Correlation based clustering of the Stockholm Stock Exchange

Author: Fredrik Rosén
Supervisor: Tor Brunzell, PhD
Abstract

This thesis presents a topological classification of stocks traded on the Stockholm Stock Exchange based solely on the co-movements between individual stocks. The working hypothesis is that an ultrametric space is an appropriate space for linking stocks together. The hierarchical structure is obtained from the matrix of correlation coefficient computed between all pairs of stocks included in the OMXS 30 portfolio by considering the daily log-arithmic return. The dynamics of the system is investigated by studying the distribution and time dependence of the correlation coefficients. Average linkage clustering is proposed as an alternative to the conventional single linkage clustering.

The empirical investigation shows that the Minimum-Spanning Tree (the graphical representation of the clustering procedure) describe the reciprocal arrangement of the stocks included in the investigated portfolio in a way that also makes sense from an economical point of view. Average linkage clustering results in five main clusters, consisting of Machinery, Bank, Telecom, Paper & Forest and Security companies. Most groups are homogeneous with respect to their sector and also often with respect to their sub-industry, as specified by the GICS classification standard. E.g. the Bank cluster consists of the Commercial Bank companies FöreningsSparbanken, SEB, Handelsbanken and Nordea. However, there are also examples where companies form cluster without belonging to the same sector. One example of this is the Security cluster, consisting of ASSA (Building Products) and Securitas (Diversified Commercial & Professional Services). Even if they belong to different industries, both are active in the security area. ASSA is a manufacturer and supplier of locking solutions and SECU focus on guarding solutions and security systems and cash handling.

The empirical results show that it is possible to obtain a meaningful taxonomy based solely on the co-movements between individual stocks and the fundamental ultrametric assumption, without any presumptions of the companies business activity. The obtained clusters indicate that common economical factors can affect certain groups of stocks, irrespective of their GICS industry classification. The outcome of the investigation is of fundamental importance for e.g. asset classification and portfolio optimization, where the co-movement between assets is of vital importance.

Keywords: Stock market; Correlation; Clustering; Minimum-Spanning Tree; Taxonomy
# Table of Contents

1 **Introduction**
   1.1 Background .......................................................... 1
   1.2 Objective and research questions ................................. 2
   1.3 Purpose of the study ............................................... 3
   1.4 Delimitation .......................................................... 3

2 **Theoretical framework** .............................................. 4
   2.1 Covariance and Correlation between stocks ...................... 4
   2.2 Distance between stocks ........................................... 6
   2.3 Ultrametricity and hierarchical clustering ....................... 7

3 **Method** ..................................................................... 10
   3.1 Example of hierarchical clustering of stocks .................. 10
   3.2 Software - Hierarchical Clustering Toolbox ..................... 13

4 **Results and analysis** ................................................ 15
   4.1 The correlation structure .......................................... 15
   4.2 Correlation based clustering ....................................... 16
      4.2.1 Minimum-Spanning Tree for the OMXS 30 portfolio ...... 17
      4.2.2 Clusters within the OMX 30 portfolio ..................... 18

5 **Conclusions** ............................................................ 23

6 **Ideas for future research** .......................................... 24

7 **References** ............................................................ 25

**Appendices** .................................................................. 27
1 Introduction

1.1 Background

Financial markets can be regarded as complex systems. The macroscopic patterns in finance, such as exchange rates, stock prices etc. are made up by collective behaviour of companies and individuals. All these markets reflect the behaviour of many agents, interacting in a highly non-linear way (Bonanno et al. 2001). Complex systems are common in many areas of research, such as biology, chemistry and physics. The uniqueness of the financial markets lies in the huge amount of data collected, making the system well defined. Data exist down to the single bid and ask of a financial asset (quote). Quotes of the Stockholm Stock Exchange have been recorded on a day-to-day basis from the 19th century and electronically, on a tick-by-tick basis, from the 1st of June 1990 (Bernhardsson 2002). This enormous amount of data collected around the world makes financial markets one of the most well documented complex systems, allowing detailed statistical analysis of the system characteristics.

Recently, the concepts and techniques from theoretical physics have been adapted to analyse and describe financial systems (Mantegna & Stanley 2000). Similarities between traditional subjects in physics and the financial market make techniques originally constructed for this area easily transferred to the field of theoretical finance.

One of these interdisciplinary areas is the presence of zero lag cross-correlation. This is present in many complex systems, such as e.g. in spin glass theory (Mezard et al. 1987). The financial market is no exception and the zero lag cross-correlation between the price movements of different stocks is a well-documented fact (Lo 1991, Mantegna 1997).

The problem of quantifying cross-correlation is important, not only from the point of view of understanding collective behaviour between the constituents of a complex system, but also from the point of view of estimating the risk of an investment portfolio (Plerou et al. 2001). Correlation between assets in a portfolio is of fundamental importance when one attempts to diversify investments, for example when trying to reduce exposure to sector- or industry-specific shocks.

The importance of correlation in portfolio optimization was first addressed by Markowitz (1959) in his Capital Asset Pricing Model (CAPM). The use of cor-
relation also plays a fundamental role in more recent techniques in theoretical finance, such as Value at Risk (Embrechts et al. 1999) and Arbitrage Pricing Theory (Campbell et al. 1997).

Strong correlation between certain groups of stocks would indicate that the financial market is affected by common economic factors. Recently, techniques such as the Random Matrix Theory (RMT), originally developed for quantum mechanics, have been used to study these effects (Plerou et al. 2001). Several of these studies have shown that stocks have a collective behaviour, thus supporting the idea that the financial market is affected by common economic factors.

The major taxonomy used to classify different types of stocks used in Sweden (and most other developed markets) is Standard & Poor’s Global Industry Classification Standard (GICS) (MSCI 2002). The GICS classifies companies at four different levels: sectors, industry groups, industries and sub-industries. This classification is based solely upon each company’s principal business activity and does not explicitly measure the co-movements of the stocks included in each sector or group.

