Lappeenranta-Lahti University of Technology LUT
School of Engineering Science
Computational Engineering and Technical Physics
Technomathematics

Henna Pekkala

ANALYSIS OF FACTORS AFFECTING THE DEVELOPMENT OF LIQUIDITY AND ITS FORECAST WITH HIERARCHICAL CLUSTERING ON PRINCIPAL COMPONENTS

Master's Thesis

Examiners:	D.Sc. (Tech.) Matylda Jabłońska-Sabuka, M.Sc. (Econ.) Teppo Salmi
Supervisors:	D.Sc. (Tech.) Matylda Jabłońska-Sabuka, B.Sc. (Econ.) Pekka Kotovaara, M.Sc. (Econ.) Teppo Salmi, Prof. Lasse Lensu, B.Eng. Laura Sainio

TIIVISTELMÄ

Lappeenrannan-Lahden teknillinen yliopisto LUT
School of Engineering Science
Laskennallinen tekniikka ja teknillinen fysiikka
Technomathematics

Henna Pekkala

Maksuvalmiuden kehitykseen ja sen ennustamiseen vaikuttavien muuttujien analysointi pääkomponenttien hierarkkisella klusteroinnilla

Diplomityö

2021

50 sivua, 15 kuvaa, 1 taulukko

Tarkastajat: TkT Matylda Jabłońska-Sabuka, KTM Teppo Salmi

Hakusanat: kassavirtaennuste, maksuvalmius, toimialaluokitus, factominer, monityyppisen datan faktorianalyysi, pääkomponenttien hierarkkinen klusterointi

Tässä työssä tutkittiin maksuvalmiuden kehitystä sekä kassavirtaennusteen ja -toteutuman eroa, jotta saataisiin laajempi kuva yrityksien taloudellisesta nykytilasta. Päämääränä oli tutkia mahdollisia yhteisiä tekijiöitä toimialaluokituksessa, yrityksen koossa tai iässä. Pääkomponenttien hierarkkinen klusterointia suoritettiin käyttämällä monityyppisen datan faktorianalyysia yli 5000 yrityksen anonymisoituun viikottaiseen dataan sekä pienempiin otantoihin samasta datasta. Tulokset eivät tuottaneet yksikäsitteisiä johtopäätöksiä, kuinka yrityksen koko tai ikä vaikuttaa maksuvalmiuden kehitykseen tai mitkä toimialat kukoistavat tai kohtaavat vaikeuksia. Mahdollisia syitä selkeiden tuloksien puuttumiselle on maksuvalmiuden epävakaus lyhyellä ajanjaksolla, muut tuntemattomat tai mittaamattomat tekijät, jotka vaikuttavat kassavirtojen kehitykseen sekä erilaiset reaktiot COVID-19:n aiheuttamiin rajoituksiin ja pandemian vaikutukset kansainväliseen talouteen.

ABSTRACT

Lappeenranta-Lahti University of Technology LUT
School of Engineering Science
Computational Engineering and Technical Physics
Technomathematics

Henna Pekkala

Analysis of Factors Affecting the Development of Liquidity and its Forecast with Hierarchical Clustering on Principal Components

Master's Thesis

2021

50 pages, 15 figures, 1 table

Examiners: D.Sc. (Tech.) Matylda Jabłońska-Sabuka, M.Sc. (Econ.) Teppo Salmi

Keywords: cash flow forecast, liquidity, industrial classification, factominer, factor analysis of mixed data, hierarchical clustering on principal components

In order to get a glimpse of the current state of the companies on a broader level, the development of liquidity and the difference between the cash flow forecast and its realized values were examined to find possible common factors between line of industry, company size or age. Hierarchical Clustering on Principal Components with the help of Factor Analysis of Mixed Data were performed on weekly data of over 5000 anonymized companies and smaller sets of the same data. The results did not offer conclusive remarks on how the size or the age of the company affects the development liquidity or which industries were succeeding or facing harder times. Possible reasons for the absence of clear results were the volatility of liquidity in short time span, other unidentified or unmonitored variables affecting the development of companies' cash flow and different reactions to the restrictions as well as the bigger international impact on the economy caused by COVID-19.

PREFACE

As this project comes to a close so does an important part of my life. Back in 2015, when I was applying to universities, I did not anticipate to form this kind of attachment to the one I ended up in. For the unforgettable years I want to thank all the friends along the way, Polytekninen Willimiesklubi, Student Union of LUT-University LTKY, Student Association for Students of Computational Engineering and Technical Physics and the staff of LUT-University.

I want to thank each and everyone involved in this project and for making it possible for me to do my thesis in the first place. Special thanks to the examiners Matylda, for the invaluable guidance and endless patience with my master's thesis, and Teppo, for all the advice so far on data management, data utilization and career paths and the genuine interest in helping out. I also have gratitude for Pekka, for helping me understand the business aspects of this study, and Lasse, for all the insightful questions and comments. Aatu-Ville, Inka, Karoliina, Elmer and Laura, you inspired and supported me to keep on working on my thesis, even on the days when the thought of completing it sounded extremely distant.

Heartfelt thanks to all of my friends and family for sharing countless laughs and always being there for me when I needed you. I am excited to see what kind of adventures we will come up with in the future.

Lappeenranta, May 28, 2021

Henna Pekkala

CONTENTS

LIST OF SYMBOLS AND ABBREVIATIONS

I	Total number of categories across all the variables
I_q	Number of individuals who belong to cluster q
J	Maximum number of variables
Q	Selected maximum number of clusters
SS_B	Sum of squares between different clusters
SS_W	Sum of squares within clusters
$\Delta(Q)$	Increase of inertia between clusters $Q-1$ and Q
$\boldsymbol{A}$	Data table
$\boldsymbol{D_c}$	$\mathrm{diag}(\boldsymbol{c})$, column totals as a diagonal matrix
$\boldsymbol{D_r}$	$\mathrm{diag}(\boldsymbol{r})$, row totals as a diagonal matrix
$\boldsymbol{F}$	Factor scores matrix (principal component scores matrix)
$\boldsymbol{G}$	Column factor scores
$\boldsymbol{U}$	Left singular vectors can be found in the columns of this matrix
$\boldsymbol{V}^\top$	Right singular vectors can be found in the rows of this matrix
$\boldsymbol{Z}$	Probability matrix
$\boldsymbol{\Sigma}$	Diagonal matrix with the singular values
$\boldsymbol{c}$	Column totals
$\boldsymbol{f_s}$	Coordinates of the projection of supplementary row
$\boldsymbol{g_s}$	Coordinates of the projection of supplementary column
$\boldsymbol{i_s}$	Supplementary row
$\boldsymbol{j_s}$	Supplementary column
$\boldsymbol{r}$	Row totals
$\overline{x}_j$	Mean of variable j
$\overline{x}_{qj}$	Mean of variable j in cluster q
i	Individual
i_m	Number of categories in a variable
j	Variable
k	Number of clusters
m	Number of columns
n	Number of individuals
q	Cluster
x_{iqj}	Value of variable j of individual i in cluster q
CA	Correspondence Analysis
FAMD	Factor Analysis of Mixed Data
HCPC	Hierarchical Clustering on Principal Components
MCA	Multiple Correspondence Analysis

PCA	Principal Component Analysis
SaaS	Software as a Service
SMEs	Small to Medium-sized Enterprises
SVD	Singular Value Decomposition

1 INTRODUCTION

1.1 Background

Managers, employees, suppliers, external investors, loan creditors, tax agencies and the list goes on. There are undoubtedly interest from multiple parties to understand how a company is performing. Unsurprisingly, there are multiple different measures to do just that: valuate business performance.

Balance sheets and profit and loss statements usually come to mind when talking about examining the profitability of a company. Balance sheet is great for showing how a company finances its activities, how it is investing and what are company's assets and liabilities, while profit and loss statement discloses the profit that was earned, how it was calculated and distributed. While these method provide a good overall picture of how the company is faring, they have their flaws. Some factors require approximation and opinions, like the value of a project that is not yet finished and the provisions for doubtful debts, and some parts of these financial statements can be subjected to modifications and even manipulation. (Fight 2006)

The financial statements mentioned above are usually produced annually. In the fast pace of companies' daily lives during uncertain times, this seems to be quite slow way to evaluate the performance of a company, especially performances of small to medium-sized enterprises (SMEs). It is much more meaningful to analyze the snapshot glimpses of the current states of the companies, during the challenging times caused by COVID-19. In order to get an idea on how industries are faring right now and might do in the near future, cash flows and their forecasts are examined.

Cash flow forecasting is especially important for small to medium-sized enterprises (SMEs). In addition to keeping the business running, liquidity is needed in expanding the business, which is certainly important for companies of this size. Insufficient liquidity is the most common reason for not being able to expand business (Fight 2006). The financial well-being of SMEs should not be overlooked, since they make up a significant portion of every country's economic activity. In 2019, SMEs in Finland constituted over 59.6% of value added and they were the employer to 65.2% of all the people employed (The European Commission Directorate-General for Internal Market, Industry, Entrepreneurship and SMEs 2019).

Most of the bankruptcies happen, because cash runs out due to poor management, so planning is crucial. With careful planning, the company can prepare for different scenarios. Forecasting, like any other way of predicting the future, can be uncertain and should not be taken as a fact how things will go, but rather as an opportunity to find weaknesses or plan upcoming investments. If cash flow predictions are not done or done carelessly, a company can find themselves facing a problematic situation financially, which can put the whole company in a compromising situation.

There is understandably an interest to constantly improve in utilizing cash flow data. The aim of this work is to study patterns in companies' cash flow situation and identify possible common factors influencing their performance. Moreover, the aim is to study how well the case study system forecasts cash flows and verify whether any groups of customers seem consistently more difficult to predict accurately.

1.2 Objectives and delimitations

Because the used data must be first identified and accumulated, there are naturally a few limitations to the amount of data available. In this study, only week-level data that has accumulated over the course of three months is used.

The research questions of the thesis are as follows:

1. Is there variation in forecast and realized cash flow based on the line of business, location, turnover or the age of the company?
2. Is there another dimension that explains the variation in companies' cash flows?
3. Does the cash flow forecast have trouble predicting some segment of companies accurately?
4. How the different lines of business cope with the uncertain times caused by COVID-19?

1.3 Structure of the thesis

First, the second and third chapter introduce the methods proposed to analyse the data. Chapter 2 introduces the different principal analysis methods and the third chapter covers clustering methods utilized in this study. Next, chapter 4 introduces the R-package, which is used to perform the calculations. Then, in chapter 5, the data as well as its limitations and scope is discussed. From there, in chapter 6, the results of the experiments are given in detail. Finally, discussion on the results in chapter 7 as well as conclusion in chapter 8 are the closing chapters of the study.

1.4 Cash Flow Forecast

Keeping track of the cash flow is crucial for any business. Without any easily transferable assets company could go bankrupt shockingly quickly. This is why keeping track of and forecasting cash flows are such important lines of study. With cash flow forecasts, companies can prepare for the future invoices and plan for investments without the fear of suddenly finding themselves in a situation, where they have overdrawn bank account.

The importance of cash flow is recognised widely and many argue that it is at least one of the most important indicators on whether the company is doing well (Fight 2006). As all of the incoming and outgoing cash is recorded, it gives a realistic picture on the liquidity of the company. Unlike profit-based analysis, which can be painting an optimistic picture of the company's future or even be modified and manipulated to look good, the company's cash flow can be more reliable way of telling how the company is faring (Fight 2006). Analysing cash flow at intricate level, makes it possible for the management to react and take corrective actions much more quickly compared to month-end financial results (Kaufman 2014). Even though cash flow analysis can be broken down to be analyzed by different divisions or product lines in greater detail, it is still quite intuitive tool.

There are many definitions for different cash flows, but every one of them demonstrate at least some part of how the capital flows through the company. They are metrics of the cash inflows and outflows from different perspectives. Cash flow can be broken down to a few more specific sections, usually to operative, investment and finance. Operative cash flow are the standard daily inflows and outflows of capital. Paying salaries, taxes and suppliers as well as generating revenue. Operative cash flow keeps count on the simplest purpose of a company: to generate cash by operating profitably (Bhandari et al. 2013). Investment cash flow consists of the capital equipment and intellectual property of the company, otherwise known as the needed infrastructure. Cash flow from financing activities cover taking and paying loans as well as the money generated from share issues or spent on the buybacks. (Ryland 2020)

Of course, it all gets even more interesting when, instead of examining how the company has been doing or is doing now, the focus is shifted to how the company will do in the future or if it even has one. That is why predicting the development of the cash flow in the future, forecasting, is such an interesting topic. With well made cash flow forecast company can plan on when to make investments, apply for a loan or ask for an opportunity to delay payment well in advance. Good planning not only lowers risks and costs, but also ensures the capability of paying invoices.

With forecasts, though, one must remember that the future cannot be fully predicted. This is why cash flow forecast can be seen as more of a tool to identify critical weak spots in the future operations and help the company to prepare for what lies ahead (Fight 2006). Sensitivity analysis on cash flow forecast can further explain weaknesses by exposing parts of the company that cannot keep up if the sales increase or decrease.

In this study, boarder definition of cash flow is used. The cash flow of a company is healthy, if the net inflow covers the expenses of producing goods and maintaining the assets (Fight 2006). More specifically, the development of liquidity and the difference between the forecast and realized value are being studied in later chapters.

