the linear regression using the AIC criteria

Variables	Coefficients	p-value
log(ratio)	≈ 0.00	0.10
log(CMC)	< 0.01	$< 8.56 * 10^{-11} ***$

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 12: Coefficients of the linear regression without intercept

The variable Market Capitalization is very significant even at level $\alpha = 1\%$. The CMC is positively linked to the returns meaning that the higher the CMC the higher the returns. However, the results that can be found in Appendix D are not accurate at all :

RMSE	15.28
R^2	0.71
p-value	$< 2.2 * 10^{-16}$

The p-value is very low, meaning that the model is significant. Also, the R^2 is high, meaning that the linear regression fits the data well. However, the RMSE is very high, implying that the estimation are not accurate. This model does not seem to be very efficient.

To sum up, the linear regression is more time consuming that the other methods because assumptions need to be checked beforehand. Furthermore, the results are not more accurate than the other methods so it will not be chosen over the other methods.

None that the preceding methods do not take the temporal aspect into account. The following part describes a method that does and it is expected to work better than all the preceding ones.

6.5 Fama-Mac Beth regression

This part describes the results obtained with the Fama-MacBeth regression.

6.5.1 Assumptions

Linearity assumptions :

These refer to the assumptions that need to be valid in order to conduct a linear regression. They have been tested in the previous part 6.4.1.

Random sampling of observations :

The observations used for the model have to be randomly picked. In this study, they have been picked according to the amount of available data. The amount of available data being random for each company, it implies that this assumption is valid.

Conditional mean of errors :

The conditional mean of errors inferior to $2.72 * 10^{-}19$ which can be interpreted as negligible hence equal to 0.

6.5.2 Coefficients and estimations

The variables have been scaled to a million -apart from the *FTE* variable - in order to have higher coefficients.

Variables	**Coefficients**	**p-value**
Intercept	0.64	0.40
log(FTE)	-0.04	0.65
log(BV)	-0.02	0.44
log(CMC)	0.01	0.58

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 13: Results of the Fama-Mac Beth regression

None of the coefficients is significant. The intercept is extremely high which will probably lead to very high returns.

Company	R^2	Actual return	Model implied return
Allianz SE	0.74	6.80	-1.57
AXA SA	0.69	10.04	0.31
AG SpA	0.66	6.47	1.90
Sampo PLC	0.64	6.46	9.75
Aviva PLC	0.63	3.32	6.61
Prudential PLC	0.52	-0.05	9.75
ZI Group	0.50	6.92	3.72
WW AG	0.49	2.25	11.70
Aegon NV	0.48	6.86	5.65
PhGH PLC	0.44	4.06	11.50
UIG AG	0.37	4.66	4.88
USA SpA	0.34	8.71	10.00
SCA SpA	0.33	5.54	17.30
Tryg A/S	0.28	1.82	14.50
HG PLC	0.15	2.71	31.30
Topdanmark A/S	0.12	5.54	19.20
Chesnara PLC	0.11	5.64	26.00
Admiral Group PLC	0.07	2.61	13.50
PeGH PLC	0.04	5.98	30.00

Table 14: Results of the Fama-Mac Beth regression

The return results are in %

This method was expected to work best but it does not seem to explain the returns correctly at all. However, some studies show that this method gives biased estimators even when all the assumptions are met **?**. The intercept value is very high, implying that a company with 0 for all the variables will be expected to have a 64% return. The estimated values are either extremely high or extremely low.

RMSE	12.46
R^2	0.74

Although the R^2 is high, the RMSE is the highest compared to the other tested methods. This method clearly does not perform well to estimate the returns of a company. As the numbers seemed very high, the same model was tested but without an intercept and the results can be found below :

None of the coefficients are significant so the results are not expected to be accurate :

Variables	Coefficients	p-value
log(FTE)	≈ 0	0.95
log(BV)	≈ 0	0.94
log(CMC)	0.01	0.60

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 15: Results of the Fama-Mac Beth regression

Company	R^2	Actual return	Model implied return
Allianz SE	0.74	6.80	17.70
AXA SA	0.69	10.04	17.60
AG SpA	0.66	6.47	17.00
Sampo PLC	0.64	6.46	17.30
Aviva PLC	0.63	3.32	16.80
Prudential PLC	0.52	-0.05	17.80
ZI Group	0.50	6.92	17.70
WW AG	0.49	2.25	14.90
Aegon NV	0.48	6.86	16.61
PhGH PLC	0.44	4.06	16.30
UIG AG	0.37	4.66	15.10
USA SpA	0.34	8.71	16.20
SCA SpA	0.33	5.54	15.00
Tryg A/S	0.28	1.82	17.00
HG PLC	0.15	2.71	13.70
Topdanmark A/S	0.12	5.54	16.40
Chesnara PLC	0.11	5.64	14.70
Admiral Group PLC	0.07	2.61	17.00
PeGH PLC	0.04	5.98	13.80