From an economic point of view it would be of interest to have a classification, clustering stocks based only on their correlation with other stocks. As mentioned, the co-movement plays an important role in asset allocation and a clustering based on this could, for example in the case of asset diversification, be more useful than a clustering based solely on specific business activities.

A method for correlation-based clustering was first introduced by Mantegna (1999) to give a meaningful taxonomy and hierarchical structure based solely on the correlation between individual stocks. This research, based on stocks present in the S&P 500 index (U.S. equities market), supports the fact that common economic factor are present inside the financial market, resulting in co-movement of certain groups of stocks.

### 1.2 Objective and research questions

Most of the current research on correlation between individual assets focuses on the large stock exchanges, such as the New York Stock Exchange (NYSE) and other world dominating markets. To my knowledge there has not been any research conducted regarding the inter-stock correlation and the underlying hierarchical structure for the Stockholm Stock Exchange (SSE).
The overall research interest of this study is to investigate the correlation structure within SSE and to derive an hierarchical structure based solely on the co-movements between individual stocks. Within this main objective, the study will focus on these questions:

- What degree of correlation and anti-correlation is present between pairs of time series of stock price movements?
- Are there, based on correlation between individual stocks, indications of common economic factors affecting specific groups of stocks?

### 1.3 Purpose of the study

The purpose of this study is to derive a hierarchical structure with a meaningful economic taxonomy based solely on the co-movement of individual stocks traded on the Stockholm Stock Exchange. Groups of closely related stocks (clusters) identified from the hierarchical structure will be analyzed and compared to Standard Poor’s Global Industry Classification Standard.

### 1.4 Delimitation

The investigation is limited to the stocks included in the *OMX Stockholm 30 Index* (as of 2006-08-01). The index includes the 30 stocks that have the largest volume of the trading on the Stockholm Stock Exchange. The underlying hypothesis of the investigation is that an ultrametric space is an appropriate space for linking stocks together.
2 Theoretical framework

2.1 Covariance and Correlation between stocks

In real life, many measured variables are related. Consider for example the two variables weight, $X_1$, and length, $X_2$, measured on a population consisting of random individuals. Screening this data will probably indicate, assuming the population being of reasonable size, that there is a relationship between the measured quantities; tall persons normally tend to be heavier than the short persons. Two statistical measurements that help to assess the relationship between two random variables are covariance and correlation (Schaeffer & McClave 1995). The covariance is defined as

$$COV(X_1, X_2) = E[(X_1 - \mu_1)(X_2 - \mu_2)],$$

where $\mu_1 = E(X_1)$, i.e. the expected value of $X_1$, and $\mu_2 = E(X_2)$. If $X_2$ tends to be large when $X_1$ is large and small when $X_1$ is small, then $X_1$ and $X_2$ will have a positive covariance. If, on the other hand, $X_2$ tends to be large when $X_1$ is small and large when $X_1$ is small, then $X_1$ and $X_2$ will have a negative covariance. The covariance measures the direction of the association, but the value is unit-dependent and this makes comparison difficult. To get an independent measure of the association strength, the covariance must be normalized with respect to the variance of the measured variables. This value, the correlation between two variables, reflects the degree to which the variables are related. The most common measure of correlation is the Pearson’s Product Moment Correlation. The coefficient of correlation between two variables $X_1$ and $X_2$ is in this case given by

$$\rho_{i,j} = \frac{COV(X_1, X_2)}{\sqrt{V(X_1)V(X_2)}},$$

where $V(X_i)$ is the variance of variable $X_i$. Pearson’s correlation reflects the degree of linear relationship between two variables. It ranges from +1 to -1. A correlation of +1 means that there is a perfect positive linear relationship between the variables; a correlation of -1 means that there is a perfect negative linear relationship (anti-correlation) between variables and 0 indicate that there is no correlation.

Co-movement between individual stocks and between stocks and specific market indices plays an important role in finance. Fig. 1 shows a comparison
Correlation based clustering of the Stockholm Stock Exchange

between the daily (logarithmic) closing prices for Holmen/SCA (left) and Nokia/Drott (right) over the year 2002.

Figure 1 A comparison between the daily (logarithmic) closing price for Holmen/SCA (left) and Nokia/Drott (right) during 2002. The correlation between Holmen and SCA is estimated as 0.56 and the correlation between Nokia and Drott is estimated as 0.18.

The advantage of working with a logarithmic instead of a linear scales is evident when comparing stock quote changes. The daily logarithmic price changes for a stock \( i \) is defined as

\[
S_i \equiv \ln Y_i(t) - \ln Y_i(t-1). \tag{3}
\]

It can be showed, using the algebraic rules for exponents, that Eq. 3 is equivalent to \( \ln \left( \frac{Y_i(t)}{Y_i(t-1)} \right) \), i.e. the logarithmic return. This equivalence implies that \( S_i \) is independent of price scale, i.e. a stock quote change from 10 - 11 SEK will result in the same logarithmic difference as a change from 500 - 550 SEK. This makes comparison between stocks at different price scales possible.

The correlation coefficient between the stocks in Fig. 1 is determined over the 250 trading days of 2002 using Eq. 2. The correlation between Holmen and SCA is estimated as 0.56 and the correlation between Nokia and Drott is estimated as 0.18. The result agrees with the visual impression of the compared time series; the time evaluation of \( \ln(Y) \) is more coherent for Holmen and SCA compared to Nokia and Drott.

The difference in co-movement is hardly surprising, given the fact that both Holmen and SCA:s main business activity is paper production (GICS industry classification: Paper and Forest Products) and they are thus likely to be
affected by common economical factors. Nokia (Communications Equipment) and Drott (Real Estate) do not belong to the same industry sector and are thus less likely to have similar movements.

