DEGREE PROJECT IN MATHEMATICS,
SECOND CYCLE, 30 CREDITS
STOCKHOLM, SWEDEN 2017

Cluster analysis of European banking data

FELIX MOLIN

KTH ROYAL INSTITUTE OF TECHNOLOGY
SCHOOL OF ENGINEERING SCIENCES

Cluster analysis of European banking data

FELIX MOLIN

Degree Projects in Financial Mathematics (30 ECTS credits)
Degree Programme in Applied and Computational Mathematics
KTH Royal Institute of Technology year 2017
Supervisors at Finansinspektionen (FI): Gunnar Dahlfors
Supervisor at KTH: Boualem Djehiche
Examiner at KTH: Boualem Djehiche

TRITA-MAT-E 2017:76
ISRN-KTH/MAT/E--17/76--SE

Royal Institute of Technology
School of Engineering Sciences
KTH SCI
SE-100 44 Stockholm, Sweden
URL: www.kth.se/sci

Abstract

Credit institutions constitute a central part of life as it is today and has been doing so for a long time. A fault within the banking system can cause a tremendous amount of damage to individuals as well as countries. A recent and memorable fault is the global financial crisis 2007-2009. It has affected millions of people in different ways ever since it struck. What caused it is a complex issue which cannot be answered easily. But what has been done to prevent something similar to occur once again? How has the business models of the credit institutions changed since the crisis? Cluster analysis is used in this thesis to address these questions. Banking-data were processed with Calinski-Harabasz Criterion and Ward's method and this resulted in two clusters being found. A cluster is a collection of observations that have similar characteristics or business model in this case. The business models that the clusters represents are universal banking with a retail focus and universal banking with a wholesale focus. These business models have been analyzed over time (2007-2016), which revealed that the credit institutions have developed in a healthy direction. Thus, credit institutions were more financially reliable in 2016 compared to 2007. According to trends in the data this development is likely to continue.

Keywords

Business models, modelling, Europe, credit institution, cluster analysis and financial crisis.

Abstrakt

Kreditinstituten utgör en central del av livet som det ser ut idag och har gjort det under en lång tid. Ett fel inom banksystemet kan orsaka enorma skador för individer likväl som länder. Ett nutida och minnesvärt fel är den globala finanskrisen 2007-2009. Den har påverkat millioner människor på olika vis ända sedan den slog till. Vad som orsakade den är en komplex fråga som inte kan besvaras med lätthet. Men vad har gjorts för att förebygga att något liknande händer igen? Hur har affärsmodellerna för kreditinstituten ändrats sedan krisen? Klusteranalys används i denna rapport för att adressera dessa frågor. Bankdata processerades med Calinski-Harabasz Kriteriet and Wards metod och detta resulterade i att två kluster hittades. Ett kluster är en samling observationer med liknande karakteristik eller affärsmodell i detta fall. De affärsmodeller som klustrena representerar är universella banker med retail fokus samt universella banker med wholessale fokus. Dessa affärsmodeller har analyserats över tid, vilket har avslöjat att kretitinsituten har utvecklats i en hälsosam riktning. Kreditinstituten var mer finansiellt pålitliga 2016 jämfört med 2007. Enligt trender i datan så är det troligt att denna utveckling forsätter.

Keywords

Affärsmodeller, modellering, Europa, kreditinstitut, klusteranalys och finanskris.

Acknowledgements

I would like to thank my supervisor Gunnar Dahlfors at the Swedish Supervisory Authority for giving me the chance to do this project, his guidance and for the hours he put in. I would also like to thank my supervisor Boualem Djehiche from KTH for always being eligible. This enabled the work to run smoothly and for that, I am grateful.

Contents

List of Figures

List of Tables

1 Introduction

In this introduction a background is first presented followed by a description of the problem at hand. The purpose to achieve and the goals to accomplish by solving the problem are also described in this section. How the problem was solved is explained in the subsection Methods and some things that were ruled out of the project can be read about in Delimitations. Finally, the Subsection Disposition is giving a smaller overview of the upcoming sections.

1.1 Background

There are thousands of credit institutions (c.i.'s) in Europe today. They are an important cog in the wheel that enables our modern society to function the way it does. Most people have their savings in a bank account, have a mortgage loan on their home and are using some other banking-service. Thus it is imperative to individuals that the c.i.'s are reliable. However, there is more to it than financial security to individuals. To have a sustainable economy in a country, these institutions have to be able to withstand extreme circumstances. If they fail to deal with these circumstances properly, entire countries might have to suffer or even larger areas.

After the latest global financial crisis 2007-2009 the business models of credit institutions have changed (more about the crisis in Subsection 2.5). This is a result of new regulations but it is also self-inflicted. Many institutions were too naive before the crisis and got to pay a high price for that. To prevent such a catastrophic event from repeating itself in the future the institutions have to be well prepared. A way to investigate their preparations is to analyze how the latest financial crisis affected the c.i.s business models and also what measures has been taken to avoid one from happening again.

1.2 Problem

The initial idea for this thesis was to make the most covering research ever made about how the financial crisis has affected European banking business models. To clarify, more observations was going to be used compared to other similar projects. In this thesis, credit market companies would also be accounted for and not just banks. Finding the needed number of observations proved to be a hard task and the outcome of this was that the number of observations got fewer than desired. The character of the observations were still unknown but presumably consisting of mainly large banks. This is nevertheless a very interesting area to investigate and the research is still purposeful. Larger banks naturally hold most of the assets on the banking market. Hence, even though not all business models were covered, much of the financial activities were. The content of the sample might have changed but the main questions to answer still remained; how were the business models affected by the financial crisis and are credit institutions less exposed to threats today than before the crisis?

1.3 Purpose

The purpose of this thesis is to gather information about the effect of the latest global financial crisis. The information is concerning business models of European c.i.'s. This particular study separates itself from other studies by the choice of variables, which can lead to a better understanding of this event. The idea for the project was initially thought of by Gunnar Dahlfors at the Swedish Financial Supervisory Authority (FSA). Therefore, the purpose is also to contribute to FSA's work by meeting their demands.

1.4 Goals

The goal of this thesis is to be able to answer the following questions about European banking business models:

- How many different business models are there?
- What business models are there?
- How has the business models developed over time?
- How has the business models changed since the crisis?
- Are the c.i.s in a more stable position today than in 2007?
- Is it likely that the c.i.'s of today could withstand a financial crisis?

1.5 Methods

This thesis is about clustering yearly data and studying clusters over a time period. When clustering the data a stopping method had to be used first. This method found the number of clusters and nothing more. The stopping number found was used in a clustering method that actually divides the data into clusters or collections of observations. The development of the clusters were analyzed over time, which means that the same clusters had to be found each year. To making sure of this, three different methods were used. When looking at this amount of data it can be hard to see patterns. Therefore figures were produced that simplifies the evaluation of the data. One detailed description of this is given in Section 4.

The data was collected from annual reports with help from the FSA of Sweden. When selecting the variables, a lot of factors had to be put into consideration. This is a difficult process for which Principal Component Analysis could be a helpful tool. A detailed description of everything concerning the data is given in Section 5.

1.6 Delimitations

The clusters are assumed to be making equally risky investments within the different investment areas. To clarify, e.g. all equity instruments are considered equally risky. As the number of observations increase this should be more true. It is hard to say for sure if this is a good assumption our not.

1.7 Disposition

In Section 2, theory concerning the project will be presented. Different business models are described as well as the global financial crisis that was investigated. Further, the Literature overview in Section 3 is showing and discussing facts about studies already made in the same area. After this, the selected methods for each stage in the process are written about in Section 4. How to chose the data is an essential part of this project and can be read about in Section 5. Next, section the results are displayed in Section 6. Lastly the results are being discussed in Section 7 and conclusions are made in Section 8 .

2 Background

In this section, theory is presented. This theory is necessary to understand before reading further.

The project is about business models of credit institutions. But what is a c.i. and what is a business model? Read about this in the upcoming Subsections 2.1-2.3. Before any cluster can be analyzed it is vital to understand the existing business models, see Subsection 2.4. With this knowledge the clusters found can be recognized and categorized. It is also beneficial to look at the financial crisis and try to understand why it happened, see Subsection 2.5.

2.1 What is a credit institution?

The definition of a credit institution is according to the Capital Requirements Regulation (CRR) "an undertaking the business of which is to take deposits or other repayable funds from the public and to grant credits for its own account" European Banking Authority (2017). What this means is basically that a credit institution is a company that is allowed to accept customers deposits and use them on their own initiative.

What is the difference between a bank and a c.i.? A bank actually is a credit institution. What separates the terms is that a c.i. also comprises credit market companies. A credit market company is financial institution controlled by a bit different set of rules than a bank.

