

LUND UNIVERSITY

MASTER'S THESIS

**Analysing Customer Behaviour in the
FX Market Using Order Flow Data and
Machine Learning Techniques**

Author:
Lovisa THORDIN

Supervisors:
Erik LINDSTRÖM
Martin RICHTER

Examiner:
Magnus WIKTORSSON

*A thesis submitted in fulfilment of the requirements
for the degree of Master of Science*

in the

Mathematical Statistics
Centre for Mathematical Sciences

February 2015

LUND UNIVERSITY

Abstract

Faculty of Engineering
Centre for Mathematical Sciences

Master of Science

Analysing Customer Behaviour in the FX Market Using Order Flow Data and Machine Learning Techniques

by Lovisa THORDIN

This thesis has two main objectives related to trading foreign currencies. First, it is investigated how the customer order flow of Nordea is related to currency price changes. Second, the goal is to find a new way of grouping customers that can give additional insights in the trading behaviour of different customers.

The study uses order flow data which consists of spot and forward transactions made in Norwegian Kronor and Swedish Kronor during a period of nearly three years. The counterparties with whom Nordea is trading foreign exchange are divided into the customer groups asset managers, banks, corporates, hedge funds and private clients. As a measure of the dependence between order flow and exchange rate movements an index is introduced. The index tells how the spot exchange rate moves before and after a trade is executed. To examine the trading behaviour for a customer group the indices for all trades done by that group are weighted with the traded volume.

Grouping customers in a new way is addressed by using Machine Learning techniques in the field unsupervised learning, called clustering. The applied clustering algorithms are the K -means, the Fuzzy C -means and the Self-Organizing Map. The customers are clustered according to four different features calculated from the order flow data.

The main findings include that there are differences in the dependence between exchange rate changes and the order flow from different customer groups. A contrarian behaviour is found for the corporates and private clients, while the asset managers and hedge funds tend to hold a trend-following trading style. The results from the clustering do not provide a better way of grouping clients than the previous one, but contributes with a deeper understanding of the trading behaviour of different customers.

Acknowledgements

First of all I would like to thank Nordea Markets for giving me the opportunity to write this thesis. It has been a great and valuable experience to work in an professional environment with inspiring and engaged people. A special thanks to Martin Richter, the FX Quant team, for extraordinary support and patience throughout this project, and for all the given time and advice. I would also like to thank Lars Staune and John Hydeskov in the FX Quant team for including me in their team during this fall. In addition I am very grateful to everyone in the FX group at Nordea Markets who has been sharing useful knowledge during this process.

I would like to thank my supervisor at the department of Mathematical Statistics at Lund University, Associate Professor Erik Lindström, for his valuable feedback and help during this project. Finally I would like to thank my family and friends for all encouragements in writing this thesis.

Thank you,

Lovisa Thordin

Contents

Abstract	i
Acknowledgements	ii
Contents	iii
List of Figures	v
Abbreviations	vi
1 Introduction	1
1.1 Background	1
1.2 Problem Presentation	3
1.3 Outline	3
2 The Foreign Exchange Market	4
2.1 Trading in the FX Market	5
2.1.1 Instruments	6
2.1.2 Market Participants	7
2.1.3 Prime Brokerage Relationship	8
2.2 Analysing Exchange Rates	8
2.3 The Role of Nordea	9
3 Order Flow	10
3.1 What is Order Flow?	11
3.2 Academic Review	11
3.2.1 Disaggregated Customer Flow	12
3.2.2 Information Content	13
3.2.3 Forecasting Power of Order Flow	14
4 Machine Learning Techniques	16
4.1 Clustering	17
4.1.1 Requirements and Possible Problems	17
4.1.2 Quality Measures	18
4.2 Clustering Algorithms	19
4.2.1 K-means	20
4.2.2 Fuzzy C-means	21

4.2.3	Hierarchical Clustering	22
4.2.4	Self-Organizing Maps	24
5	Data	26
5.1	Trading data	26
5.1.1	Data Filtering	26
5.1.2	Initial Counterparty Labels	28
5.1.3	The Bank Flow	29
5.1.4	Descriptive Statistics of the Data	29
5.1.5	Prepare Data for Clustering	31
5.2	The Spot Exchange Rate Data	33
6	Methods	34
6.1	Order Flow	34
6.1.1	Calculate Index	35
6.1.2	Confidence Interval for the Index	36
6.2	Clustering	37
6.2.1	Choose Features	37
6.2.2	Cluster Evaluation and Interpretation	39
7	Results	41
7.1	Order Flow	41
7.1.1	Customer Group Indices	43
7.1.2	Bank Indices	47
7.2	Clustering	48
7.2.1	Clustering of the NOK Flow	50
7.2.2	Clustering of the SEK Flow	58
7.2.3	Clustering of the Asset Managers' and Corporates' Flow	64
8	Conclusions	70
8.1	Order Flow	70
8.2	Clustering	71
8.3	Summary	71
9	Discussion	73
A	Correlations and Regressions	76
A.1	Methods	76
A.1.1	Correlation Coefficients	76
A.1.2	Regressions	77
A.2	Results	77
B	Features for Initial Customer Groups	80
	Bibliography	83

List of Figures

5.1	NOK Traded Volume	30
5.2	SEK Traded Volume	31
7.1	NOK Customer Flow.	42
7.2	SEK Customer Flow.	43
7.3	NOK Disaggregated Customer Flow.	44
7.4	SEK Disaggregated Customer Flow.	45
7.5	NOK Customer Indices	46
7.6	SEK Customer Indices	47
7.7	NOK Bank Indices	48
7.8	SEK Bank Indices	49
7.9	NOK Scatter Plot	51
7.10	SEK Scatter Plot	52
7.11	NOK Number of Clusters	53
7.12	NOK Silhouette Plots	53
7.13	NOK K-means Clusters	54
7.14	NOK K-means Cluster Indices	55
7.15	NOK FCM Clusters	56
7.16	NOK FCM Cluster Indices	57
7.17	NOK SOM Clusters	58
7.18	NOK SOM Cluster Indices	59
7.19	SEK Silhouette Plots	59
7.20	SEK K-means Clusters	60
7.21	SEK K-means Cluster Indices	61
7.22	SEK FCM Clusters	62
7.23	SEK FCM Cluster Indices	63
7.24	SEK SOM Clusters	64
7.25	SEK SOM Cluster Indices	65
7.26	SEK AM Clusters	66
7.27	SEK AM Cluster Indices	67
7.28	SEK CO Clusters	68
7.29	SEK CO Cluster Indices	69
B.1	NOK Group Scatterplot	81
B.2	SEK Group Scatterplot	82

Abbreviations

BIS	Bank of International Settlement
DKK	Danish Krone
EBS	Electronic Broking System
ECN	Electronic Communication Network
FCM	Fuzzy C-Means
FX	Foreign Exchange
m EUR	million euro
NOK	Norwegian Krone
OTC	Over The Counter
RFQ	Request For Quote
SEK	Swedish Krona
SOM	Self-Organizing Map
USD	US Dollar

Chapter 1

Introduction

1.1 Background

The foreign exchange (FX or forex) market is the second largest financial market in the world¹ [Lindström, Madsen, Nygaard, 2015]. It is made up of banks, commercial companies, central banks, investment management firms, hedge funds, and retail forex brokers and investors. These agents buy, sell, exchange and speculate on currencies. The FX market is not a single market, but is constructed of a global network of computers that connect participants from all parts of the world. The most active trading centres are geographically located in London and New York, but also in other places such as Singapore and Hong Kong. Participants in the market are spread all over the globe. Copenhagen is the primary city from where liquidity in the Scandinavian currencies is provided by Scandinavian banks.

This thesis is written in cooperation with the FX Quant team at Nordea Markets in Copenhagen. Nordea is one of the largest banks in Scandinavia (next to SEB and Danske Bank). With a variety of customers trading foreign currencies Nordea accounts for approximately 20% of the market in the Scandinavian currencies. The tasks of the FX Quant team comprise supporting the quantitative side in the FX business at Nordea from a trading perspective.

Movements in the FX market are rapid, the rates are updated several times each second. When trading strategies for horizons from one day and above are created different

¹Interest rates account for the largest market value.

macroeconomic factors such as interest rates, GDP trends, and volatility in the market are usually considered. There can be different strategies for different horizons and currencies, with the overall purpose naturally being to profit, with bounded risk, from trading. There is always a need for analysing the market, as well as important events in the world economy, and to develop the trading strategies.

Since August 2013 the FX Quant team at Nordea is doing research in the area of *Order Flow*, which represents a microeconomic viewpoint of analysing exchange rates. This is a relatively new field in the academic literature pioneered by Martin D. D. Evans and Richard K. Lyons in 2002 [Evans, Lyons, 2002]. The discovery was that exchange rates are connected to the order flow, as there is empirical evidence of a positive correlation between the log of the spot exchange rate changes from one day to another and the net of buys and sells of currencies. These findings catalysed a number of articles and papers on the topic, further investigating this relationship.

Several researchers soon found that the order flow from different customer groups had different correlations with exchange rates. As an example, flow from leveraged financial institutions was showed to be positively correlated with log spot changes in exchange rate, while flow from non-financial corporates turned out to have the opposite relationship. Information like this is of high interest to banks that have a large client base since the way that customers trade with them might give a clue to in which direction the currency prices are heading. It is still unknown if it is possible to actually profit from this knowledge.

As mentioned earlier, Nordea has a significant share of the market in the Scandinavian currencies, commonly referred to as "Scandies". These are the Swedish Krona (SEK), the Norwegian Krone (NOK), and the Danish Krone (DKK). Implicitly it is valuable for Nordea to examine the behaviour in their own order flow, to learn about the price impact of the trades of their customers. The order flow data used in this thesis consists of information about trades done in NOK and SEK during 2012 to 2014. It is especially interesting to investigate the flow from different customer groups, to see what differences that could be found in their trading behaviour. In the previous research performed by Nordea, the customers have been labelled after what type of client it is. The groups labels are banks, asset managers, corporates, private clients, hedge funds, central banks and others. There might be other ways of separating the clients that can give new

information about the trading behaviour and price impact. This issue will be addressed in this thesis using tools from a field called *Machine Learning*.

1.2 Problem Presentation

With the background presented in previous section, this project aims at answering the following questions:

1. What information can be found in the customer order flow of Nordea?
2. Will Machine Learning techniques when applied to order flow data of Nordea result in another customer grouping which
 - (a) is different from the typical one (corporates, hedge funds, banks, private clients etc.)?
 - (b) contains information about customer behaviour and the impact on the underlying exchange rate?
3. Can Machine Learning techniques provide information about the different customers trading through a prime brokerage relationship with other banks?

1.3 Outline

The three following chapters in this thesis will give background knowledge and theory about the main topics of the thesis; Chapter 2 will make the reader familiar with the FX market while Chapter 3 describes some previous results about order flow. In Chapter 4 the Machine Learning techniques used in the thesis will be explained.

The order flow data set used for all computations will be presented in Chapter 5, thereafter the used methods will be presented in Chapter 6. All results on order flow and clustering methods are exposed in Chapter 7. Further the conclusions will be given in Chapter 8, with a discussion of the findings and steps further of the thesis in Chapter 9.

Chapter 2

The Foreign Exchange Market

The foreign exchange market is considered to be the most liquid global market today. The daily turnover in April 2013 was \$5.3 trillion according to a survey made by Bank for International Settlements [BIS, 2013]. This is up from \$4.0 trillion in April 2010 and \$3.3 trillion in April 2007. It is a decentralised market comprising two distinct groups of participants: dealers and end-user customers [Della Corte, Rime, Sarno, Tsiakis, 2011]. End-users always trade currencies with dealers while dealers can trade with either customers or other dealers, the latter referred to as the *interdealer market*. Since most dealers are banks it is also called the *interbank market*.

The huge trading volume, the wide geographical dispersion and the continuous operation with trading hours between 22:00 GMT on Sunday (Sydney) to 22:00 GMT on Friday (New York) are unique characteristics of the forex market. Other properties are the variety of factors that affect exchange rates and the low margins of relative profit compared to other markets of fixed income. These are some of the arguments why the FX market has been referred to as the market closest to the ideal of perfect competition.

As mentioned, the global FX market has been growing in turnover over the last decade. The US dollar is the world's dominant vehicle currency being on one side of the transaction in 87% of all deals initiated in April 2013. The euro is the second most traded currency, although its market share has decreased by almost 6 percentage points to 33% between 2010 and 2013. The Japanese yen is also among the most traded currencies globally. Other currencies such as Mexican peso, Chinese renminbi and Russian ruble

show upward trends since they have been changed from pegged to free floating [BIS, 2013].

2.1 Trading in the FX Market

Foreign exchange is an over-the-counter (OTC) market where brokers/dealers negotiate directly with one another without any central exchange or clearing house. A majority of the trading nowadays takes place at different Electronic Communication Networks (ECNs) and other electronic platforms provided by banks and brokers. Deals are still made by phone as well, but not at all to the same extent as ten years ago when this was the main channel for trading FX.

An agent trading in the FX market chooses a currency pair for trading, for instance euro against US dollars, denoted EURUSD. Here euro is called the "base" currency and US dollar is the "price" currency¹. Buying a currency pair means paying an amount in the price currency to receive an amount in the base currency determined by the "ask" price quote. Selling a currency pair means receiving an amount in the price currency by paying the corresponding amount in the base currency according to the "bid" quote. The prices of a currency pair are always given in the price currency value of one unit of the base currency. To avoid arbitrage, the ask price should always be higher than the bid price. The difference between the bid and ask prices is called the *spread* and can be given in *pips* or *basis points*. The spreads vary between currency pairs and give indications on the liquidity in the market.

The foreign exchange market has two tiers. In one tier, customers trade with dealers at banks, in the other dealers trade with each other. The customer-dealer trades are only observed by the parties to the trades and, since there are no disclosure requirements in foreign markets, banks do not report them [Bjønnes, Osler, Rime, 2009].

There are two possible routes that interdealer trading can take in the FX market. Most commonly dealers trade via the ECNs which are firms that provide traders with an electronic platform that allows them to buy and sell foreign currencies. The other alternative for dealers is to trade through phone deals where they call each other directly and request quotes just like regular customers. The interdealer market is virtually

¹Other names for price currency are settlement currency or term currency.

unrecognizable from its form in the 1990s and earlier. Direct interdealer trading is almost non-existent, today trades happen almost exclusively through the ECNs such as the Reuters Dealing system and Electronic Broking Services (EBS) [Foster, Rosov, 2014]. Their matching engines perform limit checks and match orders, usually in less than 100 milliseconds per order. By trading through an ECN, a currency trader generally benefits from greater price transparency, faster processing, increased liquidity and more availability in the marketplace. Banks also reduce their costs as there is less manual effort involved in using an ECN for trading. To trade with an ECN, one must be a subscriber or have an account with a broker that provides direct access trading.

Retail customers normally do not have access to the ECNs but trade with banks through the bank's sales desk or through an online portal that most banks operate. Customers can also trade via *brokers*, who serve as agents of the customers in the broader FX market, by seeking the best price in the market and dealing on behalf of the customers. The brokers charge a fee in addition to the price obtained in the market.

There are different ways and instruments to use when trading currencies, the most common will be described in section 2.1.1. Depending on the goal and purpose of trading there can be advantageous choices of contracts for the trades. FX swaps were the most actively traded instruments in April 2013, \$2.2 trillion per day. After swaps the second most actively traded instrument in April 2013 was spot trades, accounting for \$2.0 trillion per day [BIS, 2013].

2.1.1 Instruments

The list below contains the definitions of the most common instruments used when trading in the FX market. The contracts are different in the forming of conditions and maturity.

Spot transactions Single outright transactions involving the exchange of two currencies at a rate agreed on the date of the contract for a value or delivery (cash settlement) within two business days.

Outright forwards Transactions involving the exchange of two currencies at a rate agreed on the date of the contract for value or delivery (cash settlement) at some time in the future (more than two business days).

Foreign exchange swaps Transactions involving the actual exchange of two currencies on a specific date at a rate agreed at the time of the conclusion of the contract, and a reverse exchange of the same two currencies at a date further in the future at a rate (generally different from the rate on the agreement day) agreed at the time of the contract.

Currency swaps Contracts which commit two counterparties to exchange streams of interest payments in different currencies for an agreed period of time and/or to exchange principal amounts in different currencies at a pre-agreed exchange rate at maturity.

OTC Options Option contracts that give the right to buy or sell a currency with another currency at a specific exchange rate during a specific period. These contracts are non-standardised and the participants can choose the characteristics themselves.

Source [[BIS, 2013](#)].

2.1.2 Market Participants

The forex market has the most varied client base of any other product, with shifting objectives (hedging, risk taking), different horizons (intraday to ten years), and various styles (rational, irrational). The market participants can be grouped in different ways, here six main categories will be briefly described.

Corporates Importing and exporting firms that need to exchange money from payments in other countries.

Real Money Investors Asset managers like for instance pension funds, mutual funds and insurance companies who are typically unleveraged institutions that trade FX as part of their investments.

Hedge Funds Leveraged financial institutions who use advanced investment strategies to generate high returns through a portfolio of different investments.

Central Banks The entities responsible for overseeing the monetary system for nations, implementing goals such as currency stability.

Banks Large enough banks act as dealers in the FX Market and trade both with customers and other dealers, the latter referred to as the Interbank Market.

Regional Banks Smaller banks with no access to the Interbank Market act as customers to larger banks.

High Net Worth Individual Wealthy individuals who choose to invest significant amounts in the FX market.

The BIS Survey from 2013 shows that reporting dealers in the Interbank market accounts for 39% of the total turnover, while a smaller part, 9%, comes from non-financial customers such as corporates and high net worth individuals. The largest share, 53% of the turnover belongs to other financial institutions including real money investors (11%), hedge funds and propriety traders (11%) and regional banks (24%).

2.1.3 Prime Brokerage Relationship

A prime brokerage relationship is a certain agreement between large and highly rated banks and their clients (often institutional funds, hedge funds and other proprietary trading firms), through which the clients can trade currencies in the bank's name. The bank then has the role of a prime broker who enables its clients to conduct trades, subject to credit limits, with a group of pre-determined third-party banks. It may grant the client access to electronic platforms normally available only to large dealers. The prime broker becomes the counterparty to both the client and the third-party bank. Dealers were requested to report how much of their total turnover was attributed to transactions conducted in a foreign exchange prime brokerage relationship for the first time in the 2013 Central Bank Survey [[BIS, 2013](#)].

2.2 Analysing Exchange Rates

Anyone who has done studies related to exchange rates has most likely been taught that the best model for explaining exchange rate movements is a random walk, [[Meese, Rogoff, 1983](#)]. In their paper from 1983, Meese and Rogoff state that the random walk model explains and forecasts exchange rates better than economic models, and nobody

really managed to prove them wrong after that. Many have tried to explain exchange rate fluctuations with macroeconomic fundamentals but it has turned out to be a challenging task. In the beginning of the 21st century the microstructure models for exchange rates got a boost with the pioneering article covering *order flow* by Evans and Lyons [[Evans, Lyons, 2002](#)]. This is also the starting point of the work of this thesis.