It is important to remember that the correlation coefficient only reflects the degree of linear relationship between the two stock price movements. It gives no indication of the causal relation between the measured variables.

2.2 Distance between stocks

The most well known distance metric that we encounter in every-day life is the three-dimensional Euclidean distance (Anton 1994). This distance is the physical distance between two points in space that one would measure with a ruler.

Would it be possible to, in the same way as we have a distance between physical object, define a distance between two synchronous time-series? Given such a measure, it would be valid to talk about a “distance between individual stocks”. For a distance function, $d_{ij}$, to be a valid metric distance, the following four properties must hold:

\begin{align*}
(i) & : d_{ij} \geq 0 \\
(ii) & : d_{ij} = 0 \iff i = j \\
(iii) & : d_{ij} = d_{ji} \\
(iv) & : d_{ij} \leq d_{ik} + d_{kj}
\end{align*}

One measure of the degree of similarity between two time-series is the correlation coefficient (as defined in Eq. 2). This measure does hoverer not fulfill the properties of Eq. 4. To obtain a metric distance based on the correlation coefficient that fulfill the properties of Eq. 4, Mantegna (1997) proposed the distance function

$$d_{ij} = \sqrt{2 (1 - \rho_{ij})}.$$  

The mapping between the correlation coefficient and the distance function is showed in Fig. 2. $d_{ij}$ ranges from 0, for totally correlated stocks, to 2 for totally anti-correlated stocks. For uncorrelated stocks, the distance is $\sqrt{2}$. The distance function $d_{ij}$ fulfills all four properties of Eq. 4 and qualifies consequently as a metric distance. Eq. 5 is just one of many possible Euclidean distance functions and it is chosen because it is the one that is most commonly used in the literature.
2.3 Ultrametricity and hierarchical clustering

The distance function between a synchronous evolving pair of assets introduced in the previous section will form the basis for further discussion of taxonomy.

The knowledge of a distance function makes it possible to, without any prior knowledge of specific groups, decompose a set of $n$ objects into subsets of closely related objects (clusters). Cluster analysis is a common technique in multivariate data analysis and it can be applied in various ways (Johnson 1998).

The first step in performing a cluster analysis is to make an assumption about the topological space linking the objects together. The working hypothesis used by e.g Mantegna (1999) is that an ultrametric space is an appropriate topological space for linking $n$ stocks. The ultrametric distance $\hat{d}_{ij}$ must satisfy property (i)-(iii) of Eq. 4, while property (iv), the triangular inequality, is replaced by the stronger inequality

$$\hat{d}_{ij} \leq \max \left[ \hat{d}_{ik}, \hat{d}_{kj} \right].$$

(6)

So what does this difference between a regular metric space and a ultrametric space really mean? Fig. 3 shows a visual interpretation of the triangular inequality. The property is, according to Eq. 4, defined as $d_{ij} \leq d_{ik} + d_{kj}$, i.e. the distance between $i$ and $j$ is always less or equal to the distance between $i$ and $j$, passing through some intermediate point $k$. Another way of stating this would be to say that the shortest distance between two points in space
is always a straight line (cf. Fig. 3). Eq. 6 puts an even tighter constrain, in an ultrametric space, the distance between the points \( i \) and \( j \) is always less or equal then the maximum of: the distance between \( i \) and any other point \( k \) and the distance between \( j \) and any other point \( k \).

Ultrametric spaces provide a natural way of describing hierarchical structured complex systems, since the concept of ultrametricity is directly connected to the concept of hierarchy (Mantegna & Stanley 2000). For a more in-depth explanation of ultrametricity, please refer to Ramell et al. (1986).

One of the easiest ways of performing the hierarchical clustering and find the ultrametric distances is to obtain the Minimal-Spanning Tree (MST) from the metric distances that link together the objects to be clustered. The MST is a concept from graph theory and one of the most famous algorithms used to determine the MST is Kruskal’s algorithm (West 1996). The algorithm is conceptually described by Tola et al. (2005) in the following way:

Assume that we have a list \( D \) consisting of distances between pairs of elements (e.g. stocks) in the system to be clustered (e.g. a portfolio of stocks). Arrange all the distances \( d_{ij} \) (the distances between element \( i \) and element \( j \)) in \( D \) in increasing order. Different elements are iteratively included in clusters, starting from the first two elements of the distance measure ordered list. At each step, when two elements or one element and a cluster or two clusters \( p \) and \( q \) merge in a wider single cluster \( t \), the similarity or distance between the new cluster \( t \) and cluster \( r \) is determined as

\[
s_{ij} = \min \{d_{pr}, d_{qr}\}.
\] (7)

This definition, where the distance between groups is defined as the distance between the closest pair of elements, is called single linkage clustering (nearest neighbor). An alternative way of linking together separate clusters is called average linkage clustering. In this case the distance between groups is defined
as the average distances between all pairs of objects. The distances from both single and average linkage clustering obey the ultrametricity criteria. The two clustering techniques are visualized in Fig. 4. The individual steps in the clustering procedure and the final result can be summarized diagrammatically in a tree form called dendrogram.

\[ d_{ij} = \min\{d_{pr}, d_{qr}\} \]

\[ d_{ij} = \text{avg}\{d_{pr}, d_{qr}\} \]

Figure 4 The difference between single linkage (left) and average clustering (right).

The above description of ultrametricity and hierarchical clustering is intentionally kept at a somewhat general level. To show the use for the technique within the scope of this thesis, a more straightforward example is presented in the next section.
3 Method

3.1 Example of hierarchical clustering of stocks

This section presents and an example of hierarchical clustering applied to stock quote data.