2.2 The importance of credit institutions

Credit institutions are vital to society and has been for a long time. They enhance productivity and contributes to economical growth in many fields like agriculture, industry and construction that are highly dependent of loans. These are cases where credit is moved to where it is needed the most, which enable it to grow. Almost every person have savings, housing loans and investments through an institution. This makes it possible for people to move to new places where their competence and skills are needed. The economy grows and development in general. Mismanagement of institutions can lead to inflation, deflation and recession. Credit that is miss-placed stop economic growth and people get unemployed. To keep all of this from happening the institutions are controlled by regulations.

2.3 What is a business model?

Oxford dictionaries explains a business model like this "A plan for the successful operation of a business, identifying sources of revenue, the intended customer base, products, and details of financing.". Thus it is the strategy a company has to be successful in the future. The strategy will hopefully turn out well and secure a sustainable future for the company. There are endless number of business models fitting all kind of markets. In this project only models concerning the banking market will be analyzed and discussed.

A business model can be described in many ways. The interpretation made in this thesis is that a business model only can be illustrated by factors that the company can control. This becomes quiet reasonable when it is thought of. Because how can a strategy be simulated by variables, which are not controllable? A few of the previous studies in Section 3 seems to have failed to see this.

2.4 Known banking business models

In this subsection most common existing business models for credit institutions are being described. To be able to recognize and categorize the clusters found in the results the existing models need to be understood.

The business models:
- Universal banks
- Savings banks
- Retail banks
- Publicly owned banks
- Securities banks
- Wholesale banking

2.4.1 Universal banks

The Financial Times describes the definition of a universal bank as "A universal bank is a financial service conglomerate combining retail, wholesale and investment banking services under one roof and reaping synergies between them" Financial Times (2017). Thus these are banks that have a vast range of services for both households and companies. They usually have a large percent of the market where they are located. The reason why universal banks often are large is because large banks have the economy to sustain many services. Net interest is the main income source.

2.4.2 Savings banks

The Merriam Webster dictionary's definition of a savings bank is "a bank organized to hold funds of individual depositors in interest-bearing accounts and to make long-term investments (as in home mortgage loans)" MW (2017). They are very focused on deposits from customers and loans to customers. They can be run by the government with the purpose to support a local area.

2.4.3 Retail banks

This type of bank that primarily focus is providing services to individual customers rather than to businesses Financial Times (2017).

2.4.4 Publicly owned banks

These are banks owned by the public, hence controlled by the government. The advantage of being publicly owned is that it does not have to pay dividend or compensate any investors. The earnings are collected by the government and can be used to do good in society. The interest of a publicly owned bank does not lay in making a huge profit, it is rather to be an asset to society. The revenue is made on interest and other services like any bank Finansinspektionen (2017).

2.4.5 Securities banks

Securities can be stocks, bonds and options. A security bank is an institution that offers to trade securities for individuals or companies. The main part of their income is generated from net commission. Finansinspektionen (2017).

2.4.6 Wholesale bank

Unlike retail, wholesale banking is focusing on services to larger corporations and other institutions Cambridge dictionary (2017).

2.5 The Global Financial Crisis of 2007-2009

What defines a financial crisis is that financial assets of some sort drop in value rapidly. This can occur in a broad variety of situations and cause different amounts of damage.

Between the years 2007-2009 a global financial crisis took place that was the worst of its kind since the Great depression in 1930. The global financial crisis originated in USA, which is one of the greatest economies in the world. Naturally their economical problems affected a lot of other countries that were making affairs with them. These other countries quickly got economical problems themselves and in turn affected their affair partners as well. A chain-reaction was created and the financial crisis spread to such an extent that it became global United Nations (2011).

Institutions got support from their governments to survive but some did declare bankruptcy. It affected a vast number of people, millions lost their jobs and even ended up on the streets. A crisis of this magnitude does not blow over in a year or two. What followed was the global "Great recession" and "The European dept crisis". The European dept crisis is still not over and is affecting countries using the euro.

Why did this crisis ever take place? This a very complex issue and there was no single event but a series of them that triggered the crisis. A press release from the Financial Crisis Inquiry Commission states the following reasons FCIC (2017). The press release only concerns USA.

- Widespread failures in financial regulation, including the Federal Reserve's failure to stem the tide of toxic mortgages.

- Dramatic breakdowns in corporate governance including too many financial firms acting recklessly and taking on too much risk.

- An explosive mix of excessive borrowing and risk by households and Wall Street that put the financial system on a collision course with crisis.

- Key policy makers ill prepared for the crisis, lacking a full understanding of the financial system they oversaw.

- Systemic breaches in accountability and ethics at all levels. Mortgage-holders took out loans they never intended to pay; lenders made loans they knew the borrowers could not afford.

This report describes a financial system prepped with problems in every corner. The financial institutions acted recklessly and took too much risk. The financial regulations failed to see this which enabled the institutions to proceed. At the same time were the politicians ill prepared for a crisis and could not handle the situation.

3 Literature overview

This part of the report discusses similar projects made. A general guidance and help to draw conclusions is provided by doing this. An idea of what to expect is created, which can eliminate possible mistakes. Facts about the literature can be seen in Table 1, 2 and 3.

3.1 Data

Facts about the data used in the literature is presented and discussed in this subsection.

3.1.1 Time interval

The time frame studied in the literature is for the most part covering 2007-2009. This was the time when the financial crisis took place. The only study that stands out in this matter is Farnè and Vouldis (2017), which is only looking at one quarter of 2014. This means that the researches analyzed contains potentially useful information. See Table 3 for exact information about the time intervals.

3.1.2 Variables

The number of variables and choice of variables differs a lot between the studies. The reasons for this is that different methods were used to process the variables and that the issue is complex. An other factor could be that it can be hard to find the preferred data. Consistent in every study is that the used variables are presented as fractions. This is to deal with the present heteroskedasticity in the data set. The observations are clearly heavily heteroskedastic due to the fact that banks varies greatly in size. In Farnè and Vouldis (2017), 1039 variables are used. This is a huge number. They rely on the constructed program to figure out which ones are worth keeping. The program saved 382 variables and excluded the rest for being redundant. The problem with this strategy is capturing what a business model is. In Ayadi and Groen (2015) 5 variables were more or less picked by hand from the sample of 760 variables and trust were put in their own expertise. Inspections were also made to see if the selection were reasonable. One point they made, which is interesting, is that a business model is something that is chosen. This means that the selected variables should be of the sort that the institutions can regulate. Another way to select variables is done in Cronqvist and Smed (2016) where they initially have selected a set of 19 variables. The variables are processed with Principal Component Analysis which returns 5 principal components. These principal components cover 84% of the characteristics in the data. This is a way to lower the number of variables and eliminate any linear correlation in the set. This procedure excludes much of the noise in the data at the price of some characteristics.

Variables simulating economic activities can be put into the categories; income, investments, assets and liabilities. Some of the studies, Cronqvist and Smed (2016) and Lucas, Schaumburg, and Schwaab (2017), uses variables from all of the categories. Doing this one have to be a bit careful. Income is the result of an investment, which means that one could capture two variables that basically is the same thing. What would happen is that the model captures how well the investments are doing, which is not a part of a business model. A special variable that is used in Cronqvist and Smed (2016) and Ayadi and Groen (2014) is "Tangible assets". This variable is a bit tricky but should probably not be used. It is a measure of the value of the c.i.'s material things. Material things like cars can be rented against a fee, but there are also a lot of material assets that do not generate an income. Hence it is difficult to say how much of these Tangible assets that are investments that will return an income. Perhaps this was the reason why it was not used in Ayadi and Groen (2015).

3.2 Methods

There are two steps that every study have used. The first one is to find out how many clusters there are. The second one is dividing the data into clusters. The methods used can be seen in Table 1.

Project	Stopping method	Clustering method
Cronqvist and Smed (2016)	Calinski-Harabasz Criterion	Wards's method
Ayadi and Groen (2014)	Calinski-Harabasz Criterion	Wards's method
Ayadi and Groen (2015)	Calinski-Harabasz Criterion	Wards's method
Farné and Vouldis (2017)	Calinski-Harabasz Criterion	Wards's method
Lucas, Schaumburg and Schwaab (2017)	Calinski-Harabasz Criterion	Wards's method
Ferstl and Seres (2012)	Calinski-Harabasz Criterion	Wards's method
Hryckiewicz and Kozlowski (2016)	Calinski-Harabasz Criterion	Wards's method

Table 1: Used methods in the projects.

3.2.1 The stopping method

To decide the stopping number three different methods have been used. They can all be seen in Table 1. The most common one is Calinski-Harabasz Criterion, which has been used in five studies. Some of them like Lucas, Schaumburg, and Schwaab (2017) test several methods for reinsurance e.g. Silhouette. In Hryckiewicz and Kozlowski (2016) their experience of the banking sector settled the number of clusters. Ferstl and Seres (2012) used an bootstrapping algorithm.