2 MULTIVARIATE DATA ANALYSIS

For any data set there are multiple different kinds of variables that can be utilized in finding similarities between individuals. These variables can be divided between ordinal or categorical variables. With limited number of numerical continuous variables, the traditional feature extraction like principal component analysis (PCA) is rendered quite ineffective as it cannot take advantage of the categorical features. There are reasons to believe, that in this study the categorical variables, for example the industrial classification of the company, might explain the development of liquidity for some of the companies, as the COVID-19 pandemic has hit hardest to the catering and tourism industry while some other lines of business, like those who provide software solutions online, might even have fared a bit better than how they would have without the pandemic.

2.1 Principal Component Analysis

Studying the differences between a handful of individuals with two different variables is usually a task which does not require very advanced methods. These variables, for example height and weight, can be plotted to two corresponding axes to find the underlying possible correlations of the variables or the homogeneity of some individual groups. When dealing with more than three variables, though, this kind of analysis gets complicated very quickly. In other words, a tool is needed to find the similarities and dissimilarities between individuals across all of the variables. (Francois Husson et al. 2017)

PCA is, without a doubt, among the most popular methods to reduce dimensions and perform multivariate statistical analysis. It takes a table consisting of observations (rows) and variables (columns), which are usually dependent and inter-correlated, and performs orthogonal transform to get new, linearly independent, set of variables. These new variables are principal components. In other words, the goal of this method is to identify and extract meaningful information from the original data table and transform them into principal components. Principal components can be used to determine the similarities between the observations and the relationships between the variables. (Abdi and Williams 2010)

Before performing PCA, the data in the input table should be centered and standardized, because the measures can have different units or variances. Typically, the first is done by centering the columns around zero, so that the mean of the column is zero. Standardization is usually done by dividing the measures with the variables norm. The result is an unit norm.

Understanding singular value decomposition (SVD) is crucial when talking about PCA since it is very closely related to and often used in the finding the components of PCA. The data table $\boldsymbol{A}$ is matrix of size $n \times m$, with n observations which are described by m variables. The SVD of matrix $\boldsymbol{A}$ is

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top, \tag{1}$$

where $\boldsymbol{U}$ is $n \times m$ matrix with the columns representing left singular vectors, $\boldsymbol{\Sigma}$ is $m \times m$ diagonal matrix with singular values on the diagonal and $\boldsymbol{V}^\top$ is $m \times m$ matrix with the rows representing the right singular vectors. (Wall et al. 2003)

The $n \times m$ factor scores matrix, or principal component scores, $\boldsymbol{F}$ is obtained as

$$\boldsymbol{F} = \boldsymbol{U}\boldsymbol{\Sigma}. \tag{2}$$

This leaves $\boldsymbol{V}$, sometimes called the projection or loading matrix, which is a matrix containing coefficients of the linear combinations that were used to compute the principal component scores $\boldsymbol{F}$. As can be seen here,

$$\begin{aligned} \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top &= \boldsymbol{A} \\ \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top\boldsymbol{V} &= \boldsymbol{A}\boldsymbol{V} \\ \boldsymbol{U}\boldsymbol{\Sigma} &= \boldsymbol{A}\boldsymbol{V} \\ \boldsymbol{F} &= \boldsymbol{A}\boldsymbol{V} \end{aligned} \tag{3}$$

multiplying the observations $\boldsymbol{A}$ with $\boldsymbol{V}$, results in the values of the principal components of the observations' projections. This also makes it possible to insert observations, not used in the PCA, in order to get factor scores for supplementary observations.

The first principal component is the linear combination of the variables, which explains the most of the variance, or inertia, of the data. The next principal component is found by examining the linear combination, which captures the most of the remaining inertia. The remaining principal components follow this pattern until all of the variance can be explained.

For two-dimensional data, one can think of fitting the first principal component to the data in a way that minimizes the squared differences between the line and the observations.

The second principal component is then placed orthogonally to the first and that way it captures the most of the remaining variance.

2.1.1 Circle of correlations

Usually PCA results are based on the distances between points in a low-dimensional map of first few principal components. In this study, the focus is on how different factors explain the current state and the ability to forecast the state of the liquidity of companies. That is why it makes more sense to examine the loadings, the correlations of the principal components and the variables, rather than the contributions, the correlations of the principal components with the observations of individual companies.

Interestingly, the sum of squared loadings for a variable for all principal components equals one. This tells the proportion of the explained variance by each principal component for a variable. This fact can be taken advantage of when aiming to visualize the results with circle of correlations. As the sum of the squared loadings of the components is one, they should fall inside the unit circle in the two dimensional plot with the first and the second principal components as the x- and y-axis, respectively. If a variable cannot be fully explained by the two principal component visualized, the variables might not reach the perimeter of the unit circle, but end somewhere inside it, depending on how much of the squared loadings are explained by the chosen components. (Abdi and Williams 2010)

Just as observations, also the relationship of supplementary variables to the original set can be calculated. The loadings for the supplementary variable can be calculated from the same or part of the same data set. The supplementary variables can be added to the circle of correlations with the the squared loadings for respected components. It should be noted that the squared loadings of the supplementary variables do not sum up to one.

2.1.2 Number of selected components

There is no definitive way to determine the 'perfect' number of principal components to represent the data. However, there are a few very plausible options in choosing the cut-off point: elbow test from scree-plot or eigenvalues that are greater than one. (Abdi and Williams 2010)

The elbow test is done by examining the scree-plot, which depicts the principal components on the x-axis and the respective explained variance on the y-axis. The cut-off point can be determined by discovering a point where the curve becomes noticeably flatter.

Another way of determining the cut-off point is to look at the eigenvalues. If the data has been centered and standardized, all the principal components, whose eigenvalues are greater than one, can be selected. If centering and standardization has not been done, however, the eigenvalues should be greater than the average. (Abdi and Williams 2010)

2.2 Multiple Correspondence Analysis

PCA only works on quantitative variables. Unfortunately, not everything in life can be explained in numbers or line up in a two dimensional order. There are a lot of labels that cannot be transformed to numbers as it could imply some kind of order among them when there is none. This is where multiple correspondence analysis (MCA) provides answers. To put it simply, it is a generalized version of PCA, where, instead of numerical variables, it examines the relationship of categorical values. In practice, MCA converts categorical values of a variable for the same number of columns of an indicator matrix. After the conversions on that said matrix, standard correspondence analysis (CA) is performed. (Abdi and Valentin 2007)

The input data table $\boldsymbol{A}$ has n observations which are described by m categorical variables. Each variable has their own i_m number of categories, with sum of the categories of all the variables being I. Thus, the indicator matrix is of size $n \times I$. CA calculates factor scores for the columns and rows. Usually the factor scores are scaled as such, that their variances are equal to their eigenvalues. In order to get the factor scores, the probability matrix $\boldsymbol{Z}$ needs to be found by taking advantage of the grand total of the table, , and the original table $\boldsymbol{A}$, $\boldsymbol{Z} = N^{-1}\boldsymbol{A}$. From $\boldsymbol{Z}$, row totals $\boldsymbol{r}$ and column totals $\boldsymbol{c}$ can be acquired by multiplying $\boldsymbol{Z}$, or its transpose for the column total, by a conformable vector of ones. These totals can be transformed to diagonal matrix, which are denoted as $\boldsymbol{D_r}$ and $\boldsymbol{D_c}$. The factor scores can then be derived with the help of SVD with the following equation

$$\frac{1}{\sqrt{\boldsymbol{D_r}}}(\boldsymbol{Z} - \boldsymbol{r}\boldsymbol{c}^\top)\frac{1}{\sqrt{\boldsymbol{D_c}}} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top \tag{4}$$

The row factor scores can be obtained as

$$\boldsymbol{F} = \frac{1}{\sqrt{\boldsymbol{D_r}}}\boldsymbol{U}\boldsymbol{\Sigma} \tag{5}$$

while the column factor scores, $\boldsymbol{G}$, can be calculated as

$$\boldsymbol{G} = \frac{1}{\sqrt{\boldsymbol{D_c}}}\boldsymbol{V}\boldsymbol{\Sigma} \tag{6}$$

Similarly to PCA, also with MCA, supplementary observations or variables can be introduced to the mix. The projection to factors is done using the transition formula. To get the coordinates $\boldsymbol{f_s}$ of a supplementary row, the transition is calculated as

$$\boldsymbol{f_s} = (\boldsymbol{i_s}^\top \mathbf{1})\boldsymbol{i_s}^\top \boldsymbol{G}\Sigma^{-1}, \tag{7}$$

where $\boldsymbol{i_s}$ is the supplementary row. For a supplementary column the calculation of the coordinates $\boldsymbol{g_s}$ is as follows

$$\boldsymbol{g_s} = (\boldsymbol{j_s}^\top \mathbf{1})\boldsymbol{j_s}^\top \boldsymbol{F}\Sigma^{-1}, \tag{8}$$

where $\boldsymbol{j_s}$ is the supplementary column.

It should be noted, that because every single category in the variables means new dimensions in the indicator matrix, it is common for the results of MCA to have smaller percentage of explained variance than that of PCA. In addition, the scree-plot curve produced by MCA is much smoother than the curves provided by PCA. (Pagès 2014)

2.3 Factor Analysis of Mixed Data

It is not a new idea to combine quantitative and qualitative variables in factorial analysis. There have been methods for converting the variables to some form or another in order to perform one of the better known factorial analysis methods on the data. If not enough attention is paid, converting the numerical values to bins or the values in categorical columns to a set of numbers, one might lose some valuable information along the way or distort the results.

Factor analysis of mixed data (FAMD) has been developed to take advantage of both quantitative and qualitative data, i.e. mixed data, by combining the fundamentals of PCA and MCA in a suitable way. By using FAMD, the inertia can be maximized in a projected cloud of quantitative and categorical variables. The most important part of FAMD is the appropriate balancing of different types of variables. This is especially important when the vector with the highest explained variance of the (remaining) dimensions is being calculated. Specifically, the qualitative variables with different number of categories need to be appropriately balanced with each other and the quantitative variables. (Pagès 2014)

The combination of quantitative and qualitative variables can be done by using appropriate coding in MCA or specific metric in PCA. With unstandardized PCA, the FAMD could be implemented in the following way: first, the quantitative variables must be centered and reduced. Then, the qualitative variables are to be transformed to indicator columns and divided by the square root of the proportion of the individuals, who possess the category in question. As the results, the PCA presents the representation of the individuals and quantitative variables.

The FAMD excels in two particular cases, compared to the more commonly known method of converting quantitative values to categorical bins before utilizing MCA. When there are few qualitative variables compared to quantitative variables and when there are low number of individuals. Even though the amount of individuals is large in this study, the clear majority of the variables are quantitative variables. Although there are mentions that the FAMD performs especially well with small number of individuals compared to the other methods, there are no suggestions that it would not fare well with larger data sets.

3 CLUSTERING METHODS

There is a number of different clustering methods that could be used to try to separate the data. In order to avoid the results being skewed by extreme values, it is important to perform preprocessing on the data to make it consistent and normalized.

3.1 K-means

K-means is a well-known clustering method which separates the data into k number of clusters. The variable k must be given to the algorithm beforehand. K-means algorithm starts by picking starting points for the cluster centroids and then starts to iteratively move towards a point where the distances to cluster members are minimized. The cluster centroids and their members are reassigned until the centroids do not move anymore. This algorithm is more commonly known as Lloyd's algorithm. (Zhao et al. 2018)

3.1.1 Choosing the number of clusters

If the number of clusters, k, is not provided, it can be determined by performing the partition for different number of ks and calculatining which number of k divide the data best. There are multiple methods to determine the number of clusters that best fit the data Sum-of-Squares, Silhouette, Calinski-Harabasz and Davies-Bouldin just to name a few. These criterions have different emphasis on different parts of the data, like density of the clusters or clear separation Sum-of-squares, for example, is calculated as follows

$$\frac{SS_B}{SS_W} \times k, \tag{9}$$

where SS_B is the sum of squares, or in other words overall variance, between the different clusters, SS_W is the sum of squares within the clusters and k is the number of clusters. (Baarsch et al. 2012)

3.2 Hierarchical Clustering on Principal Components

As principal component analysis methods can be seen as a way to denoise and preserve the structure of the data, they are often used as a preprosessing step before clustering.

Hierarchical clustering methods complement principal analysis methods very well. In hierarchical clustering on principal components (HCPC), instead of only using principal analysis methods to remove the noise and restructure the data, the chosen number of principal components are used in computing the hierarchical clustering.

HCPC identifies clusters, as the name suggests, by taking advantage of the principal components. This method mixes two methods by performing factorial analysis with Ward's hierarchical classification and K-means, a posterior clustering process. (Argüelles et al. 2014)

Traditional hierarchical clustering can have different distance measures and agglomeration criterions. There are numerous distance measures that could be used, Euclidean and Manhattan to name a few, and a few agglomeration criterions, for example single-link, complete-link and Ward's. As HCPC takes advantage of principal component methods, hierarchical clustering and K-means, partitional clustering, the input and output variables should be uniform and able to be treated similarly in different steps. Hierarchical clustering covers multiple different methods to do partitioning, but when referring to HCPC, it can be assumed that the used distance measure is Euclidean and agglomeration criterion is Ward's. These choices allow the combination of HCPC both with K-means and principal component methods. The first, is traditionally calculated with Euclidean distance and, like principal component methods, Ward's criterion is based on inertia, or multidimensional variance. (F. Husson et al. 2010)

The Ward's method examines the growth of the intra-inertia of the clusters. It tries to minimize the reduction of inter-intertia between different clusters as it chooses which clusters to aggregate next, until there is only one cluster with all the individuals or it reaches Q number of clusters. This is based on Huygens theorem, which decomposes the total inertia to be the sum of inter-inertia (between inertia) and intra-inertia (within inertia). In other words,

$$\sum_{j=1}^{J}\sum_{q=1}^{Q}\sum_{i=1}^{I_q}(x_{iqj}-\overline{x}_j)^2 = \sum_{j=1}^{J}\sum_{q=1}^{Q} I_q(\overline{x}_{qj}-\overline{x}_j)^2 + \sum_{j=1}^{J}\sum_{q=1}^{Q}\sum_{i=1}^{I_q}(x_{iqj}-\overline{x}_{qj})^2, \tag{10}$$

where j is a variable with maximum of J, q is a cluster with maximum of Q, i denotes individuals, I_q being the number of individuals who belong to cluster q, x_{iqj} is the value of variable j of individual i in cluster q, $\overline{x}_j$ is the mean of variable j and $\overline{x}_{qj}$ is the mean of variable j in cluster q. (F. Husson et al. 2010)

The number of clusters is determined by the growth of inertia. If the growth of inertia from $Q-1$ clusters to Q clusters is greater than the inertia from Q clusters to $Q+1$, Q number of clusters is selected. In other words, Q is selected so that the following statement is minimized:

$$\frac{\Delta(Q)}{\Delta(Q+1)}, \tag{11}$$

where $\Delta(Q)$ is the increase of inertia between $Q-1$ and Q clusters.