As previously stated, clustering analysis is performed without any prior knowledge of predefined groups. The only starting point for this example is the daily logarithmic price changes for five stocks traded on the SSE during the year 2003. The stocks selected for the example are: Föreningssparbanken (FSPA), Drott/Fabege (FABG), Ericsson (ERIC), SEB (SEB) and Nokia (NOKI) (Fig. 5).

Figure 5 The companies included in the example - Föreningssparbanken (FSPA), Drott/Fabege (FABG), Ericsson (ERIC), SEB (SEB) and Nokia (NOKI). The companies are in parenthesis identified by their tick symbol.

Using Eq. 2, the correlation matrix $\rho_{ij}$ for the five stocks is estimated as

<table>
<thead>
<tr>
<th></th>
<th>FSPA</th>
<th>FABG</th>
<th>ERIC</th>
<th>SEB</th>
<th>NOKI</th>
</tr>
</thead>
<tbody>
<tr>
<td>FSPA</td>
<td>1</td>
<td>0.25</td>
<td>0.43</td>
<td>0.73</td>
<td>0.48</td>
</tr>
<tr>
<td>FABG</td>
<td>1</td>
<td>0.24</td>
<td>0.23</td>
<td>0.19</td>
<td></td>
</tr>
<tr>
<td>ERIC</td>
<td>1</td>
<td>0.49</td>
<td>0.59</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SEB</td>
<td>1</td>
<td>0.53</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NOKI</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The matrix displays the correlation values between all the stocks (identified with their tick symbol). Due to symmetry reasons, only the right side of the matrix is presented. The associated distance matrix $d_{ij}$ is calculated using Eq. 5 as

<table>
<thead>
<tr>
<th></th>
<th>FSPA</th>
<th>FABG</th>
<th>ERIC</th>
<th>SEB</th>
<th>NOKI</th>
</tr>
</thead>
<tbody>
<tr>
<td>FSPA</td>
<td>0</td>
<td>1.22</td>
<td>1.07</td>
<td>0.73</td>
<td>1.02</td>
</tr>
<tr>
<td>FABG</td>
<td>0</td>
<td>1.23</td>
<td>1.24</td>
<td>1.27</td>
<td></td>
</tr>
<tr>
<td>ERIC</td>
<td></td>
<td>0</td>
<td>1.01</td>
<td>0.91</td>
<td></td>
</tr>
<tr>
<td>SEB</td>
<td>0</td>
<td></td>
<td>0.97</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NOKI</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The MST associated with the distance matrix is obtained using Kruskal’s algorithm (single-linkage clustering), as described in the previous section. The two stocks separated with the shortest distance is FSPA and SEB ($d=0.73$). The next shortest distance is observed between NOKI and ERIC ($d=0.91$). There are now two separate regions in the MST, FSPA-SEB and NOKI-ERIC (Fig. 6A). The next two closest stocks, connecting the two regions, are SEB and NOKI ($d=0.97$). At this point, the MST is: FSPA-SEB-NOKI-ERIC (Fig. 6B). The next pair of stocks are SEB and ERIC. This connection is not considered since both stocks already have been sorted. The only remaining stock is FABG, which is closest to FSPA ($d=1.22$). This last link concludes the clustering and the final MST is observed in Fig. 6C.

![MST diagram](image)

Figure 6 The MST associated with the distance matrix obtained by Kruskal’s algorithm for the example of the five companies: Föreningssparbanken (FSPA), Drott/Fabege (FABG), Ericsson (ERIC), SEB (SEB) and Nokia (NOKI).

In principle, obtaining the MST for a portfolio of $N$ stocks can be seen as way to find the $N - 1$ most relevant connections, reducing it from the $\frac{N(N-1)}{2}$.
connections in the distance matrix. In this case, the number of connections has been reduced from 10 to 4.

The indexed hierarchical tree (dendrogram) associated with the MST is showed in Fig. 7. The vertical axis shows the ultrametric distance between the stocks and the horizontal axis shows the tick symbols. The tree clearly shows that there are two groups of stocks (clusters) in this selected portfolio. In the first cluster we have the telecom companies Ericsson and Nokia. The second group consist SEB and Föreningssparbanken, both active in the bank sector. Out of these five companies, the real estate company Drott/Faberge is the one that is least connected with the others.

If the clustering procedure instead had been performed using average-linkage clustering, the distance between groups had been defined as the average distance between all the stocks in the two groups, not just the two closest. E.g. the distance, $\hat{d}$, between FSPA-SEB and NOKI-ERIC (Fig. 6A) would have been defined as 1.01 instead of 0.97.
3.2 Software - Hierarchical Clustering Toolbox

A software tool, *Hierarchical Clustering Toolbox*, was developed\(^1\) in order to carry out the investigations presented in this thesis. A screen dump of the front window is showed in Fig. 8.

![Hierarchical Clustering Toolbox](image)

Figure 8 A screen dump of the main window of the Hierarchical Clustering Toolbox software. The software is composed of three main parts: Download Stock Quotes, Calculate Distance Matrix and Calculate Hierarchical Structure.

The software is composed of three main parts: *Download Stock Quotes*, *Calculate Distance Matrix* and *Calculate Hierarchical Structure*. The three parts of the software are explained below.

**Download Stock Quotes**

The stock quotes included in OMXS 30 are downloaded from the OMX homepage\(^2\) (provided that the computer is connected to In-

---

\(^1\)All development was done in MATLAB\(^\text{®}\) - www.mathworks.com

\(^2\)The data is obtained from http://www.se.omxgroup.com/slutkurser/
ternet) by clicking on Load OMX. The data is saved as excel files (SYMBOL200X.XLS) in separate folders, one for each selected year.

**Calculate Distance Matrix**

*Distance Matrix* calculates the correlation matrix (Eq. 2) and the associated distance matrix (Eq. 5) for the downloaded stock quote data. The correlation coefficients are estimated from the daily log-arithmetic returns (adjusted for dividends and splits). The matrices can be saved to separated files.