3.2.2 The clustering method

Four different methods have been used to extract the clusters from the data. They are presented in Table 1. The most common one, the Ward's method, is used in four of the projects. Hryckiewicz and Kozlowski (2016) and Ferstl and Seres (2012) used k-medoid method respective k-centroid method, which are closely related. The idea is that the data-point closest to a cluster's medoid/centroid is added to that cluster. The distance is measured as the Euclidean distance. These methods can be separated by the fact that k-medoid uses the most centered data-point in the cluster (medoid), while k-centroid uses the center point of the clusters (centriod). The last project, Lucas, Schaumburg, and Schwaab (2017), stands out and used a mixture model. This is done assuming that the banks do not change cluster over time. It is a justifiable assumption considering the results from other projects.

3.3 Results

Results from previous studies are discussed in this section.

3.3.1 Number of clusters

The number of clusters found in each project were very much alike. The lowest number of clusters was four and the highest six. All numbers can be seen in Table 3. Why did they not find the same number of clusters? This is due to the fact the they choose different samples, variables, methods and also region to study.

General in all the studies, the number of clusters were constant over time and there were not many institutions that swished between clusters either. There is no case where new clusters appeared over time. These are indicators that the actors on the financial market held on to their business models after the crisis hit.

3.3.2 Found Business models

The business models found are not the same in every studies. This is expected because there is a lot that separates the projects e.g. region investigated, types of institutions used, number of institutions. However, some business models can be linked to one another. They are presented below and gathered in Table 2.

Universal banks
This business model was clearly stated to be found in two studies, Ferstl and Seres (2012) and Lucas, Schaumburg, and Schwaab (2017). Another pair of studies which have got a cluster with mainly large banks are Ayadi and Groen (2014) and Ayadi and Groen (2015). These clusters are called "Investment banks". It is very likely that these clusters are similar to the universal bank clusters found in the before-mentioned studies. In the remaining three projects, only large banks are being observed. Large banks are often universal banks. Thus, the clusters found are special cases of universal banks. E.g. the cluster "Model D" in Ferstl and Seres (2012) which contains universal banks with strong retail.

Savings banks
This type of banks are representing a cluster in Cronqvist and Smed (2016) and in Farnè and Vouldis (2017). Even if they are called traditional commercial banks in the last mentioned paper.

Retail banks
A retail cluster is clearly stated to be found in Ayadi and Groen (2014), Ayadi and Groen (2015) and Lucas, Schaumburg, and Schwaab (2017).

Wholesale banks
Wholesale is a found cluster in Ayadi and Groen (2014), Ayadi and Groen (2015) and Farnè and Vouldis (2017).

Cronqvist and Smed (2016)	Universal banks	Savings banks	Leasing companies	Service focused credit institutions	Non-deposit funded credit institutions	Other credit institutions
Ayadi and Groen (2014)	Investment banks	Diversified retail	Focused retail banks	Wholesale banks	-	-
Ayadi and Groen (2015)	Investment	Wholesale	Focused retail	Diversified retail 1	Diversified retail 2	-
Farné and Vouldis (2017)	Securities holdings banks	Traditional banks	Complex commercial banks	Wholesale	-	-
Lucas, Schaumburg and Schwaab (2017)	Large universal banks	Fee-focused banks	Domestic diversified lenders	International diversified lenders	Domestic retail lenders	Small international banks
Ferstl and Seres (2012)	Model A	Model B	Model C	Model D	Model E	-
Hryckiewicz and Kozlowski (2016)	Investment	Trader	Specialized	Diversified	-	-

Table 2: Found business models.

3.4 Factual summary

Project	Number of c.i.s	Area researched	Time investigated	Number of variables	Clusters found
Cronqvist and Smed (2016)	165	Sweden	2000-2013	19	6
Ayadi and Groen (2014)	145	EEA + subsidaries in non-EEA countries	2006-2013	6	4
Ayadi and Groen (2015)	2542	EEA + Switzerland	2005-2014	5	5
Farné and Vouldis (2017)	365 (banks)	Euro area	Last quarter of 2014	1039	4
Lucas, Schaumburg and Schwaab (2017)	208 (banks)	Europe	2008-2015	13	6
Ferstl and Seres (2012)	234 (banks)	Europe	2005-2011	5	5
Hryckiewicz and Kozlowski (2016)	458 (large banks)	65 World countries	2000-2012	7	4

Table 3: Collection of facts about the studies.

3.5 What can be expected?

The results gotten in this thesis should be pretty similar to the ones presented here in the studies. At least compared to the studies that investigates large banks.

3.5.1 Choice of methods

The variables that are chosen are going to be fractions to loose the heteroskedastisity. They are going to be based on investments, income, assets, liabilities and perhaps risk in some way.

The best way to find the clusters is reasonably to use Calinski-Harabasz Criterion combined with Ward's method. This is reasonable due to the fact that it is commonly used and with good results.

3.5.2 Predicted results

From the results mentioned above conclusions about what to expect can be drawn. The number of clusters is probably going to between 4-6, perhaps lower due to the sample of observations used in this thesis. Business models that are likely to be found are versions of Universal banks.

4 Methods

The methods used to divide the credit institutions into clusters was a stopping method and a clustering method. They both serve an essential purpose in the analysis and are written about in Subsection 4.1. This is a research concerning the change of specific models over time. Subsection 4.2 tell how it is verified that the same model is found each year. In Subsection 4.3 it is described how the analysis of the clusters was done.

4.1 Methods used to cluster the data

The stopping method is used to find the number of clusters existing and the clustering method divides the data into clusters. To get a better understanding of the mathematical aspects see the Mathematical background in Section 10.

4.1.1 Stopping method

To find the number of clusters in the data a Stopping method was used. The method selected was Calinski and Harabasz criterion due to the fact that almost every study in the Literature overview used it with good results. To be able to compare the results and confirm its validity, Davies-Bouldin index and Silhouette were also used. These are all the methods available in Matlab.

4.1.2 Clustering method

It was possible to divide the observations (credit institutions) into any desired number of clusters. The desired number in this case was the found stopping number. Ward's method is selected as clustering method. The combination between Calinski-Harabasz Criterion and Ward's method is well known and has been used frequently in related studies. To back up the results k-means method is also computed.

4.2 Correlation between found clusters

To ensure that the business model found in some year is considerably related to the business model found in adjacent years, three measures were taken. The distribution among the clusters were looked at (Table 7), how many c.i.s that stay in the same cluster each year (Table 8) were analyzed and the data was also visually evaluated (Table 9).

4.3 Analyzing the clusters over time

This part is about the methods used to analyze the development and changes within the clusters from 2007 until 2016. The first step in the process was to find the average c.i. for each cluster each year. It was done by calculating the average of the variables. These found average c.i.s represent a typical c.i.s to each business models each year. The development and change over time is now easy to find. This information is presented in three different ways; pie charts, plots and tables. The pie charts are presenting data from 2007 and 2016, which returns a visually clear picture of the total changes. The plots are showing the yearly development from 2007 to 2016. The tables are presenting data from 2007 and 2016 and also the total development. Given all this information conclusions can be made.

5 Data

This section is concerning the data, the region it covers, how the variables were selected, reparation of missing values and also information about left out variables and institutions.

5.1 Region investigated and type of c.i.

The data collected comes from 169 randomly picked European credit institutions. Of this sample only 66 c.i.s were usable due to bad coverage. Larger banks are more likely to have better coverage than smaller ones, which means that the sample in this thesis probably consists of mostly large banks.

Investigating an area this big means that a business model might not perform equally well depending on location. Thus using any type of income as a variable might be a problem.

5.2 Process of selecting variables

When selecting variables the first thing to do is to pin point exactly what is meant to be achieved. The goal was to simulate business models of credit institutions. So, what is a business model? The conclusion in Subsection 2.3 tell us that it is a strategy to earn money. A strategy is something that is controlled and selected carefully by the companies. Thus it is reasonable to say that the variables selected for this project should be of the kind that the institutions can control themselves. There are of course other ways to address this, like letting a program decide which variables that are most significant. In Farnè and Vouldis (2017) over 1000 variables were used from the start. The most significant are kept by the program. The problem with this solution is that you will get a result that is based on variables that might not have anything to do with the chosen business models. The path that was chosen in this project was selected after performance. Results were gotten with and without PCA. Not using PCA proved to have an slight edge in the sense that the clusters were more stable.

The chosen variables should cover as much as possible of the c.i.s financial activities. Hence there need to be an understanding of how they get their income and where the money invested came from initially. The income can basically be divided into three sections; net interest, net finance and provision. Net interest is classical banking and usually makes up the biggest part of the income. Because net interest is the biggest income source it is a important that the chosen variables represent this part well. Net finance is smaller and is income from investment into securities. Provision is money gotten from doing services to customers.