2.3 The Role of Nordea

Nordea is one of the largest banks in the Nordic region providing services in FX trading. As any other bank Nordea is profiting from being in the middle between trading agents matching their trades, and receiving the spread that customers pay to trade. The aim and purpose is the same as for any other business, namely to support customers with good service and provide competitive prices in the market, while managing risk that comes with trading. Apart from trading with customers, Nordea is also participating in the interdealer market where they both take liquidity from larger banks and provide liquidity to smaller banks. They also have a large client base, mostly made up by corporations of various kinds and sizes, to whom they provide currency prices. Customers can trade with Nordea through contact with the sales team or via electronic platforms.

Chapter 3

Order Flow

There is little evidence that standard macroeconomic model of exchange rates can explain short-term exchange rate movements beyond the impact of news announcements. The literature primarily suggests that macroeconomic fundamentals such as money supplies, prices and income levels can explain exchange rate movements over horizons in excess of two years [Marsh, O'Rourke, 2005]. In their work from 2002, Richard K. Lyons and Martin D. D. Evans provide a new approach to the analysis of exchange rate movements [Evans, Lyons, 2002]. They augment the traditional macro analysis with a micro-based viewpoint, in which the focus lies on the study of order flow. A contemporaneous relationship between the daily order flow and exchange rate movements was discovered and launched the interest for further investigations in the area of order flow. Traditionally exchange rates have been considered to follow a random walk and the analysis of movements has been limited to this assumption. The microstructural study opens a new way of looking at exchange rate movements.

In this chapter order flow will be explained and a summary of results from former academic research will be given. It should be mentioned that this thesis only includes parts of the literature on order flow. There is a diversity in the approaches and results in this field, where not all are accounted for here.

3.1 What is Order Flow?

Order flow is defined as the net of buyer-initiated and seller-initiated trading orders; it is a measure of net buying pressure. Some articles consider order flow as the number of trades where the flow is +1 (−1) for a buy (sell) of the currency pair [Bjønnes, Osler, Rime, 2009, Evans, Lyons, 2002]. Others will look at the signed volume of the trades [Della Corte, Rime, Sarno, Tsiakis, 2011, Fan, Lyons, 2003] which is also the case in this study. The initial research uses order flow from the interdealer market [Evans, Lyons, 2002], while the latter research is focusing on customer order flow [Marsh, O'Rourke, 2005, Menkhoff, Sarno, Schmeling, Schrimpf, 2013].

3.2 Academic Review

Since the foundation of order flow analysis in the foreign exchange market was laid in 2002 several papers have been written on the topic. A common approach throughout the articles studied for this thesis is to aggregate the order flow over different time horizons such as one day or one month and let x_t denote the volume order flow at time t . One method to measure the relationship between exchange rates is to perform regressions like

$$\Delta s_t = \beta_0 + \beta_1 x_t, \quad (3.1)$$

where Δs_t is defined as the difference between the logarithms of the spot exchange rates at time t and the time h steps before or after the trade. This can be written

$$\Delta s_t = \begin{cases} \log s_t - \log s_{t-h}, & \text{for } h \text{ steps before the trade,} \\ \log s_{t+h} - \log s_t, & \text{for } h \text{ steps after the trade.} \end{cases} \quad (3.2)$$

The sign of the estimated β_1 coefficient reveals the direction of the dependence and the estimates are tested at the chosen significance level. Some academic literature uses the excess returns instead of spot rate changes [Evans, Lyons, 2006, Menkhoff, Sarno, Schmeling, Schrimpf, 2013]. Excess returns are defined as $er_{d+h} \equiv \log(s_{d+h}) - \log(s_d) + \Delta i_d^h$ where s_d is the quote at day d and Δi_d^h is the interest differential on day d for h deposits. Order flows from each segment are aggregated from day d to day $d + h - 1$ [Evans, Lyons, 2006]. The conclusions that can be made are in principle the same.

There exist some limitations of research in this area due to lack of proper data sets. The earliest results that showed a positive contemporaneous correlation between order flow and exchange rate changes made with smaller data sets have been confirmed in later research dealing with larger data sets. The focus is as time passes shifted from the interdealer trades to the customer trades. The reason for this shift is that the customer orders catalyse the market, implying that customer flow is the ultimate driver of interdealer flow. Data on customer order flow is sensitive information for banks and was not available in the beginning but as this data became available researchers could begin studies on customer order flow [Fan, Lyons, 2003]. It is a stylised fact that dealers open and close their trading day with zero inventory positions. Considering customer order flow aggregated over one trading day, as in most of the papers, the expected net flow is zero. This is an approximation since the foreign exchange market never closes. Daily customer order flow from a representative bank should therefore be only randomly different from zero and uncorrelated with exchange rate movements. However, the entire market can not be represented by one single bank [Marsh, O'Rourke, 2005].

To fully understand the connection between order flow and exchange rate dynamics, one would need access to the entire market with all buys and sells of currencies. Because of the way the FX market is constructed, no single market participant possesses the full overview of the market. A large share of the market is needed before any significant correlation can be expected to be found between the order flow data and changes in currency prices.

After confirming the contemporaneous relationship between order flow and exchange rate changes the question is naturally arisen whether there is predictive content in order flow as well. Further one might ask if it is possible for banks to exploit this possibly existing information to create profitable strategies in trading FX.

3.2.1 Disaggregated Customer Flow

Short after the initial results on order flow it was discovered that different components of the end-user order flow have different correlations with exchange rate movements. The order flow was divided into different customer groups such as non-financial corporates, unleveraged financial institutions (e.g. mutual funds) and leveraged financial institutions (e.g. hedge funds) [Evans, Lyons, 2006, Fan, Lyons, 2003], or alternatively

asset managers, hedge funds, corporates and private clients [Menkhoff, Sarno, Schmelting, Schrimpf, 2013]. In treating all order flow equally one assumes that agents are symmetrically heterogeneous, which means that they differ but in the same way. However, trades in the foreign exchange market come from categories of agents that are quite different, they differ in motivations, attitudes towards risk and horizons [Evans, Lyons, 2006]. Performing regressions like

$$\Delta s_t = \beta_0 + \beta_1 x_t^{Corp} + \beta_2 x_t^{Unlev} + \beta_3 x_t^{Lev} + \beta_4 x_t^{Other} + u_t, \quad (3.3)$$

where the flow is disaggregated according to customer groups shows some interesting behaviour. Flows from profit-seeking financial institutions are positively correlated with exchange rates movements, while flow from non-financial corporates are typically negatively correlated. The trading motivation for the category "other financial institutions" is not clear and results of the regressions for this group are mixed.

3.2.2 Information Content

Several papers have attempted to find the cause of the contemporaneous relationship between exchange rate changes and net order flows. Marsh and O'Rourke investigate three different explanations of the correlations. First there may be private information contained in the order flow that is relevant for the valuation of a currency in a non-transitory way. The authors find evidence that a measure of the degree of informedness of customers corresponds to the size of the correlation between order flow and exchange rates. The second explanation is that there are transitory liquidity effects on exchange rates, coming from that risk-averse dealers may need to be compensated for holding unwanted inventory. This explanation is discounted since otherwise equivalent order flow from different counterparties have different correlations with exchange rates. Finally they reverse the causality and argue that changes in exchange rate induce flows, so called feedback trading. It can not be rejected that feedback trading could be an explaining factor [Marsh, O'Rourke, 2005].

Bjønnes, Osler and Rime address the information hypothesis in the interbank market by dividing banks into groups according to their grade of informedness. They suggest that information in the FX market comes from customers, implying that banks with the most customers (large banks) should have the most information. Partitioning banks after the

contemporary size ranking by *Euromoney* results in the the four groups "Immense", "Big", "Medium" and "Small" banks. The findings include that order flow from banks of all sizes have positive price impact that is achieved within roughly five trades and persists thereafter, and this price impact is more substantial for larger banks [Bjønnes, Osler, Rime, 2009].

3.2.3 Forecasting Power of Order Flow

The earlier literature on order flow is focusing on *explaining* exchange rate movements [Evans, Lyons, 2002, Fan, Lyons, 2003]. If order flow is proven to contain information about future movements it opens the possibility to create strategies and profit from that knowledge, and this would certainly be of more interest to practitioners. Evans and Lyons investigate the issue on a monthly basis, by regressing excess returns between month t and $t + 1$ on Citibank's customer order flows during month $t - 1$. They find that the six customer segments in their flow account for approximately 19 percent of the variation in future excess returns. There are large differences between the estimated coefficients across customer flow segments. The main findings include that customer flows predict returns because they are correlated with the future market-wide information flow that dealers use to revise their currency prices. They also conclude that one third of flow's power to forecast exchange rates one month ahead comes from flow's ability to forecast future flow [Evans, Lyons, 2006].

In on of the papers in this study the question about the source of the forecasting power of order flow is addressed. The authors compare the performance of a trading strategy based on order flow with seven other strategies used for trading foreign exchange. They implement a multi-currency dynamic asset allocation strategy and relate the excess portfolio returns from the order flow strategy to the ones from public information strategies. The first finding is that the order flow strategies strongly outperform the standard carry trade strategy¹. Furthermore, about two thirds of the excess portfolio returns generated from conditioning on order flow can be replicated using a combination of strategies based on publicly available information. Hence variables included in these strategies such as interest rates, money supply, output, external imbalances and momentum which capture information that is publicly available play an important role in determining the

¹The carry strategy is based on the assumption of a random walk and is used as a benchmark here. It borrows in low interest rate currencies and lends in high interest rate currencies.

net demand for currencies as observed in FX order flow. The authors conclude that the information conveyed by customer order flow is a particular aggregation of public information, where customers collect and interpret information in their own way [[Della Corte, Rime, Sarno, Tsiakis, 2011](#)].

Chapter 4

Machine Learning Techniques

Machine Learning is a subfield in statistics and computer science that deals with algorithms that can learn from data. It studies how to automatically learn to make accurate predictions based on past observations. One distinguishes between **supervised learning**, where an algorithm is trained on a training set to predict values for a test set, and **unsupervised learning**, where no labels are predefined on the data set. Supervised learning is further divided into *regression* and *classification*, depending on whether the predicted output is quantitative or qualitative. Both tasks have a lot in common and can be viewed as different methods for function approximation [[Hastie, Tibshirani, Friedman, 2009](#)].

Machine Learning techniques are cheap and flexible and can be applied to several areas of interest, such as image analysis, pattern recognition and biology. They have the advantages that they are often more accurate than human-crafted rules (since data driven) and don't need a human expert or programmer. Humans are often incapable of expressing what they know but can easily classify examples that can be used in an automatic method to search for hypotheses explaining data. Some disadvantages of Machine Learning algorithms are that a lot of labelled data is needed. It is also usually impossible to get perfect accuracy in the results because of the many error sources.

The following section will describe *clustering* and a few of its techniques, which fit in the category of unsupervised learning.

4.1 Clustering

The overall goal of clustering is to find groups, so called "clusters", in data where the data points within each cluster are more similar to each other than to points in other clusters. There are multiple factors to consider when grouping data, for instance in what way similarity between data points is measured, the optimal number of clusters and which algorithm to use. Each user and application can have different aims for the clustering, such as finding representatives for homogeneous groups, finding natural clusters and describe their unknown properties or detecting outliers. The clustering algorithms considered in this thesis are among the most commonly used and are presented in Section 4.2.

One can easily confuse classification within supervised learning with the unsupervised clustering described in this thesis. The difference lies in that in classification the classes are predefined and the target is to classify new points into the existing clusters, while unsupervised clustering gives an unknown grouping of objects where relationships in the data are to be discovered [Bishop, 2006].

In this thesis the objects to be clustered are the customers trading currencies with Nordea. The goal is to group the customers in new way that can hopefully add information about their trading behaviour and impact on currency prices.

4.1.1 Requirements and Possible Problems

A clustering technique is required to handle scalability, different types of attributes and have the ability to deal with noise and outliers. It should also be able to handle a large dimensionality and be insensitive to the order of the input parameters. The result must be interpretable and usable and it should be possible to discover clusters of arbitrary shape.

Current techniques do not address all requirements adequately. Different scaling leads to different clusterings, which must be accounted for when preparing the data. High dimensionality and many observations can be problematic because of time complexity. The effectiveness of the method depends on the definition of the distance and if an

obvious distance measure does not exist it must be defined, which can be complex. The result of the clustering can be interpreted in many different ways [Matteucci, 2003].

4.1.2 Quality Measures

There is a number of ways to measure and validate the quality of clustering techniques for different problems. Two of them, the Calinski-Harabasz criterion and the Silhouette value, will be used in this thesis and are explained here.

Calinski-Harabasz Coefficient

The Calinski-Harabasz Coefficient (CH) is the ratio between the normalised between-cluster sum of squares SS_B and the normalised within-cluster sum of squares SS_W . It is also called the variance ratio criterion (VRC) and is defined as

$$VRC_k = \frac{SS_B/(k-1)}{SS_W/(N-k)}, \quad (4.1)$$

where N is the number of observations and k is the number of clusters. The CH coefficient can be used to determine the number of clusters in a data set, and works well in many situations, [Mooi, Sarstedt, 2014]. Well-defined clusters have a large between-cluster and a small within-cluster variance. The larger the CH coefficient, the better the data partition.

Silhouette Coefficient

The silhouettes can be useful when the proximities are on a ratio scale (as in the case of Euclidean distances) and when one is seeking compact and clearly separated clusters. Let $a(i)$ denote the average distance or dissimilarity of object i to all other objects within the same cluster as i . Further let $b(i)$ denote the lowest average dissimilarity of i to all objects of any other cluster, in other words the distance to the second most similar cluster of i . The silhouette coefficient is defined

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}. \quad (4.2)$$

and measures how well object i matches the clustering at hand. From the above definition it is easily seen that $-1 \leq s(i) \leq 1$ for each object i , [Rousseeuw, 1987].

4.2 Clustering Algorithms

There is a variety of clustering algorithms that can be applied to any data set. The resulting clusters depend on the used algorithm. In the following subsections the K -means, Fuzzy C -means, Hierarchical and Self-Organizing Map clustering algorithms will be explained. For this some notation will first be introduced.

Suppose $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ represents the data set consisting of N observations of D -dimensional data. On matrix form this becomes

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_N \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1D} \\ \vdots & \ddots & \vdots \\ x_{N1} & \cdots & x_{ND} \end{bmatrix}. \quad (4.3)$$

Here N is typically far larger than D , $N \gg D$. The dissimilarity or distance measure between two data points is denoted d , the most used one is the Euclidean distance defined as

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{k=1}^D (x_{ik} - x_{jk})^2}. \quad (4.4)$$

The Euclidean distance is a special case of the Minkowski metric, with $p = 2$. The Minkowski metric for D -dimensional data is for $p \geq 1$ defined as

$$d_p(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_{k=1}^D |x_{ik} - x_{jk}|^p \right)^{\frac{1}{p}} \quad (4.5)$$

[Matteucci, 2003].

Determining the optimal number of clusters in a data set is usually a challenging task. The clusters should be well-separated, meaning that the within-cluster distances should be small and the between-cluster distances should be larger. Many of the clustering algorithms require that the number of clusters is specified by the user.

4.2.1 K-means

The most common and simplest of all clustering methods is the K -means algorithm. It is intended for situations in which all variables are of the quantitative type. The main idea of the K -means is to define K cluster centres and assign each point in the data set to the cluster centre most similar or closest to the point. It is an exclusive clustering algorithm, meaning that the clusters are disjoint sets of data points. The initial step involves choosing a number of clusters K and defining the same number of initial cluster centres, $\{\mathbf{c}_k\}_{k=1}^K$. In the following step new cluster centres are calculated as the mean of the points in that particular cluster. Next, each point forgets the previous cluster and is assigned to the new closest cluster centre. This could be done by either *batch* updates or *online* updates. In the batch update all points are assigned to their respective cluster centre at once and thereafter the new centres are computed, while the online version assigns each point individually and cluster centres are recomputed after each assignment. The process is repeated until the cluster centres have converged so that no more points change cluster. Finally, the algorithm aims at minimizing an objective function, in this case the squared error function, defined as

$$J = \sum_{i=1}^N \sum_{k=1}^K r_{ik} \|\mathbf{x}_i - \mathbf{c}_k\|^2, \quad (4.6)$$

where the variable $r_{ik} \in \{0, 1\}$ indicates if the point belongs to cluster k or not;

$$r_{ik} = \begin{cases} 1, & \text{if } k = \arg \min_j \|\mathbf{x}_i - \mathbf{c}_j\|^2 \\ 0, & \text{otherwise.} \end{cases} \quad (4.7)$$

The objective function represents the sum of the squared distances from each point to its assigned cluster centre and the goal is to find values for the $\{r_{ik}\}$ and the $\{\mathbf{c}_k\}$ so as to minimize J . Each phase of step 2 and 3 below reduces the value of J , implying that convergence is assured. However, it may converge to a local minimum rather than the global minimum. The main steps are summarized in the box below.

***K*-means Algorithm**

1. Choose K initial points in the object space to represent the cluster centres.
2. Assign each point to the cluster that has the closest centre.
3. Recalculate the positions of the K cluster centres.
4. Repeat steps 2 and 3 until the centres no longer move. The metric to be minimized can be calculated from the resulting separation of the points.

[Bishop, 2006, Hastie, Tibshirani, Friedman, 2009, Matteucci, 2003]

4.2.2 Fuzzy C-means

In opposite to K -means, the Fuzzy C -means (FCM) algorithm allows points to belong to more than one cluster which is the case in so called soft or overlapping clustering. This is done by including a partition matrix U containing the degree of membership of the data points to each cluster, such that the objection function becomes

$$J_m = \sum_{i=1}^N \sum_{k=1}^C u_{ik} \|\mathbf{x}_i - \mathbf{c}_k\|^2, \quad 1 \leq m < \infty, \quad (4.8)$$

where $u_{ik} \in [0, 1]$ is the degree of membership of data point \mathbf{x}_i in cluster k , \mathbf{c}_k is the D -dimensional cluster centre and $\|\cdot\|$ is any norm. The parameter m determines the "fuzziness" of the clustering, a large m gives smaller memberships u_{ik} which results in fuzzier clusters. If m is set to the limit value 1 the membership will converge to 0 or 1 and hence a crisp partitioning is obtained (gives the K -means algorithm). A property of the membership matrix U is that each row sums to one, $\sum_{k=1}^C u_{ik} = 1$. The columns in the membership matrix could be interpreted as the probabilities of the points being in the cluster represented by each column.

In the process of optimizing the objective function, the parameters u_{ik} and \mathbf{c}_k are updated in the following way:

$$u_{ik} = \frac{1}{\sum_{j=1}^C \left(\frac{\|\mathbf{x}_i - \mathbf{c}_k\|}{\|\mathbf{x}_i - \mathbf{c}_j\|} \right)^{\frac{2}{m-1}}}, \quad (4.9)$$

$$\mathbf{c}_k = \frac{\sum_{i=1}^N u_{ik}^m \mathbf{x}_i}{\sum_{i=1}^N u_{ik}^m}. \quad (4.10)$$

As for the K -means, there might be an issue of the objective function not converging to its global minimum. The main steps of the algorithm are included in the box below.