**Calculate Hierarchical Structure**

*Hierarchical Structure* calculate the Minimum-Spanning Tree using Kruskal’s algorithm. The dendrogram (single or average) is plotted in a separate window.

The main functions included in the software can be found in Appendix I.
4 Results and analysis

All presented results are based on the stocks included in the OMX Stockholm 30 Index (OMXS 30) as of 2006-08-01. The index includes the 30 stocks that have the largest volume of the trading on the Stockholm Stock Exchange. The company name, stock symbol and GICS industry classification of the stocks are listed in Appendix I. If nothing else is stated, information concerning the business activities of the investigated companies (in addition to the GISC classification) is taken from the latest annual report.

4.1 The correlation structure

The clustering technique adapted in this thesis relies on the correlation between the evaluated stocks. The spread and time dependence of the correlations are thus of fundamental interest. Fig. 9 shows the probability density function $P(\rho_{ij})$ observed between stocks in the OMXS 30 portfolio between 2001 and 2005. The density was estimated using a kernel smoothing method (Bowman & Azzalini 1997) to compensate for the limited number of $\rho_{ij}$.

![Figure 9 Probability density function $P(\rho_{ij})$ of the correlation coefficients $\rho_{ij}$ observed between stocks in the OMXS 30 portfolio between 2001 and 2005.](image)

The probability density is rather bell shaped and there is a clear overweight toward positive $\rho_{ij}$. To show the dynamics of the system, Table 1 summarizes the minimum and maximum values of the $\rho_{ij}$ observed between stocks in the portfolio during 2001 to 2005. The results were estimated from daily logarithmic returns using Eq. 2.
Table 1 Minimum and maximum correlation coefficients $\rho_{ij}$ observed between stocks in the OMXS 30 portfolio during 2001 to 2005.

<table>
<thead>
<tr>
<th>Time period</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>2001</td>
<td>-0.05</td>
<td>0.75</td>
</tr>
<tr>
<td>2002</td>
<td>-0.13</td>
<td>0.80</td>
</tr>
<tr>
<td>2003</td>
<td>0.03</td>
<td>0.76</td>
</tr>
<tr>
<td>2004</td>
<td>-0.04</td>
<td>0.73</td>
</tr>
<tr>
<td>2005</td>
<td>-0.25</td>
<td>0.62</td>
</tr>
<tr>
<td>2001-2005</td>
<td>-0.03</td>
<td>0.72</td>
</tr>
</tbody>
</table>

The maximum correlation, $\rho_{ij} = 0.80$, is observed between Nordea (Commercial Banks) and SEB (Commercial Banks). The high degree of correlation is not surprising since the companies both belong to same sector. The largest anti-correlation, $\rho_{ij} = -0.25$, was observed between Atlas Copco (Machinery) and Vostok Nafta (Oil, Gas & Consumable Fuels). Between 2003 and 2005, the price of e.g. oil\(^3\) increased by almost 100%. Vostok Nafta benefits from a high energy price, for Atlas Copco the situation is the opposite. As a manufacturing company, both the internal production costs the financial situation of a majority of their customers are negatively influenced by a high energy price. Even if the same situation could be argued for other companies as well, it is one plausible explanation of the observed anti-correlation.

The maximum correlation observed over the total time period 2001-2005 was between Atlas Copco and Sandvik, both Machinery companies. The results show that there is a dynamic behaviour of the degree of correlation between stocks for the observed time period.

### 4.2 Correlation based clustering

Two stocks included in the index portfolio have been omitted from the cluster analysis: Investor and Atlas Copco ser. A. Investor is an investment company and several of its core investments (ABB, ATCO, ERIC, SEB and AZN) are themselves included in OMXS 30 portfolio. This would, to some degree, implicate that Investor would behave as an index for the core investments, introducing “circular” correlations, and the stock is thus omitted. Two Atlas

\(^3\)Spot price of Light Crude Oil - www.futuresource.com
Copco stocks are included in the index portfolio. The elementary correlation between these stocks is of no interest and only the stock with the largest trading volume is included.

The results are based on the daily logarithmic returns during the five years ranging from 2001 to 2005. The stock quotes used to estimate the correlation coefficients were the daily close price (adjusted for dividends and splits). The clustering procedure was preformed as in the example in Section 3.1 using the software developed for this purpose.

\subsection{Minimum-Spanning Tree for the OMXS 30 portfolio}

Fig. 10 shows the MST obtained for the OMXS 30 portfolio during the years 2001 to 2005. The color of the line segments connecting the stocks indicates the distance. The indexed hierarchical tree associated with the MST is showed in Fig. 11. Under the hypothesis that an ultrametric space is an appropriate topological space for linking stocks, the MST shows the 27 most relevant connections out of the 378 connections in the distance matrix.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{mst.png}
\caption{MST obtained for the OMXS 30 portfolio during the years 2003-2005. The color of the line segments connecting the stocks indicates the distance: \( d \leq 0.8, \quad 0.8 < d \leq 0.9, \quad 0.9 < d \leq 1.0, \quad 1.0 < d \leq 1.1, \quad 1.1 < d \leq 1.2 \) and \( \cdots \quad d \geq 1.2 \).

Studying the connections present between the stocks in Fig. 10, the MST seems to describe the reciprocal arrangements of the stocks included in the OMXS
30 portfolio in a way that also makes sense from an economical point of view. Several groups of closely related stocks can be observed.