Table 16: Results of the linear regression

The results are in %

The results using this method seem very high compares to the actual returns :

RMSE	11.60
R^2	0.72

The RMSE using this method is lower than the previous one but is still very high compared to the other methods' RMSEs. Although the R^2 is high -equal to 0.72 - this method does not perform well at all. The Fama-

MacBeth regression does not seem to be an efficient estimation method of the returns in the present case.

6.6 Fama-French Three and Five Factor Models

The following table describes the coefficients of the Fama-French Three and Five Factor Models. This part does not correspond to an estimation method but shows how the CAPM can be expanded with additional factors. Those factors do not correspond to an estimation. They show how a company's characteristic influences the expected return. For example, if the *SMB* factor is equal to 0.2, it means that a small company has returns higher by 0.2 percentage points than big ones.

The coefficients are quite high compared to other studies. This is due to the fact that the data set is not very big. This means that the averages are taken on a small sample so each company can have a big impact on the coefficients. Furthermore, studies usually use average monthly returns. In this study, return over a 6-month period were used, this means that only 2 values are taken into account instead of 30 in the other studies. This is why the interpretations need to be taken into account carefully.

Period	**SMB**	**HML**	**RMW**	**CMA**
31/12/2021	-0.73	1.04	1.12	6.45
30/06/2021	1.61	2.66	9.61	2.34
31/12/2020	0.2	-2.24	-1.93	1.57
30/06/2020	-1.38	1.41	-0.01	-2.14
31/12/2019	-0.14	5.82	-0.77	2.13
30/06/2019	0.17	2.88	1.65	6.34
31/12/2018	1.64	-1.63	-0.1	4.7
30/06/2018	-0.26	0.28	3.2	0.96
31/12/2017	2.77	0.8	7.02	8.17
30/06/2017	1.99	2.29	6.57	10.37
31/12/2016	-1.14	9.65	-3.58	2.39
30/06/2016	1.5	-4.55	12.41	11.29
31/12/2015	0.17	-1.65	5.41	12.95
30/06/2015	-0.41	0.79	5.09	1.6
31/12/2014	-3.61	-5.58	8.15	-1.29
30/06/2014	0.38	-8.63	10.78	-1.29
31/12/2013	-2.1	-2.51	7.36	5.66
30/06/2013	-0.37	-8.74	8.25	14.43
31/12/2012	-2.89	-5.44	8.15	-2.24
30/06/2012	NA	1.2	8.52	NA
Average	-0.14	-0.61	4.85	4.84

Table 17: Fama-French 3 and 5 factor models results

The results are in %

The *SMB* factor goes against expectations Rosenberg & Reid (1985). Smalls firms are expected to have higher average returns but in our case, small firms have on average 0.73% lower returns than large firms. The same can be said for the *HML* factor, high book-to-market ratio companies gave on average 0% lower returns than companies with a low book-to-market ratio. The results of the two other factors are in line with our expectations : companies with robust profitability had on average returns 4.85% higher than companies with weak profitability over the time period. Finally, companies with conservative investment have on average 4.84% higher returns than the ones with aggressive investment.

In the case of an unlisted company, the factors can be added or subtracted according to where the company lies in terms of size, book-to-market ratio, etc. However, in the case of a small sample, the figures can get big and disrupt the expected returns. It would be preferable to use those additional factors with a higher number of observations.

6.7 Comparison of the results

The following table sums up the RMSEs associated to each of the estimation method.

Method	RMSE	
Comparable Companies Analysis	3.06	
K-means clustering	2.34	
Similarity Metric $k = 3$	4.11	β estimation
Similarity Metric $k = 5$	2.24	
Similarity Metric $k = 7$	2.45	
Similarity Metric with weights	2.31	
Cross sectional regression	4.40	
Optimal Cross sectional regression	3.31	
Cross sectional regression without intercept	15.28	Returns estimation
Fama-MacBeth regression	12.46	
Fama-MacBeth regression without intercept	11.60	

Table 18: RMSEs of the different estimation methods

The approach consisting in estimating the systematic risk always performs better on average than the ones estimating the returns. The Fama-MacBeth regressions performed very bad and adding a temporal aspect did not increase the accuracy of the estimations. In the present case, estimating the returns through an estimation of the systematic risk gives better results with an average RMSE equal to 2.76 versus 9.41.