Dividing the clusters can be done in three ways: simply cutting the tree with Q clusters, performing K-means with Q clusters or combining the two by using the cut hierarchical tree as the initial partition for the K-means. The result from K-means is then considered as the final result. This method can slightly modify the clusters and improve the results as the ratio of the inter- and intra-inertia is normally better.

4 FACTOMINER-PACKAGE

There are a few solutions for the algorithms described above: widely used FactoMineR package for R and Prince-library for Python. The easy implementation of FactoMineR-package and its extensive function-catalogue ultimately led to its utilization.

FactoMineR is an R package that is developed for multivariate data analysis. It is versatile as it can handle data with different structures, different types of variables as well as supplementary variables (Lê et al. 2008). Factoextra-package provides a lot of useful functions and plots of the results of FactoMineR's functions. The functions provided by FactoMineR are very trivial to implement. For *FAMD*-function, only the data frame to perform the function on needs to be given as a parameter, but it is possible to pass more specific arguments like the the number of dimensions kept, indexes of supplementary variables or individuals and whether or not to visualize the results. The algorithm takes care of scaling and centering of the variables as well as the handling of possible supplementary variables or individuals. The results of this multivariate data analysis can be examined as a data frame or with the help of multiple visualization tools from the factoextra-package. The variables' effects on all of the dimensions can be visualized with corrplot-package's *corrplot*-function. (Lê et al. 2008)

HCPC-function is as simple to implement as *FAMD*. It only needs results of the chosen principal component method as an input, but also accepts more specific parameters like, the number of clusters, minimum and/or maximum number of clusters and whether the graph should be displayed, to name a few. The package manual introduces variable *kk*, which is the number of clusters used for K-means as a preprocessing step. Usually, K-means is performed at the end to get slightly more robust results than that which a hierarchical clustering method can provide on its own. If the algorithm is used as a preprocessing step, the usual consolidation can not be performed. Using K-means as the preprocessing step is considered useful for large data sets, however, with the drawback being inability to use some of the visualizing functions on the results.

The variables associated with the clusters give some interesting insights on how the data is partitioned. Statistically significant quantitative variables can be examined with `res.hcpc$desc.var$quanti` the mean of the category and contrasting that to the overall mean of the whole dataset. Similarly, these variables' intra-cluster standard deviation can be compared to the standard deviation of the whole data set. P-value determines if the variable is significantly linked to the cluster and v-test, which is derived from p-value, gives insight if the variable is overrepresented (v-test is positive) or underrepresented (v-test is negative) in the cluster.

For the description of the relationship with the categorical variables and the clusters, a number of interesting values can be accessed through `res.hcpc$desc.var$category$`. The *HCPC*-algorithm describes these variables with Cla/Mod, Mod/Cla, Global, p-value and v-test. The last two values are defined the same way as with quantitative variables. Cla/Mod explains the percentage of the individuals with a particular categorical variable that belong to the cluster in question. Mod/Cla describes the percentage of the individuals who belong to this group in question and have a certain categorical variable. Global is the overall percentage of the individuals with a certain categorical variable.

5 DATA

The data used for calculations consists of approximately 5260-5469 individuals per week, with a slight variation from week to week. For each row there is six quantitative variables and two qualitative variables with 16 and 18 categories. The data is acquired from two different sources and steps like anonymization and aggregation are done before any analysis is done. The schema depicted in Figure 1.

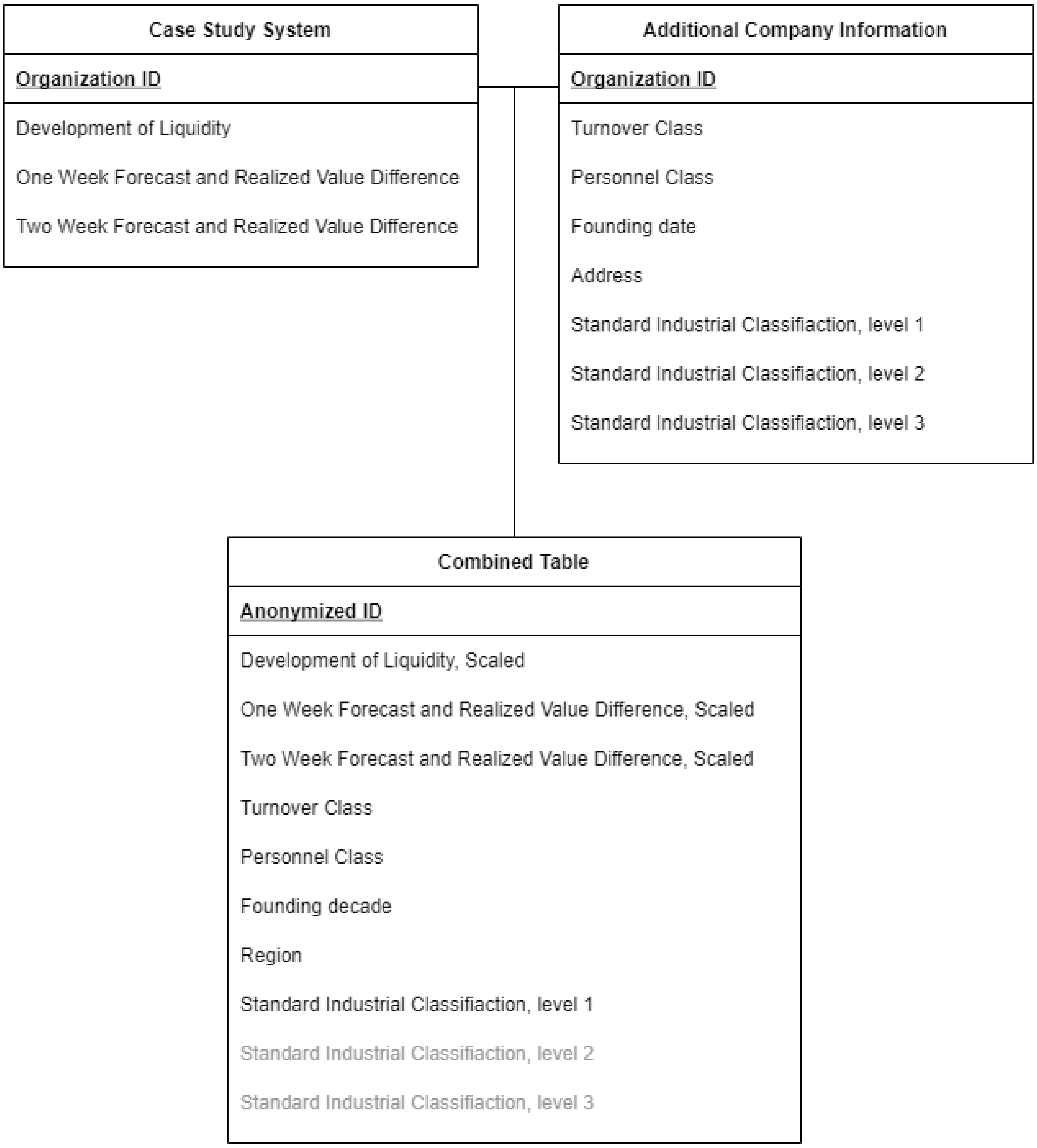
Case Study System
Organization ID
Development of Liquidity
One Week Forecast and Realized Value Difference
Two Week Forecast and Realized Value Difference
Additional Company Information
Organization ID
Turnover Class
Personnel Class
Founding date
Address
Standard Industrial Classifiaction, level 1
Standard Industrial Classifiaction, level 2
Standard Industrial Classifiaction, level 3
Combined Table
Anonymized ID
Development of Liquidity, Scaled
One Week Forecast and Realized Value Difference, Scaled
Two Week Forecast and Realized Value Difference, Scaled
Turnover Class
Personnel Class
Founding decade
Region
Standard Industrial Classifiaction, level 1
Standard Industrial Classifiaction, level 2
Standard Industrial Classifiaction, level 3

Figure 1. Schema. Anonymization, aggregation and scaling are done after joining the two information sources.

5.1 Case Study System

There are multiple tools for monitoring and forecasting company's cash flow. In this study, aggregated and anonymized data were provided by an accounting system. This system is a software as a service (SaaS) product and it aims to automate as much as possible of the paperwork that comes with running a business.

The case study system used in this work has a cash flow forecast that takes advantage of the accounting data of the company in order to produce a cash flow forecast for each individual company. Sent and received invoices are marked on the cash flow forecast diagram on their due date, if the user does not change the date manually. The user can also make custom entries to the flow as a form of cash inflow or outflow. The system's cash flow forecast calculates liquidity and operative cash flow for the company. It takes into account a number of different entries that affect the liquidity: sales and purchase invoices and orders as well as their payments, salaries and trip expenses.

There is lot of potential value in conducting extensive analysis on liquidity and cash flow forecasts. First, it is clear that when the customers are succeeding and content, also the software is succeeding. If the customers are satisfied with the product, they can recommend it to their friends and also be willing to try more of the same provider's products or are more inclined to purchase additional add-ons. Furthermore, if a company goes bankrupt, they will most definitely also cease to be a customer. With analysis on the development of liquidity and cash flow forecasts, the company could offer insights of the data straight to the customer and help them understand how to manage and act on their cash flow forecast. Customer churn might also experience a positive impact when the users would find the software not only useful for performing day-to-day accounting tasks but also helping the company to survive and succeed.

From the insights extracted from the cash flow analysis, a dashboard could be created to help the users understand their financial situation better and offer deeper analysis as an add-on. Prompts could be developed to encourage people to take advantage of their excessive working capital or warn about incoming problems with current spending and propose to start looking for a loan. The data could also be very valuable in the hands of an in-house analysts, who could compose monthly barometers on how the companies fare based on different lines of businesses, geographical locations or size groups. This might improve the image and the appeal of the company as a chosen service compared to the competitors.

5.2 Scope and preprocessing

There are multiple different usage types of the case study system and many users might have very different habits, tasks and frequency of use. These things can depend on multiple things, such as the nature of the company and the user's job description. Some users use the system daily and might take advantage of the more detail-oriented tools, while some might only use a very bare-bones version of the system. In this study, companies, that might find use of the cash flow forecast and have enough data for the recorded liquidity to reflect reality, are being examined. These companies can be narrowed down by examining companies, that take advantage of multiple features of the system, which in turn helps the system produce a more reliable forecast. The companies that are studied use the service frequently enough, so that the company's cash flow should reflect reality.

The time frame was set to be on a weekly basis. This enables some aggregation, but still produces very informative data even in shorter time span. The daily or otherwise more precise distribution of the cash inflows and outflows do not give a lot of extra information, as a company cannot be viewed as very stable if they do not have reliable weekly cash flows.

During the preprocessing, the data acquired from the case study system is enriched. The cash flow forecast and liquidity data are matched with up-to-date company information provided by a third party, which has compiled publicly available information of the companies from different sources like The Business Information System (YTJ), Finnish Register of Associations and Finnish Trade Register. These sources are all governed by Finnish Tax Administration. Finnish Patent and Registration Office is also co-governing The Business Information System.

Anonymity and respecting the companies' privacy are topics that are taken very seriously in the case study system and in this study, which is why anonymization is an important step in the preprocessing of the data. Anonymization of the companies makes sure that the cash flow forecasts or the liquidity information cannot be traced to the company, as only the big picture is of interest in the context of this study. In other words, the main focus is on the overall performance of the forecasts and the possible common factors that contribute to liquidity during the second spring of COVID-19 restrictions in Finland.

5.3 Variables and scaling

The variables used are listed in Table 1. Since the categorical data is converted to appropriate number of binary columns, the principal component method ends up with many more columns than described in the table. On top of the six quantifiable variables, 16 additional columns from Standard Industrial Classification level one and 18 additional columns representing the regions of Finland are acquired, resulting in 40 columns. In further calculations this number can be different, depending on the levels of the Standard Industrial Classification that are being included.

Even though the FAMD-algorithm handles the centering and standardizing of the data before the calculations, there are few variables that need to be scaled unconventionally. In order to emphasize the difference between the individual bins representing the different classes of personnel and turnover, new values were calculated to represent the bins better. The original discrete ordinal values were replaced with the mean values of each bin. The resulting turnover values were then used to scale the variables acquired from the case study system. In this way, the effect of the difference between the forecast and realized value or development of liquidity can be better highlighted. Without scaling, one could not tell how the changes affect the companies. One thousand Euro difference can be considered a minor miscalculation for a bigger company, while the same amount can be devastating for a smaller company.