There is also a factor of risk involved in banking businesses models. This is important to take into account because of the nature of this thesis. The investments made and liabilities gotten can give some information about level of riskiness in the models. Investments can differ a lot in riskiness just by looking at e.g. different stocks. The data gotten is not granular enough for us to capture this but there are other ways. Some riskiness gets captured by looking at the amount of stocks and different types of given loans. Loans given to different sources differ in risk. Received deposits also say something about risk because a higher cost for a loan will make it harder to return a profit from the investment. The amount of liabilities provide information about how stable the firm is and how resilient it is against potential losses.

Coverage have a great impact on the results, some missing values can change them completely. When selecting which final variables to use and observations to keep this has to be taken into account. If a variable have bad coverage it cannot be used and the same applies to the observations. Data can sometimes be repaired. E.g. if some observation is missing one data-point, one might consider trying to repair this.

5.3 Selected variables

Like so often when dealing with data it can be hard to get exactly what is sought. Then some sort of compromise had to be made.

The chosen variables are largely covering the financial activities of c.i.s. They are displayed in Table 4 along with some discarded variables. A description of the selected variables is presented below this table. They are all presented as fractions to return an understanding of how the cash is divided between investments, where income is generated and also a sense of how risky the business models are.

Selected variables	Discarded variables
Net fee and commission income/Operating income	Total dept/Total assets
Retail loans/Net loans to customer	Deposits from customers/Total deposits
Corporate loans/Net loans to customers	Deposits from corporations/Total deposits
Total equity instruments/Total assets	Net interest income/Operating income
Total dept instruments/Total assets	Trading income/Operating income
Total sub-oriented dept/Total assets	-
Net loans to customer/Total assets	-
Total liabilities/Total assets	-
Total deposits from customers/Total assets	-

Table 4: Selected variables and some discarded variables

- Net fee and commission income/Operating income:
The income generated from providing different services to customers.

- Retail loans/Net loans to customer:
Loans given to individuals.

- Corporate loans/Net loans to customers:
Loans given to corporations.

- Total equity instruments/Total assets:
Value of ownership in other firms (value of stocks owned).

- Total dept instruments/Total assets:
A type of loan given by the institutions. It is possible to transfer the ownership of the contract. They are traded frequently.

- Total sub-oriented dept/Total assets:
A type of loan or security that is riskier than a retail loan or corporate loan. The higher risk is created due to the fact that if the borrower defaults it has a lower rank than other loans. The extra risk give rise to a higher return.

- Net loans to customer/Total assets:
A summation of retail loans and corporate loans given to customers.

- Total liabilities/Total assets:
The total liabilities that a c.i has.

- Total deposits from customers/Total assets:
Assets that comes from deposits from customers.

5.4 Reparation of missing values

The reparation took place where there were missing values. To fill the gaps the mean of the two adjacent data-points were used. Thus it was only done when one values was missing and not two in a row. This means that missing values at the edges of the interval years 2007 and 2016 was not repaired and the c.i. had to be excluded instead.

5.5 Outliers and left out institutions

Variables that are left out can be seen in Table 4. There are two reasons why these variables where left out. The first one is data coverage which is concerning Retail deposits, Corporate deposits and Total dept. Trading income and Net interest income were discarded because the information they possessed were covered by other variables.

6 Results

In this section results are presented and discussed to some degree.

6.1 How many clusters were found?

Two clusters where found each year using the Calinski-Harabasz Criterion and the Silhouette. Davies-Bouldin index on the other hand returned spread answers and do not seem to work well with the used data set. The results are presented in Table 5 below.

Year	Calinski-Harabasz Criterion	Davies Bouldin Index	Silhouette
2007	2	10	2
2008	2	10	2
2009	2	6	2
2010	2	9	2
2011	2	7	2
2012	2	7	2
2013	2	10	2
2014	2	8	2
2015	2	10	2
2016	2	7	2

Table 5: The number of clusters found.

6.2 The average c.i. of each cluster

Table 6 is presenting the average c.i.s found in year 2007. This presentation of the clusters is used to easily display how the c.i.s characteristics look like. Because the average clusters are gotten by finding the average of each variable, Corporate loans + Retail loans does not equal exactly one.

Variables	Net fee and commission income	Corporate loans	Retail loans	Total equity instruments	Total dept instruments	Total suboriented dept	Total liabilities	Total deposits from customers	Net loans to customers
Cluster 1	0.2760	0.3138	0.6837	0.0383	0.0694	0.0151	0.9072	0.4024	0.7731
Cluster 2	0.2206	0.6527	0.3746	0.0225	0.1323	0.0220	0.9274	0.5257	0.6187

Table 6: Average c.i. of cluster 1 and 2 from the year 2007.

6.3 Is the same business models found each year?

To find out if the same business models are found each year, three methods are being used. Each one can offer some confirmation about the issue.

6.3.1 Distribution between the clusters

In Table 7 it can be seen that the number of c.i.s is almost constant trough the years. This is an indicator that the dynamic in the data is almost the same every year. The same characteristics dominate.

Year	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
Cluster 1	0.561	0.545	0.545	0.561	0.561	0.576	0.561	0.576	0.636	0.636
Cluster 2	0.439	0.455	0.455	0.439	0.439	0.424	0.439	0.424	0.364	0.364

Table 7: Distribution of c.i.s between the two clusters.

6.3.2 C.i.s changing cluster

This section describes the data in Table 8, which display how many c.i.s that changed cluster during a year. The overall change is also presented. The years 2008-2009 and 2014-2015 experienced the biggest changes of 6.06%. 2008-2009 is the last year after the financial crisis so it is natural with some change in character. The reason for the change in 2014-2015 is unknown. The over all change is only 10.61%. The found clusters consist of pretty much the same c.i.s. every year. The conclusion that can be made is that there is a strong correlation between the business models found each year.

Year	07-08	08-09	09-10	10-11	11-12	12-13	13-14	14-15	15-16	07-16
Change	0.0152	0.0606	0.0455	0.0303	0.0152	0.0455	0.0455	0.0606	0	0.1061

Table 8: Data on how much c.i.s are changing cluster.

6.3.3 Visual evaluation

What can visually be evaluated is a that there is a strong connection between the clusters found each year. Table 9 presents the average c.i. from each cluster for year 2007 and year 2008.

Variables	Net fee and commission income	Corporate loans	Retail loans	Total equity instruments	Total dept instruments	Total suboriented dept	Total liabilities	Total deposits from customers	Net loans to customers
Cluster 1 (2007)	0.2760	0.3138	0.6837	0.0383	0.0694	0.0151	0.9072	0.4024	0.7731
Cluster 2 (2007)	0.2206	0.6527	0.3746	0.0225	0.1323	0.0220	0.9274	0.5257	0.6187
Cluster 1 (2008)	0.2654	0.3194	0.6804	0.0235	0.0730	0.0155	0.9241	0.3906	0.7833
Cluster 2 (2008)	0.2299	0.6548	0.3633	0.0134	0.1276	0.0206	0.9340	0.5280	0.6237

Table 9: Data on the average c.i from years 2007 and 2008

6.4 Which business models are identified?

Looking at Table 9 it is understood that the institutions have multiple income sources. Banks like these are usually called universal banks. This statement is strengthened by the knowledge that the sample is likely to consist of large banks. What separates the two clusters is given loans. Cluster 1 is investing more into Retail loans, while cluster 2 is investing more into Corporate loans. Hence, cluster 1 is consisting of retail oriented universal banks and cluster 2 is consisting of corporate/wholesale oriented universal banks. An observation that shall be made is that both models are heavily investing into loans but cluster 1 distinguishes itself by investing 77% of its assets compared to the 62% that cluster 2 invests.

6.5 Changes within the clusters

In this subsection pie charts, plots and tables are going to be reviewed. The figures are displaying the change within the clusters over time.

Figure 1 is showing how the assets are used. Just like it was noticed earlier, investment into loans dominate. Other is a category consisting of assets not covered by the chosen variables. It is likely to mainly consist of cash and tangible assets.

Cluster 1: Looking at this figure, not much has happened the last 10 years. A bit more is invested into Total dept instruments and less into Suboriented dept.

Cluster 2: More changes are experienced here. 8% less is invested into Net loans to customers. Instead investments have been made into Total dept instrument and Other. This means that Cash + Tangible assets has increased.