The k-means clustering and the similarity metric are the methods that perform best. Note that when using other companies in order to estimate a TC's systematic risk, it is better to use between 5 and 7 companies. Using less than 5 gives significantly less accurate results. Instead of just choosing companies by hand, it could be interesting to use the similarity metric method and the k-means clustering before comparing the results. This would give two quite accurate estimations that, if close to one another, can imply that the estimation is probably accurate.

When estimating the returns, it seems that the method that performed best is the Cross Sectional regression using the AIC criteria. It is surprising, as adding the temporal aspect was expected to give similar or better results but not worse ones. Unfortunately, no information was found to explain why there is such a gap between the normal and time-series models' accuracies.

The Fama-French Three and Five factor models are supposed to give more precise expected returns than the CAPM, as more factors are included. However, it is better to use them with a high number of individuals. Dividing the population in three groups with only 19 companies gives small subgroups leading to sensitive factors. Hence, the coefficients are very sensitive to each of the company's returns and do not result in accurate expected returns. It is also important to keep in mind that those coefficients are relative to the data set meaning that a company considered big in this study could be considered small if the available data was different. This implies that it is preferable to use those models with as many observations as possible.

7 Conclusion and limits of the study

In this paper, different estimation methods used to determine the cost of equity of an unlisted company are investigated. Indeed, it is extremely important for companies to valuate themselves but it is not always easy in the case of private companies. Furthermore, the existing literature on the subject and the diversity of the methods are quite poor. In this study, new methods were tested and also compared in order to find the best complexity/efficiency trade off. Two approaches were considered throughout the analysis :

- Estimating the cost of equity through an estimation of the systematic risk
- Estimating the cost of equity through an estimation of the returns

Finally, additional factors were included in order to see their impact on the expected returns.

First, I focused on estimating the systematic risk through different techniques. The first method, consisting in finding similar companies to the TC, is the easiest one in terms of required Machine Learning skills. Even though the results are not incredibly bad, this method is very time-consuming which led to the challenge of finding a method that performs better and that is quicker. The k-means clustering method is most definitely the most rapid one of this paper and it gives estimations that are significantly better than the previous method. However, this Machine Learning technique only takes quantitative variables into account resulting in a loss of information as qualitative variables seemed very important too. This led to the construction of a similarity metric that aimed to be an automation of the Comparable Companies Analysis. This technique, being a code written by myself, allows a flexibility that lacked using the k-means clustering. All types of variables can be taken into account and variables can be added or deleted from the process any time. Furthermore, it also allows to change the significance of each criteria according to the our choices. In the end, the similarity method giving a higher importance to the qualitative variables with $k = 5$ is the one that performs best.

Concerning the methods focusing on the returns, the goal was to find a linear relationship between the returns and other variables such as the FTEs of a company, its book value, etc. Unfortunately, all those methods performed quite poorly even when taking the temporal aspect into account.

To sum up, estimating the cost of equity of an unlisted company gives better results using an estimation of the systematic risk. Furthermore, around 5 companies need to be used in order to do a reliable estimation. Finally, in the case of a small sample, it is preferable not to use additional factors. However, some improvements could be made in order to enrich the study.

As mentioned in the paper, the number of individuals used in this study was quite small. It would be interesting to conduct it again with a sample of at least 40 companies and see the impact on the estimations. Furthermore, using more variables or more time periods would also probably make the estimations more accurate but it is not easy to find complete market data before 2012. Finally, an additional estimation technique could have given interesting results. It consists in finding a relationship between the Restricted Tier 1's (RT1) yield to call and the returns of companies in a time-series. However, due to lack of data and time, this lead was not studied.

References

Abudy, M., Benninga, S., & Shust, E. (2015). The cost of equity for private firms. *The Journal of Financial and Quantitative Analysis*, *37*, 431–443.

ACCA. (2020). *Capm : theory, advantages, and disadvantages* (Working Paper). Association of Chartered Certified Accountants. Retrieved from https://www.accaglobal.com/gb/en/student/exam-support-resources/fundamentals-exams-study-resources/f9/technical-articles/CAPM-theory.html

Alhassane Garba, A., Sene, B., & Mendy, P. (2019). *Essai sur les modèles financiers appliqués à la brvm* (Research paper). Cheikh Anta Diop University. Retrieved from http://colloque.supdeco.sn/pdf/articles/abdoulaziz-garba.pdf

Amézola, L., & Dolz, M. (2017). *A 5-factor risk model for european stocks* (Master Thesis). Hautes Etudes Commerciales. Retrieved from https://upcommons.upc.edu/bitstream/handle/2117/114352/Research%20paper%202017_M_Dolz_L_Amezola.pdf?sequence=1&isAllowed=y

Beitone, A., Cazorla, A., Dollo, C., & Drai, A.-M. (2012). *Dictionnaire de science économique, 3ème édition revue et augmentée* (3rd ed.). Colin, Armand.