FCM Algorithm

1. Determine the number of clusters, C , and initialize the $N \times C$ matrix $U = [u_{ik}]$.
2. Calculate the centre vectors $\{\mathbf{c}_k\}$ according to (4.10) for all clusters.
3. Update U , using (4.9).
4. If $\|U^{(n+1)} - U^{(n)}\| < \varepsilon$ then stop, otherwise return to step 2.

[Matteucci, 2003]

4.2.3 Hierarchical Clustering

Clustering with Hierarchical methods creates clusters at different levels, so that no specified number of clusters is needed. The two basic strategies for hierarchical clustering are *agglomerative* and *divisive*, where agglomerative methods start with N clusters with one single observation and ends up with all objects comprised in one large cluster and the divisive strategies work the opposite direction. The agglomerative bottom-up direction is the most widely applied and will be the focus in the rest of this section.

The agglomerative algorithm is an iterative process where at each level two selected clusters are recursively merged together into one single cluster, going from k to $k - 1$ clusters from one level to the next. The two groups with the smallest intergroup dissimilarity are chosen for merging. An appropriate, disjoint clustering of the data is represented at each level of the hierarchy. The user needs to decide which level fits the best for the specific purpose such that points within the clusters are sufficiently more similar to points within their own cluster compared to points in other groups.

The distances or (similarities) between two different clusters denoted C_k and C_l , can be calculated in three different ways¹. *Single linkage* (SL) agglomerative clustering

¹There are other methods as well but these three are the most commonly used.

calculates the distance between two groups as the shortest distance between two points from separate clusters, also called nearest-neighbour technique

$$d_{SL}(C_k, C_l) = \min_{\mathbf{x}_i \in C_k, \mathbf{x}_j \in C_l} d(\mathbf{x}_i, \mathbf{x}_j). \quad (4.11)$$

Complete linkage (CL) agglomerative clustering takes the intergroup distance to be the greatest distance between any two points from different groups,

$$d_{CL}(C_k, C_l) = \max_{\mathbf{x}_i \in C_k, \mathbf{x}_j \in C_l} d(\mathbf{x}_i, \mathbf{x}_j). \quad (4.12)$$

Group average (GA) clustering uses the average distance between the clusters

$$d_{GA}(C_k, C_l) = \frac{1}{N_{C_k} N_{C_l}} \sum_{\mathbf{x}_i \in C_k} \sum_{\mathbf{x}_j \in C_l} d(\mathbf{x}_i, \mathbf{x}_j), \quad (4.13)$$

where N_{C_k} and N_{C_l} are the respective number of observations in each group. Note that in all three cases, the distance or similarity between two points is calculated as for the other algorithms, for instance using the Euclidean metric given in (4.4). All distances between clusters are collected in the proximity matrix, denoted D_m for step m , which initially is of dimension $N \times N$ and given by $D_0 = [d(\mathbf{x}_i, \mathbf{x}_j)]$. The level of the clustering at step m is denoted $L(m)$ and explains the distance between clusters at different levels. When two clusters K and L are merged together into cluster KL , the proximity between the new cluster and the old cluster is defined according to the chosen linkage method. A step by step description of an agglomerative algorithm is given in the box below.

Agglomerative Hierarchical Algorithm

1. Define the N data points as N distinct clusters.
2. Find the least dissimilar pair of clusters according to (4.11), (4.12) or (4.13), denote these clusters K and L .
3. Set $m = m + 1$ and merge clusters K and L into a single cluster. Set the level of this clustering to $L(m) = d(K, L)$ according to chosen linkage method.
4. Update the proximity matrix, D_m by deleting the rows and columns corresponding to clusters K and L .
5. If all objects are in one cluster then stop, otherwise return to step 2.

[Hastie, Tibshirani, Friedman, 2009, Matteucci, 2003]

4.2.4 Self-Organizing Maps

Describing Self-Organizing Maps (SOMs) without giving an introduction to Neural Networks might be a challenge, but an attempt will be made in this section². In this unsupervised learning method of a neural network, the neurons in the network can be viewed as representing the cluster centres. The principal goal of a SOM is to transform an incoming signal pattern of arbitrary dimension into a one or two dimensional discrete map, also called a *constrained topological map*. The original SOM algorithm was online, meaning that observations are processed one at a time, and later a batch version was developed.

Consider K neurons placed in a two-dimensional rectangular grid and denote these $m_j \in \mathbb{R}^D$, where D is the dimension of the input data. The neurons are given integer coordinates in $Q_1 \times Q_2$, where $Q_1 = \{1, \dots, q_1\}$, $Q_2 = \{1, \dots, q_2\}$ and it follows that $K = q_1 q_2$. The SOM tries to bend the two-dimensional plane so that the neurons approximate the data points as well as possible. In this way the model creates a mapping of the data points onto the neurons in the two-dimensional grid.

²More information about Neural Networks can for instance be found in Chapter 11 in *The Elements of Statistical Learning*, [Hastie, Tibshirani, Friedman, 2009].

The online algorithm processes the N observations (inputs) \mathbf{x}_i one at a time. The closest neuron m_j is found in Euclidean distance in \mathbb{R}^D , and then for all neighbours m_k of m_j , move m_k toward \mathbf{x}_i via the update³

$$m_k \leftarrow m_k + \alpha(\mathbf{x} - m_k), \quad (4.14)$$

where α is called the learning rate. The neighbours of m_j can be defined in some different ways. The simplest approach uses Euclidean distance and determines the neighbourhood around neuron m_j by a threshold r , always including m_j itself. Notice that the distance between neurons is defined in the space $Q_1 \times Q_2$ of integer topological coordinates of the neurons. The update 4.14 has the effect that the neurons are moved closer to the data and also that a smooth two-dimensional spatial relationship between neurons is maintained. The performance of the SOM method depends on the learning rate α and the distance threshold r . Over a few thousand iterations α is typically decreased from say 1.0 to 0.0. Similarly r is decreased linearly from its initial value R down to 1. In this way the final mapping of observations to the neurons is created.

If the distance r is taken small enough so that each neighbourhood contains only one point, then the spatial connection between neurons is lost and the SOM can be shown to be an online version of the K -means algorithm, [Hastie, Tibshirani, Friedman, 2009].

In a batch version of the SOM algorithm each m_j is updated via

$$m_j = \frac{\sum \mathbf{w}_k \mathbf{x}_k}{\sum w_k}, \quad (4.15)$$

where the sum is taken over points \mathbf{x}_k that are mapping neighbours m_k of m_j . The weight function may be rectangular, that is, equal to 1 for the neighbours of m_k or may decrease smoothly with the distance between the neurons. If the neighbourhood size is small enough to only include one neighbour and the weights are rectangular, this reduces to the K -means algorithm as described in Section 4.2.1.

³A more sophisticated way would be to modify the update step according to the distance between the neurons.

Chapter 5

Data

5.1 Trading data

The order flow data used in this thesis covers all currency exchange transactions during a period of two years and ten months, from January 2nd 2012 until November 1st 2014. It is divided into two data sets each containing all trades done with one of the currencies being Swedish Krone (SEK) and Norwegian Krone (NOK), respectively. Further on the order flow data will simply be referred to as "the data sets". To each trade there is information about different parameters given in Table 5.1. Some information is given from the data base of Nordea, while other is calculated from existing parameters.

The data set in this thesis could contain errors such as of wrongly booked data, for instance that timestamps are not exact, the traded spot rate is incorrect or that changes have been made manually in the trade. A large amount of time has been used to analyse and prepare the data set as carefully as possible.

5.1.1 Data Filtering

By filtering the data set in certain ways, much of the noise can be removed without losing the significant information. The number of small trades is large and these trades do not have any significant spot impact, hence all trades of less than 100,000 EUR are removed from the data sets. In that way the data becomes easier to handle and the calculation time is shortened. The most active trading hours for Nordea are Monday to

TABLE 5.1: Information given in the order flow data set for each trade.

Basic Information	
timestamp	The exact time stamp when the trade enters the system given in milliseconds. This is not necessarily equal to the execution time, depending on the how the trade is executed (phone deals can have a delay of up to 20 minutes).
cust_tid	Customer identification number which is unique for each counterparty.
base_curr	The base currency.
price_curr	The price currency.
base_amount	Signed traded amount given in the base currency.
price_amount	Signed traded amount given in the price currency.
trade_rate	The spot price at which the trade was carried out.
channel_id	The channel through which the trade is executed.
trade_type	Tells if the trade is a spot trade or an outright forward.
routing_portfolio	Says where the trade is booked.
Calculated Information	
BuySell	1 if the trade is a buy of the currency pair and -1 if it is a sell of the currency pair.
eur_amount	Traded amount converted into million EUR.
group	One of the customer labels defined in Section 5.1.2.
ccypair	The traded currency pair.
ccy_is_base	Logical variable that gives 1 if the actual currency is the base currency and 0 if it is not.
eur_amount_net_ccy	Traded amount in million EUR signed according to the actual currency.
BuySell_ccy	Gives 1 if the actual currency is bought and -1 if it is sold.
round_trade	Logical variable that is 1 if the traded amount is larger than one million euros and done in even millions either in the base currency or price currency, and 0 otherwise.
new_direction_ccy	Logical variable which is 1 if trade is in a new direction of the actual currency compared to the last trade done by the same counterparty, and 0 if it is in the same direction.

Notes: The basic information is fetched from the data base. The remaining information is calculated from the basic information. The "actual currency" refers to NOK and SEK for the two separate data sets.

Friday between 07:00 and 17:00 GMT+1, since the bank is seated in the Nordic region. Trades occurring in the weekends are removed since that flow is so small compared to weekdays. There is also much less activity during holidays, hence trades at these dates were removed as well¹.

¹Holidays in Sweden, Norway, USA and Germany were the ones taken into account, defined by Bloomberg.

The order flow of Nordea can be divided into the customer order flow and the flow in the interbank market. Trades done with banks and prime brokered clients belong to the interbank market, while the remaining counterparty groups are regarded as customers. One could argue that the smaller banks should count as customers as well, if they are not big enough to have direct access to the interbank market. Putting the same label on all banks is a simplification made here. For most of the analysis in this thesis the trades initiated by Nordea have been removed from the data set. With help from people in Nordea working with the order book it is possible to distinguish some of these trades, since it is not initially given in the dataset which side is initiating the trade. It is done by looking at the trading channel for the trades where the counterparty is a bank. For some deals, for instance phone deals with other banks, it is impossible to know who commenced the trade. There is at least one trading channel where all deals that are going through are for certain initiated by Nordea. Both the trades with unknown initiator and the trades where Nordea is known to be the initiator are removed from the flow.

5.1.2 Initial Counterparty Labels

The counterparties in the data sets are grouped according to the type of firm or institution it is. The unique counterparties are identified with the variable `cust_tid`, and the ones which are considered significant in trading volume are classified before the analysis is made. Each counterparty trading a significant amount is labelled with one of the following labels;

AM Asset Managers, i.e. financial institutions including e.g. insurance companies and mutual funds.

BA Banks of all sizes.

CB Central Banks.

CO Non-financial corporates of all sizes.

HF Hedge funds and other leveraged financial institutions.

OT Others, i.e. customers that do not fit into any of the other groups.

PB Customers trading via a larger bank through a prime brokerage agreement.

PC Private clients and speculative persons.

XX Non-labelled customers trading less than the threshold amount.

The FX Quant team has put a lot of effort into putting the right labels on the customers. There are no routines today in Nordea where information about the type of a customer is stored, meaning that this has to be done manually and requires a large amount of time. Unfortunately not all customers are labelled, only the ones with the largest trading volume.

5.1.3 The Bank Flow

When banks trade with each other it is usually referred to as "the interbank market" and this flow is distinguished from the customer flow in the data. From Nordea's perspective, the trades done with banks, i.e. trades labelled "BA", could be of four different types;

1. Trades done by the Algo trading at Nordea, referred to as "Our hedges - Algo" (OHAL).
2. Trades initiated by the traders at Nordea, referred to as "Our hedges" (OH).
3. Trades initiated by other banks, referred to as "Their hedges" (TH).
4. Trades done with a prime brokerage agreement in a bank's name where the customer behind is unknown, referred to as "Prime Brokerage" (PB).

The trades that are initiated by Nordea are identified by which trading channel they are made through and the ones of these done by algo trading are found by a certain value of the routing portfolio parameter. One of the questions in this thesis addresses the issue of finding out what types of clients that are behind the prime brokered trades.

5.1.4 Descriptive Statistics of the Data

Approximately 80% of the volume traded with Nordea in the period from January 2012 to October 2014 are by customers and the rest is flow from the interbank market. How

the customer flow is divided between the pre-labelled groups for the NOK and SEK flow respectively can be seen to the left in Figures 5.1 and 5.2. The right pie chart in these figures show the total traded volume per currency pair.

NOK: Traded Volume

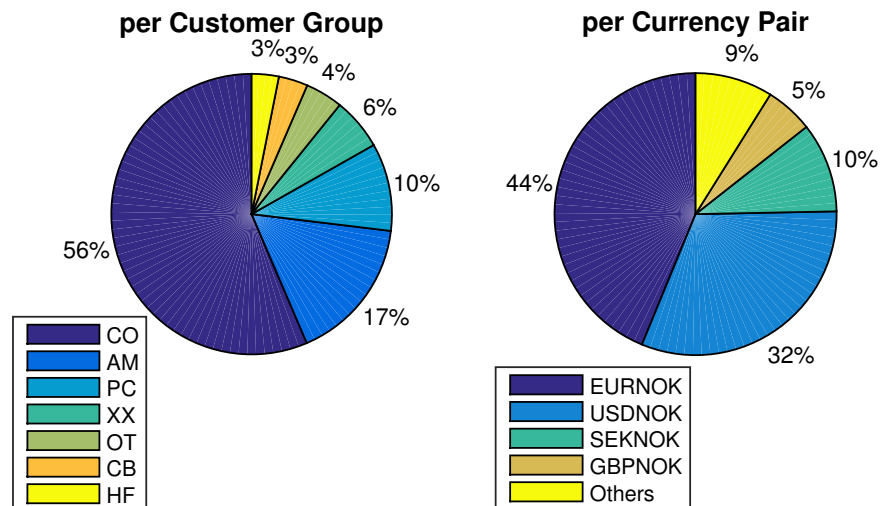


FIGURE 5.1: These pie charts show how the total traded volume in NOK is divided between the pre-labelled customer groups (left) and currency pairs (right). The corporates make up the largest group, followed by the asset managers, private clients, unlabelled clients, others, central banks and hedge funds. The trades in euro, US Dollar, Swedish Krona and British Pounds account for 91% of the total volume traded in NOK.

Slightly more than half of the volume is traded with corporations in both currencies. The second largest group is the asset managers, followed by private clients. The unlabelled clients are trading smaller amounts, and together they account for 5 – 6% of the flow. Hedge funds, Central Banks and "Other" customers compose a small client base for Nordea. Since the corporates and asset managers are the largest groups the focus will be put on these customers in this thesis, since the most significant results are expected to be produced for these groups.

SEK: Traded Volume

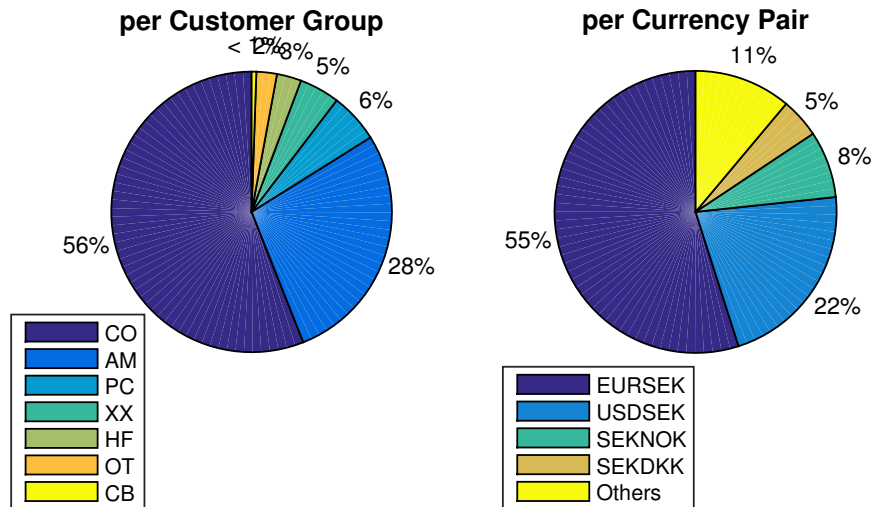


FIGURE 5.2: These pie charts show how the total traded volume in SEK is divided between the pre-labelled customer groups (left) and currency pairs (right). The corporates make up the largest group, followed by the asset managers, private clients, unlabelled clients, hedge funds, others and central banks. The trades in euro, US Dollar, Norwegian Krona and Danish Krona account for 89% of the total volume traded in SEK.

5.1.5 Prepare Data for Clustering

The customers are identified with a customer identification number that is unique for the customer. The data set containing all trades is grouped according to this ID number using Matlab's `grpstats`. This function is instructed to calculate the mean and sum of some given variables of the trades. After discussions with the Quant Team and another team at Nordea, six features are chosen to be considered for the clustering, namely

Buy/sell-ratio A constant between 0 and 1 that is calculated by taking the absolute value of the sum of the net amount divided by the sum of the total amount traded by the customer, which can be written $b = \left| \frac{\sum_j V_j^{signed}}{\sum_j V_j} \right|$. The net amount is positive (negative) if it the currency is bought (sold). A value close to 1 means that the customer is either always buying or always selling the currency, while a value close to 0 means that the customer equally buys and sells the currency.

Traded volume The total volume traded by the customer in any direction given in million euros.

New direction for currency For each trade this variable can take the values 0 or 1, depending on whether the trade is in the same or opposite direction as previous trade on a currency basis. For each customer the mean of this variable for all trades done by the customer is taken.

Round number trading If the traded amount is in even millions in either the base or price currency this parameter is one, otherwise it is zero. For each customer the average of this value for all trades is taken.

High volatility trading Periods with high volatility are defined as intervals of one hour before and one hour after large changes in volatility have occurred. The volatility at minute t is estimated as the standard deviation of the log spot changes the previous 30 minutes. The probability of high volatility in the given period is estimated as one event per week, giving $p = 1/(5 \cdot 24 \cdot 60)$, and large changes are given from the empirical upper and lower quantiles with probability $p/2$ of the volatility changes. The result is a logical variable for each trade that is 1 if the trade is done in a period of high volatility and 0 if it is not.

Trading platform A discrete variable which takes on the value -1 if the customer trades mostly on a Single bank platform, 1 if the customer does mostly Request for Quote and 0 if the customer trade more on other platforms.