4.2.2 Clusters within the OMX 30 portfolio

To make a clear definition whether or not a group of stocks form a cluster, the maximum distance between two stocks (or groups of stocks) must be chosen. Subjectively choosing \( d \leq 0.95 \) result in three main clusters (cf. Fig. 11)

- Atlas Copco, Sandvik and SKF
- FöreningsSparbanken, SEB, Sv. Handelsbanken and Nordea
- Ericsson and Nokia

The first cluster consists of the Industrial Machinery companies ATCO, SAND and SKF. The second cluster consists of the Commercial Bank companies FSPA, SEB, SHB and NDA. The third cluster include the Communication Equipment companies ERIC and NOKI.

The hierarchical tree in Fig. 11 is (as the MST representation) based on single linkage clustering. An alternative way of linking together separate clusters is, as presented in Section 2.3, average linkage clustering. In this case the distance between groups is defined as the average distances between all pairs of objects and not just the as the distance between the two closest objects. The hierarchical tree in Fig. 12 shows the result of the average linkage procedure. Choosing \( d \leq 1.04 \) result in the five main clusters

- **Industrial cluster**: Atlas Copco, Sandvik, SKF, Electrolux, Volvo and Autoliv
- **Bank cluster**: FöreningsSparbanken, SEB, Sv. Handelsbanken and Nordea
- **Security cluster**: ASSA ABLOY and Securitas
- **Telecom cluster**: Ericsson, Nokia, TeliaSonera and Tele2
- **Pulp & Paper cluster**: Holmen, SCA and Stora Enso
The hierarchical arrangement of stocks is in principle the same as for single linkage clustering. However, compared to the results based on single linkage, more groups of closely related stocks can be identified and the clusters are more clearly separated from each other. The identified clusters also seem relevant comparing them to the arrangement of stocks in the MST (Fig. 10). These subjective observations together indicate that average linkage clustering is more appropriate to describe the hierarchical arrangement within the OMXS 30 portfolio than single linkage clustering. Based on this, the rest of the analysis focus on these clusters.

**Industrial cluster**

This cluster consists of the Machinery companies ATCO, SAND, SKF and VOLV, together with ALIV (Auto Components) and ELUX (Household Durables). VOLV is an auto manufacturer and ALIV manufactures automotive safety systems. It is reasonable to believe that the mutual dependences between these two companies make them appear in the same cluster, even though they belong to different sectors. ELUX manufactures and sells household appliances.

**Bank cluster**

This cluster consists of the Commercial Bank companies FSPA, SEB, SHB and NDA. All these companies connect at a distance $d \leq 0.8$, making it the tightest cluster. The fact that it includes only banks also makes it very homogeneous. The close relation between the bank companies is not surprising from an economical point of view since they are all directly influenced by economical factors such as e.g. interest rates. The interest rate is set by the central bank, as the main tool of monetary policy, and cannot be controlled by individual actors.

**Security cluster**

This cluster consists of ASSA (Building Products) and SECU (Diversified Commercial & Professional Services). Even if they belong to different industry groups, both are active in the security
area. ASSA is a manufacturer and supplier of locking solutions and SECU focus on guarding solutions, security systems and cash handling.

**Telecom cluster**

This cluster consist of two sub-groups, one with the Communication Equipment companies ERIC and NOKI and one with the Integrated Telecommunication Service companies TELSN and TEL2. That these two groups of companies together form a cluster is not surprising since they all are active within the telecom area, the service providers depend on the technical equipment suppliers and vice versa.

**Paper & Forest cluster**

This cluster consist of the Paper & Forest Products companies HOLM, SCA and STE. The main business of these companies is paper production and they are all directly influenced by common factors, such as pulp prices and the global supply/demand for their products.

One interesting observation is that, irrespective of how the clustering is performed, the companies that are furthest away from all other companies are Swedish Match (Tobacco), Boliden (Metals & Mining), Vostok Nafta (Oil, Gas & Consumable Fuels) and AstraZeneca (Pharmaceuticals) (cf. Fig. 10). This makes sense since the main business activities of these companies differ significantly from the other companies in the portfolio.

The clustering of companies with similar business activity is in accord with previous studies (Mantegna 1999, Mantegna & Stanley 2000, Bonanno et al. 2001). However, to my knowledge, average linkage clustering has previously not been evaluated as an alternative to the conventional single linkage clustering in studies concerning clustering based on stock quote data.
Figure 11 The indexed hierarchical tree based single linkage clustering (nearest neighbor) obtained for the OMXS30 portfolio during the years 2001-2005. Choosing $d \leq 0.95$ result in three main clusters. The first cluster consists of the Industrial Machinery companies ATCO, SAND and SKF. The second cluster consists of the Commercial Bank companies FSPA, SEB, SHB and NDA. The third cluster include the Communication Equipment companies ERIC and NOKI.
Figure 12 The indexed hierarchical tree obtained based on average linkage clustering for the OMXS 30 portfolio during the years 2001-2005. Choosing $d \leq 1.04$ result in the five main clusters. The first cluster consists of the Machinery companies ATCO, SAND, SKF and VOLV, together with ALIV (Auto Components) and ELUX (Household Durables). The second cluster consists of the Commercial Bank companies FSPA, SEB, SHB and NDA. The third cluster consists of ASSA (Building Products) and SECU (Diversified Commercial & Professional Services). The forth cluster consist of two subgroups, one with the Communication Equipment companies ERIC and NOKI and one with the Integrated Telecommunication Service companies TELSN and TEL2. The fifth cluster consist of the Paper & Forest Products companies HOLM, SCA and STE.
5 Conclusions

The working hypothesis for this investigation was that an ultrametric space is an appropriate space for linking stocks together. The empirical investigation did not contradict this assumption since the MST describe the reciprocal arrangement of the stocks included in the OMXS 30 portfolio in a way that also makes sense from an economical point of view. Several groups of linked stocks were observed. The empirical investigation indicates that average linkage clustering is more appropriate for the evaluated portfolio of stocks compared to single linkage clustering.