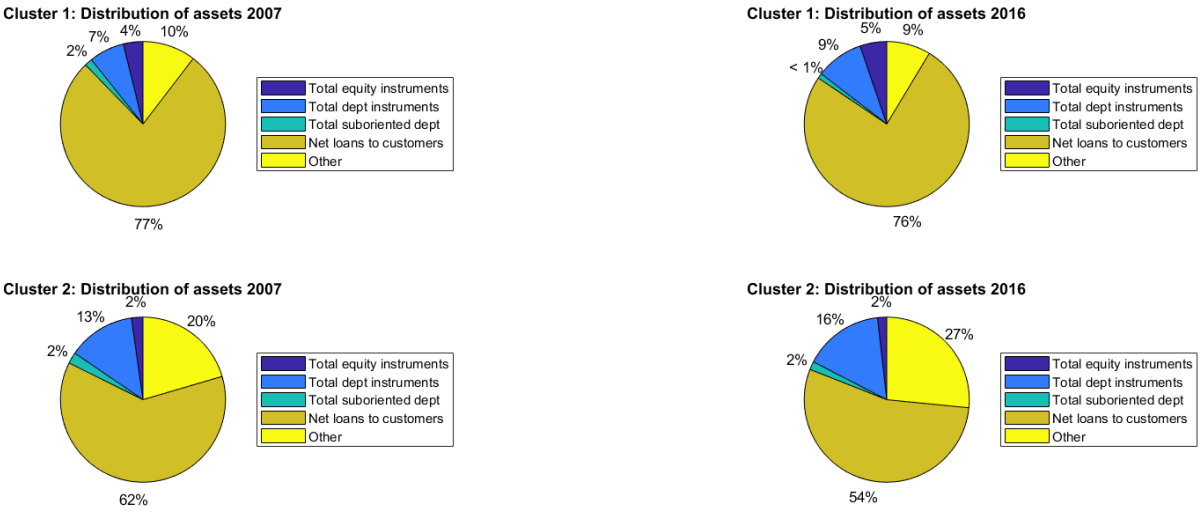

Figure 1: Investments done in the years 2007 and 2016.

The distribution of income from the years 2007 and 2016 can be seen in Figure 2. These are the results of the investments. Keep in mind that only Net fee and commission was used to cluster the data. Other is an unknown income.

Cluster 1: Net interest income is increasing. This is interesting considering that almost no changes were made in Net loans to customers, Total dept instruments or Total suboriented dept. Net fee and commission income is decreasing some.

Cluster 2: The changes made is actually very similar to the ones made in cluster 1. It is even more surprising how much the Net interest income has gone up in this case due to the cut in investments into this area presented in Figure 1. Net fee and commission income is increasing a bit.

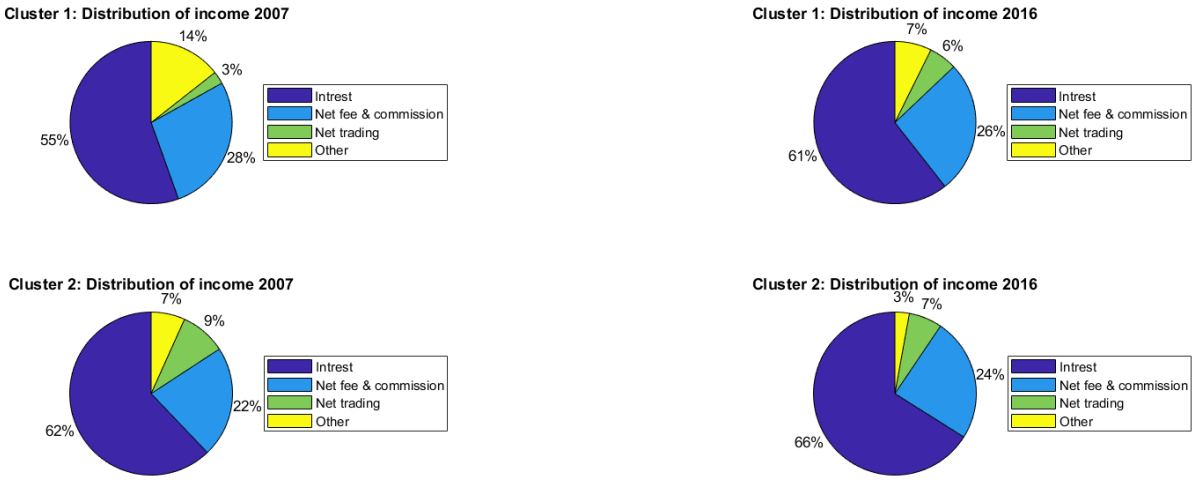

Figure 2: Distribution of income in the years 2007 and 2016.

To get an better understanding of why the Net interest income has increased, Figure 3 was produced. It displays how the given loans are divided between retail and corporate loans.

Cluster 1: Invests a bit more into retail after than before the crisis.

Cluster 2: The change is similar to the one seen in cluster 1, just larger (8%).

To summarize, the change in investments into loans is likely to have caused the increase in Net interest income.

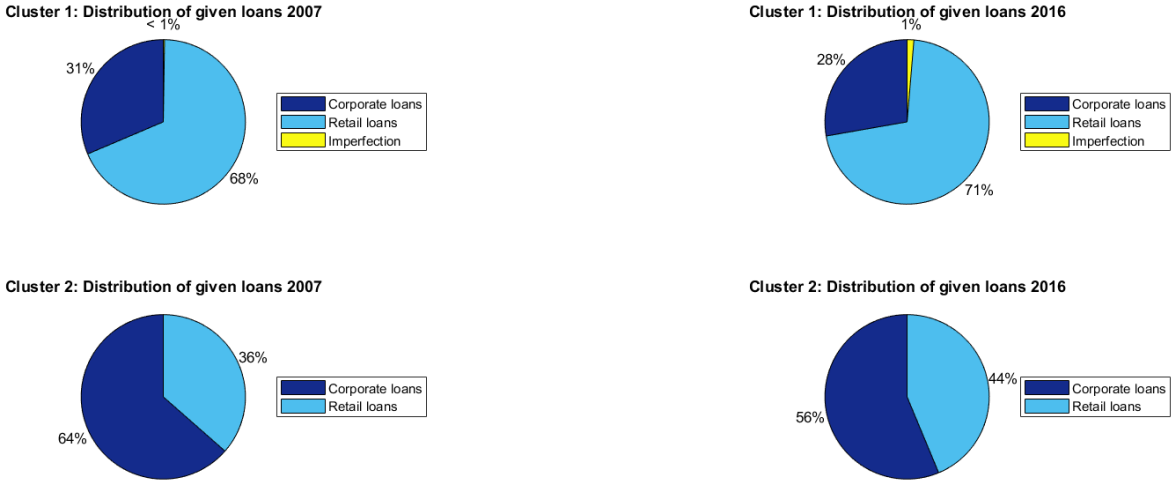

Figure 3: Given loans in the years 2007 and 2016.

The following plots are showing the yearly development since 2007 to 2016. This will enable us to see patterns and say if some trends are likely to continue in the following years. Coming up first is Figure 4 that can tell how the investments got affected by the crisis.

Cluster 1: It is noticed that Total equity instruments where heavily affected 2008 and experienced an extreme downfall. The possessed equity seems to have dropped in value quickly, which is strengthened by the loss made in Trading income in Figure 5. Total dept instrument increased soon after the crisis and somewhat stayed in this position until 2016. This trend is likely to proceed. Concerning the other investments there are no clear abrupt changes, they are more slowly developing over time. Suboriented dept show no sign of stopping its trend of declining and is probably going to keep doing this in the future. Net loans to customers on the other hand have stopped its trend of slightly decreasing and might even rise some in the years to come.

Cluster 2: This cluster was affected in a very similar manner compared to cluster 1. The differences are that Total equity instruments are seeing an up-going trend the last year and that Net loans to customers has experienced a down-going trend.

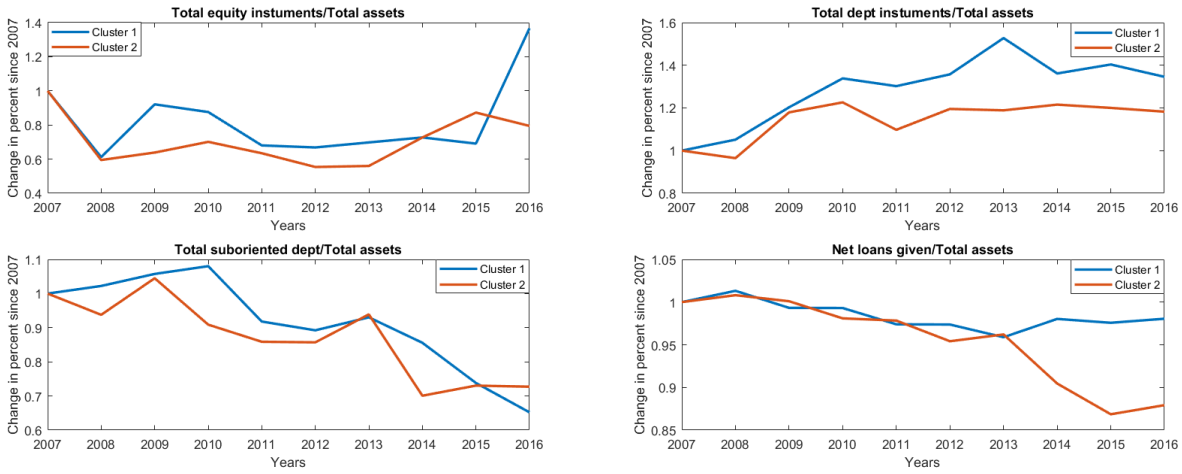

Figure 4: Development of financial instruments from year 2007 to 2016.