5.4 Filtering and outlier detection

In order to protect the anonymity of the companies, the distribution of the companies was examined. As well as companies with missing or unknown values, also underrepresented categories were excluded from the calculations. This meant that the region of Ahvenanmaa, companies founded before 1960 and certain underrepresented standard industrial classifications and personnel sizes were excluded.

As liquidity and cash flows, not to mention their scaled counterparts, are computed continuous values, they are bound to be distributions of different values. These distributions have peaks around zero and tails that extend far out to both ways, with a number of extreme values that might distort the results one way or another. In order to minimize the effect of the extreme values, they can be identified as outliers by utilizing the empirical rule. The distributions are standardized by calculating the z-score

$$z = \frac{x - \mu}{\sigma}, \tag{12}$$

where x is the sample value, μ is the sample mean and σ is the sample standard deviation. From these standardized distributions it can be concluded, that any values that are over three times the standard deviations away from the mean are outliers. In the case of standardized distribution that means any values under -3 or over 3.

The example standardized distributions on a logarithmic scale are depicted in Figure 2, for development of liquidity, Figure 3, for the difference between one week forecast and realized values, and Figure 4, for the difference between two week forecast and realized values. Outliers have been already removed in these distributions.

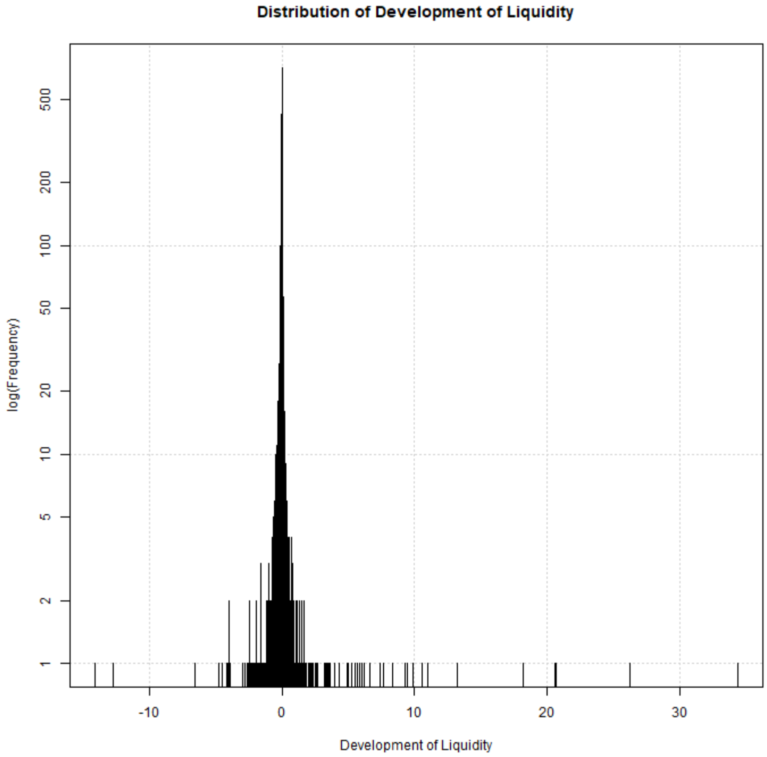

Figure 2. Standardized distribution of development of liquidity of the variables included in calculations for week 6 on a logarithmic scale

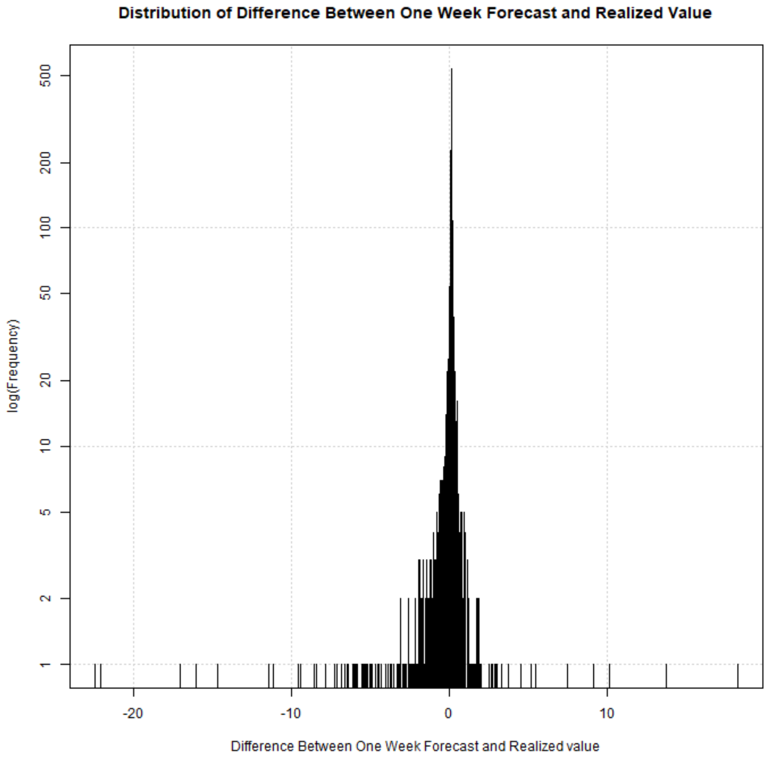

Figure 3. Standardized distribution of difference between the one week forecast and realized value of the variables included in calculations for week 6 on a logarithmic scale

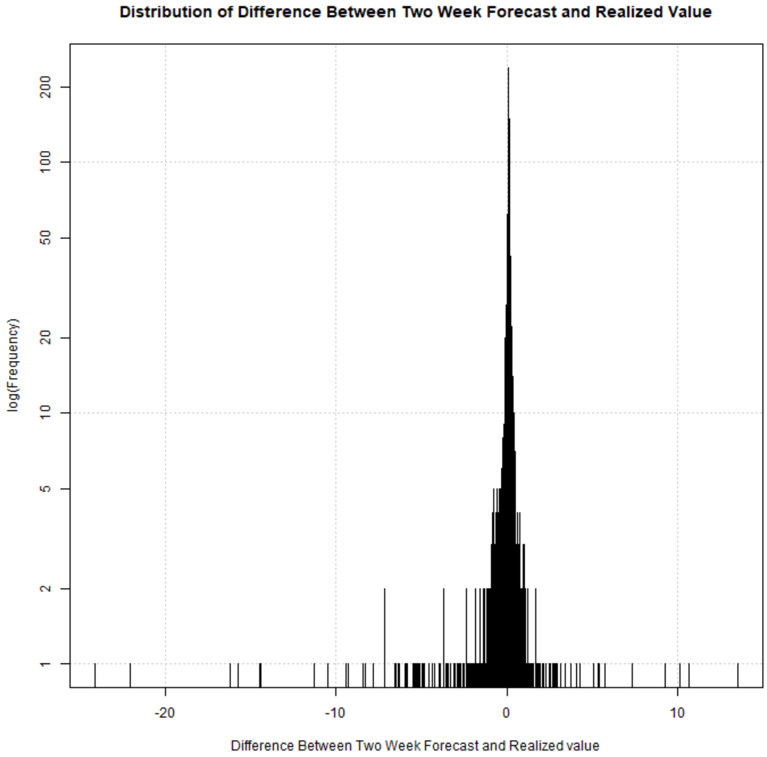

Figure 4. Standardized distribution of difference between the two week forecast and realized value of the variables included in calculations for week 6 on a logarithmic scale

Table 1. The data variables

Variable Name	Data Type	Description
Development of Liquidity	Real, Continuous	Development of liquidity from the previous week in proportion to the turnover class of the company
Difference Between One Week Forecast and Realized Liquidity	Real, Continuous	Difference between the forecast value from one week ago and the actual realized value in proportion to the turnover class of the company
Difference Between Two Week Forecast and Realized Liquidity	Real, Continuous	Difference between the forecast value from two weeks ago and the actual realized value in proportion to the turnover class of the company
Personnel Class with Scaled Proportions	Real, Continuous	Original personnel classes were distributed in the following bins marked by numbers from one to seven: 1=0-4 employees, 2=5-9, 3=10-19, 4=20-49, 5=50-99, 6=100-249, 7=250-499. These were scaled to match the proportions by using the mean of each bin.
Turnover Class with Scaled Proportions	Real, Continuous	Original turnover classes were distributed in the following bins marked by numbers from one to nine: 1 = not known, 3 = 1-199k, 4 = 200-399k, 5 = 400-999k, 6 = 1000-1999k, 7 = 2000-9999k, 8 = 10000-19999k, 9 = 20000k or more. These were scaled similarly to the personnel class values.
Founding decade	Integer, Discrete	Founding decade of the company. Companies founded in the same decade are grouped together. For example, the companies founded in years 1960-1959 are assigned the decade 1960.
Region	String, Categorical	Regions of Finland. Consists of 19 different regions of which 18 were utilized.
Standard Industrial Classification Level One	String, Categorical	The broadest level of the division. Consists of 22 classes of which 16 are utilized.
Standard Industrial Classification Level Two	String, Categorical	The companies are divided in more detail, consists of 80 classes. Only used in the industry specific calculations.
Standard Industrial Classification Level Three	String, Categorical	The companies are divided in even more detail, consists of 216 classes. Only used in some of the industry specific calculations.

6 RESULTS

The analysis is done in three parts. First, the data is being examined on a weekly basis, for three different weeks. Then, the same calculations are done with the company's industrial classification provided as a supplementary value. Lastly, smaller sets, which are made up of only one industry, are examined.

6.1 All companies

All companies were examined on a weekly basis, with an interest in finding out if any correlation exists between the line of business, the size of the company or the turnover class and the predicted and realized liquidity. Even though the calculations for each week were done separately, the results of different weekly snapshots did not differ significantly.

6.1.1 Performing FAMD on the data

After performing FAMD on the data, the companies can be examined on the first two dimensions. Since the two dimensions only explain, depending on the week, approximately 9.6-11% of the variance, the observations seem to be quite homogeneous and one cannot really point out clearly separated clusters using two-dimensional visualization as can be seen from Figure 5. Identifiable clusters can still be found by taking into account more than just the first two principal components with HCPC.

In order to figure out the appropriate number of principal components, the scree-plot and eigenvalues are examined. Looking at the scree-plots in Figure 6, it can be stated that the elbow points do not provide very satisfying options for the cut-off points. The elbow-points can be identified too early at the start and just before the last dimensions. The first few dimensions do not explain the variance of the data sufficiently, whereas discarding only a couple of the last dimensions does not fully utilize the advantages of the dimension reduction technique. This is why, in order to decide on the number of dimensions, those, whose eigenvalue are over one, are examined. Depending on the week, this is approximately 18 dimensions. Unfortunately, this method does not provide very satisfactory results, as these dimensions account only for approximately 56% of the variance. This still leaves almost half of the variance outside. Lowering the threshold from 1 to 0.95, approximately 26 dimensions and 77% of the explained variance is being retained. This

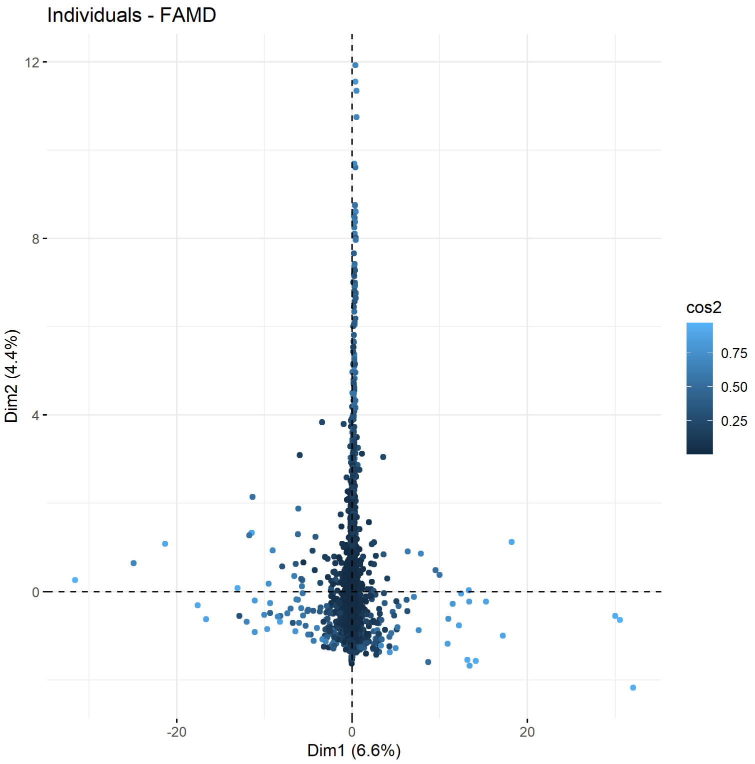

Figure 5. Individuals plotted in the first two dimensions in week 14. Since the dimensions only explain 11% of inertia, clusters are not identifiable on a two-dimensional plane.

amount of explained variance seems satisfactory.

Overall, the results of the principal component analysis from each week are quite similar, with the greatest differences being in the first dimension and towards the end, in the last dimensions. These differences indicate that the values, that change from week to week, like the difference of the liquidity from last week as well as the difference of the forecast and the realized liquidity, have an impact to the first and last dimensions. This can be further proved in the correlation circle, presented in Figure 7, where the difference between the forecast and the realized value for both one and two weeks have the highest quality of representation in the first dimensions. Difference to last week's liquidity is not of as high quality, but it is negatively correlated to the difference of the forecasts and the realized liquidity. This has quite straightforward explanation: if the difference from last week is negative, the forecast for that week has probably been too positive. The quality of representation for the development of liquidity from last week is interestingly also the most prone to change from week to week.