These six features are considered to be of interest when investigating the behaviour and price impact of different clients. Nothing is pre-known about the relations, but they are intuitive aspects that might have an impact. If the buy-sell ratio for a customer is close to zero it is said that the flow is non-directional, if it is close to one the flow is considered to be directional. Intuitively, corporates who receive a large amount of one foreign currency would be more directional as they need to exchange that flow into their native currency. The buy/sell-ratio is most likely connected to the new direction for the currency. If the flow is directional, the customer will probably not change the direction of its trades in the actual currency. It is an intuitive thought that a customer's total trading volume will affect the price impact of that customer, which is why it is interesting to consider. The trading platform is important since it gives an indication about the

price sensitivity of the customer. Customers trading on a single bank platform are often not that sensitive to prices since there is often more latency and randomness and not so efficient trading. Clients who request for quotes are normally requiring better prices with narrower spreads. Some customers tend to trade more or less in periods where the volatility is high, for instance around economic releases. Including an estimate of to which extent this behaviour is found for different customers might give an indication of how important this behaviour is.

Clustering is performed without including trades initiated by Nordea. It means that the only bank trades kept in the flow are trades initiated by the the other bank, these correspond to the group "Their hedges" defined in Section 5.1.3. Before a clustering algorithm is applied the customers who trade less than ten times in the covered period are removed. If the customer trades too few times the buy-sell ratio does not make sense since it is dependent on previous trades. Outliers corresponding to customers that trading a relatively large volume are also removed. Otherwise they will have a too large effect on the result when calculating distances between points.

5.2 The Spot Exchange Rate Data

Spot prices for the main currency pairs at each minute during the covered time period are collected from Reuters (EURSEK, EURNOK and EURGBP) and EBS (EURUSD). The mid prices are calculated as the mean of the bid and ask prices. Prices for the remaining currency pairs are obtained out of these prices. As an example, the NOKSEK rate is obtained by dividing the EURSEK rate by the EURNOK rate. Mid prices are considered to be a good approximation instead of using the bid and ask prices in the studies of this thesis.

Chapter 6

Methods

In this chapter the methodology used in this thesis will be described. The first section defines the techniques used for the order flow study and the second section presents the way clustering is performed.

6.1 Order Flow

Inspired by the academic literature about order flow the relationship between exchange rate changes and the order flow of Nordea is investigated. The focus of this thesis for examining the dependence will be to use an *index* defined in Section 6.1.1, which is a method developed at Nordea. Other methods are Pearson's correlation coefficient and multiple linear regression, these are carried out in Appendix A. In opposition to earlier studies, the data set used here contains information about every unique trade and not just the order flow aggregated over some horizon. The advantage is that it is actually possible to see what happens around the trade time, information which is lost if the unique trades are summed up. It could for instance happen that the price moves a lot within one day and the aggregated order flow does not say when trades are made with respect to the large price move.

When performing regressions and calculating indices, the four currency pairs with the largest traded volume for each currency are considered. Recall the charts from Figures 5.2 and 5.1 showing that the most important pairs are

NOK with EUR, USD, SEK and GBP,

SEK with EUR, USD, NOK and DKK.

Removing the least traded currency pair only removes a small part of the total volume and the significant part of the flow is still kept.

6.1.1 Calculate Index

Various methods could be used to measure the impact of the order flow on the spot exchange rate. In this thesis it is of particular interest to investigate the price impact of different customer groups, including both the pre-labelled groups and the ones formed by the clustering techniques. For this an index is constructed which gives a measurement of the currency price changes before and after a trade is executed. The index is similar to the yield decay which is more commonly used in financial analysis. By defining the index as below it is interpreted as change of price in percentage.

The index for each trade and time horizon h is defined as

$$I_{j,h} = \left(\frac{s_{j,h}}{s_{j,0}} \right)^{b_j}, \quad (6.1)$$

where $s_{j,0}$ denotes the traded spot price for trade j given from the trade data. $s_{j,h}$ denotes the spot mid price at time horizon h after the trade, or before if h has a negative sign. For trade j it also holds that

$$b_j = \begin{cases} 1, & \text{if the customer buys the currency} \\ -1, & \text{if the customer sells the currency.} \end{cases} \quad (6.2)$$

For this thesis horizons of half-hours are considered. Now let $\{C_g\} = \{\text{AM}, \dots, \text{XX}\}$ be the different customer groups, then the index at horizon h for customer group C_g is given by

$$I_{C_g,h} = 100 \sum_{j \in C_g} \lambda_j (I_{j,h} - 1). \quad (6.3)$$

The choice of the weights λ_j in this thesis is

$$\lambda_j = \frac{V_j}{V_{C_g}}, \quad (6.4)$$

where V_j is the traded amount in trade j and V_{C_g} is the total traded volume of customer group C_g in million euros. Another option could be to equally weight all trades. The index for each group can be interpreted as the percentage change in the spot rate between the horizon and the trading time for an average trade done by a member of the considered group.

Note that this index is computed with pure spot prices, not with logarithms of the spot prices as for the regressions. By doing so, the share of the actual spot price return is given and by multiplying with a factor of 100 the percentage changes are obtained which makes the results interpretable. The expressions are approximately the same since according to the Maclaurin series of $\log(1 + x)$ it holds that

$$\left(\frac{s_{j,h}}{s_{j,0}} - 1 \right) \approx \log s_{j,h} - \log s_{j,0}. \quad (6.5)$$

6.1.2 Confidence Interval for the Index

To determine if the spot impact of a customer group is significant, confidence bounds are created at a 99% confidence level. If an index is inside the confidence bounds there is no significant spot impact at that time horizon for the current group. Two different methods are used to calculate the confidence bounds.

The first method takes the empirical quantiles of the indices where the mean is subtracted for all trades at a specific time stamp. The quantile is then multiplied with 100 and divided by the square root of the number of trades, n . Let the quantiles at levels 0.005 and 0.995 respectively be denoted $z_{0.005}^{emp}$ and $z_{0.995}^{emp}$. Then the confidence interval at confidence level 0.99 is given by

$$\text{Int}_{0.99}^{emp} = \frac{100}{\sqrt{n}} [z_{0.005}^{emp}, z_{0.995}^{emp}]. \quad (6.6)$$

In this way, the confidence interval is not necessarily symmetric around zero, taking into account that the data can contain skewness. The skewness could either come from the spot rate or from the customer.

The second way to obtain a confidence interval is by using the quantiles of the standard normal distribution with zero mean and unit standard deviation instead. The standard deviation of the indices for each time stamp is also taken into account. Denote the

quantiles $z_{0.005}^{norm}$ and $z_{0.995}^{norm}$, then it holds that $z_{0.005}^{norm} = -z_{0.995}^{norm}$ since the normal distribution is symmetric around zero. Let σ_{I_h} denote the standard deviation of the indices at horizon h . The confidence interval at confidence level 0.99 will be symmetric around zero and is given by

$$\text{Int}_{0.99}^{norm} = \frac{100 \cdot \sigma_{I_h}}{\sqrt{n}} [z_{0.005}^{norm}, z_{0.995}^{norm}]. \quad (6.7)$$

The purpose of calculating two different confidence intervals is to examine if there is a very skewed behaviour in the indices. The skewness can either come from that currency prices are skewed or that there are large outliers among the customers.

6.2 Clustering

Applying clustering techniques to trading data in the way it is done in this thesis has not been done previously, neither by the FX Quant team at Nordea nor at any other publicly known place. The purpose of this thesis is to investigate if some useful results can be obtained by using these types of techniques and algorithms. The algorithms described in Section 4.2 are some of the most common ways to perform clustering, and they are implemented in Matlab which is the program used in this project. Further this section gives an explanation how the final features of the data were chosen (Section 6.2.1) and thereafter presents how the resulting clusters are evaluated and interpreted (Section 6.2.2).

6.2.1 Choose Features

The data set is prepared according to Section 5.1.5 and consists of N observations each representing one customer¹. As mentioned in referred section, a few customers are considered to be outliers because of their large trading volume and are removed before clustering. Besides, some attributes do not make sense if the customer is trading too few times and therefore customers trading less than ten trades in the given period are excluded in the clustering.

To achieve a proper clustering, the features of the data points should be as little dependent as possible to be well separated. The dependence is measured by calculating the

¹Regular customers, prime brokered clients and banks are included.

correlation coefficients between the six features in the data set. As could be expected, the buy/sell-ratio and the mean value of new direction for the currency are highly negatively correlated, with a correlation coefficient of -0.82 . The dependence affects the clustering in a negative way, which is why the variable for the new direction for currency is removed. The Trade platform has a significant non-zero correlation with all other features, so it is removed as well. The remaining four features are the total volume, the buy-sell ratio, round trades and high volatility. The values of the different features are large in variation, which is handled by normalising the data by subtracting the mean value and divide by the standard deviation, performed with Matlab's `zscore`. Correlation coefficients and corresponding P-values between the four final features are given in table 6.1. There are still significant correlations in the data, but not to the same extent as with all six features.

TABLE 6.1: Correlation coefficients for SEK data variables.

NOK	Buy/Sell	Total Volume	Round Trades	High Volatility
Buy/Sell	1 (1)	-0.180 (0)	-0.122 (0)	-0.117 (0)
Total Volume	-0.180 (0)	1 (1)	0.263 (0)	-0.019 (0.441)
Round Trades	-0.122 (0)	0.263 (0)	1 (1)	0.061 (0.016)
High Volatility	-0.117 (0)	-0.019 (0.441)	0.061 (0.016)	1 (1)
SEK	Buy/Sell	Total Volume	Round Trades	High Volatility
Buy/Sell	1 (1)	-0.284 (0)	-0.201 ($1.110e-16$)	0.003 (0.917)
Total Volume	-0.284 (0)	1 (1)	0.267 (0)	0.018 (0.456)
Round Trades	-0.201 ($1.110e-16$)	0.267 (0)	1 (1)	0.040 (0.106)
High Volatility	0.003 (0.917)	0.018 (0.456)	0.040 (0.106)	1 (1)

Notes: Correlation matrix of the data variables for clustering, represented as columns in the matrix X of Equation (4.3). P -values for the null hypothesis of no autocorrelation are reported in parenthesis.

6.2.2 Cluster Evaluation and Interpretation

As the final features for the customers are chosen the next step is to determine the optimal number of clusters. The Matlab function `evalclusters` is used to evaluate the best clustering solution for three different methods using the Calinski-Harabasz criterion defined in Equation (4.1). The function maximises VRC_k with respect to k , where k is set to range from 1 to 20.

After applying each clustering technique to the data it is measured how well the result fits the data by calculating silhouette values defined in Equation (4.2). The Matlab function `silhouette` computes and plots the silhouette values in a silhouette plot. Studying the silhouette plots gives an indication of how well each point is placed in the clusters. By comparing them between different methods it can be determined which method that performs the best.

The clustering of the data is performed with the different techniques described in Section 4.2 and the results of different algorithms are compared by studying the clusters that are created. Indices are calculated for all clusters to display the spot price impact. In this way connections between the features and the price impact is exposed.

The Matlab function for K -means is the most developed one. To make sure that it is not converging to a local minimum the algorithm is run several times, which can be done by an option in Matlab. The the total sum of distances is given after each replication to make sure that the best result is repeated more than once and hence is more likely to be the global minimum.

In studying the clustering results, a closer look at in what way the customers trading through prime brokerage agreement are clustered will be taken.

Matlab Clustering Functions

The Machine Learning techniques used in this thesis have been applied to the data using Matlab's Machine Learning Toolbox, Fuzzy Logic Toolbox and Neural Network Toolbox. There are other options for tools to use, for instance R and Octave. Matlab was chosen since it has many powerful and built-in tools that are relatively easy to handle and it

is the most used computing program at both Nordea and LTH. The needed clustering algorithms are

kmeans

fcm

linkage

selforgmap.

Chapter 7

Results

In this chapter the results of the order flow and cluster analysis are presented with tables and graphs. Some extra results can also be found in the Appendix.

7.1 Order Flow

The aggregated accumulated net customer order flows in million euros are plotted in Figures 7.1 and 7.2 for NOK and SEK respectively, with the normalised volume on the left y -axis. The different lines represent the total flow and the flow split into the most important currency pairs, and the corresponding spot exchange rates with euro are given as well, with the rate on the right y -axis. If the order flow and the changes in spot rate are correlated, they would be expected to follow the same pattern. The SEKEUR rate does not seem to align with the total SEK order flow, but the flow for the euro seems to follow the same trend more. The customers of Nordea tend to sell Swedish Kronor against the US Dollar during the period, which is seen from the negative slope of the yellow curve. The net buying pressure seen in the total flow comes mainly from heavily buying SEK against other currency pairs. The same net buying behaviour is there for the Swedish Krona against the Norwegian Krona.

Studying the NOK flow instead, it looks like the NOKEUR rate is negatively correlated with the total flow. The movement more corresponds to the flow from the euro in this period, which is not strange since it is the rate that is compared and the total flow is widely affected by the USD trades. Nordea's customers tend to buy Norwegian Kronor

against the US dollar, but sell against the other currencies. The explanation to this behaviour is that Norway has a large income in US Dollar which is related to their oil production. Some of the USD flow goes back to the Norwegian economy and it is seen in the flow of Nordea as the customers are selling USD against NOK. Sweden has a broader corporate base and the net buying of SEK that is seen is due to a surplus in the trade balance of the customers of Nordea. Naturally the flow is in currencies more related to the SEK, like the EUR and NOK, rather than the USD. Sweden does not have an oil production like the one in Norway and is selling SEK against USD.

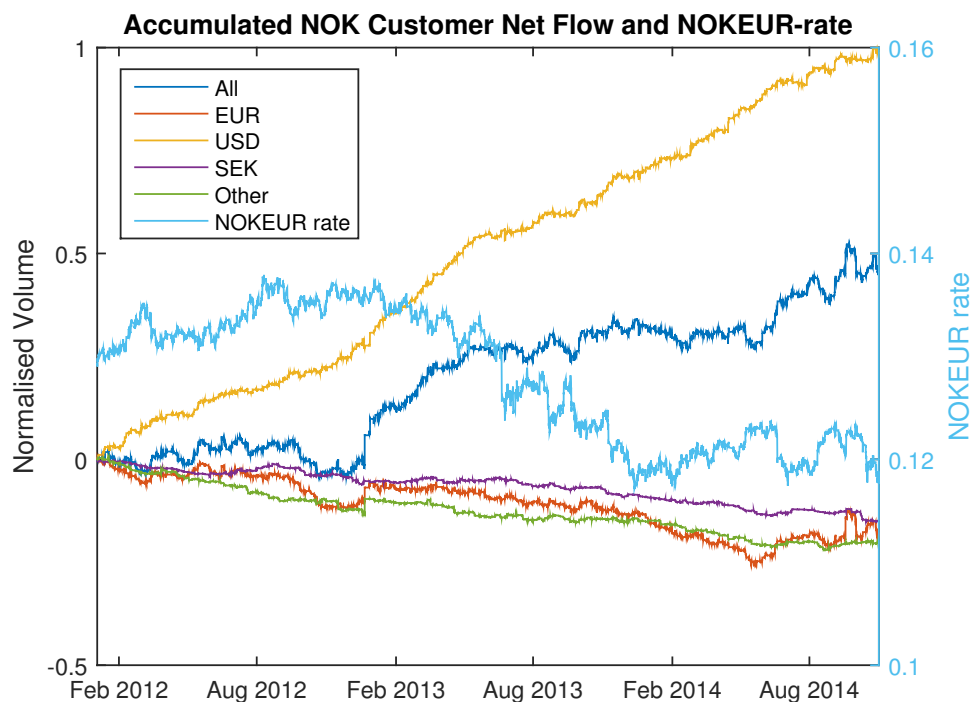


FIGURE 7.1: Accumulated NOK customer net order flow and NOKEUR spot rate. The NOKEUR rate moves in the opposite direction as the total flow, but is more aligned with the flow in euro only. Nordea's customers tend to buy NOK against USD but sell NOK against all other traded currencies.

The accumulated net order flow for the four customer groups asset managers, corporations, hedge funds and private clients are given in Figures 7.3 and 7.4. The blue lines showing all currency pairs indicate that hedge funds and private clients are buying Norwegian Kronor in about half the period and after that are more selling. Corporations are continuously more buyers of NOK, especially against the US Dollar, while asset managers' behaviour is closer to a random walk. It is again related to the Norwegian oil production, where it is assumed that it is mainly corporates who take part of the large

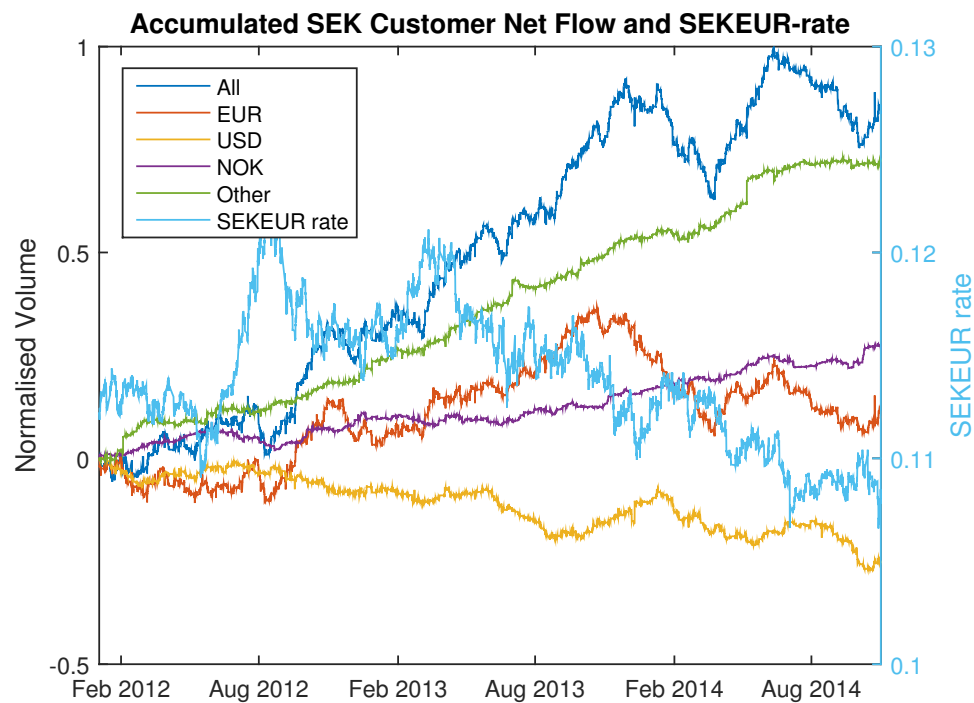


FIGURE 7.2: Accumulated SEK customer net order flow and SEKEUR spot rate. The SEKEUR rate is more aligned with the flow in euro than the total flow. The customers of Nordea are net buyers of SEK, the trend is most clear for the other currency pairs. The only currency that is bought against SEK is the USD.

USD surplus. It is clear that the behaviour is different across the different currency pairs.

Figure 7.4 shows that the non-financial customers have been buying Swedish Kronor. Corporates buy SEK against all the other currencies except from USD and private clients have most of their flow in EURSEK. In the middle of 2012 the hedge funds were heavily buying euro and US Dollars and selling Norwegian Kronor against SEK. A large sell of SEK towards EUR occurred in the middle of 2014 for the hedge funds and in the beginning of the same year for the asset managers.