In Figure 5 the development of the income is presented.

Cluster 1: The income fluctuate a lot. Net interest is tending to become more stable with time. It is also noticed that a massive loss is made in trading in 2008. This is in line with previous discoveries about Total equity instruments.

Cluster 2: Net interest income is stabilizing in this case as well. The loss in trading is not as big in this case but they are still loosing money. Hence, there ought to have been a difference in riskiness of the Total equity instruments possessed by Cluster 1 and Cluster 2 at that time. Net fee and commission income is experiencing an up-going trend that probably will continue the following years.

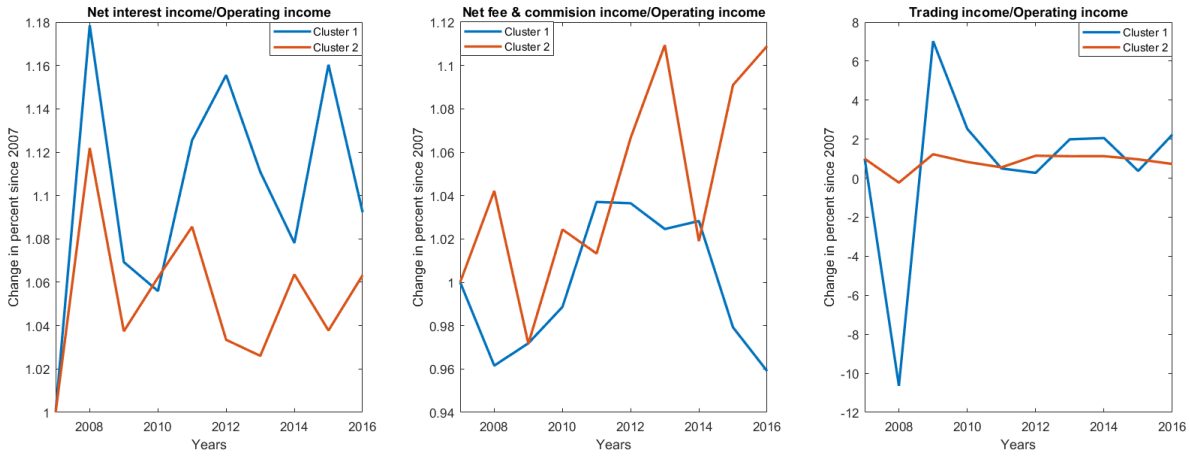

Figure 5: Income development from year 2007 to 2016.

Figure 6 is describing the development of loans given.

Cluster 1: The development is steady and do not show any signs of stopping.

Cluster 2: The same conclusion can be made in this case even though the development is a bit more aggressive.

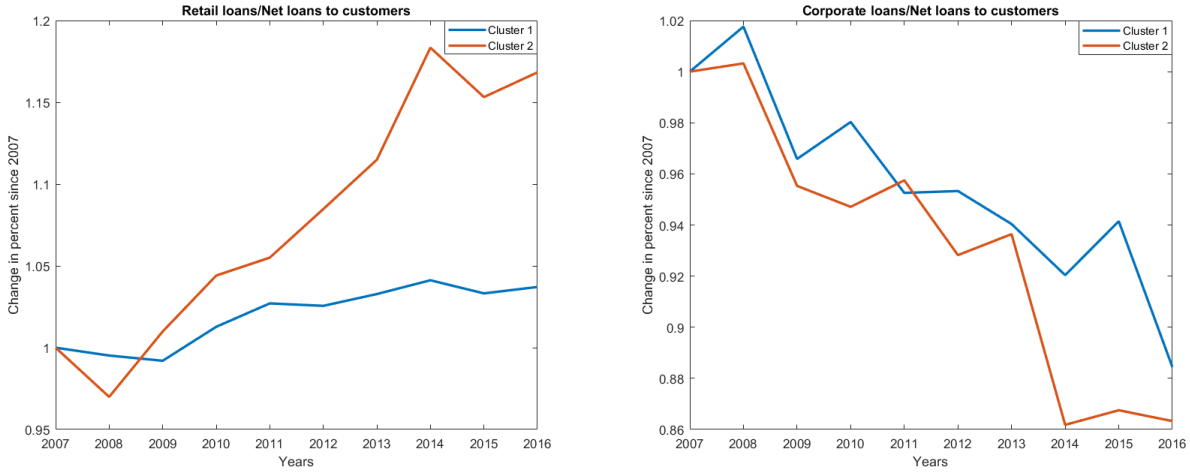

Figure 6: Development of given loans from the year 2007 to 2016.

Figure 7 is presenting the two remaining variables, Total liabilities and Total deposits from customers.

Cluster 1: The Total liabilities peaked in 2008 but has ever since then been going down slowly and will probably keep on doing this. Total deposits from customers have been growing steadily and there is now sign that this development will slow down.

Cluster 2: The same conclusions can be made for this cluster even though there have been a bit more fluctuations.

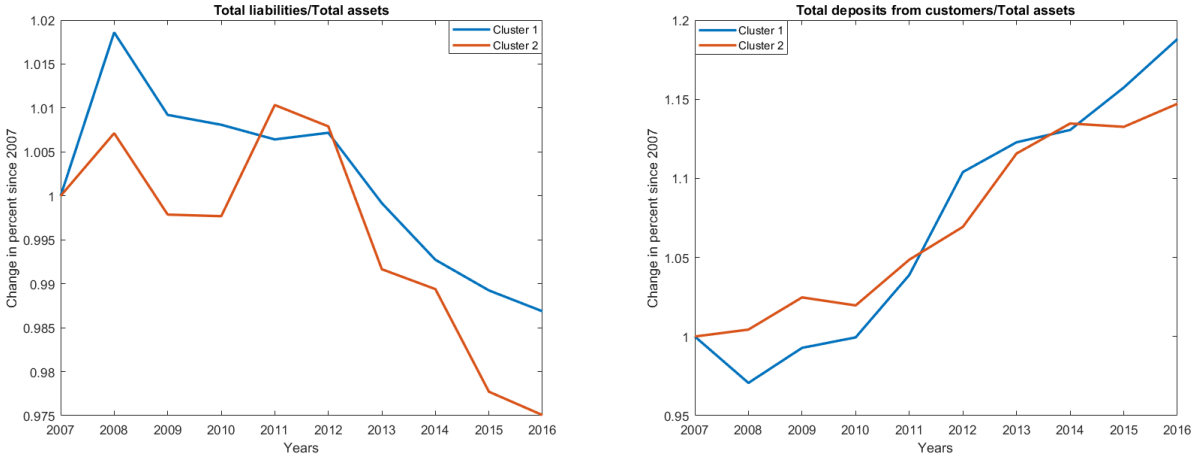

Figure 7: Development of liabilities and deposits from customers from the year 2007 to 2016.

To summarize what has happened between 2007-2016 data is presented in Table 10 and Table 11. This way is enabling us to see the all the development at the same time and very accurately. In Table 10 one can see the overall development och in Table 11 data from 2007 and 2016 is presented.

Variables	Net fee and commission income	Corporate loans	Retail loans	Total equity instruments	Total dept instruments	Total suboriented dept	Total liabilities	Total deposits from customers	Net loans to customers
Cluster 1	0.9590	0.8844	1.0371	1.3666	1.3459	0.6521	0.9869	1.1880	0.9805
Cluster 2	1.1089	0.8633	1.1683	0.7947	1.1824	0.7272	0.9751	1.1471	0.8792

Table 10: Data on the development from 2007 to 2016.

Variables	Net fee and commission income	Corporate loans	Retail loans	Total equity instruments	Total dept instruments	Total suboriented dept	Total liabilities	Total deposits from customers	Net loans to customers
Cluster 1 (2007)	0.2760	0.3138	0.6837	0.0383	0.0694	0.0151	0.9072	0.4024	0.7731
Cluster 1 (2016)	0.2647	0.2776	0.7091	0.0524	0.0935	0.0099	0.8953	0.4780	0.7580
Cluster 2 (2007)	0.2206	0.6527	0.3746	0.0225	0.1323	0.0220	0.9274	0.5257	0.6187
Cluster 2 (2016)	0.2447	0.5635	0.4376	0.0179	0.1564	0.0160	0.9043	0.6031	0.5440

Table 11: Data on the average c.i for years 2007 and 2016

7 Discussion

In this section a discussion is held concerning direct affects of the financial crisis, what the development have been and how the business models are expected to look like in the near future.

7.1 Direct affects of the financial crisis

This is an interesting topic to discuss in order to understand why changes were made in the business models.