Turnover class and Personnel class are clearly strongly correlated, and very well explained by the second dimension. The decade in which the company was founded is not as well depicted on the plane, but since there are so many individuals, the quality of represen-

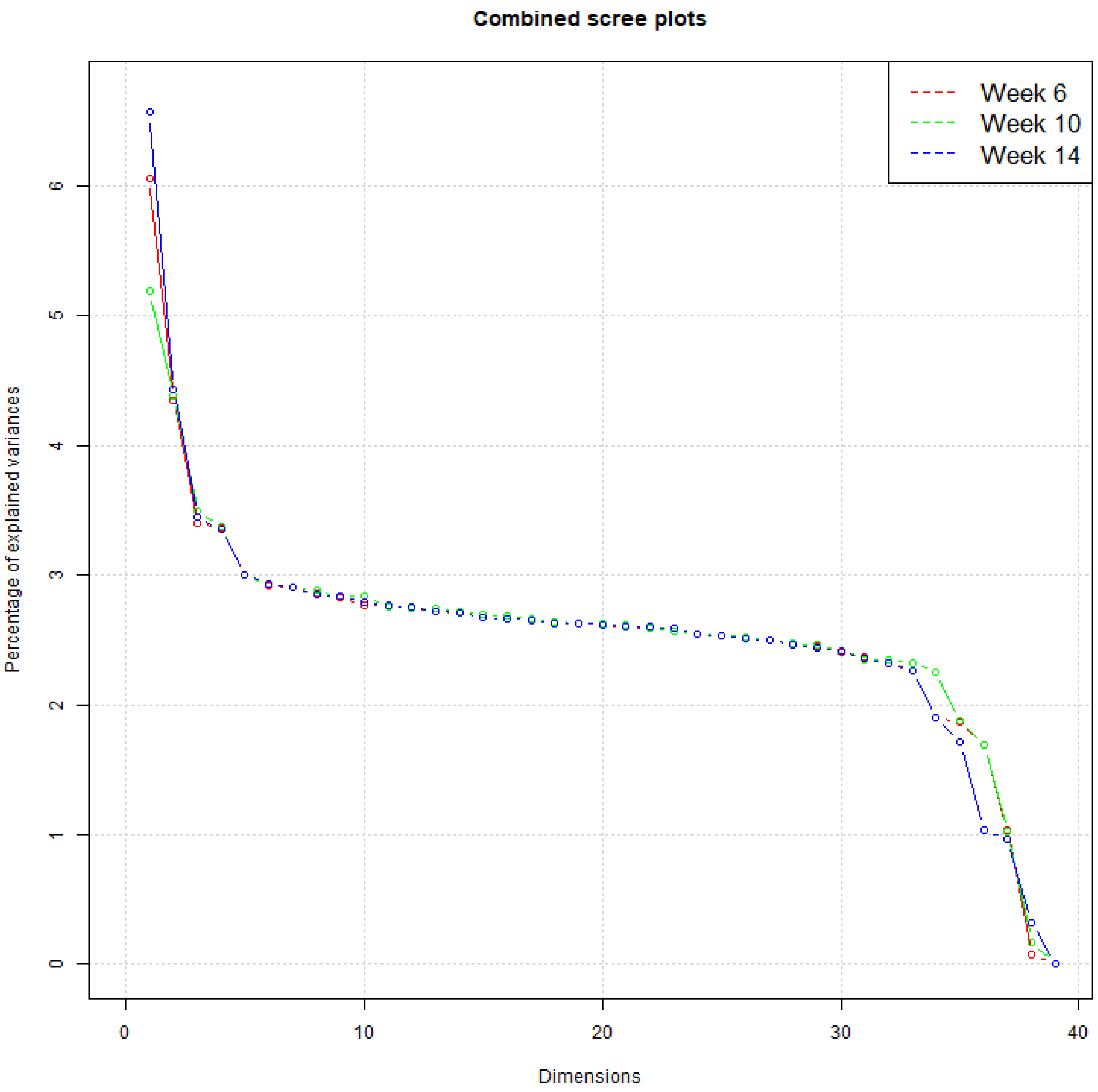

Figure 6. FAMD produced very similar results for the scree-plots for different weeks. With the greatest differences being in the first dimension and towards to the end, which can be explained by the weekly noise.

tation of 0.2 is still quite significant. The founding decade is negatively correlated with personnel and turnover class, which is logical since the companies that were founded earlier tend to have bigger turnover or personnel size.

Overall, the direction of the variables change very slightly from week to week. The biggest change is in the change of the liquidity from the previous week. It is important to remember that only about 10% of variance is explained by these axes.

As can be seen from Figure 8, the qualitative variables have more discrepancy from week to week, while the values of cos^2 are very weak, therefore, too much speculation should

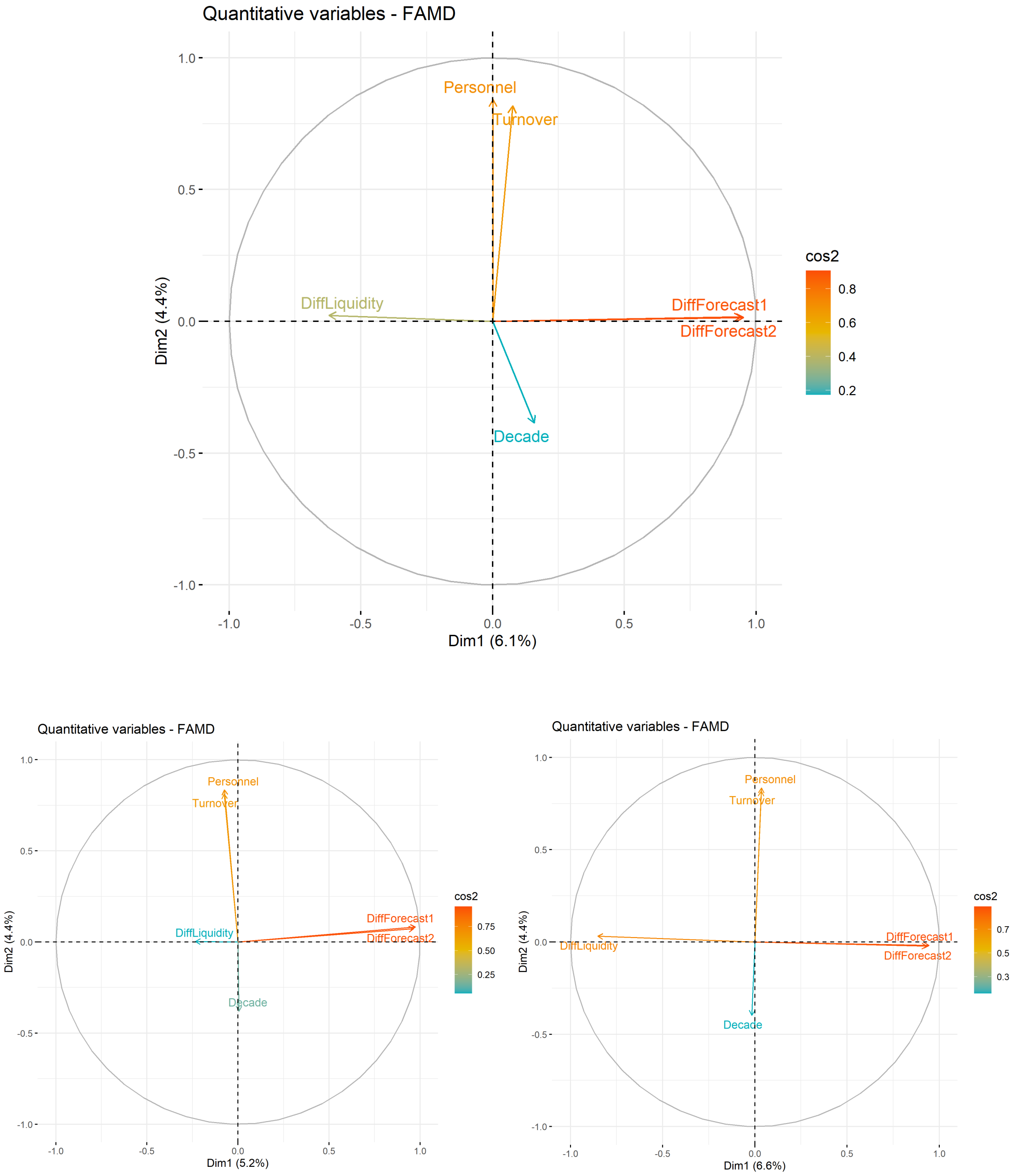

Figure 7. Correlation circles for weeks 6, 10 and 14. For each week, the differences between the forecasts and the realized value have the highest quality of representation and correlate strongly. Development of the company's liquidity has the most differences with the cos^2-value from week to week and it is negatively correlated to the forecast differences. Turnover class and personnel class are highly correlated and contribute highly to the second dimension.

not be done based on the figure. The most contributing factors in these dimensions are 'other service activities', 'professional, scientific and technical activities', 'manufacturing', 'transportation and storage' and the region of Pohjois-Karjala. Towards the upper part of the figure the individuals are more correlated with primary production and 'manufacturing’ as well as ’transportation and storage’ and ’education’. The regions associated with this direction are Pohjois-Karjala, Varsinais-Suomi, Pohjois-Pohjanmaa, Satakunta and Kainuu to name a few. In the lower part of the figure industries like ’professional, scientific and technical activities’, ’financial and insurance activities’ and ’information and communication’ as well as regions such as Keski-Pohjanmaa and Kymeenlaakso can be found.

It is crucial to keep in mind when analysing the results in two dimensions, that the first principal dimensions only explain approximately 10% of the variance, so as not to draw too generalized conclusions. The low percentage of the variance explained is common for MCA results, but it also can imply that the data is too volatile for this kind of analysis or that the right variables are not selected for the analysis. The variables and their degree of association between all of the dimensions can be examined with the help of corrplot-package. As can be seen from Figure 9, the difference in the forecasts and realized liquidity are clearly the dominating variables on the first dimension with difference in liquidity from the previous week having a different impact from week to week on the first dimension. The personnel class affects the second dimension the most, closely followed by the turnover class and further back the founding decade The categorical values, the line of business and the region, seem to have weaker quality of representation on multiple dimensions.

Stronger representation of the variables on the last dimensions imply that these variables introduce a lot of noise to the data. If the variables have stronger presence towards the last dimensions, it might be worth trying to analyze the data without them. Biggest accounts of noise are from founding decade, difference in liquidity from previous week and personnel and turnover class. None of them are too centered towards the last dimensions.

6.1.2 HCPC on the data

The results from FAMD are fed to the HCPC-algorithm which divides the data, depending on the week, into 26-32 clusters. The maximum number of clusters is set to be half of the number of individuals and the algorithm uses Equation (11) to determine the optimal number of clusters. Also, a number of initial number of clusters for pre-clustering with K-means is given as parameter for the algorithm. This is set to be 50, as that is significantly bigger than the number of clusters that the HCPC-algorithm usually provides and this number of clusters still produces clear dendrogram plots. Visually, the clusters are not identifiable on a two-dimensional plane, as they are mostly on top each other. As the first

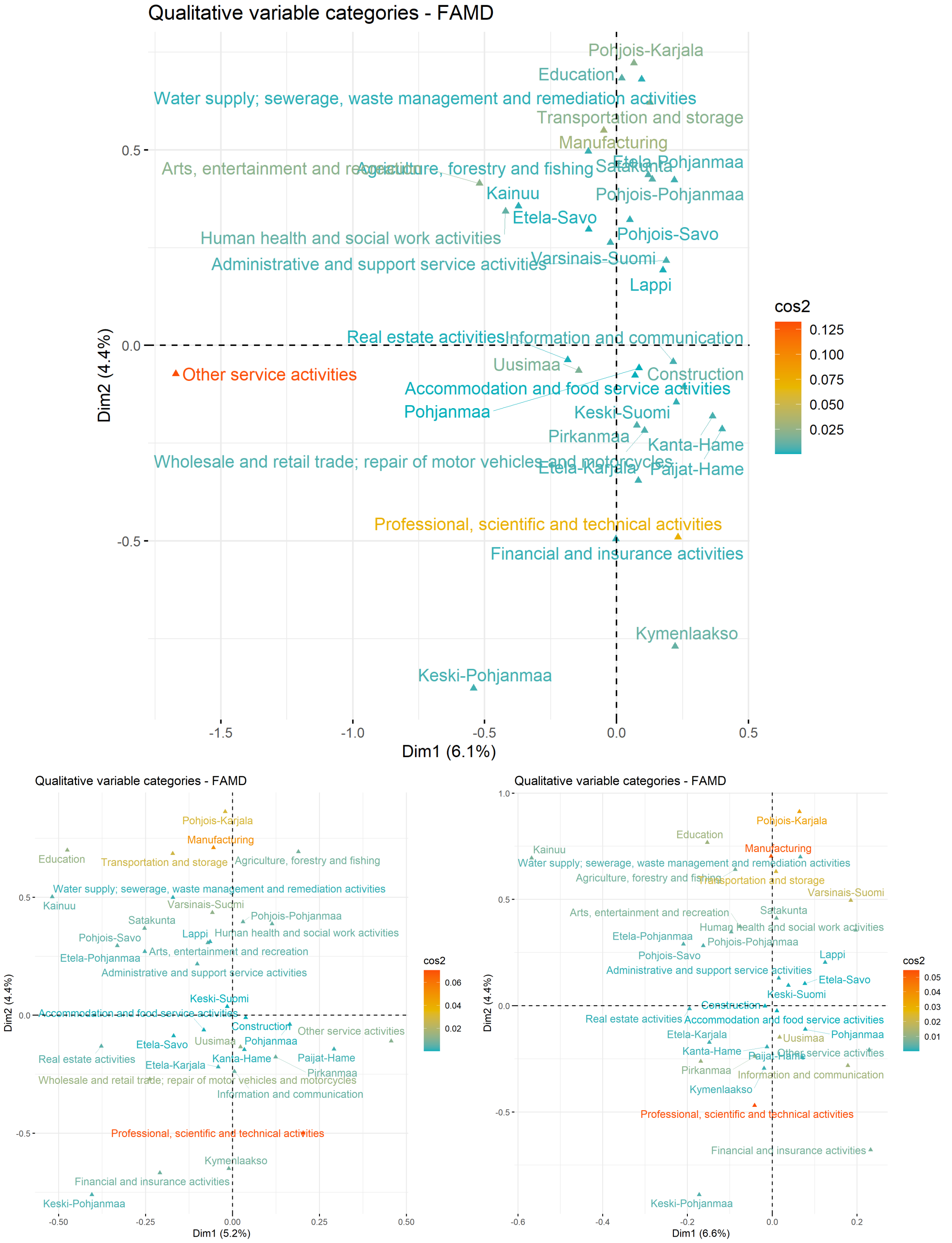

Figure 8. Qualitative variables of weeks 6, 10 and 14. The quality of representation varies from week to week but is ultimately quite weak.