7.1.1 Customer Group Indices

Indices are calculated for each trade according to (6.1). The indices are then volume-weighted for each group according to Equation (6.3) with weights as in (6.4), giving the wanted customer group indices. The time intervals used are 2 days of full half-hours before and after the trade. In Figures 7.5 and 7.6 the indices are plotted for

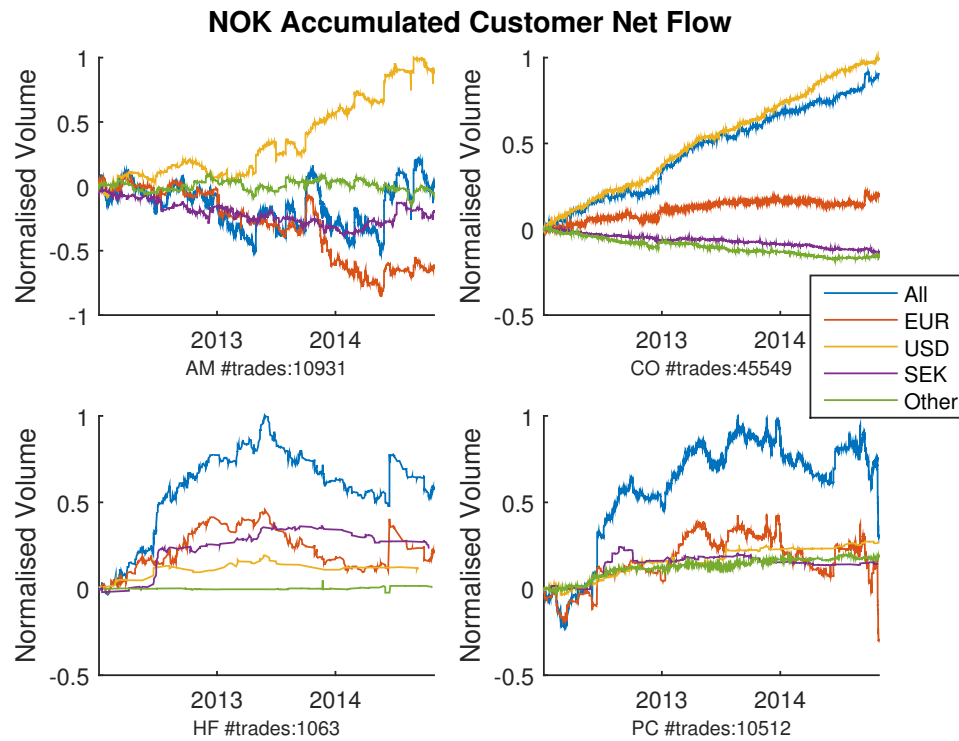


FIGURE 7.3: Accumulated NOK customer net order flow disaggregated into different customer groups. The corporates are heavily buying NOK, especially against USD. The hedge funds and private clients bought a lot in the middle of 2012 and after that the trend is evened out. The asset managers are the only ones buying EUR and they tend to sell NOK against SEK as well, but buy against USD. The hedge funds are the only net buyers of NOK against SEK. Both AM and HF sell a lot of EUR against NOK at once in the middle of 2014.

asset managers, corporations, hedge funds and private clients. The x -axis represents the fixings, with 0 at the trade time, negative values before the trade and positive values after the trade. Time points -1 and 1 represent the full half hour closest prior to or post trade time. The blue lines represent the 99% empirical confidence intervals for the indices calculated as in (6.6) and the orange dots are the volume-weighted indices at each time step. If the index lies outside the confidence bounds it is considered to be significant. The y -axes show the percentage change in the spot currency price. Note that the scales of the y -axes vary between the plots.

The figures show the spot price impact from the four customer groups. Corporations and private clients are contrarians for both NOK and SEK, meaning that they buy when the price goes down and sell when the price rises. This result is intuitive since it is a natural behaviour to buy cheap and sell expensive. Their price impacts after the trade differ in the way that the corporations have a negative spot impact while the price is

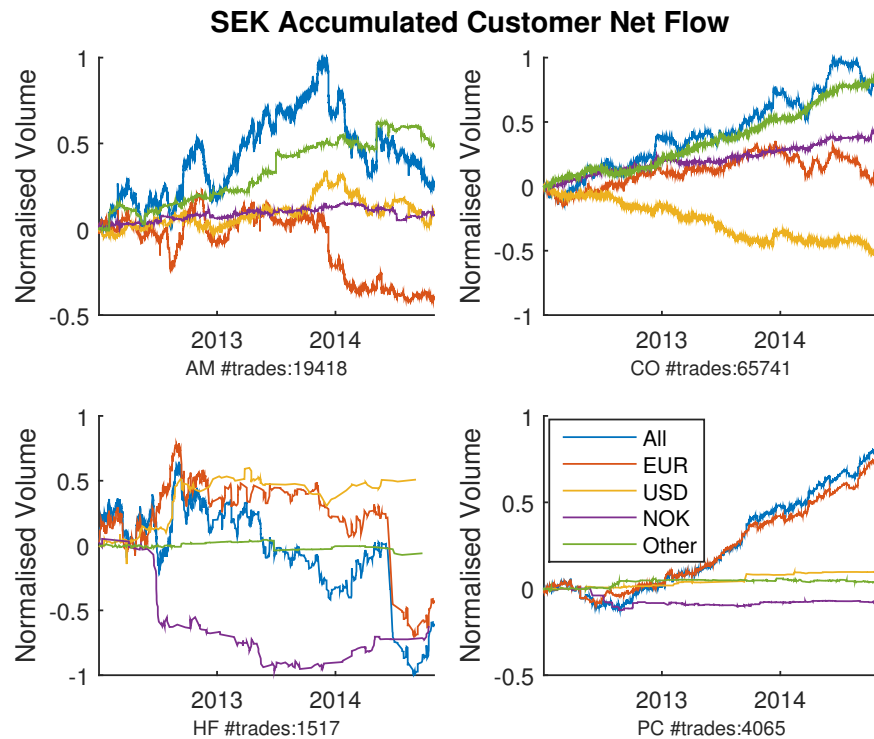


FIGURE 7.4: Accumulated SEK customer net order flow disaggregated into different customer groups. The asset managers are selling SEK against EUR but are buying against the other currencies. The corporates are selling SEK against USD but are buying against the other currencies. The hedge funds are selling SEK against NOK but buy against the other currencies except from in the end of 2014 where there is a large sell of SEK against the EUR. Private clients have been constantly increasing their buys of SEK against EUR from 2013 while the flow in other currencies is rather non-directional.

not affected when the private clients have traded. The corporates and private clients, which are considered to be non-financial customers, are expected to have other motives than earning money when they trade currencies and the indices show that at least the corporations are losing some money after the trade. The asset managers seem to be trend-following for the SEK, but show no clear pattern in their behaviour for NOK. The spot impact they show in Swedish Kronor is very small but still significant. The hedge funds are the winners in this game, since the spot price seems to move with them after they have traded. The pattern is more significant for the SEK than the NOK. If there were more hedge funds trading with Nordea, perhaps this behaviour would be more significant.

The results are aligned with the previous literature about customer order flow and confirm the negative correlation between corporates and private clients with spot changes

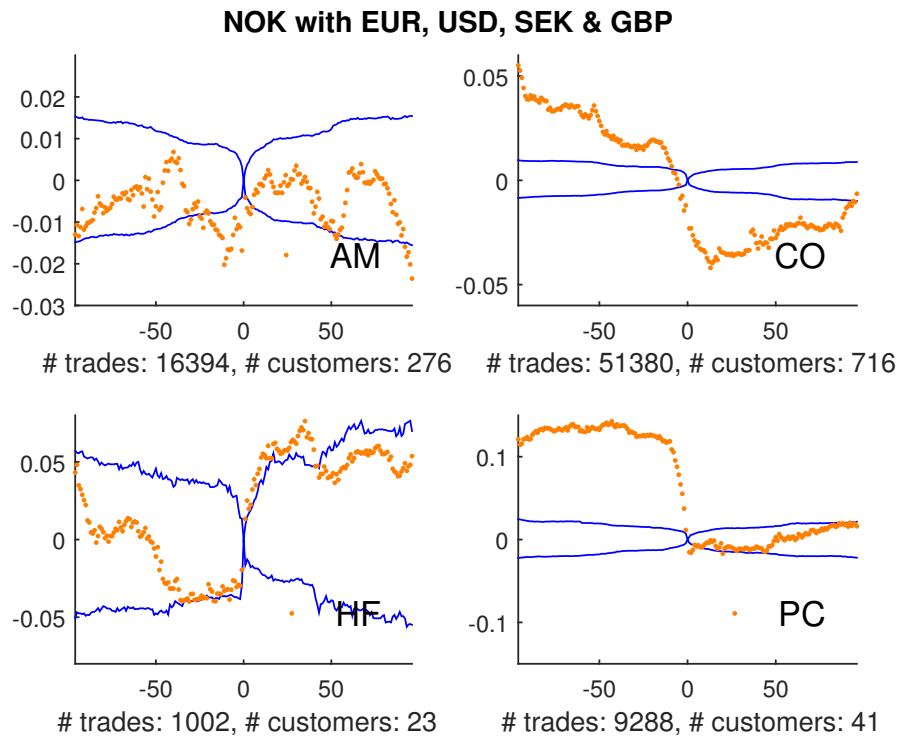


FIGURE 7.5: The orange dots are the volume-weighted indices for all NOK trades done by the customer groups asset managers (AM), corporations (CO), hedge funds (HF) and private clients (PC) at 30 minutes intervals 2 days prior to and post trade time. The blue solid lines are the 99% empirical confidence intervals. AMs show a small trend-following behaviour during the last hours before trading, while COs and PCs are contrarians before trade time. The indices of the HFs are mostly insignificant but they tend to be trend-following with the spot moving with them after they trade. COs show a negative spot impact after trade time, while AM and PC do not show any significant impact.

and the positive correlation with the financial institutions such as the asset managers and the hedge funds. It also adds information that is lost when the order flow is aggregated over a daily horizon as in many papers. With aggregated flow it can not be distinguished whether the negative correlation between the flow from non-financial customers and exchange rates comes from a trading behaviour or if they are simply always losing money on the trades. In other words it can not be distinguished whether the spot moves against them before or after the trade. In a similar way it is impossible to tell if the positive correlation from financial customers' flow is due to that they are informed about the prices or if it is their trading style. The indices are calculated with respect to the actual trade time and reveal that it is the contrarian way of trading that accounts for the negative relation and that hedge funds are trend-following and seem to be informed about the future movements.

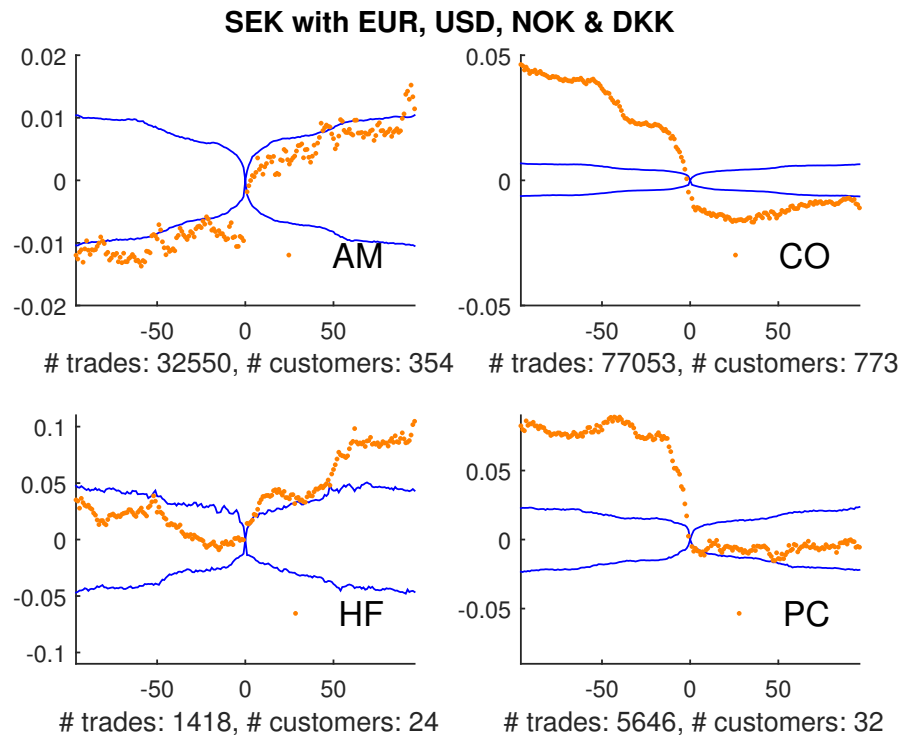


FIGURE 7.6: The orange dots are the volume-weighted indices for all SEK trades done by the customer groups asset managers (AM), corporations (CO), hedge funds (HF) and private clients (PC) at 30 minutes intervals 2 days prior to and post trade time. The blue solid lines are the 99% empirical confidence intervals. The AMs seem to be trend-following prior to trade but show no impact after the trade. The HFs show no clear behaviour before the trade but have a significant positive impact on the spot price. The COs are contrarians and tend to lose money on their trades. The PCs seem to trade after the spot has changed more than 5% the last hours, but show no price impact after the trade.

7.1.2 Bank Indices

The study of the interbank market is presented in Figures 7.7 and 7.8 where the indices are plotted for the partitioning of the bank flow into "Our Hedges", "Our Hedges Algo", "Prime Brokerage" and "Their hedges". For the first two groups it is Nordea who is initiating the trades and the customers are the counterparties they are trading with (normally larger banks). In these cases the indices are calculated from Nordea's perspective. The two latter groups are seen from their perspective since they are initiating the trades.

The two upper plots of both figures show a significant contrarian behaviour when Nordea is trading both NOK and SEK. Nothing significant is happening after the trade time, except for a small positive price change in when SEK is traded as OH. Other banks seem

to trade purely random in NOK but trend-following in SEK and there is no significant spot impact.

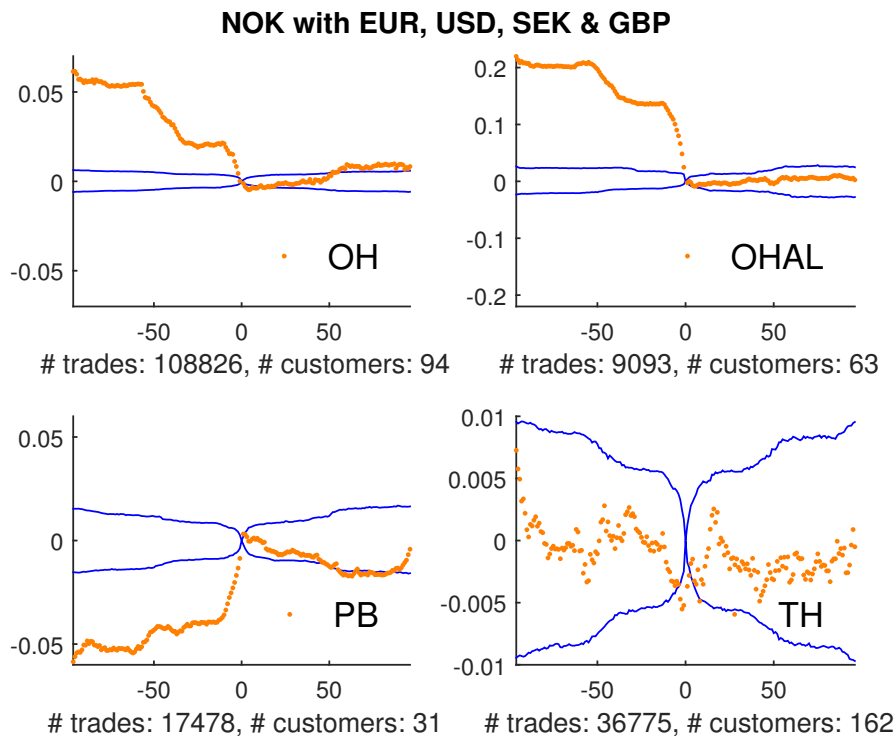


FIGURE 7.7: Volume-weighted indices at 30 minutes intervals 2 days prior to and post trade time for all NOK trades done by Nordea (OH and OHAL), prime brokered customers (PB) and other banks' hedges (TH). OH and OHAL show similar patterns, namely trading after a significant price move and with not much impact after the trade. The PB clients show the opposite behaviour. The pattern for TH is purely random.

The indices of the clients trading through a prime brokerage relationship show the same significant trend-following pattern for both NOK and SEK. There is a small tendency that the spot moves against them after they trade, but it is statistically insignificant.

7.2 Clustering

The data points that are clustered correspond to the customers (including interbank customers), each associated with four different attributes. The final Swedish and Norwegian data set contain 1666 and 1565 unique customers respectively. How these are divided between the different customer groups is shown in Table 7.1. A large number of customers are undefined clients, meaning that their traded volume is not large enough

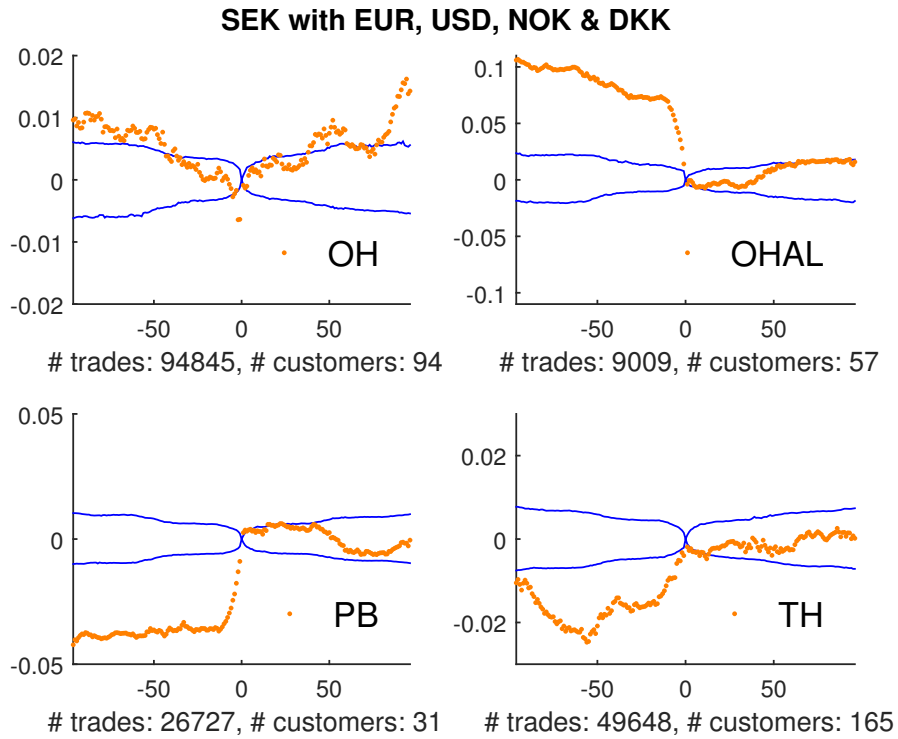


FIGURE 7.8: Volume-weighted indices at 30 minutes intervals 2 days prior to and post trade time for all SEK trades done by Nordea (OH and OHAL), prime brokered customers (PB) and other banks' hedges (TH). The patterns differ between OH and OHAL in that OH has a smaller contrarian (almost random) behaviour and a positive price impact while OHAL is clearly contrarian and has no really significant spot impact. PB clients are trend-following as well as TH, but have no impact on the price after the trades.

for being classified earlier. All observations are plotted in Figures 7.9 and 7.10, where the features are pairwise combined in scatter plot matrices. Histograms of the four features are seen on the diagonal, and the values seem to be distributed in similar ways for both NOK and SEK.