Average linkage clusters resulted in five main clusters, consisting of Machinery, Bank, Telecom, Paper & Forest and Security companies. Most groups are homogeneous with respect to their sector and also often with respect to their sub-industry, as specified by the GICS classification standard. E.g. the Bank cluster consists of the Commercial Bank companies FöreningsSparbanken, SEB, Handelsbanken and Nordea. However, there are also examples where companies form cluster without belonging to the same sector. One example of this is the Security cluster, consisting of ASSA (Building Products) and Securitas (Diversified Commercial & Professional Services). Even if they belong to different industries, both are active in the security area.

The MST and resulting clusters follow directly from the correlation coefficients between all pairs of assets, the spread and dynamic behavior of these coefficients are thus of fundamental importance. The probability density function of the coefficient showed that the distribution is approximately bell shaped and that there is a clear overweight toward positive $\rho_{ij}$. The minimum and maximum correlations also vary on a year-to-year basis, indicating that the obtained clusters also could show a dynamic behavior.

The empirical results show that it is possible to obtain a meaningful taxonomy based solely on the co-movements between individual stocks and the fundamental ultrametric assumption, without any presumptions of the companies business activity. The obtained clusters indicate that common economical factors affect certain groups of stocks, irrespective of their GICS industry classification. The outcome of the investigation is of fundamental importance for e.g. asset classification and portfolio optimization, where the co-movement between assets is of vital importance.
6 Ideas for future research

The outcome of this thesis shows the potential with correlation based clustering of stock quote data. The empirical investigation was limited to the stocks included in the OMXS 30 portfolio, there is however no limitations on the amount of data the clustering technique can handle. The developed software can easily be extended to handle other portfolios.

The correlation coefficients showed a dynamic behavior on a year-to-year basis. Based on this, it is likely that also the hierarchical structure would show a dynamic behavior and it would be interesting to study how the resulting clusters change over time. To do this in a structured way, the maximum distance between two stocks for them to form a cluster should be chosen objectively, either by setting a fix maximum distance or by determining it from the spread between the maximum and minimum distances.

One interesting and current topic for future research is the ongoing process to create an integrated Nordic Baltic market on the exchange side. As a result of this, on October 2, 2006 the current exchange list structures for Sweden, Denmark and Finland will be replaced with the Nordic list\(^4\). Companies on the Nordic list will be presented in a common manner and divided into segments. Companies will be presented first by market capitalization and then by industry sector following the GICS. There will be three new market capitalization segments: Nordic Small Cap, Nordic Mid Cap and Nordic Large Cap. This process will introduce new Sector indexes. It would be interesting to compare the classification and resulting indexes to the clusters obtained by correlation based clustering for the entire Nordic market. Such a comparison could give new insights into the reciprocal arrangement of the stocks in comparison to the new GICS based indexes.

Correlation based clustering has recently also been investigated as a tool for portfolio optimization. E.g. Tola et al. (2005) uses the clustering procedure to filter the part of the covariance matrix which is less likely to be affected by statistical uncertainty (remember that Kruskal’s algorithm can be seen as way to find the \(N - 1\) most relevant connections from the \(\frac{N^2}{2}(N - 1)\) of the original matrix). The filtered information is then used to build portfolios. The result from the investigation showed that improvements were obtained but there is still much left to be done to refine the technique and to identify new areas of application.

\(^4\)The Nordic list - http://www.se.omxgroup.com
7 References


APPENDIX I

The main functions in the Hierarchical Clustering Toolbox software.

function downloadOMX(STOCKLIST,startYear,endYear,folder)
% Function: Download stock data from OMX
% INPUT: STOCKLIST - List the stocks included OMXS 30 (symbol,name and OMXid)
% startYear - from year (e.g. 200X)
% endYear - to year
% folder - name of folder to save data
% Fredrik Rosén (c) 2006
% Download OMX stocks
mkdir(folder)
for i=1:length(STOCKLIST)
    address = [url_main,STOCKLIST(i).id,'&InstrumentType=1&From=',startYear,...
    '-01-01','&todate=',endYear,'-12-31']
    fid = urlwrite(address,[folder,'\',STOCKLIST(i).symbol,folder,'.xls']);
end

function corrmatrix = getCorrMatrix(STOCKLIST,year)
% Function: Estimate correlation-matrix from the log-reurns
% INPUT: STOCKLIST - List the stocks included OMXS 30 (symbol,name and OMXid)
% year - what year?
% OUTPUT: corrmatrix - Correlation matrix
% Fredrik Rosén (c) 2006

daelist = getDateList(year) % Get a list of trading dates for the active year
n = length(STOCKLIST);
for i=1:n
    for j=1:n
        corrmatrix(i,j) = stockCorr(STOCKLIST(i).return,STOCKLIST(j).return,datelist);
    end
end
function corr = stockCorr(stock_a, stock_b, datelist)
size_datelist = length(datelist);
stock_correct_a = [];
stock_correct_b = [];
j=1
for (i=1:size_datelist)
    index_a = [];
    index_b = [];
    index_a = find(stock_a(:,1) == datelist(i));
    index_b = find(stock_b(:,1) == datelist(i));
    % check so that each stock was traded that day
    if(~isempty(index_a) && ~isempty(index_b))
        stock_correct_a(j) = stock_a(index_a,2);
        stock_correct_b(j) = stock_b(index_b,2);
        j = j+1;
    end
end
corr = corrcoef(stock_correct_a, stock_correct_b);
corr = corr(1,2); % get first index

function plotDendrogram(STOCKLIST, corrmatrix, type)

    d = calculateEuclidean(corrmatrix)
d_list = matrix2list(d);

    switch type
        case 'Single'
            links = linkage(d_list,'single');
titletext = 'Hierarchical clustering (Single-linkage)';
            [i j v] = mst(sparse(d));% Calculate links using kruskals
            links_mst = [i,j,v];
        case 'Average'
            links = linkage(d_list,'average');
titletext = 'Hierarchical clustering (Average-linkage)';
        end
dendrogram(links) % Plot the dendrogram in new window

function d = calculateEuclidean(roh)
% Calculate distance matrix
d = real(sqrt(2*(1-roh)));

function d = matrix2list(roh)
% Convert matrix to list
n = length(roh)
k=1;
for i=1:n
    for j=(i+1):n
        d(k) = roh(i,j);
        k = k+1;
    end
end
APPENDIX II