and second dimensions only explain 5-7% and 4.4% of the variance respectively, it is not too surprising that the clusters cannot be easily identified. Some of the functions, that offer visualizations of the clusters, do not work if pre-clustering is done with K-means. This is not too troubling, since even if the figures could be extracted, they would not offer

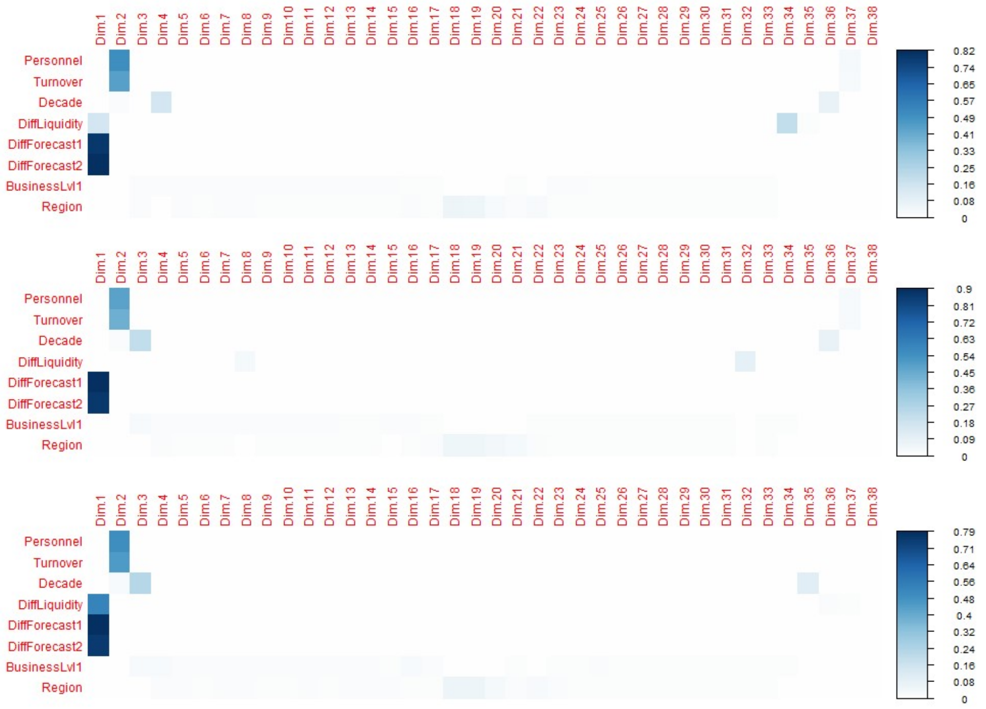

Figure 9. Visualized quality of representation on all the dimensions for weeks 6, 10 and 14.

much information as the explained variance is so low in these dimensions.

As can be seen from Figure 10, the sizes of the clusters vary quite much from week to week. Depending on the week, the balance of the distribution of the companies to a cluster may be different and there are even a few significantly bigger clusters. Here, the determination of the best number of clusters is left to the algorithm, but the distributions could be more balanced with constraints like minimum amount of clusters. The relationships of the different clusters can be understood from dendrogram plots in Figures 11, 12 and 13. The bigger clusters can be identified from the dendrograms as well and it does seem that increasing the number of clusters would produce more balanced distribution.

More interesting, however, are the variables associated with these clusters, as one of the objectives is to identify common factors that the companies have in terms of forecast and development of liquidity. The variables related to the development of liquidity and forecasts are most prominent in week 6 of all the weeks that are being examined. For example, clusters 1 and 2 have a much greater negative difference between the forecast and realized values than that of the average. Also the development of liquidity is much better than the average. These clusters are both representing older companies and are defined by the industry of ’other service activities’. Additionally, the first cluster is equally represented by ’arts, entertainment and recreation’ industry, while the 82% of the individuals in the

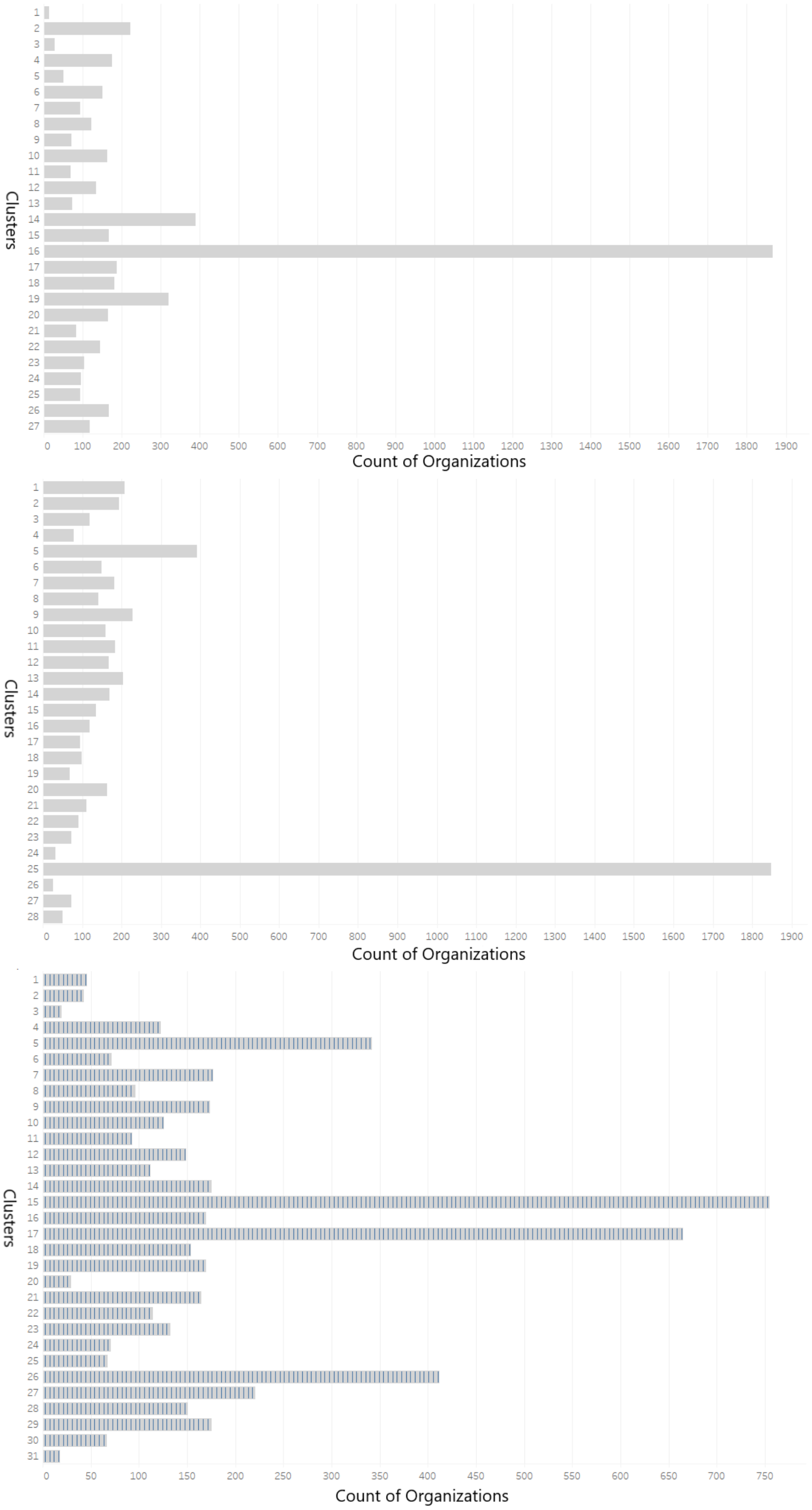

Figure 10. Distributions of individuals in clusters for weeks 6, 10 and 14.

second cluster lie in the region of Uusimaa.

Ninth cluster has bigger positive difference between one week forecast and the realized liquidity than the average. All of the companies in this cluster are in the industry of 'finan-

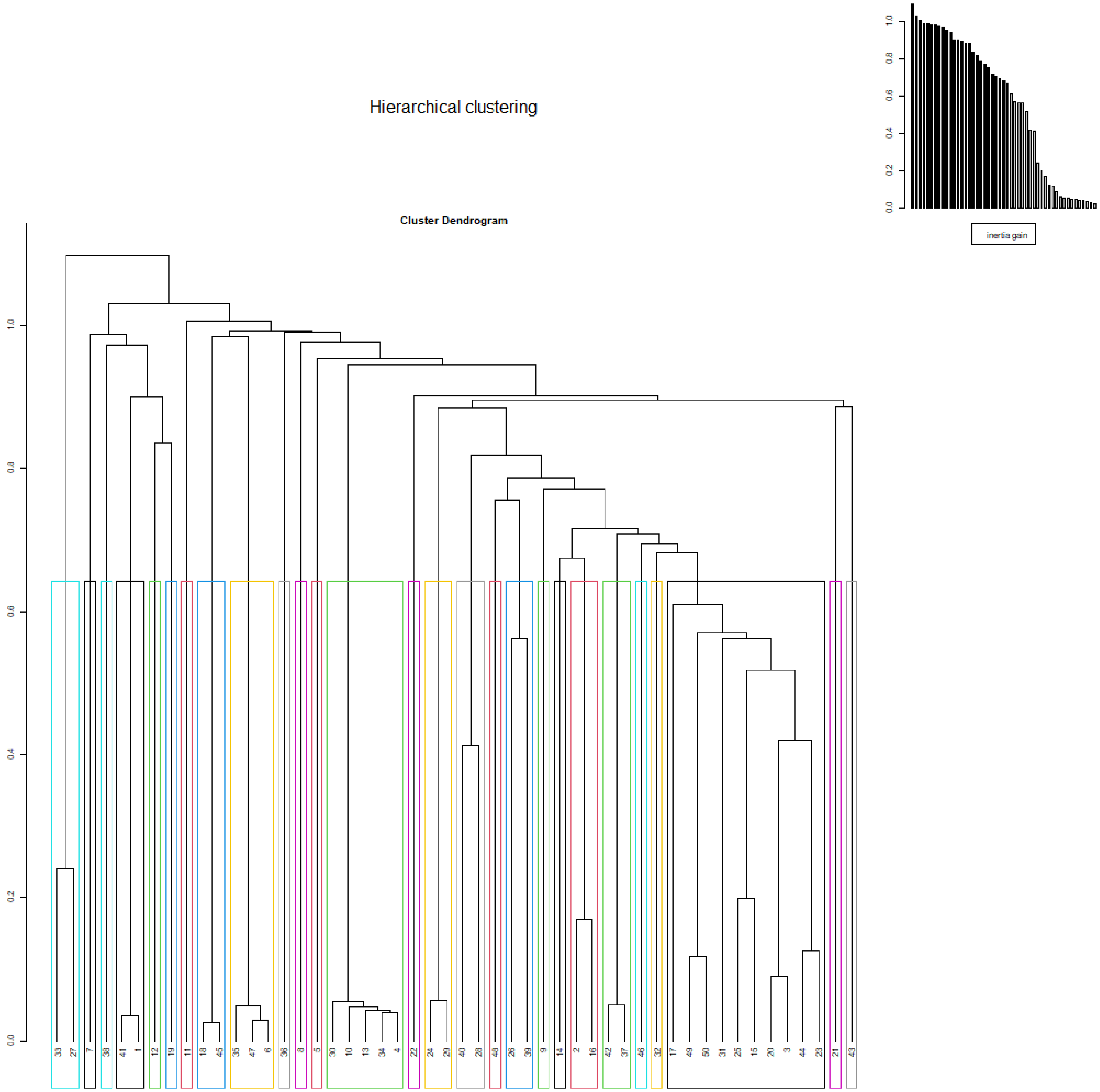

Figure 11. Dendrogram produced by the hierarchical clustering that has been performed on the K-means pre-clustering results for week 6. Top right one can see the inertia gain for each additional cluster.