Bloomberg. (2022). *Bloomberg data bases.* Retrieved from https://www.bloomberg.com

Corbière, M., & Larivière, N. (2020). *Méthodes qualitatives, quantitatives et mixtes.* PUQ.

Drobetz, W., Ivan, M., & Seidel, J. (2014). *Leverage, beta estimation, and the size effect* (Working Paper). University of Hamburg. Retrieved from https://deliverypdf.ssrn.com/delivery.php?ID=39200110407002608702412202906702200701406805706302803709201203103110401312700011203100603704912410600
pdf&INDEX=TRUE

ECB, E. C. B. (2021). *Euro short-term rate.* Retrieved from https://sdw.ecb.europa.eu/browseTable.do?org.apache.struts.taglib.html.TOKEN=c3c6b0ae178dc359b4af1fa840881a60&df=true&MAX_DOWNLOAD_SERIES=500&DATASET=0&org.apache.struts.taglib.html.TOKEN=1f6479a7fc9eff2cfffbb4cd5bce236e&org.apache.struts.taglib.html.TOKEN=08970dde9e9808917cba1abc0f94ba64&legendRef=reference&node=9698150&SERIES_MAX_NUM=50&activeTab=EST&start=31-12-2021&end=05-01-2022&submitOptions.x=0&submitOptions.y=0&trans=N&q=&type=

Eikon. (2022). *Eikon data bases.* Retrieved from https://www.refinitiv.com/en/products/eikon-trading-software

Fama, E., & French, K. (1992). The cross-section of expected stock returns. *Corporate Finance Institute*, *47*, 427–465.

Fama, E., & MacBeth, J. D. (1973). Risk, return, and equilibrium: Empirical tests. *The Journal of Finance*, *81*, 607–636.

Fama-french three-factor model. (2020). Retrieved from https://corporatefinanceinstitute.com/resources/knowledge/finance/fama-french-three-factor-model/

Favereau, O. (2015). La fin de l'entreprise privée. In *L'entreprise dans un monde sans frontières : perspectives économiques et juridiques* (p. 305-320). Dalloz, Collection les Sens du Droit. (Supiot, A.)

Guyader, A. (2012). *Régression linéaire* [Working Paper]. Retrieved from https://perso.lpsm.paris/~aguyader/files/teaching/Regression.pdf

Hodson, T. O. (2022). Root-mean-square error (rmse) or mean absolute error (mae): when to use them or not. *European Geosciences Union*, *15*, 5481—5487.

Jansen, E. T. (2019). *The five-factor model in the netherlands* (Bachelor Thesis). Erasmus University Rotterdam.

KPMG. (2021). *Corporate tax rates table.* Retrieved from https://home.kpmg/xx/en/home/services/tax/tax-tools-and-resources/tax-rates-online/corporate-tax-rates-table.html

KPMG. (2022). *Equity market risk premium – research summary.* Retrieved from https://indialogue.io/clients/reports/public/5d9da61986db2894649a7ef2/5d9da63386db2894649a7ef5

Limaiem, I. (2009). *Les facteurs du modèle de fama et french: Cas du marché des actions canadiennes* (Master Thesis). Université du Québec in Montréal. Retrieved from https://archipel.uqam.ca/2202/1/M10858.pdf

Liu, Z., Bao, J., & Ding, F. (2018). *Research on k-means clustering algorithm: An improved k-means clustering algorithm* (Working paper). Imperial College London.

Mirzayev, E. (2021). *How to calculate the beta of a private company.* Retrieved from https://www.investopedia.com/articles/personal-finance/050515/how-calculate-beta-private-company.asp

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). *Introduction to linear regression analysis.* John Wiley & Sons.

Pasquariello, P. (1999). *The fama-macbeth approach revisited* (PhD Research Paper). New York University – Stern School of Business. Retrieved from https://www.researchgate.net/publication/238733735_THE_FAMA-MACBETH_APPROACH_REVISITED

Pereiro, L. E. (2002). *Valuation of companies in emerging markets: A practical approach.* Wiley.

Reinganum, M. R. (1981). A new empirical perspective on the capm. *The Journal of Financial and Quantitative Analysis*, *16*, 439-462.

Rosenberg, B., & Reid, R., K.and Lanstein. (1985). Persuasive evidence of market inefficiency. *Journal of Portfolio Management*, *11*, 9–16.