The company name, stock symbol and GICS industry classification of the stocks included in the OMXS 30 portfolio summarized in a table.
<table>
<thead>
<tr>
<th>Stock name</th>
<th>Short name</th>
<th>Sector</th>
<th>Industry group</th>
<th>Industry</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABB LTD</td>
<td>ABB</td>
<td>Industrials</td>
<td>Capital Goods</td>
<td>Electrical Equipment</td>
</tr>
<tr>
<td>Alfa Laval</td>
<td>ALFA</td>
<td>Industrials</td>
<td>Capital Goods</td>
<td>Machinery</td>
</tr>
<tr>
<td>Autoliv SDB</td>
<td>ALIV</td>
<td>Consumer Discretionary</td>
<td>Automobiles &amp; Components</td>
<td>Auto Components</td>
</tr>
<tr>
<td>ASSA ABLOY</td>
<td>B ASSA</td>
<td>Industrials</td>
<td>Capital Goods</td>
<td>Building Products</td>
</tr>
<tr>
<td>Atlas Copco</td>
<td>ATCO</td>
<td>Industrials</td>
<td>Capital Goods</td>
<td>Machinery</td>
</tr>
<tr>
<td>AstraZeneca</td>
<td>AZN</td>
<td>Health Care</td>
<td>Pharmaceuticals &amp; Biotechnology</td>
<td>Pharmaceuticals</td>
</tr>
<tr>
<td>Boliden B</td>
<td>BOL</td>
<td>Materials</td>
<td>Materials</td>
<td>Metals &amp; Mining</td>
</tr>
<tr>
<td>Electrolux A</td>
<td>ELUX</td>
<td>Consumer Discretionary</td>
<td>Consumer Durables &amp; Apparel</td>
<td>Household Durables</td>
</tr>
<tr>
<td>Eniro</td>
<td>ENRO</td>
<td>Consumer Discretionary</td>
<td>Technology</td>
<td>Hardware &amp; Equipment</td>
</tr>
<tr>
<td>Ericsson B</td>
<td>ERIC</td>
<td>Information Technology</td>
<td>Technology</td>
<td>Hardware &amp; Equipment</td>
</tr>
<tr>
<td>FörhemSp</td>
<td>A FSP</td>
<td>Consumer Discretionary</td>
<td>Hardware &amp; Equipment</td>
<td>Consumer Durables &amp; Apparel</td>
</tr>
<tr>
<td>Hennes &amp; Mauritz B</td>
<td>HM</td>
<td>Consumer Discretionary</td>
<td>Retailing</td>
<td>Specialty Retail</td>
</tr>
<tr>
<td>Holmen B</td>
<td>HOLM</td>
<td>Materials</td>
<td>Materials</td>
<td>Paper &amp; Forest Products</td>
</tr>
<tr>
<td>Nordea Bank</td>
<td>NDA</td>
<td>Financials</td>
<td>Banks</td>
<td>Commercial Banks</td>
</tr>
<tr>
<td>Nokia SDB</td>
<td>NOKI</td>
<td>Information Technology</td>
<td>Technology</td>
<td>Hardware &amp; Equipment</td>
</tr>
<tr>
<td>Nordea Bank</td>
<td>NDA</td>
<td>Financials</td>
<td>Banks</td>
<td>Commercial Banks</td>
</tr>
<tr>
<td>Powerservice</td>
<td>PSEA</td>
<td>Consumer Discretionary</td>
<td>Hardware &amp; Equipment</td>
<td>Consumer Durables &amp; Apparel</td>
</tr>
<tr>
<td>Ricol Tieto</td>
<td>TIE A</td>
<td>Energy</td>
<td>Energy</td>
<td>Oil, Gas &amp; Consumable Fuels</td>
</tr>
<tr>
<td>Sanoma</td>
<td>SANOM</td>
<td>Energy</td>
<td>Energy</td>
<td>Oil, Gas &amp; Consumable Fuels</td>
</tr>
<tr>
<td>Skandia</td>
<td>SKAND</td>
<td>Energy</td>
<td>Energy</td>
<td>Oil, Gas &amp; Consumable Fuels</td>
</tr>
<tr>
<td>TeliaSonera</td>
<td>TSSL</td>
<td>Energy</td>
<td>Energy</td>
<td>Oil, Gas &amp; Consumable Fuels</td>
</tr>
<tr>
<td>Tele2 A</td>
<td>TEL2 A</td>
<td>Energy</td>
<td>Energy</td>
<td>Oil, Gas &amp; Consumable Fuels</td>
</tr>
<tr>
<td>Volvo B</td>
<td>VOLV B</td>
<td>Industrials</td>
<td>Capital Goods</td>
<td>Machinery</td>
</tr>
<tr>
<td>Vostok Nafta Inv SDB</td>
<td>VOST SDB</td>
<td>Energy</td>
<td>Energy</td>
<td>Oil, Gas &amp; Consumable Fuels</td>
</tr>
<tr>
<td>Wincor Nixdorf</td>
<td>WIN</td>
<td>Industrials</td>
<td>Capital Goods</td>
<td>Machinery</td>
</tr>
<tr>
<td>Yoko B</td>
<td>YOKU B</td>
<td>Energy</td>
<td>Energy</td>
<td>Oil, Gas &amp; Consumable Fuels</td>
</tr>
</tbody>
</table>