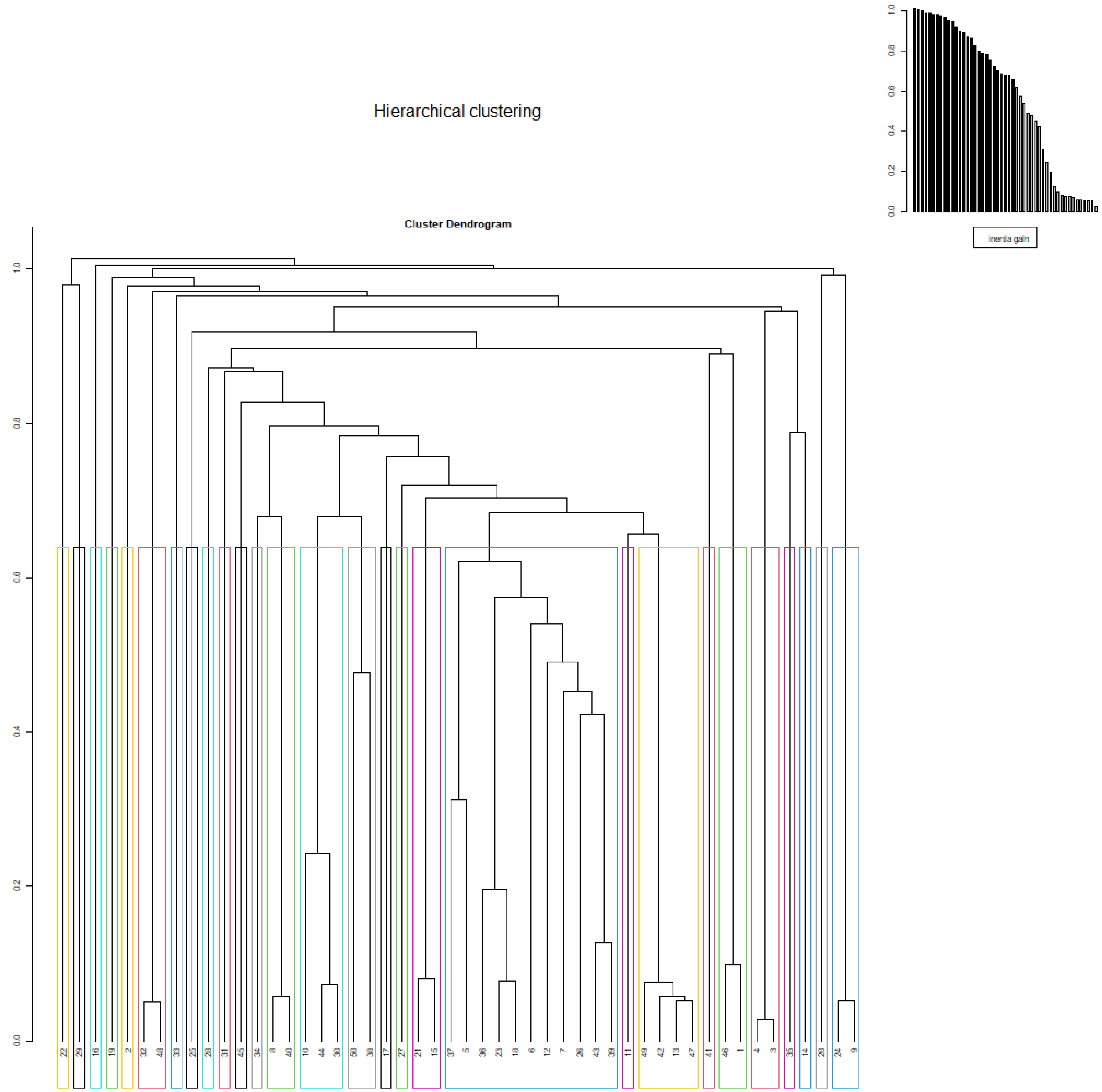

Figure 12. Dendrogram produced by the hierarchical clustering that has been performed on the K-means pre-clustering results for week 10. Top right one can see the inertia gain for each additional cluster.

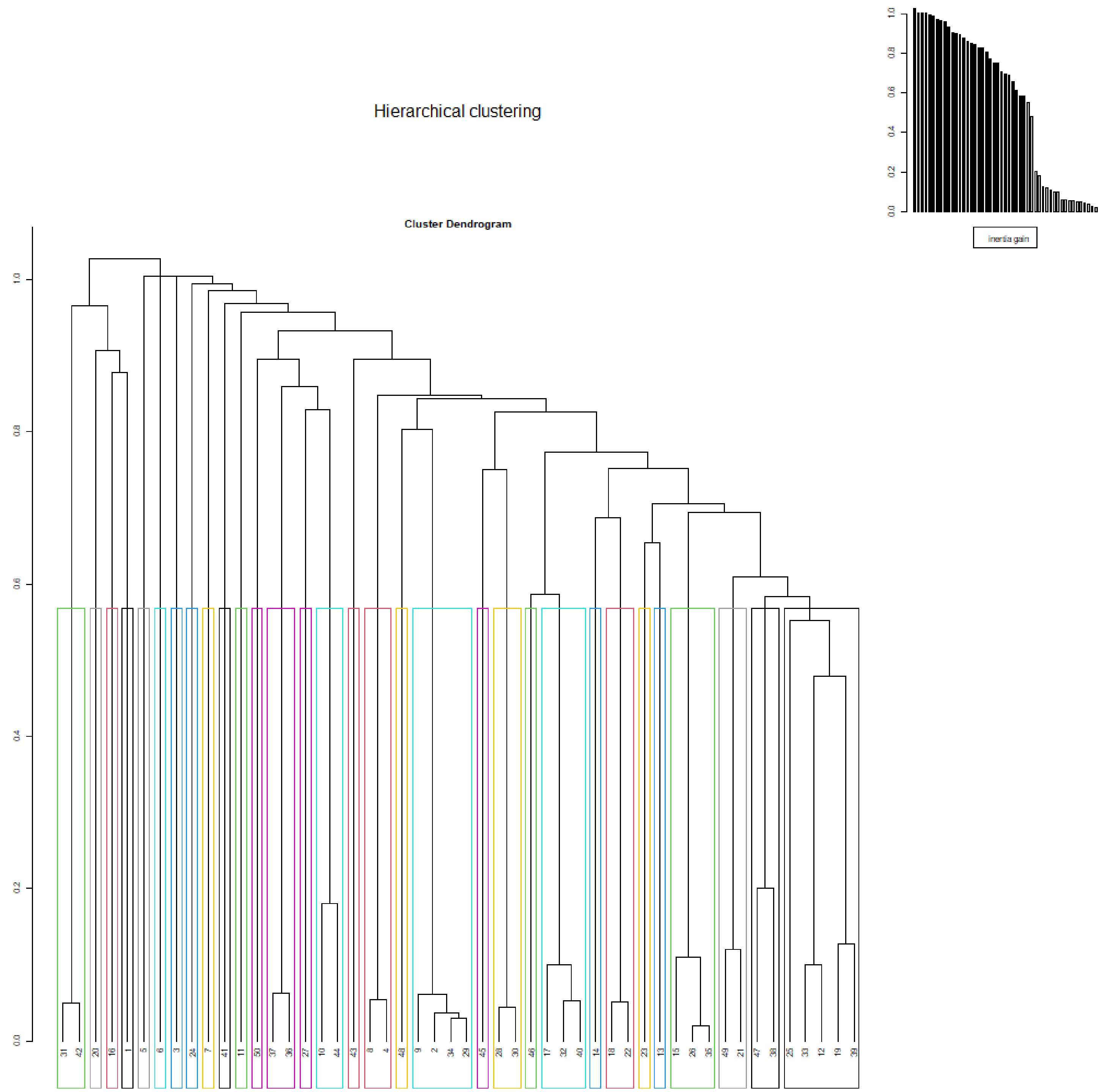

Figure 13. Dendrogram produced by the hierarchical clustering that has been performed on the K-means pre-clustering results for week 14. Top right one can see the inertia gain for each additional cluster.

cial and insurance activities’ and almost two thirds of them are in the region of Uusimaa. The cluster numbers 14 and 16 have differences between the forecast and realized values more close to zero than the sample average. In other words, the companies in these clusters have, on average, more precise forecasts. Companies in cluster 16 have also slightly worse development of the liquidity from the previous week. The 14th cluster consists only of companies in the industry of ’information and communication’ of which 87% are in the Region of Uusimaa. Cluster 16 is more of a mixed cluster since it has companies from industries like ’professional, scientific and technical activities’, ’wholesale and retail trade; repair of motor vehicles and motorcycles’, ’administrative and support service activities’, ’construction’ and ’manufacturing’ and the represented regions were the regions with the biggest cities: Uusimaa, Pirkanmaa and Varsinais-Suomi. As can be seen from Figure 10, cluster 16 is a cluster that has significantly more individuals than other clusters, which explains why so many different industries and regions are represented. Cluster number 26 has slightly too positive forecast for two weeks. This cluster consists of companies in Kanta-Häme with various industries like ’human health and social work activities’, ’administrative and support service activities’ and ’manufacturing’. Unfortunately, there are no more clusters associated with the variables linked to the liquidity and forecasts for week 6.

On the first of March 2021, due to the worsened COVID-19 situation in Finland, State of Emergency was declared by the Government (Government Communications Department 2021). A number of restrictions and recommendations were applied to companies, especially restaurants and other food and beverage service businesses and indoor sports facilities in the Uusimaa region. Looking at the clusters from week 10, it can be seen that there are not as many that have correspondence over the development of liquidity or forecasts. This indicates that there is no nationwide trend with the development of companies’ cash flow, but each company faces their own challenges and some fare better than others even when being within the same industry, region or size groups. Some similarities can be pointed out, however, like that the first cluster has too positive forecasts and that cluster number 2 has older companies that have had negative development of liquidity and forecast for two weeks has been moderate. Both of these clusters’ individuals are almost all from the region of Uusimaa with the dominating industries being ’other service activities’ and ’arts, entertainment and recreation’, respectively. Cluster 25 has had slightly better than average development in liquidity and better results than forecast. This is a cluster with considerable amount of individuals and so there are multiple regions and industries represented. Region-wise, most of the companies in this cluster are from Uusimaa and Pirkanmaa. The industries represented include ’professional, scientific and technical activities’, ’wholesale and retail trade; repair of motor vehicles and motorcycles’, ’construction’, ’administrative and support service activities’, ’transportation and storage’ and ’manufacturing’.

On week 14 of year 2021, there were even stricter restrictions in place and in the regions in the community transmission phase, as private commercial premises were forced to closed for three weeks, starting from the first of April (Eduskunta, Parliament of Finland 2021). For this week, there are even fewer clusters that can be described by the development of liquidity or the difference of forecasts and realized value. First cluster of this week has had a positive development in the liquidity from the previous week and the forecasts for this week have clearly been too moderate. The companies in this cluster are mostly from the region of Uusimaa, and ’other service activities’ and ’arts, entertainment and recreation’ industries make up approximately half of the companies. The other half does not provide industries with statistically significant number of companies. Cluster 31 is clearly the exact opposite to the first cluster. The development of liquidity has been negative, the forecasts have been way too positive and the cluster consists mostly of ’other service activities’ in the region of Uusimaa.

6.2 Supplementary variable

As the results from the analysis on the whole data set did not identify many of the clusters with their similarity of the development of liquidity or forecast difference, the calculations are performed again, only this time the Standard Industrial Classification is provided as a supplementary variable. The distributions of companies in clusters can be seen from Figure 14.

In week 6, cluster 1 has very positive development of liquidity. The cluster is made up of companies in the industry of ’other service activities’, ’arts, entertainment and recreation’, ’professional, scientific and technical activities’ and ’construction’. This cluster is quite small, though, and only 0.1-6% of all the companies in these industries belong to this class. Other common factors between the companies are that they have smaller turnover size and they are slightly older than the whole sample average. Cluster 2 has the clear majority of ’other service activities’ companies, 92% of all the companies in that industry. Their forecasts have been too modest, but development of liquidity does not appear to have common direction, as it does not appear on the list. The companies in cluster 2 have slightly bigger turnover class than the companies in cluster 1. Cluster 16 is clearly dedicated to the companies in ’arts, entertainment and recreation’. These companies seem to have, on average, slightly negative development in liquidity. Cluster 17,

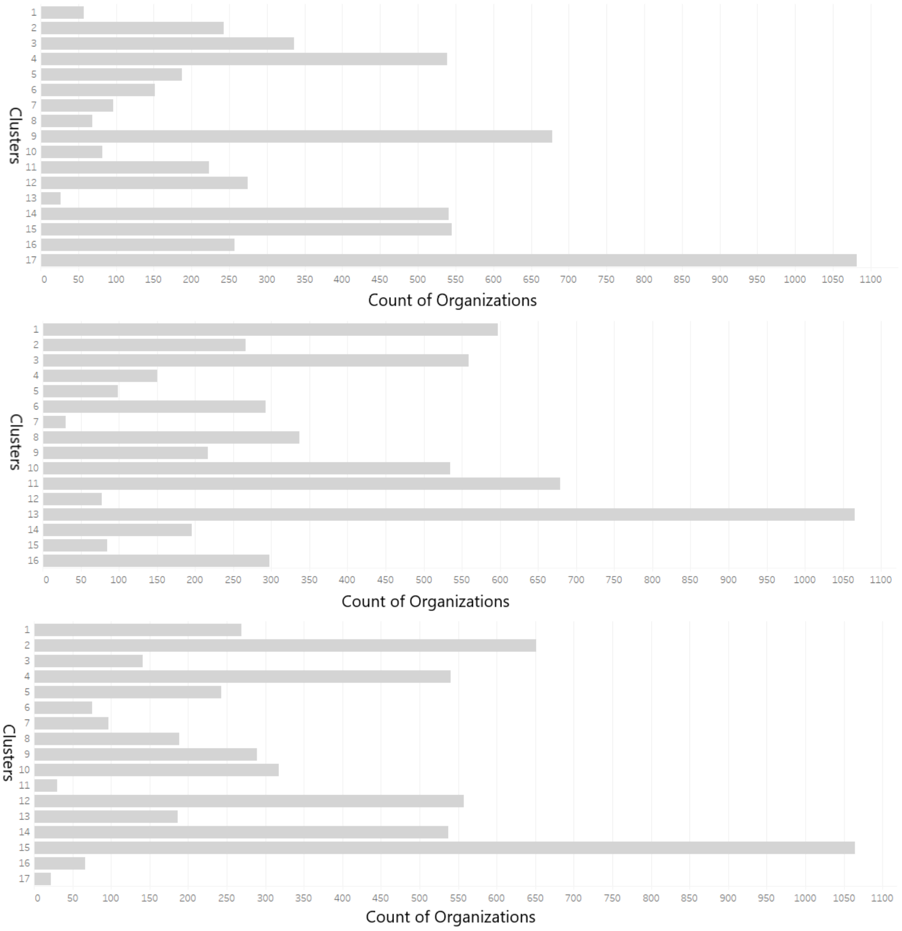

Figure 14. Distributions of individuals in clusters for weeks 6, 10 and 14. Calculations have been done by providing companies' industry as a supplementary value.

with almost all of the companies in 'professional, scientific and technical activities' industry, has slightly negative development of liquidity. 'Information and communication' in number of different regions and 'construction' in Uusimaa, Pohjanmaa, Satakunta, Etelä-Karjala, Keski-Suomi and Päijät-Häme have very accurate forecasts in clusters 9 and 15. These clusters are also newer than the average, founded on average around 2004.