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *The Journal of Finance*, *19*, 425–442.

Sinaga, K. P., & Yang, M.-S. (2020). Unsupervised k-means clustering algorithm. *IEEE Access*, *8*, 80716-80727.

Supichaya, S. (2015). *Linear regression analysis on net income of an agrochemical company in thailand* (Bachelor Thesis). Portland State University. Retrieved from https://pdxscholar.library.pdx.edu/honorstheses/131/

Appendices

A The impact of the scores on the similarity metric

A.1 Scores 1

Criteria	**Score**
Specification	100 - 100 -200
Headquarter	0 - 200
Raw values	0 - 10
2021/2012 ratio	0 - 200
Regression coefficient	0 - 200

Table 19: Scores of the similarity metric

A.2 Results scores 1

		k = 3		**k = 5**		**k = 7**	
Company	R^2	Act. return	MIR	Act. return	MIR	Act. return	MIR
Allianz SE	0.74	5.90	3.51	5.73	4.38	5.73	4.23
AXA SA	0.69	11.10	6.61	11.10	6.61	9.71	5.47
Assicurazioni Generali SpA	0.66	5.38	5.55	5.33	6.03	5.12	6.92
Sampo PLC	0.64	4.71	2.59	4.71	1.97	4.92	2.07
Aviva PLC	0.63	4.62	7.23	4.64	5.45	4.59	6.03
Prudential PLC	0.52	17.57	4.61	0.055	0.034	6.19	3.21
Zurich Insurance Group	0.50	4.06	5.27	4.06	5.43	3.72	5.18
Wuestenrot & Wuerttembergische AG	0.49	0.66	3.11	0.64	2.24	0.61	2.16
Aegon NV	0.48	8.91	6.10	7.96	5.24	7.52	5.32
Phoenix Group Holdings PLC	0.44	7.23	4.59	2.43	3.68	2.32	2.78
UNIQA Insurance Group AG	0.37	11.27	4.82	3.93	1.96	4.18	2.10
UnipolSai Assurazioni SpA	0.34	3.96	1.31	3.69	1.96	3.53	2.17
Societa Cattolica di Assurazione SpA	0.33	7.42	2.77	7.42	2.10	7.42	2.10
Tryg A/S	0.28	1.99	4.44	1.86	2.80	2.06	3.39
Hansard Group PLC	0.15	4.52	1.94	4.52	2.21	4.38	2.16
Topdanmark A/S	0.12	1.30	1.79	1.44	2.62	1.44	3.69
Chesnara PLC	0.11	2.18	4.73	2.17	3.74	2.07	3.91
Admiral Group PLC	0.07	0.61	5.06	0.61	5.06	0.56	2.03
Personal Group Holdings PLC	0.04	2.24	4.73	2.23	3.74	2.14	3.91

Table 20: Results of similarity metric according the same importance to each parameter for $k = 3$, 5 and 7

Company	R^2	CAPM	Estimated CAPM
Allianz SE	0.74	5.89	3.68
AXA SA	0.69	9.20	4.61
Assicurazioni Generali SpA	0.66	5.35	4.61
Sampo PLC	0.64	5.35	2.52
Aviva PLC	0.63	4.61	5.00
Prudential PLC	0.52	6.72	3.69
Zurich Insurance Group	0.50	3.76	5.06
Wuestenrot & Wuerttembergische AG	0.49	0.68	4.64
Aegon NV	0.48	7.32	4.61
Phoenix Group Holdings PLC	0.44	2.52	3.62
UNIQA Insurance Group AG	0.37	4.04	3.48
UnipolSai Assurazioni SpA	0.34	3.68	2.50
Societa Cattolica di Assurazione SpA	0.33	8.03	2.10
Tryg A/S	0.28	2.36	3.68
Hansard Group PLC	0.15	4.71	3.49
Topdanmark A/S	0.12	0.024	0.044
Chesnara PLC	0.11	2.24	3.68
Admiral Group PLC	0.07	0.69	3.65
Personal Group Holdings PLC	0.04	2.30	3.68

Table 21: Results of the weighing method

Method	**RMSE**
$k = 3$	4.25
$k = 5$	2.53
$k = 7$	2.37
Weighing method	2.54

Table 22: RMSEs of the third method according to the chosen k and the weighing method

A.3 Scores 2

Criteria	**Score**
Specification	100 - 100 - 200
Headquarter	0 - 100
Raw values	0 - 5
2021/2012 ratio	0 - 50
Regression coefficient	0 - 50