The Calinski-Harabasz criterion suggests nine clusters for the K -means algorithm for the NOK data set and six clusters for the SEK data set. The result of the function `evalclusters` for the NOK data set can be seen in Figure 7.11¹. The CH values are monotonically increasing up to the maximum and decreasing after that, at least for the two first methods. The values for 8 and 10 clusters are not differing much from the maximum, so these would probably be acceptable numbers of clusters as well. The picture is similar for SEK.

¹The figure for SEK will not be included since the picture is very similar.

TABLE 7.1: Clustered Customers by group

	AM	BA	CB	CO	HF	OT	PB	PC	XX	Tot
NOK										
# Customers	144	104	3	539	14	7	29	35	690	1565
% of Volume	16.0	22.8	3.3	37.5	3.1	0.7	6.9	6.6	3.1	
SEK										
# Customers	199	103	5	551	16	6	28	22	736	1666
% of Volume	18.5	16.3	0.5	44.2	2.9	1.0	8.4	5.8	2.1	

Notes: This table shows how the clustered customer observations are divided between the pre-labelled customer groups for NOK and SEK. The top row gives the number of customers in each group and the second row gives their percentage share of the total traded volume.

To examine how the spot price impact is affected by the customers according to their features, indices are calculated for the clusters formed by the algorithms. After removing trades not done in the main currency pairs, the numbers of remaining customers are 1403 for NOK and 1646 for SEK. When studying the index plots, one should keep in mind that the scales of the y -axes differ between the plots.

In the following Sections the results of the clustering of the NOK data and the SEK data will be presented. The focus will lie on the algorithms K -means, Fuzzy C -means and Self-Organizing maps and the results are produced using the techniques. Hierarchical clustering will not be included since the results are unsatisfying; using the agglomerative method results in 99% of the points ending up in the same cluster.

7.2.1 Clustering of the NOK Flow

Evaluating the number of clusters in the NOK customer data set results in nine being the optimal value. The silhouette values for the three algorithms can be seen in Figure 7.12. A comparison between the three plots gives that the performance of K -means and SOM seems to be approximately equal, while the Fuzzy C -means is giving the worst clustering.

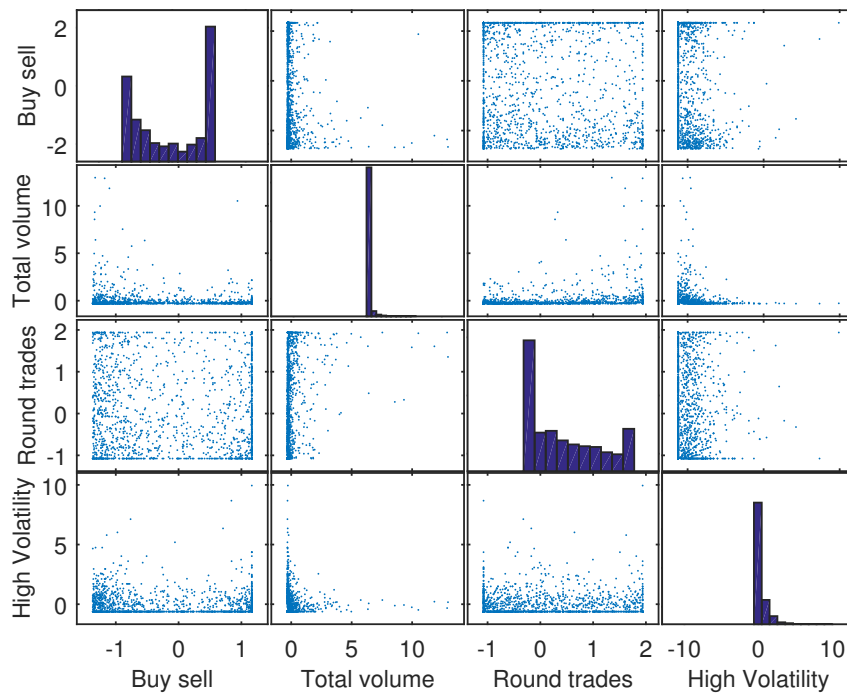


FIGURE 7.9: Scatter plot matrix of the NOK customer data that is prepared for clustering, each point representing one customer. The feature values are normalised by subtracting the mean and dividing by the standard deviation. Histograms are seen on the main diagonal. The buy/sell-ratio and the round number trading are the most spread features but cluster around the edge values. The total volume and high volatility for the customers are focused in the lowest band.

K-means

Starting off with the K -means algorithm with nine clusters the silhouette values in Figure 7.12a show that most customers are addressed to the most suitable cluster. The scatter plot matrix in Figure 7.13 visualise how the data points are clustered. Points belonging to the same cluster are plotted with the same colour and symbol. Indices are calculated for each cluster and the result is shown in Figure 7.14, where the numbers of trades and customers assigned to each cluster are given.

A significant contrarian behaviour with a negative spot impact after the trade is seen for clusters 1, 6, 8 and 9. The common features of the customers in these clusters are that they contain relatively many customers who are trading low volumes and mostly in periods with low volatility. They differ in that the clients in cluster 1 have a larger buy/sell-ratio in contrast to the others, and that clients in cluster 9 are not trading much in round numbers while the others are. All four of these clusters are dominated

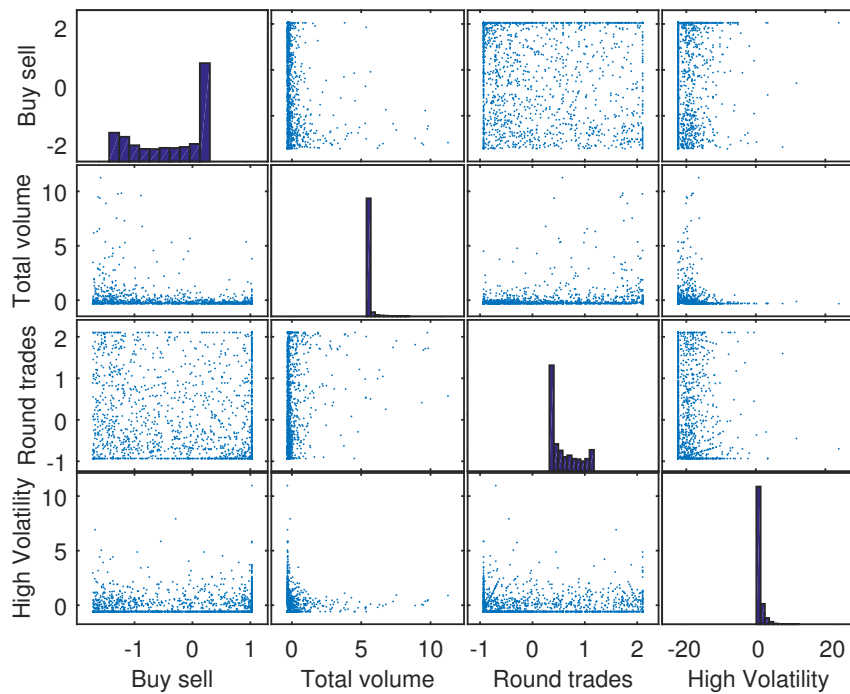


FIGURE 7.10: Scatter plot matrix of the SEK customer data that is prepared for clustering, each point representing one customer. The feature values are normalised by subtracting the mean and dividing by the standard deviation. Histograms are seen on the main diagonal. The buy/sell-ratio and the round number trading are the most spread features but are clustered around the higher and lower edge values respectively. The total volume and high volatility for the customers are focused in the lowest band.

by customers labelled CO and XX. Cluster 9 also contains some with the labels AM, BA and PB.

Clusters 2, 3, 4 and 5 show contrarian behaviour before the trade but no significant spot impact afterwards. These clusters include customers representing almost the full span in all features. There are customers trading small up to medium high volumes and with mostly low buy-sell ratios although some outliers exist. Cluster 5 and 2 contain customers who are trading more often in volatile periods and these customers all have low volumes. In cluster 3 and 4 they do not trade so much in volatile periods. There is a larger variation between customer types in these clusters as well.

The only cluster that shows a trend-following behaviour and positive spot impact is cluster 7 which are the customer with the largest trading volume, consisting of only eight customers. They all tend to trade a lot in round numbers and all but one have a low buy-sell ratio. They do not trade specifically at highly volatile periods.

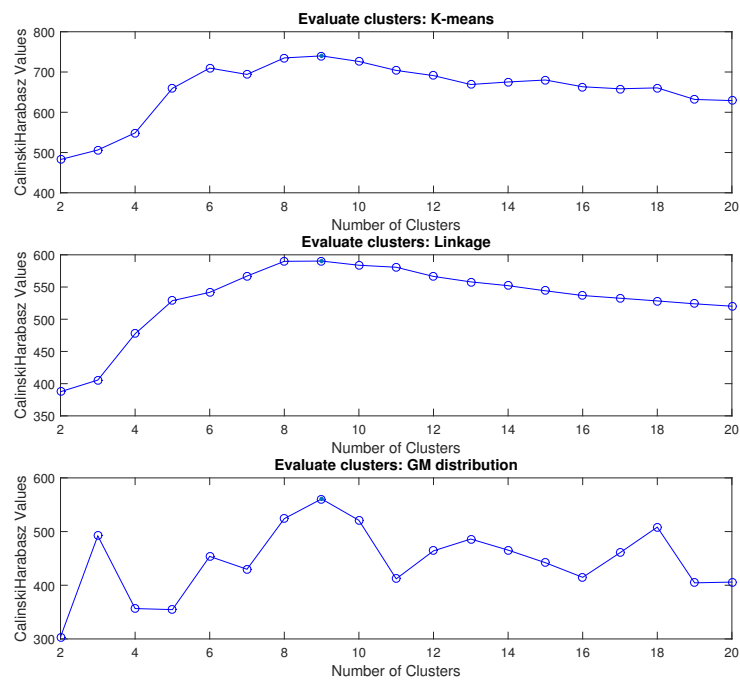


FIGURE 7.11: Shows the result of the Calinski-Harabasz criterion for the NOK data set for the three functions kmeans, linkage and fitgmdist, run with $k = 1, \dots, 20$ clusters. Nine is marked as the optimal number of clusters for all three methods.

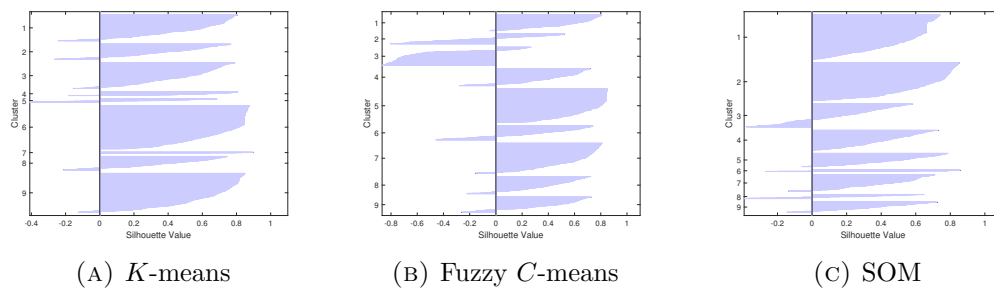


FIGURE 7.12: Silhouette plots for the three presented algorithms for the NOK data set. The Fuzzy C -means shows the least well-clustered objects. The performance of SOM seems to be slightly better than for the K -means in this case.

Half of the prime brokered clients are placed in cluster 3, the rest are primarily split between clusters 4 and 9. Hence all PB clients have a low buy-sell ratio meaning that their flow in NOK is non-directional. Figure 7.7 from previous section shows that the PB clients are trend-following on average, while the clusters they are now placed in are showing the opposite.

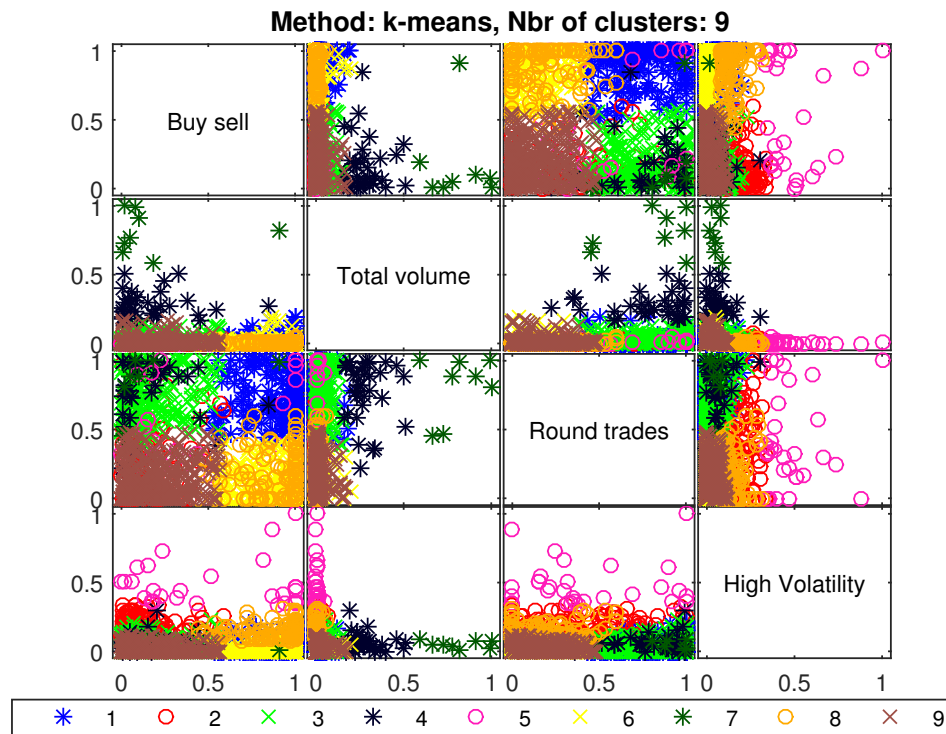


FIGURE 7.13: This scatter plot matrix shows the result of the 9 clusters with K -means for NOX. The values of the features are normalised so that they lie between zero and one.

Fuzzy C-means

The Fuzzy C -means algorithm misclassifies more of the customers if looking at the silhouette plot in Figure 7.12b. With the resulting clustering, visualised in Figure 7.15, there is no cluster which has a significant positive spot impact according to the index plots in Figure 7.16. With K -means the eight customers with largest traded volume were grouped together and now they are extended forming cluster 3 with 159 customers covering the whole span of traded volumes. This cluster is not a well-classified cluster according to the silhouette values and the result is that the significant spot impact is gone.

It is still clear that having at the same time a high buy-sell ratio, a low portion of round number trades and a relatively low volume is related to a contrarian trading behaviour and negative spot impact as for clusters 5 and 6. By only changing the proportion of round trades as for cluster 8, this behaviour is not as significant. The customers trading more at times with high volatility belong to cluster 2 and tend to trade contrarian, but

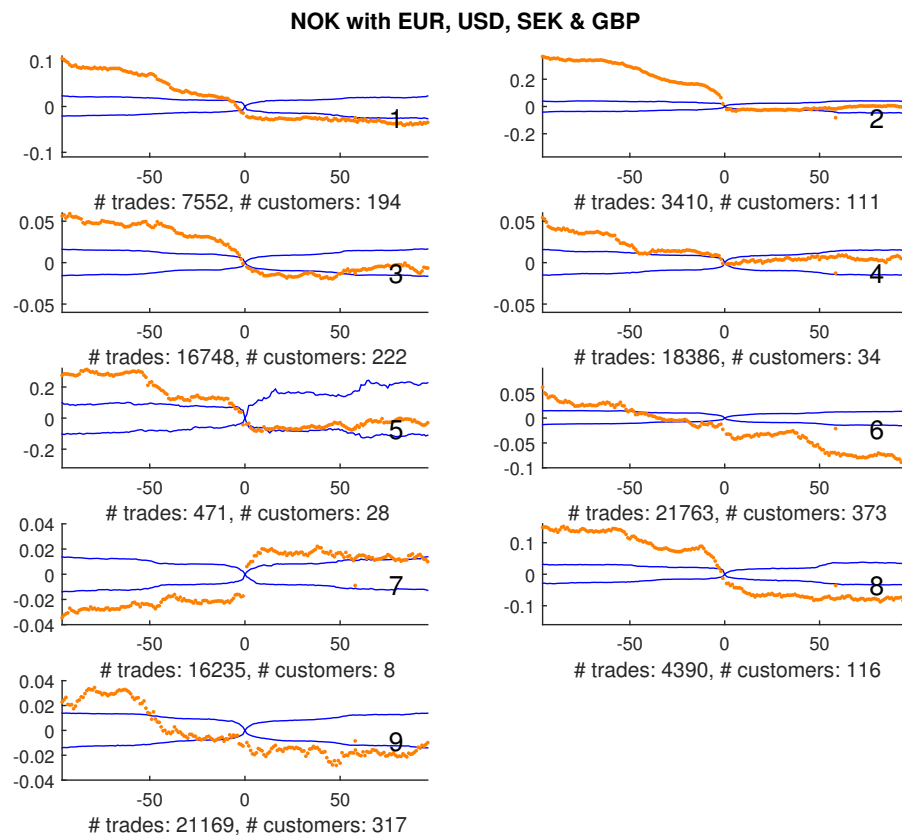


FIGURE 7.14: Volume-weighted indices for the nine clusters of the NOK flow created by K -means, together with the 99% empirical confidence bounds. All clusters more or less show a contrarian behaviour prior to trade. Clusters 6, 8 and 9 also show a significant negative spot impact after the trade. The only cluster showing a trend-following pattern is cluster 7 which consists of the eight customers trading the largest volumes in NOK.

do not affect the prices afterwards. Clusters 1, 7 and 8 do not show any significant trading behaviour at all.

About one third of the PB clients are placed in cluster 1 and another third in cluster 3. A few are also placed in cluster 7. None of these clusters show a very significant trading behaviour from the index plots.