For weeks 10 and 14, there are considerably less clusters and subsequently less clusters, that can be explained by the development of liquidity and the difference between forecasts and realized values. On week 10, almost all the companies in the industry of 'wholesale and retail trade; repair of motor vehicles and motorcycles' are in the first clusters. They are described by slightly positive development of liquidity and too moderate forecasts. Same type of cluster, cluster number 2, is formed for almost all the companies in the industry of 'arts, entertainment and recreation', although they have, on average, slightly negative development of liquidity. Interestingly, cluster 16 is formed by 94% of 'other service activities', but this cluster has clear negative development of liquidity and, complementing that, too positive forecasts. In week 14, there are identified only the positive development for the industry of ’wholesale and retail trade; repair of motor vehicles and motorcycles’ and the negative development for 5% of the companies in the industry of ’other service activities’.

6.3 Smaller set of individuals

In order to better differentiate companies from each other liquidity- and forecast-wise, the set of individuals is limited to a certain industry. The different levels of the industry classes make it possible to divide the companies even further and describe the clusters based on the standard industrial classification level that is more specific. Based on the number of the individuals in each industry examined and the number of the subcategories in that industry, a different level of detail in the form of level of industrial classification is selected.

First, some of the industries that have a significant impact on Finland’s economy are examined. ’Construction’, ’manufacturing’ and ’accommodation and food service activities’ were identified as such. Unfortunately, results do not improve by examining only one industry and depending on the week, only one or two clusters have statistical significance on the variables explaining liquidity or forecast difference. In week 14, for example, when examining only ’construction’ industry, a cluster is made up mostly of companies in the industry of ’construction of roads and railways’ is identified to have too positive forecast and negative development of liquidity.

Lastly, the industry of ’other service activities’ is examined, as it is most prominent in the analysis of the whole data set. Nonetheless, this does not translate to the deeper level of industry separation. The methods identify 8-15 clusters of which only 3-6 clusters have statistically significant representation of the development of liquidity and difference in forecasts. Week 6, for example, has the total of eight clusters of which three have commonalities with the liquidity and forecast variables. The distribution of individuals in clusters for that week is depicted on Figure 15. The first is company size-, age- and industry-wise unexplained cluster with too modest prediction and positive liquidity. The second cluster has slightly negative development for the industry of ’activities of business, employers and professional membership organisations’. The fifth cluster of the eight has smaller than average companies in the industry of ’activities of other membership organisations’. These companies have had quite accurate forecast but ever so slightly negative development in liquidity.

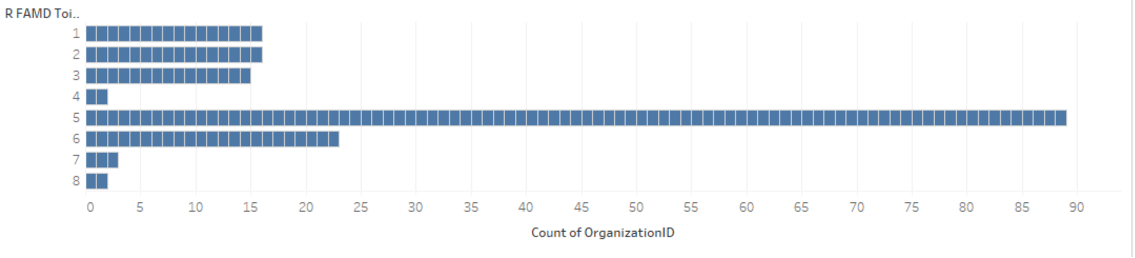

Figure 15. Distributions of individuals in clusters for week 6 for companies in the industry of ’other service activities’.

7 DISCUSSION

As can be seen from the results, it is extremely difficult to categorize companies based on the liquidity or the difference between the forecast and realized values of cash flows. There are many factors that cannot be explained by simply looking at the company sizes or the lines of business. Even seemingly identical companies might have widely different balances or massive difference between the forecast and realized value, for example, due to recently acquiring or losing a big customers. Differences in strategy, budget or management styles takes companies to a different direction. The CEO, the CFO and the board have their unique visions and plans in which way to grow the business, which can impact the liquidity directly. Mergers and acquisitions as well as loans might have a large temporary impact on the liquidity and users might not manually input values that are hard to forecast. Also, some aspects of running a business are very hard to turn into variables. How can the suitability of the business spaces, user experience of the company's website or software and innovativeness of initiatives be measured and compared? There might be a very busy street with lots of good business spaces, but just around the corner or directly above the valuable spaces, might be premises that customers do not find very often. These kinds of variables along with pure luck can explain why seemingly two very similar companies might have widely different outcomes.

Although some things are challenging to compare, there are multiple values that could improve the calculations. If more values indicating the performance of business were provided, the results could be more informative. It would be interesting to see the relationships and effects of different cash flow -based business performance measures like operating cash flow margin, quick ratio, sales growth and other variables as suggested by (Bhandari et al. 2013).

When analyzing the whole data set, qualitative values seem to be dominating the way the clusters form, since there are a lot of significant categories, but all of the clusters do not have significant quantitative variables. This is unfortunate, since the partition of quantitative values were of great interest, but they seem to be left in categorical variables' shadows. Setting the minimum number of clusters way over 30, which seems to be the point where the tree is automatically cut, different variables, including the quantitative ones, might be of greater importance in forming clusters. This might produce clusters that are more specific, but also drawing conclusions on the macro-level might be harder as the plots would become cluttered and the relationships between the clusters would be harder to follow.

Providing the standard industrial classification as a supplementary value seems to provide the most interesting results, as quite many industries are correctly grouped together and the average direction of the development of liquidity can be pointed out. It would be interesting to see, if there is a same effect if also the region was provided as a supplementary value. Then FAMD could be switched out for PCA.

The limited time frame of the data accumulation and the study also placed restrictions what could be studied. Liquidity's nature is very volatile and it can have daily, weekly, monthly, yearly or otherwise reoccurring trends. Smoothing the data and looking at the development for example on monthly-basis might give a better representation of the overall development of liquidity. The methods are not a "one size fits all" and in this volatile and large data set they might not be the best pick. If the data is accumulated more, it would be interesting to apply time series analysis to see if a development of liquidity is typical for some of the industrial classes.

One aspect that might affect the results too is that the times are very uncertain due to COVID-19. During this study, restrictions and recommendations for companies varied widely from few recommendations and restrictions to closing the premises of a number of companies. The effects of COVID-19 are surely rippling through supply chains and might even cause unnoticeable domino effects. This situation emphasizes the differences between companies. Similar companies might have different reactions to restrictions and recommendations, and the companies' ability to innovate and adapt to rapid changes play a key role in the companies' survival and success.

8 CONCLUSION

The aim of this study was to examine the development of liquidity and its forecast to see whether there were variations that could be explained by the line of industry, location, turnover size, personnel size or the age of the company or any combination of these. Factor Analysis of Mixed Data and Hierarchical Clustering on Principal Components were the methods chosen for this purpose. FAMD was selected as the principal component method due to its ability to combine quantitative and qualitative variables. HCPC was used as the clustering method, since it was based on inertia gain and offered the use of multiple principal components in clustering. The calculations were done with FactoMineR-package, that has been developed for R. Analysis was done on week-level on the whole data set, whole data set with industry provided as a supplementary value and on specific industries. Unfortunately, the results were not quite satisfactory. The descriptions of the clusters did not provide generalizable insights for the common factors in the development of the liquidity or the accuracy of the forecasts. Performing analysis with the supplementary values provided more clusters with statistical significance in the development of liquidity and forecast difference, but the values associated with these clusters were mixed, and no conclusions on the common factors between the companies that are succeeding and those that are struggling could have been properly drawn.

Examining the cash flow forecasts in this short time span and with these features does not seem to provide cohesive results. The companies' liquidity can experience quite volatile changes on a week-to-week basis which cannot be reasonably smoothed out with only couple of months worth of weeks. Furthermore, with at least a year's worth of data possible reoccurring monthly trends and seasonality in the data could be identified. The longer period of data could be used in the calculation of the future cash flow predictions. Also there are a number of different factors affecting liquidity, some of which are very challenging to incorporate in this kind of analysis. It is also possible, that analysis done on the same set of companies, but in time period prior to the COVID-19 pandemic, would have brought different results.

References

Abdi, Hervé and Dominique Valentin (Jan. 2007). “Multiple Correspondence Analysis”. In: *Encyclopedia of Measurement and Statistics*.

Abdi, Hervé and Lynne J. Williams (2010). “Principal component analysis”. In: *WIREs Computational Statistics* 2.4, pp. 433–459. DOI: https://doi.org/10.1002/wics.101.

Argüelles, M., C. Benavides, and I. Fernández (2014). “A new approach to the identification of regional clusters: hierarchical clustering on principal components.” In: *Applied Economics* 46.21, pp. 2511–2519. ISSN: 00036846.

Baarsch, Jonathan and M. Emre Celebi (Mar. 2012). “Investigation of Internal Validity Measures for K-Means Clustering”. In: *Lecture Notes in Engineering and Computer Science* 2195, pp. 471–476.

Bhandari, Shyam B. and Rajesh Iyer (2013). “Predicting business failure using cash flow statement based measures”. In: *Managerial Finance* 39.7, pp. 667–676.

Eduskunta, Parliament of Finland (Mar. 2021). *Asian käsittelytiedot HE 31/2021 vp, Hallituksen esitys eduskunnalle laiksi tartuntatautilain 58 g §:n muuttamisesta*. [Online; accessed May, 16, 2021]. URL: https://www.eduskunta.fi/FI/vaski/KasittelytiedotValtiopaivaasia/Sivut/HE_31+2021.aspx.

Fight, Andrew (2006). *Cash Flow Forecasting*. Essential Capital Markets. Butterworth-Heinemann. ISBN: 9780750661362.

Government Communications Department (Mar. 2021). *Finland declares a state of emergency*. [Online; accessed May, 16, 2021]. URL: https://valtioneuvosto.fi/en/-/10616/finland-declares-a-state-of-emergency.

Husson, F., J. Josse, and Pagès J (Sept. 2010). “Principal component methods - hierarchical clustering - partitional clustering: why would we need to choose for visualizing data?” In: *Technical Report - Agrocampus, Applied Mathematics Department*, pp. 1–17.

Husson, Francois, Sébastien Lê, and Jérôme Pagès (Apr. 2017). *Exploratory Multivariate Analysis by Example Using R*. ISBN: 9780429225437. DOI: 10.1201/b21874.

Kaufman, Margie (Aug. 2014). “Weekly Cash-Flow Analysis: Why Isn’t It a "Best Practice"?” In: *American Bankruptcy Institute Journal* 33.8, pp. 32–33, 88–89.

Lê, Sébastien, Julie Josse, and François Husson (2008). “FactoMineR: An R Package for Multivariate Analysis”. In: *Journal of Statistical Software, Articles* 25.1, pp. 1–18. ISSN: 1548-7660. DOI: 10.18637/jss.v025.i01.

Pagès, Jérôme (Nov. 2014). *Multiple Factor Analysis by Example Using R*, pp. 1–253. ISBN: 9780429171086. DOI: 10.1201/b17700.

Ryland, Philip (June 2020). “Accounting that counts: Go with the cash flow”. In: *Investors Chronicle*, p. 26.

The European Commission Directorate-General for Internal Market, Industry, Entrepreneurship and SMEs (2019). “2019 SBA Fact Sheet FINLAND”. In: URL: `https://ec.europa.eu/growth/smes/sme-strategy/performance-review_en`.

Wall, Michael E., Andreas Rechtsteiner, and Luis M. Rocha (2003). “Singular Value Decomposition and Principal Component Analysis”. In: *A Practical Approach to Microarray Data Analysis*. Ed. by Daniel P. Berrar, Werner Dubitzky, and Martin Granzow. Boston, MA: Springer US, pp. 91–109. ISBN: 978-0-306-47815-4. DOI: `10.1007/0-306-47815-3_5`.

Zhao, Wan-Lei, Cheng-Hao Deng, and Chong-Wah Ngo (2018). “k-means: A revisit”. In: *Neurocomputing* 291, pp. 195–206. ISSN: 0925-2312. DOI: `https://doi.org/10.1016/j.neucom.2018.02.072`.