Table 23: Scores of the similarity metric

A.4 Results scores 2

		k = 3		**k = 5**		**k = 7**	
Company	R^2	Act. return	MIR	Act. return	MIR	Act. return	MIR
Allianz SE	0.74	5.96	3.51	5.73	4.38	5.96	3.87
AXA SA	0.69	25.37	8.09	11.17	6.1	9.54	4.99
AG SpA	0.66	5.36	7.23	5.33	7.19	5.12	6.92
Sampo PLC	0.64	5.72	4.31	5.14	1.88	5.14	1.88
Aviva PLC	0.63	4.78	6.57	4.79	4.71	4.64	4.38
Prudential PLC	0.52	6.39	3.49	7.38	4.07	6.94	3.81
ZI Group	0.50	4.06	5.27	4.06	5.43	3.72	5.18
WW AG	0.49	0.61	2.16	0.57	2.01	0.53	4.44
Aegon NV	0.48	8.91	6.10	7.52	5.32	7.52	5.32
PhGH PLC	0.44	7.64	2.81	2.63	1.58	2.32	2.78
UIG AG	0.37	11.27	4.82	3.93	1.96	4.18	2.10
USA SpA	0.34	3.58	4.73	3.57	3.74	3.43	2.16
SCA SpA	0.33	7.12	2.59	7.12	2.64	7.42	2.10
Tryg A/S	0.28	1.99	4.44	2.19	2.78	2.06	2.62
HG PLC	0.15	3.79	1.79	4.20	2.01	4.38	2.16
Topdanmark A/S	0.12	1.6	4.32	1.44	2.62	1.44	2.52
Chesnara PLC	0.11	1.88	5.49	2.18	3.10	2.18	1.96
AG PLC	0.07	0.61	3.48	0.61	2.07	0.56	2.03
PeGH PLC	0.04	1.94	5.49	2.14	3.29	2.24	1.96

Table 24: Results of similarity metric according the same importance to each parameter for $k = 3, 5$ and 7

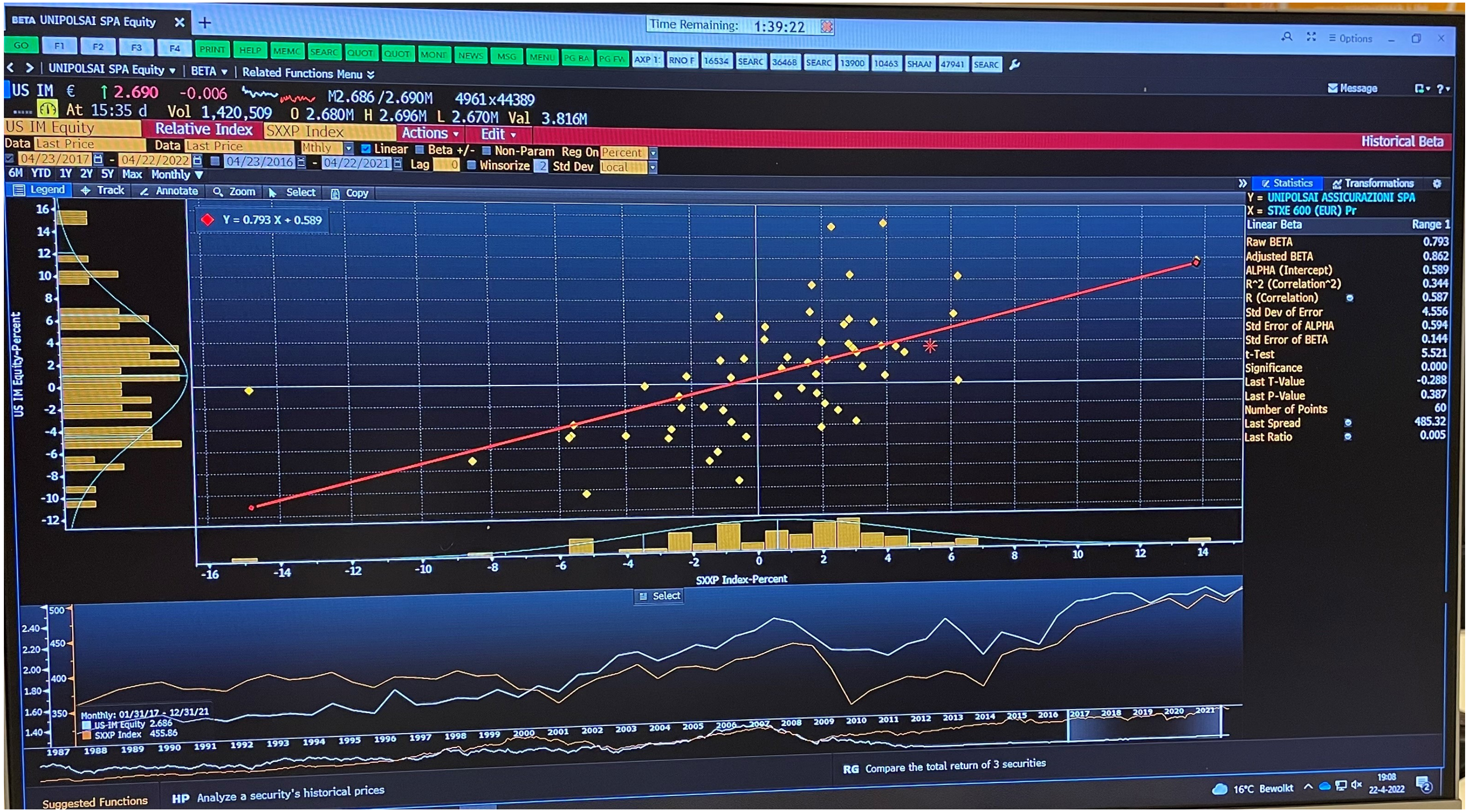
Company	R^2	Actual return	Model implied return
Allianz SE	0.74	5.92	4.06
AXA SA	0.69	9.21	4.61
AG SpA	0.66	5.30	4.57
Sampo PLC	0.64	5.35	2.52
Aviva PLC	0.63	4.57	4.95
Prudential PLC	0.52	7.11	4.31
ZI Group	0.50	3.76	5.06
WW AG	0.49	0.68	3.68
Aegon NV	0.48	7.32	4.61
PhGH PLC	0.44	2.50	3.68
UIG AG	0.37	4.03	3.48
USA SpA	0.34	3.69	2.52
SCA SpA	0.33	8.05	2.52
Tryg A/S	0.28	2.37	3.69
HG PLC	0.15	4.71	3.49
Topdanmark A/S	0.12	1.65	3.62
Chesnara PLC	0.11	2.25	3.69
Admiral Group PLC	0.07	0.69	3.68
PeGH PLC	0.04	2.31	1.96