Self-Organizing Map

SOM seems to have the least misclassified customers for NOK according to the silhouette plot in Figure 7.12c. The result of the clustering is seen in Figure 7.17 and is rather

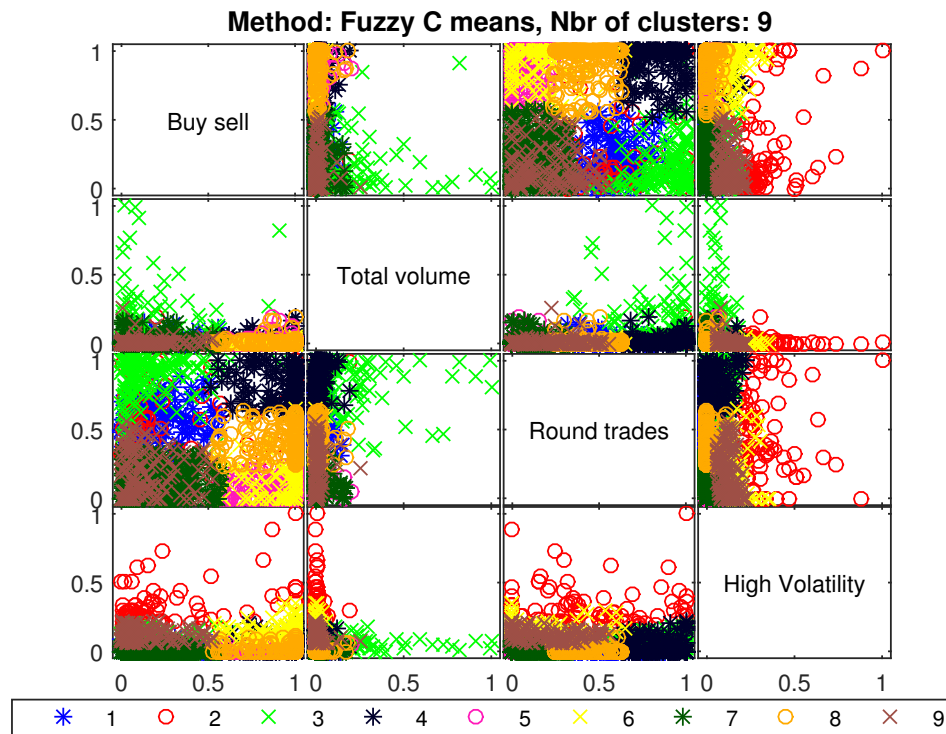


FIGURE 7.15: This scatter plot matrix shows the result of the 9 clusters with Fuzzy *C*-means for NOK. The values of the features are normalised so that they lie between zero and one.

similar to the result with *K*-means, as are the conclusions that can be made from the indices.

Again the customers with the largest trading volume are grouped together and show a small trend-following behaviour that can be seen for cluster 6 in Figure 7.18.

The customers in cluster 8 are the ones who trade in volatile periods and they show a contrarian behaviour but no price impact after the trade. They all trade low volumes and might be speculative clients who choose to trade at certain economic releases. The same contrarian trading style with no significant spot impact is seen for clusters 3, 5 and 9, whose customers are non-directional and can trade smaller to medium volumes. They are also the ones who trade a little more in volatile periods than the average customer.

The most significant negative spot impact comes again from customers who have a high buy-sell ratio together with either a large portion of round trades (cluster 7) or a small portion of round trades (cluster 1). These clusters are dominated by corporates and

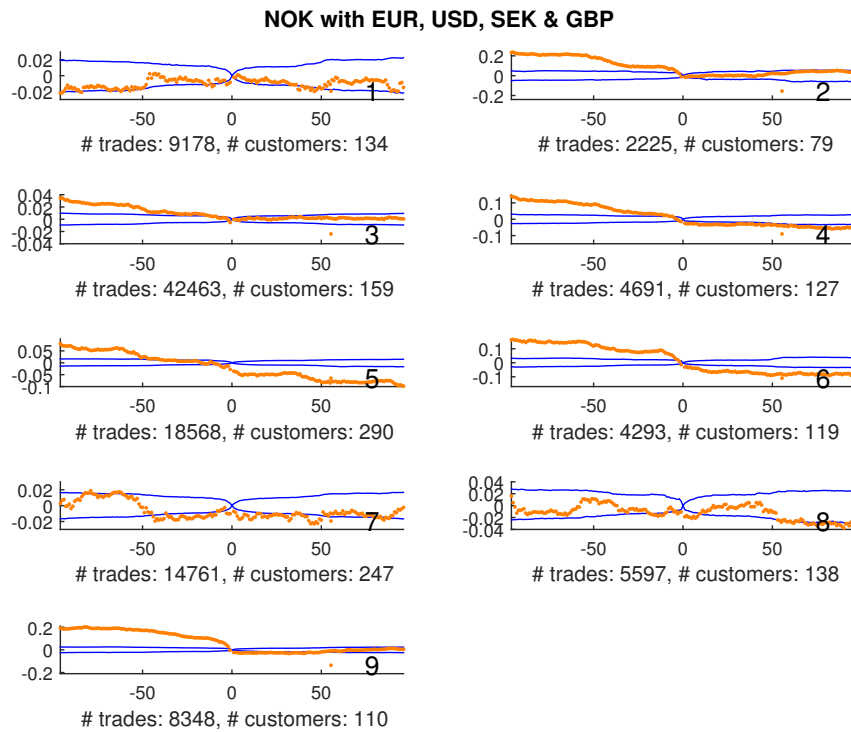


FIGURE 7.16: Volume-weighted indices for the nine clusters created by the Fuzzy C -means for the NOK flow, together with the 99% empirical confidence bounds. Most clusters show a significant contrarian trading style. The customers in clusters 1, 7 and 8 have a random behaviour.

unlabelled clients. Cluster 1 is the largest one which is why the confidence bounds are narrower and the impact larger.

Two thirds of the PB clients are assigned to cluster 3, which shows a small contradictory trading style. This cluster contains many other customer types as well.

Comparing the methods

The total volume that a customer trades seem to have an impact on the spot movement. This is seen with all methods and different number of clusters. The K -means and the SOM group the 8 and 14 customers with largest trading volume, respectively. The resulting cluster is the only cluster showing a significant trend-following trading style and a positive price impact. The customers who trade at periods with high volatility all trade lower volumes and with a contrarian behaviour. These are potentially clients

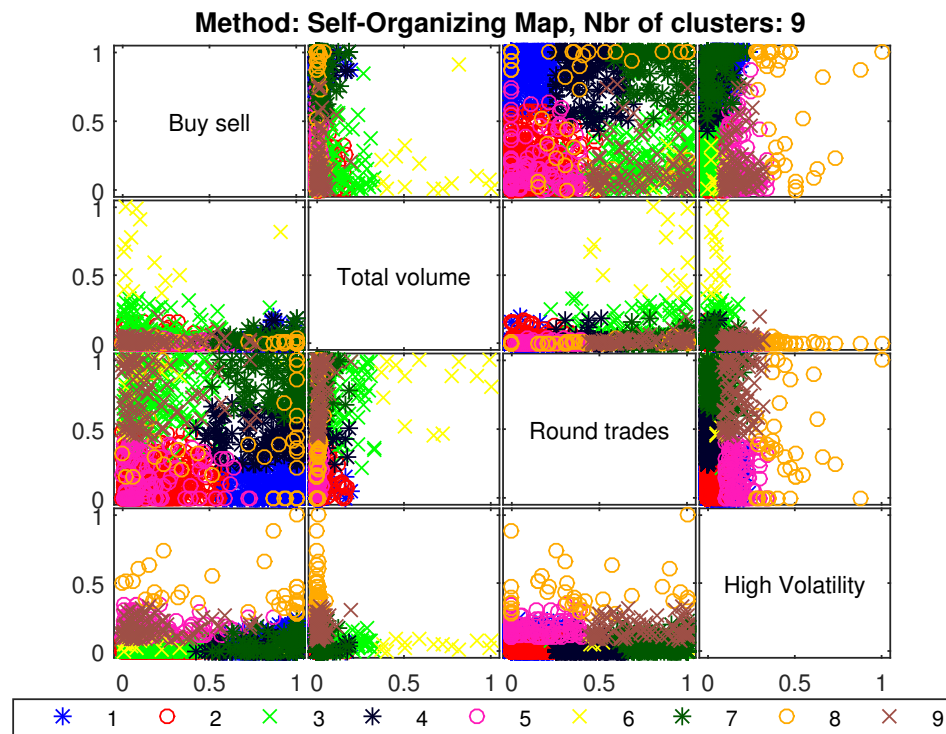


FIGURE 7.17: This scatter plot matrix shows the result of the clustering done by a self-organizing map of the customers trading NOK. The values of the features are normalised.

who choose to enter the market at economic releases and trade according to the intuitive behaviour, in other words buying cheap and selling expensive. They are not likely the most professional traders. The K -means and SOM methods are giving very similar results, while the Fuzzy C -means is the least trustworthy method for NOK.

7.2.2 Clustering of the SEK Flow

Figure 7.19 shows the silhouette plots for the three different methods carried out on the customers trading SEK. As for NOK, the Fuzzy C -means performs the worst again. The K -means and the SOM look even more similar for SEK than for NOK. The number of clusters is chosen to six, since it is the optimal value according to the Calinski-Harabasz criterion.

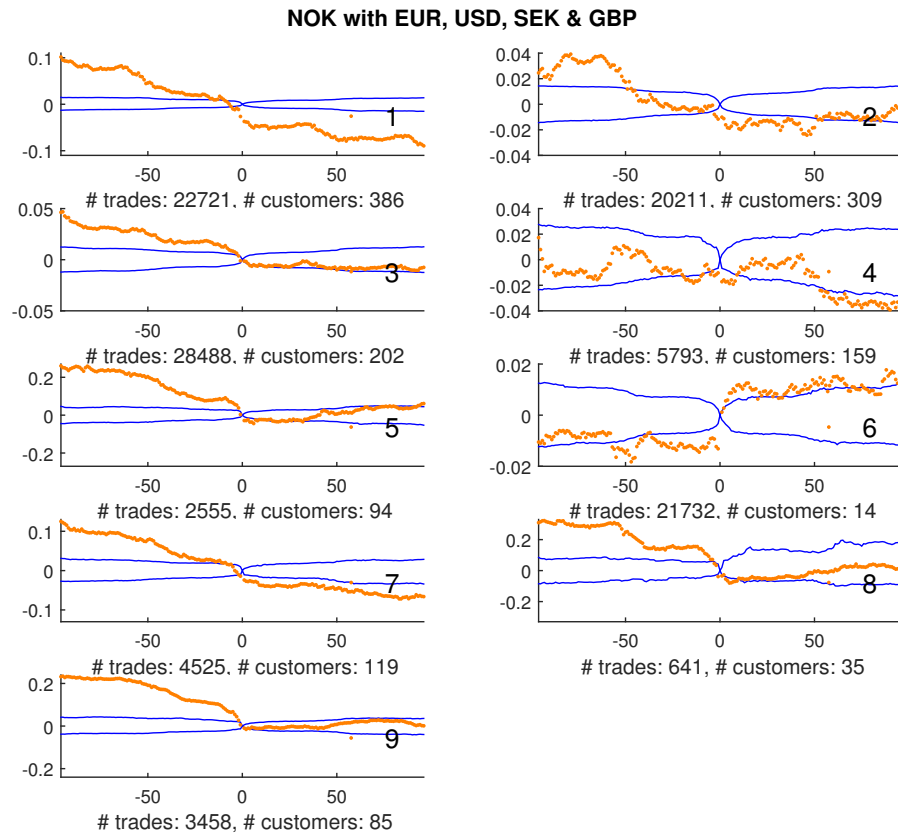


FIGURE 7.18: Volume-weighted indices for the nine clusters created by a self-organizing map for the NOK flow, together with the 99% empirical confidence bounds. A significant contrarian behaviour is seen for most clusters. The only trend-following pattern is seen for cluster 6, again consisting of the customers with the largest trading volume.

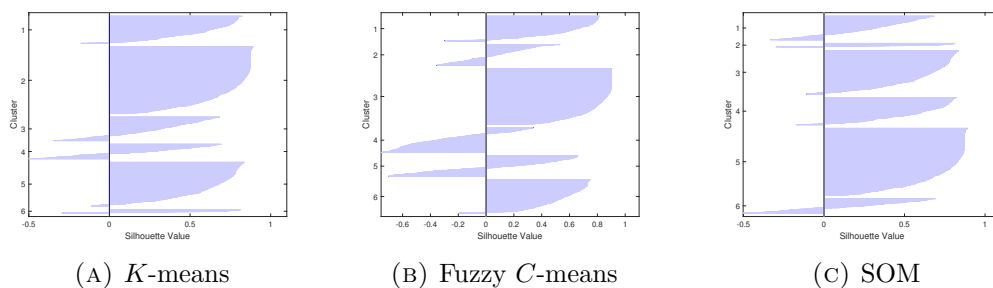


FIGURE 7.19: Silhouette plots for the three presented algorithms for the SEK data set. The Fuzzy *C*-means shows the least well-clustered objects. The performances of *K*-means and the SOM are nearly identical in this case.

K-means

The *K*-means algorithm is applied to the customers trading Swedish Kronor with six clusters as suggested by the Calinski-Harabasz criterion. The silhouette plot in Figure

7.19a shows that the clusters are overall well-defined, although there are a few negative values. Figure 7.20 shows how the observations are clustered.

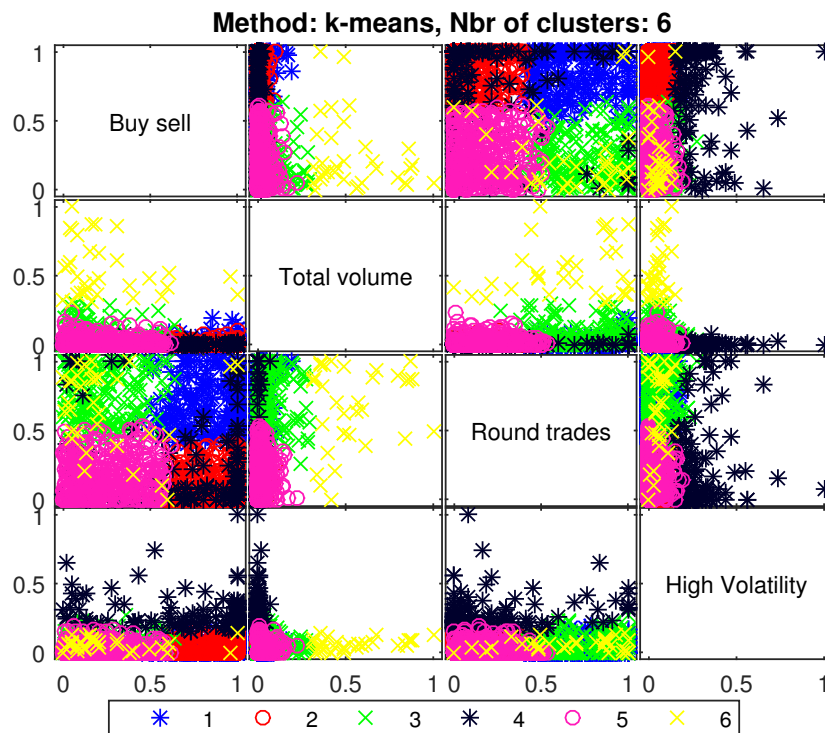


FIGURE 7.20: This scatter plot matrix shows the result of the K -means clustering of the customers for SEK, with normalised values of the features.

Five out of six clusters show a statistically significant contrarian trading behaviour according to the index plots in Figure 7.21. The final group which is cluster 5 shows no really significant spot behaviour either prior to or post trading.

Cluster 2 is the largest cluster and the index tells that the spot moves against these customers after they have traded. They are trading smaller amounts with a high buy-sell ratio, but not many round trades or trades in highly volatile periods. The customers who trade the largest volumes are placed in cluster 6 and actually seem to be contrarians for SEK, unlike this group in NOK. The price change before they trade is relatively small though, but still significant.

Since the number of customers of different types varies a lot, the flow from the smaller groups drowns a bit in the larger groups. There are many more corporate customers and a large number of unlabelled customers that affect the clustering. The banks and asset managers are also large groups. The smaller groups are hedge funds, central banks,

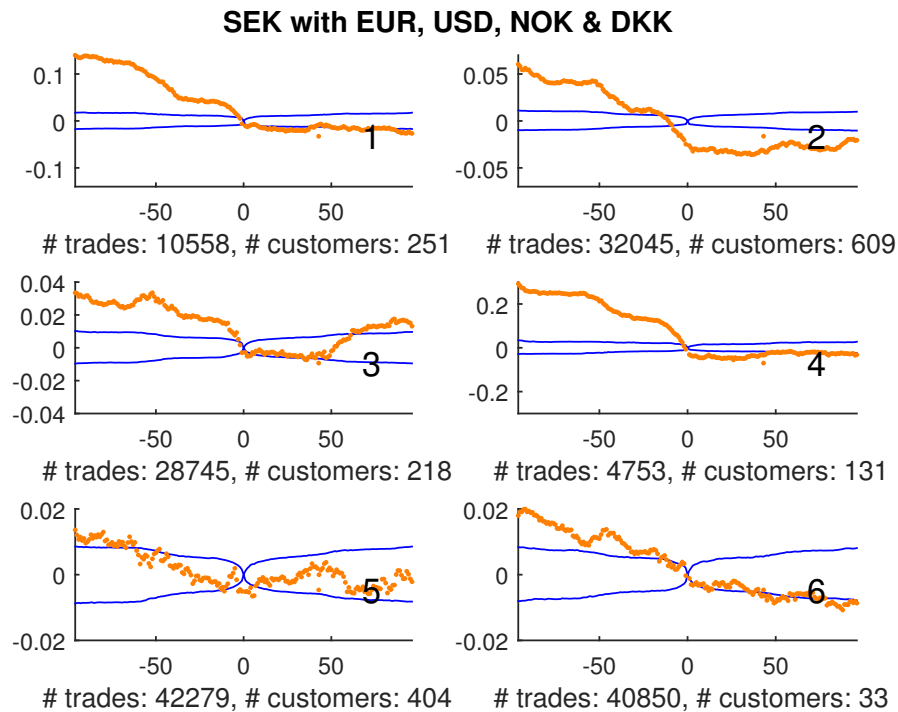


FIGURE 7.21: Volume-weighted indices for the six clusters created by K -means for the SEK flow are represented by the orange dots. The blue solid lines are the 99% empirical confidence intervals. Clusters 1 and 4 are showing the largest contrarian behaviour but not much spot impact after the trade. The customers in these cluster are trading directional. Clusters 2, 3 and 6 are also contrarians but with smaller relative changes. The clients in cluster 5 are trading more randomly.

private clients and prime brokered clients. Not too surprisingly, clusters 1, 2 and 4 are more dominated by corporations and unlabelled clients, while cluster 3 consists of a larger proportion of banks, asset managers, hedge funds and prime brokerage. Cluster 5 is the most mixed cluster, containing a lot of asset managers, banks, corporates and unlabelled clients and cluster 6 is a small cluster of the ones trading large volumes in SEK, which are mainly corporations. These observations are interesting and tell us that there is a variety in the behaviour within the previously labelled groups. Although the previous grouping seem to tell more about the price impact of different customer groups than the new clusters in this case. The results presented in Section 7.1.1 are more significant than the new clusters.

Around half of the clients trading through prime brokerage agreement are placed in cluster 3, with the rest in clusters 5 and 6. These are the three clusters showing the least contrarian behavior. As could be seen in Figure 7.8 the PB clients are on average trend-following. The PB clients are non-directional for SEK as well as for NOK.

Fuzzy C-means

As for the NOK data set, the FCM algorithm performs the worst according to the silhouette plot in figure 7.19b. The clusters in Figure 7.22 differ from the ones created by K -means for instance in that the customers with the largest volume are clustered together with customers trading low volumes. This cluster is not well-defined according to the silhouette plot.