Cluster 1: Looking at the results the obvious direct affect on the c.i.s was that Total equity instruments dropped in value very quickly. The results from this was that huge loss were made in Trading income. Thus the c.i.s took a massive hit when the crisis struck, which is also shown by the fact that Total liabilities peaked in 2008. The Net interest income was also peaking in 2008 and at the same time was the Net loans to customers pretty much constant compared to 2007. By the look of this were the interest rates increased approximately 20%. This was an act made to cover up the losses from Total Equity investments. This cannot be said for sure but it seems reasonable when looking at the results.

Cluster 2: What has happened to this cluster is very much alike what has happened to cluster 1. The value of Total equity instruments dropped and a loss in Trading income was made, smaller than the loss made by cluster 1. It is unclear why this loss where smaller in this case. The loss made resulted in a small peak in Total liabilities. To make up for this loss the interest rates were increased by approximately 10%. This is in line with the rise in Net interest income.

7.2 Total change in business model

The changes of the business models have not been too dramatic in any case.

Cluster 1: Looking at this cluster before and after the crisis it is seen that, in general, the characteristics are preserved. In year 2007 it was recognized that this business model could be called universal banks with a retail orientation. In 2016 this business model is even more retail oriented, investing almost 70% of Net loans to customers into retail. This is an increase of almost 4%, which is a lot considering how high the initial value was. The Net loan to customers has gone down a bit. Thus this positive development in Retail loans is boosted by the decrease of investments into Corporate loans. The banks have invested more into Total equity instruments and Total dept instruments and has decreased the Total liabilities. Total suboriented dept has dropped by almost 35% but this is only a drop of 0.5% in terms of Total assets. Thus this is not something that affects the overall business model much.

Cluster 2: This business model has generally experienced more change than cluster 1. As mentioned before this is the cluster containing universal banks with a corporate orientation. With this development this business model will have more focus on Retail loans than Corporate loans in a few years. Net loans to customers has seen an about 12% negative change since 2007. This money is instead used to get the Total liabilities down and also to increase Total dept instruments.

7.3 Expectation for the future

To get an reasonable assumption of how the future will turn out the trends in the plots have to be analyzed. This can reveal if the development of a variable is likely to proceed.

Cluster 1: What can be expected, investments-wise, is that Total dept instruments is staying at

this level, Total suboriented dept will keep on dropping and that Net loans to customers either stay at this level or rise some. Retail loans is going to keep on growing slowly while Corporate loans given drops. The Total liabilities will decline slowly and at the same time, will the Deposits from customers increase.

Cluster 2: Equity instruments is likely to stay about where it is, which is also true for Total dept instruments. Total suboriented dept and Net loans given on the other hand is going to keep on dropping. The development of Retail loans and Corporate loans will keep on going but in a slower rate. The Total liabilities tend to keep on shrinking, while Total deposits from customers increase.

8 Conclusion

Both Cluster 1 and 2 has been affected in a pretty similar fashion by the financial crisis. When the crisis struck their Total equity instruments dropped massively in value and they had to increase the interest rate to stay afloat. This has forced a development against more stable business models because something like this cannot happen again. The Total liabilities has gone down in both clusters, which result in more financially dependable institutions. Cluster 2 has changed more in this sense but were in a more exposed position in 2007 than cluster 1, hence it is natural. The Net interest income is getting more stable in both cases, which this is an indicator that the c.i.s are getting more under control and are approaching better positions. It is always a good sign for any business that the income is stable and secure. Thus, the business models have generally developed to become more stable, especially cluster 2.

Net loans to customers has gone down 2% in cluster 1 and 12% in cluster 2. At the same time the Net interest income has gotten a more significant role in both cases. This is possible because of the changes that has been made with given loans and the increased investments into Total dept instruments.

The changes that have been seen within the clusters during the investigated time period is for the most part going to endure for the years to come. The development will slow down and the c.i.s will become more stable. This is surely a healthy development that I am glad to present. Hence from what has been found in this project the c.i.s are more stable today than in 2007, which means that we are more likely to avoid a new financial crisis or at least be able to scale it down. It is very hard to tell if a global financial crisis can occur one more time but probably it can just because of the complexity of the financial system. There is always going to be loopholes and there is going always to be people that take advantage of them to make a quick buck.

8.1 Future studies

A potential improvement to this particular study is to develop the part of the results that analyzed trends. One could use mathematical methods to predict the outcome better. An idea for an other thesis could be to make a similar project but with a focus on some investment, like equity instruments our loans given. It would be interesting to see how the riskiness of the equity instruments has changed since before the financial crisis.

9 References

D. Arthur and S. Vassilvitskii. *The Advantages of Careful Seeding.* SODA '07: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms., 2007.

R. Ayadi and W. P. D. Groen. *Banking Business Models Monitor 2014.* Centre for European Policy Studies and International Observatory on Financial Services Cooperatives, 2014.

R. Ayadi and W. P. D. Groen. *Banking Business Models Monitor 2015.* Alphonse and Dorimène Desjardins International Institute for Cooperatives and International Research Centre on Cooperative Finance, 2015.

T. Calinski and J. Harabasz. *A dendrite method for cluster analysis vol.3.* Academy of agriculture, 1974.

Cambridge dictionary. Wholesale bank, 2017. URL `https://dictionary.cambridge.org/dictionary/english/wholesale-bank`.

E. Cronqvist and F. Smed. *Business Models in the Swedish Banking Market.* Not published, 2016.

D. L. Davies and D. Bouldin. *A Cluster Separation Measure.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 1979.

European Banking Authority. Credit institution, 2017. URL `http://www.eba.europa.eu/risk-analysis-and-data/credit-institutions-register`.

M. Farnè and A. Vouldis. *Business models of the banks in the euro area.* European Central Bank, 2017.

FCIC. Conclusions of the financial crisis, 2017. URL `http://www.fcic.gov/report/conclusions`.

R. Ferstl and D. Seres. Clustering austrian banks' business models and peer groups in the european banking sector. 2012.

Financial Times. Definition of a retail bank, 2017. URL `http://lexicon.ft.com/Term?term=retail-bank`.

Financial Times. Definition of a universal bank, 2017. URL `http://lexicon.ft.com/Term?term=universal-bank`.

Finansinspektionen. Tillsynen över bankerna. 2017.

A. Hryckiewicz and L. Kozlowski. *Banking business models and the nature of financial crisis.* Elsevier Ltd, 2016.

I. Jolliffe. *Principal Component Analysis (2ed).* Springer, 2002.

A. Lucas, J. Schaumburg, and B. Schwaab. *Bank business models at zero interest rates.* European Central Bank, 2017.

Matlab. Ward's method, 2017. URL `http://se.mathworks.com/help/stats/linkage.html`.

MW. Definition of a savings bank, 2017. URL `https://www.merriam-webster.com/dictionary/savings%20bank`.

A. C. Rencher and W. F. Christensen. *Methods of Multivariate Analysis.* Wiley, 2012.

P. J. Rousseeuw. *Silhouettes: a graphical aid to the interpretation and validation of cluster analysis.* ISES, 1987.

United Nations. *The global economic crisis: causes and transmission*. 2011. URL `http://www.un.org/esa/socdev/rwss/docs/2011/chapter1.pdf`.

J. H. J. Ward. *Hierarchical Grouping to Optimize an Objective Function*. Journal of the American Statistical Association, 1963.

10 Mathematical background

10.1 Principal Component Analysis

The information presented in this part of the thesis is inspired by Rencher and Christensen (2012).

This method can be used to convert a set of possibly correlated indicators to a smaller set of linearly uncorrelated indicators. Only the most significant indicators remain after this procedure is applied. The principal components of a data set can be found by following a few steps.

Step 1 - Gather data
Gather data and present it in a matrix $\boldsymbol{X}$, the columns represent variables and each row an observation.

Step 2 - Calculate the sample covariance matrix $\hat{\boldsymbol{\Sigma}}$ of $\boldsymbol{X}$
The components in $\hat{\boldsymbol{\Sigma}}$ is calculated with Equation 1. $\bar{\boldsymbol{X}}_{\boldsymbol{j}}$ is the mean of the variables in $\boldsymbol{X}$, thus mean of the columns. The mean is calculated with the law of large numbers which is shown in Equation 3.

$$\hat{\Sigma}_{jk} \approx \frac{1}{n-1}\sum_{i=1}^{n}(X_{ij} - \bar{X}_j)(X_{ik} - \bar{X}_k) \tag{1}$$

The sample covariance matrix is an approximation of the covariance matrix presented in Equation 2.

$$\Sigma_{jk} = E[(X_j - \bar{X}_j)(X_k - \bar{X}_k)] \tag{2}$$

$$\bar{X}_j \approx \frac{1}{n}\sum_{i=1}^{n} X_{ij} = \frac{X_{1j} + ... + X_{nj}}{n} \tag{3}$$

Step 3 - Calculate eigenvalues λ and eigenvectors $\boldsymbol{\nu}$ from $\hat{\boldsymbol{\Sigma}}$
Eigenvalues λ can be extracted from the matrix $\hat{\boldsymbol{\Sigma}}$ when solving the problem in Equation 4 where $\mathbf{I}$ is the identity matrix. This is often done using the determinant $det(\hat{\boldsymbol{\Sigma}} - \lambda\boldsymbol{I}) = 0$. The eigenvectors can also be extracted from Equation 4 by inserting the found eigenvalues into it.

$$(\hat{\boldsymbol{\Sigma}} - \lambda\boldsymbol{I})\boldsymbol{\nu} = 0 \tag{4}$$

This method to get eigenvalues and eigenvectors is possible use for low-dimensional problems. Luckily Matlab has a function for solving high-dimensional problems. This function is used in this project.