Table 25: Results of the weighing method

Method	**RMSE**
$k = 3$	5.05
$k = 5$	2.26
$k = 7$	2.31
Weighing method	2.42

Table 26: RMSEs of the third method according to the chosen k and the weighing method

B Example of a systematic risk regression



BETA UNIPOLSAI SPA Equity

UNIPOLSAI SPA Equity | BETA | Related Functions Menu

US IM € 2.690 -0.006 M2.686 /2.690M 4961x44389

At 15:35 d Vol 1,420,509 O 2.680M H 2.696M L 2.670M Val 3.816M

US IM Equity | Relative Index | SXXP Index

Data Last Price | Data Last Price | Mthly | ☑ Linear □ Beta +/- □ Non-Param Reg On Percent

☑ 04/23/2017 - 04/22/2022 □ 04/23/2016 - 04/22/2021 Lag 0 □ Winsorize 2 Std Dev Local

Y = 0.793 X + 0.589

Y = UNIPOLSAI ASSICURAZIONI SPA
X = STXE 600 (EUR) Pr

Linear Beta	Range 1
Raw BETA	0.793
Adjusted BETA	0.862
ALPHA (Intercept)	0.589
R^2 (Correlation^2)	0.344
R (Correlation)	0.587
Std Dev of Error	4.556
Std Error of ALPHA	0.594
Std Error of BETA	0.144
t-Test	5.521
Significance	0.000
Last T-Value	-0.288
Last P-Value	0.387
Number of Points	60
Last Spread	485.32
Last Ratio	0.005

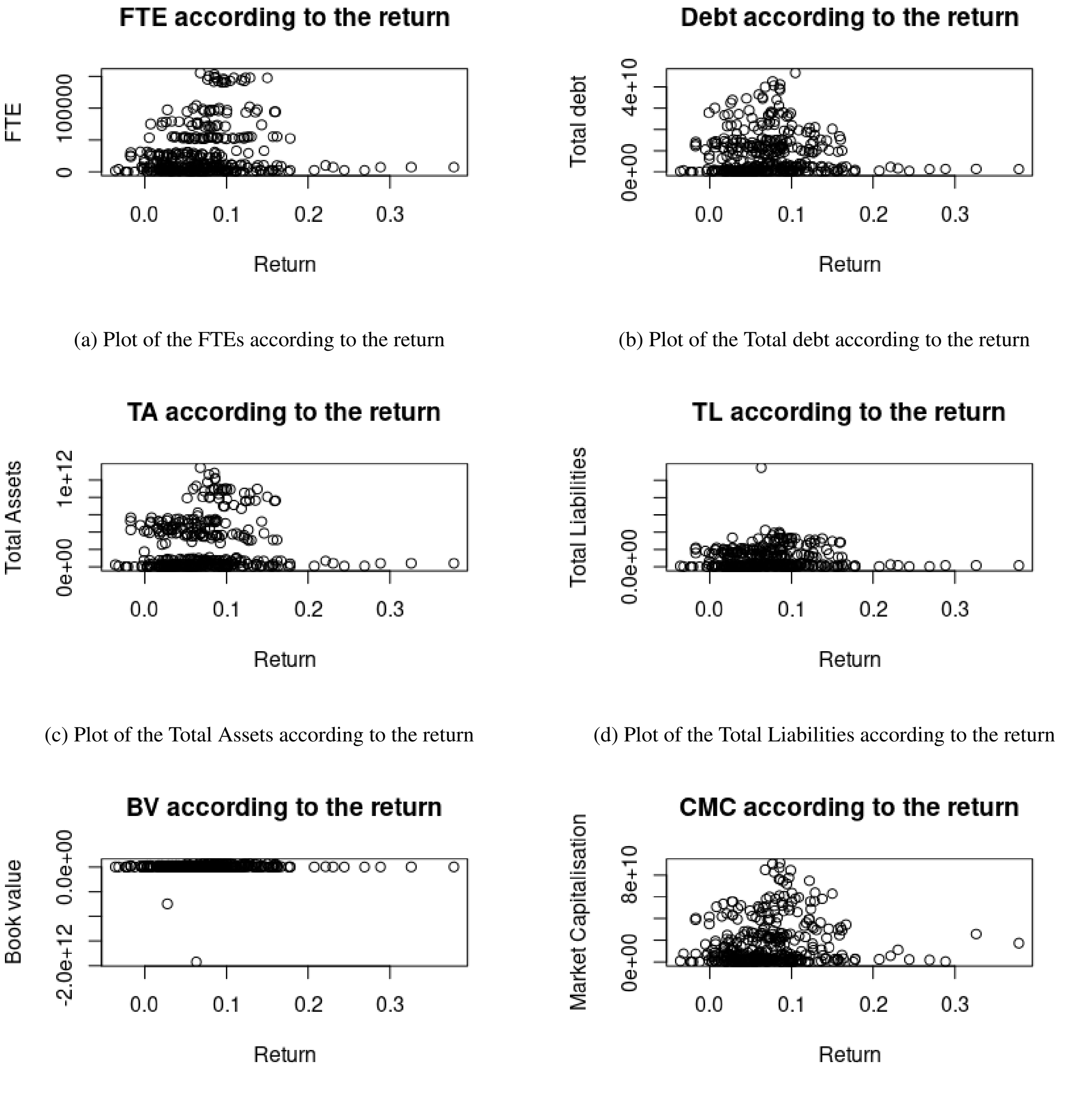
Monthly: 01/31/17 - 12/31/21
US IM Equity 2.686
SXXP Index 455.86

RG Compare the total return of 3 securities

HP Analyze a security's historical prices

Figure 8: Regression over 5 years of the monthly β for the company UnipolSai SpA

C Linearity assumption with outliers



(a) Plot of the FTEs according to the return

(b) Plot of the Total debt according to the return

(c) Plot of the Total Assets according to the return

(d) Plot of the Total Liabilities according to the return

(e) Plot of the Book Value according to the return

(f) Plot of the Market Capitalization according to the return

Figure 9: Plots of the different variables

D Linear regression results using all the variables

Company	R^2	Actual return	Model implied return
Allianz SE	0.74	6.80	-11.30
AXA SA	0.69	10.04	-11.22
AG SpA	0.66	6.47	-10.90
Sampo PLC	0.64	6.46	-10.50
Aviva PLC	0.63	3.32	-10.60
Prudential PLC	0.52	-0.05	-11.00
ZI Group	0.50	6.92	-11.10
WW AG	0.49	2.25	-9.68
Aegon NV	0.48	6.86	-10.40
PhGH PLC	0.44	4.06	-10.50
UIG AG	0.37	4.66	-9.68
USA SpA	0.34	8.71	-10.30
SCA SpA	0.33	5.54	-9.61
Tryg A/S	0.28	1.82	-10.10
HG PLC	0.15	2.71	-7.66
Topdanmark A/S	0.12	5.54	-10.20
Chesnara PLC	0.11	5.64	-9.58
Admiral Group PLC	0.07	2.61	-10.20
PeGH PLC	0.04	5.98	-7.77

Table 27: Results of the linear regression

The results are in %