Step 4 - Number of principal components
The first step of this part is usually to sort the eigenvectors. The vectors are ordered depending on the size of their eigenvalues. This is the order of significance in the data. A higher eigenvalue means a higher variance of the data. The variables with highest variance contains most information and are therefore the best representatives of the data. To decide the number of principal components that will be kept the Kaiser's rule is used. Kaiser's rule states that the the components with eigenvalues that are higher than the mean of the eigenvalues are kept Jolliffe (2002).

Step 5 - Retrieving the final data
When the principal components are found they are used to get the final data set. The matrix consisting only of the chosen eigenvectors is called $\boldsymbol{V}$. Some vector-operations are presented in Equation 5 that returns the new data set.

$$\boldsymbol{Finaldata} = (\boldsymbol{X} - \bar{\boldsymbol{X}})\boldsymbol{V}^T + \bar{\boldsymbol{X}} \tag{5}$$

This final data have got the characteristics of the principal components of the data. Thus the final data can be used to represent the original data.

10.2 The Stopping Method

What defines a good stopping method is that it favors short within-cluster-distance and long between-cluster-distance. The theory behind the three methods treated in this thesis is displayed in this section. They are Calinski-Harabasz criterion, Davis-Bouldin index, Silhouette.

10.2.1 Calinski-Harabasz criterion

The context presented here is based on the study Calinski and Harabasz (1974).

Equation 6 is showing how to find the Calinski-Harabasz criterion. SS_b is the sum-of-squares between clusters and SS_w the sum-of-squares within clusters. N is the total number of data-points and k is the number of clusters.

$$CHC = \frac{N-k}{k-1}\frac{SSb}{SSw} \tag{6}$$

The sum of squares of the within-cluster-distance and of the between-cluster-distance is presented mathematically in Equation 7 respectively Equation 8. It is easy to see that CHC gets larger when the between-cluster-distance gets larger and vice-verse for the within-cluster distance.

$$SS_w = \sum_{i}^{k} \sum_{x \in C_i} ||x - m_i||^2 \tag{7}$$

C_i is the cluster i and m_i is the centroid of cluster i and x is a data-point within cluster i. $||x - m_i||$ is the Euclidean distance between x and m_i.

$$SS_b = \sum_{i}^{k} \sum_{x \in C_j} ||x - m_i||^2, i \neq j \tag{8}$$

To receive the wanted stopping number k, Equation 6 has to be solved for $k = 2, ...$ The k that returns the highest value of CHC is the best choice.

10.2.2 Davies-Bouldin Index

Davies and Bouldin (1979) laid the foundation in this factual description.

The Davies Bouldin index is presented in Equation 9. It uses the between-cluster-distance M_{ij} and the within-cluster-distance S_i to find the number of clusters. These two are presented in Equation 10 respective Equation 11. The resulting optimal number of clusters is the number that maximizes DBI.

$$DBI = \frac{1}{N}\sum_{i=1}^{N} D_i = \frac{1}{N}\sum_{i=1}^{N} \max_{j \neq i} R_{ij} = \frac{1}{N}\sum_{i=1}^{N} \max_{j \neq i} \frac{S_i + S_j}{M_{ij}} \tag{9}$$

In Equation 10 A_i is the centriod of cluster i and $a_{k,i}$ is the kth element of A_i.

$$M_{ij} = ||A_i - A_j|| = \left(\sum_{k=1}^{n} |a_{k,i} - a_{k,j}|^2\right)^{1/2} \tag{10}$$

In Equation 11 X_m is a number containing within cluster i and T_i is the number of data points in cluster i.

$$S_i = \left(\frac{1}{T_i} \sum_{m=1}^{T_i} |X_m - A_i|^2 \right)^{1/2} \tag{11}$$

10.2.3 Silhouette

The theory behind Silhouette method in this part comes from Rousseeuw (1987).

$s(i)$ in Equation 12 is based on dissimilarities. Any object i is selected in the data set. This $i \in A$ where A is the cluster that i belongs to. If cluster A contains more data-points than just i, they are called j. Thus $j \in A$ and $j \neq i$. Now, $a_i = \frac{1}{N_A - 1} \sum_{j=1}^{N_A - 1} ||j - i||$ where N_A is the number of data-points in A. This means that $a(i) =$ is the average dissimilarity of i to all other objects of A. Other present clusters are called C and $C \neq A$. It is possible to find $d(i, C)$, which is the average dissimilarity of i to all objects of C. This can be written as $d(i, C) = \frac{1}{N_C} \sum_{z=1}^{N_C} ||z - i||$ if N_C is the number of data-points in C and z is a specific data-point in C. After finding $d(i, C)$ for all clusters $C \neq A$ it is possible to get $b(i) = \min_{A \neq C} = d(i, c)$

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{12}$$

Equation 12 returns a value $-1 \leq s(i) \leq 1$. A high value close to 1 means that a data point is a good representative of the cluster A. A value close to -1 means that the data-point is a good representative of cluster C. A number close to 0 means that the datum is close to the border between the clusters.

The highest value of the average value of $s(i)$ for the overall data decides how many clusters are appropriate.

10.3 Clustering data

Two methods for clustering data are presented in this section; Ward's method and k-means. Ward's method is selected as a main method and k-means to ensure the results.

10.3.1 Ward's method

The book Ward (1963) is about clustering data and has been used for information.

The creation of the clusters are done with Ward's method, which is an aggromerative method. This means that it will start of by n number of clusters and work its way down towards k clusters. k is the stopping number and n is the number of data-point. This is done one step at a time. The stopping point is necessary or else the algorithm will keep on working until there is only one cluster left that contains every data-point. There is also a divisive approach to Ward's method which means that the algorithm starts of with one cluster and stops when k clusters are found.

An algorithm of the used method is presented below.

Step 1 - Starting point
The algorithm starts of at a point where every observation n is seen as a cluster.

Step 2 - First action
The first action are to find the two clusters that are closest together and merge them. Hence there is n-1 clusters left.

Step 3 - Next action
Find the two clusters that are closest together and merge them. This does not mean to merge the closest data point. The centroid of each cluster has to be found to achieve this.

Step 4 - Stopping point
Step 3 is repeated multiple times until the stopping point is reach. This is a predetermined number of clusters k.

To be able to complete the steps the distance between clusters need to be calculated. Ward defines the cluster-distance as how much the sum of squares will increase if they are merged. This is described by Equation 13. $\overrightarrow{m}_j$ is the centroid of cluster j and n_j is the number of observations cluster j contains. Matlab uses the Euclidean distance between the centroids to adress the problem of finding $\Delta(A, B)$ Matlab (2017).

$$\Delta(A, B) = \sqrt{\frac{2n_A n_B}{n_A + n_B}} ||\overrightarrow{m}_A - \overrightarrow{m}_B|| \tag{13}$$

10.3.2 k-means method

The k-means++ method is used because according to Arthur and Vassilvitskii (2007), this decreases the running time and improves the results. This is also the source this part of the thesis is based upon. The only diference between k-means and k-means++ is the initial step.

The goal when solving Equation 14 is to minimize it. A stopping number k and a set of n data points $\chi \subset R^d$ is gotten beforehand.

$$\phi = \sum_{x \in \chi} \min_{c \in C} ||x - c||^2 \tag{14}$$

Step 1 - The initial centroids
In the initial step, k data-points are selected to be centroids.
1a. Find one centroid c_1, chosen uniformly at random from χ.
1b. Find a new centroid c_i, choosing $x \in \chi$ with probability $\frac{D(x)^2}{\sum_{x \in \chi} D(x)^2}$. $D(x)$ is the shortest distance from a data-point to a centroid.
1c. Step 1b. is repeated until k centroids are selected.

Step 2 - Finding the clusters
For each $i \in (1, ..., k)$, set the cluster C_i to be the set of points in χ that are closer to c_i than they are to c_j for all $j \neq i$.

Step 3 - Find the centroid in every cluster
For each $i \in (1, ..., k)$, set c_i to be the center of mass of all data-points in C_i. Find c_i with the following equation: $c_i = \frac{1}{|C_i|} \sum_{x \in C_i} x$

Step 4 - Assigning of data points
Repeat Steps 2 and 3 until the clusters no longer changes

TRITA -MAT-E 2017:76
ISRN -KTH/MAT/E--17/76--SE