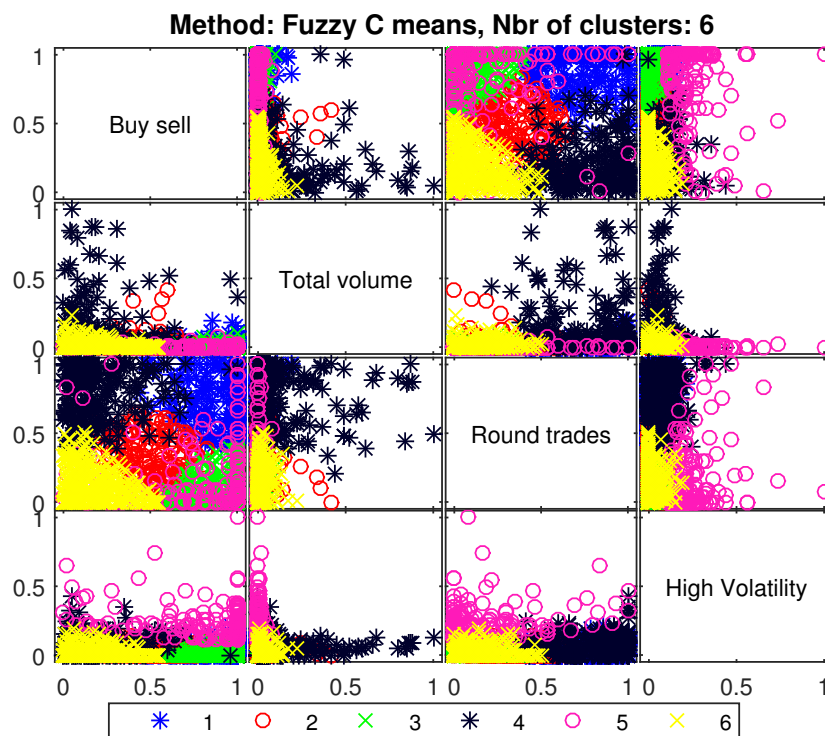


FIGURE 7.22: This scatter plot matrix shows the result of the Fuzzy C -means clustering of the customers for SEK, with normalised values of the features.

The clusters with the largest contrarian behaviour are clusters 1 and 5, corresponding to clusters 1 and 4 in the K -means. Cluster 3 shows this behaviour as well. The distinctive feature for the customers in these clusters is that they have a high buy/sell-ratio, in other words that they are directional in their trading style. Most likely directional customers are corporations that for some reason need to convert a large flow into their native currency.

A very small trend-following price change the last day before the trade is seen for cluster 6, whose customers are non-directional do not trade either large volumes, round numbers

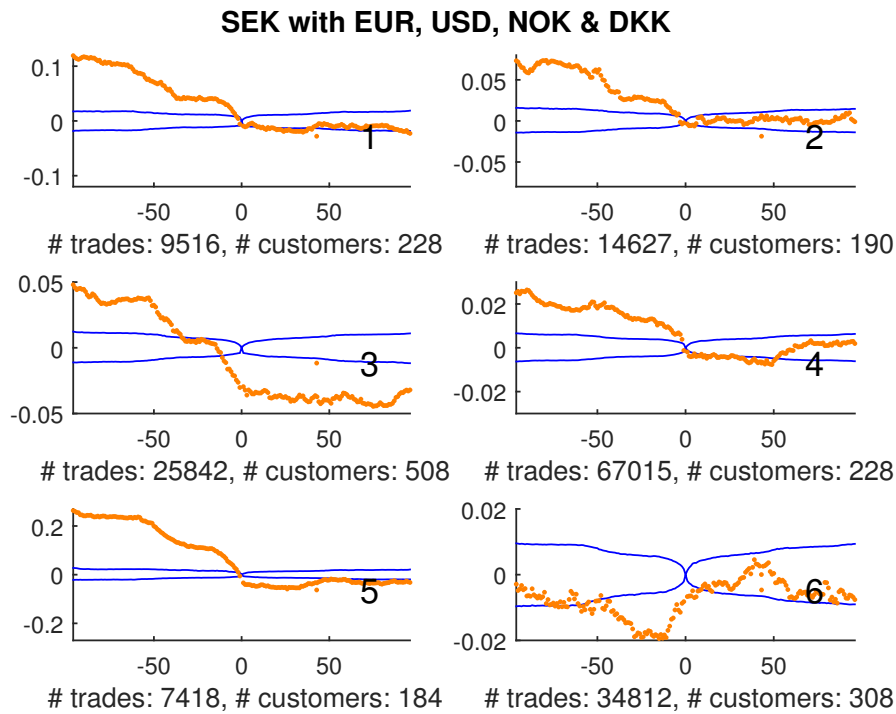


FIGURE 7.23: Volume-weighted indices for the six clusters created by Fuzzy C -means for the SEK flow are represented by the orange dots. The blue solid lines are the 99% empirical confidence intervals. All clusters except number 6 show a contrarian behaviour, which is most significant for clusters 1 and 5. Cluster 6 shows a small trend-following behaviour prior to trading, but no significant post-trading pattern.

or in highly volatile periods. The spot impact of these customers after the trade is not statistically significant.

Self-Organizing Map

The result of the self-organizing map is very similar to the result of the K -means. The formed clusters are almost identical but the cluster numbers are switched. This holds for the silhouette values in Figure 7.19c, the clusters in Figure 7.24 and the indices for the clusters in Figure 7.25. The interpretations will not be repeated here but can be found in the part about K -means. The prime brokered clients are divided into the same clusters as with K -means as well.

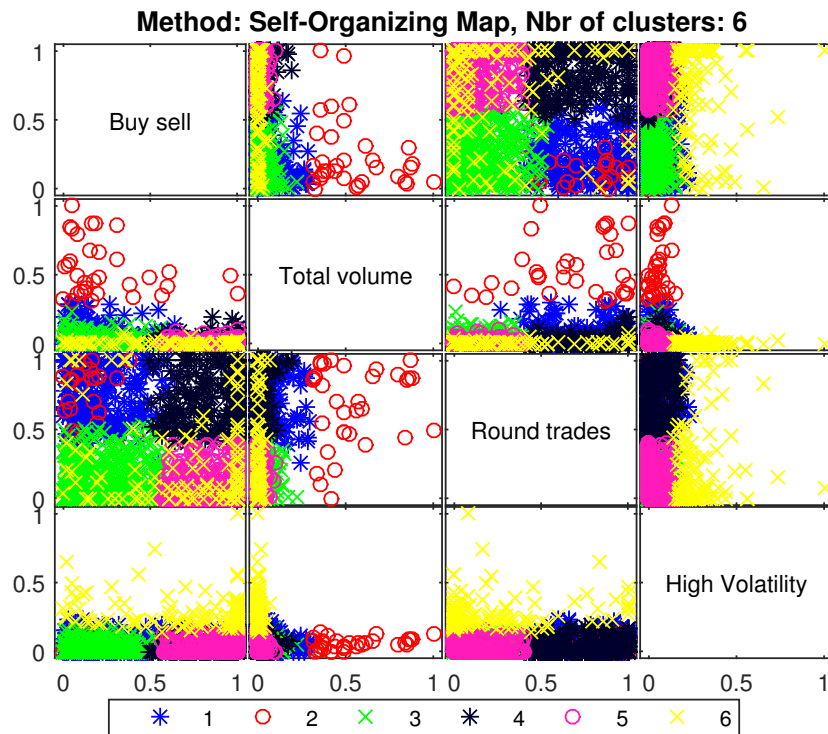


FIGURE 7.24: This scatter plot matrix shows the result of the SOM clustering of the customers for SEK, with normalised values of the features.

Comparing the methods

Since the K -means algorithm and the self-organizing map give almost the same clustering, only four customers have changed clusters, the result is considered to be rather robust. There exists an almost direct mapping between the clusters from these two methods. The Fuzzy C -means is performing poorly compared to the two other methods. The contrarian behaviour is seen among all clusters, even though it is known that there are trend-following customers as well, which is clear from Figure 7.6. The customers trading more in highly volatile periods are the ones who see the largest price change before trading, followed by customers trading in one direction and often in round numbers. The customers who are showing the random patterns are in the opposite corner, that is trading non-directional and not in round numbers.

7.2.3 Clustering of the Asset Managers' and Corporates' Flow

This part of the analysis is done only for the data set with trades in Swedish Kronor and with the K -means algorithm. It is made to illustrate the varieties in the trading

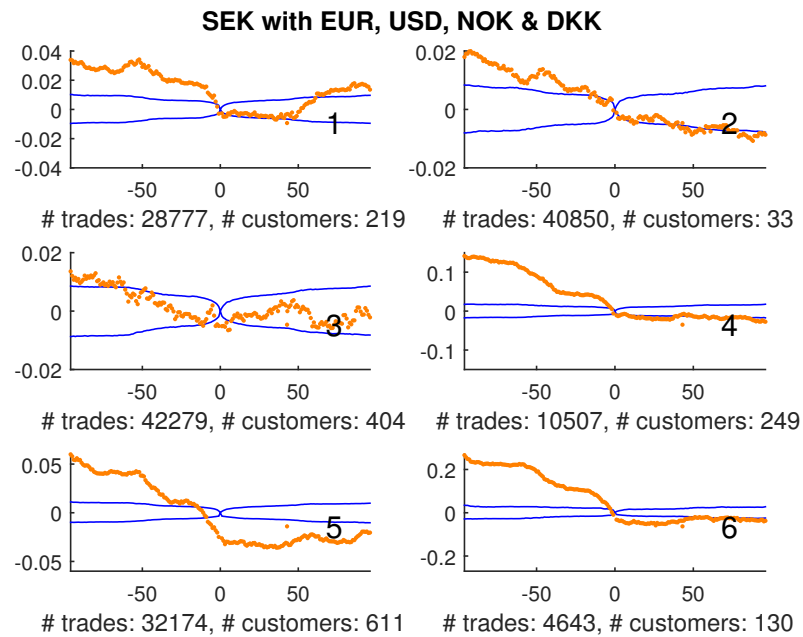


FIGURE 7.25: Volume-weighted indices for the six clusters created by K -means for the SEK flow are represented by the orange dots. The blue solid lines are the 99% empirical confidence intervals. Clusters 4 and 6 are the most contrarian and also the directional customers. Clusters 1, 2 and 5 are contrarians as well but to a smaller extent. Cluster 3 does not show any significant pattern. These clusters can be almost exactly mapped into the clusters created by K -means

styles within the intuitive customer groups.

Asset Managers

The Calinski-Harabasz criterion suggests ten clusters with the K -means algorithm. The silhouette values are good with only a few negative, the smallest is above -0.25 . The result of the clustering of the asset managers can be seen in Figure 7.26.

Figure 7.27 suggests that the spot impact differ between customers in the group asset managers. The clustering performed with K -means shows that only one cluster has a significant positive spot impact, that is cluster 2. The customers in this cluster are identified with a low trading volume and that they are in between directional and non-directional. They do not trade specifically at high volatility. It can also be seen that some asset managers are more contradicting in their trading behaviour, for instance the customers in cluster 3 and 9. These customers tend to be more directional and trade more in round numbers. The only cluster that has a negative spot impact after the

trade consists of only two customers, namely the ones that trade the most in highly volatile periods. Two customers are too few to be able to detect a certain trading style. However, most of the clusters do not show any significant behaviour according to their indices at all.

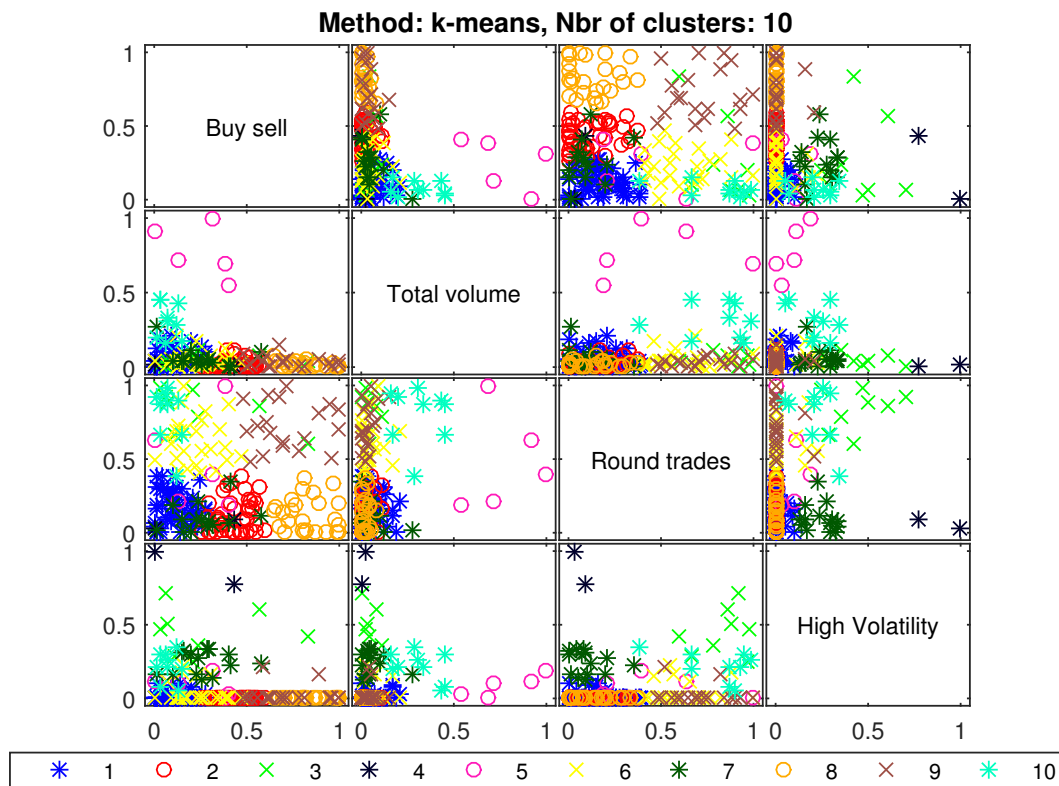


FIGURE 7.26: The result of the K -means clustering of the asset managers with ten clusters. The values of the features are normalised.

Corporations

Six clusters will give the best fit for the clustering of the corporate clients according to the Calinski-Harabasz coefficient. The silhouette values are telling that most customer observations are placed in the best cluster with the K -means algorithm.

All clusters show the contrarian behaviour in volume-weighted indices in Figure 7.29. However, there is some variation in how much the spot changes before they trade, the customers in cluster 3 trade after the largest price change and cluster 6 comes second. The remaining clusters see smaller changes prior to the trade time. Most clusters don't have a significant spot price impact after the trades, but the most significant is the

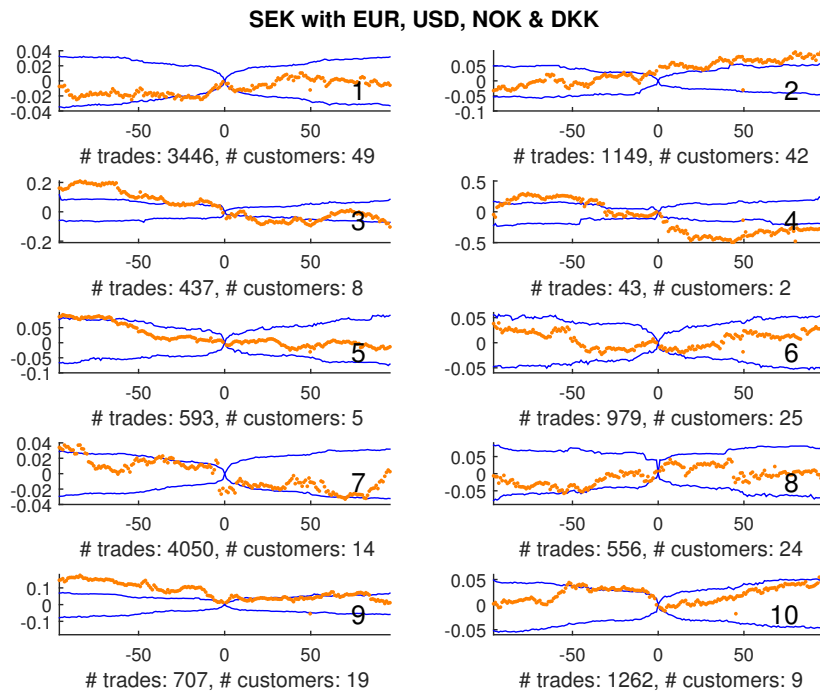


FIGURE 7.27: The orange dots are the volume-weighted indices of the clusters of the K -means clustering of the asset managers. The blue solid lines show the 99% empirical confidence bounds. Not much significant information can be found, except that the customers in clusters 3 and 9 are a little more contrarian than the rest. These are the customers that are directional and trade more in round numbers. Cluster 2 seems to have a positive spot price impact after the trade.

negative price impact from cluster 2. Those corporations are rather directional in their trading and do not trade round numbers or in volatile periods. This clustering does not distinguish any trend-following patterns among the corporate clients, but it shows that some are more strongly contrarian than others.

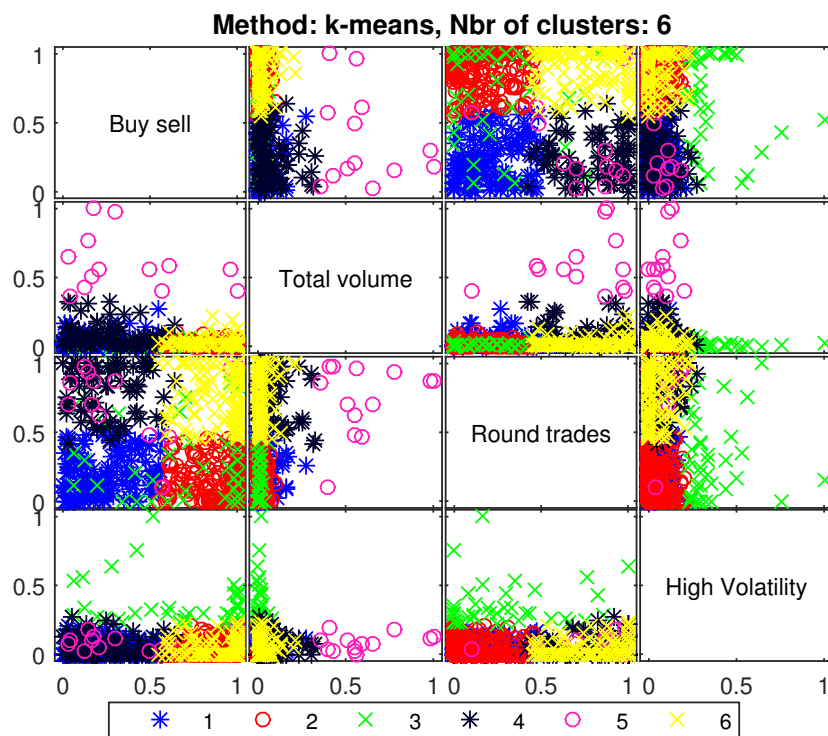


FIGURE 7.28: The result of the K -means clustering of the corporate clients alone. The values of the features are normalised.

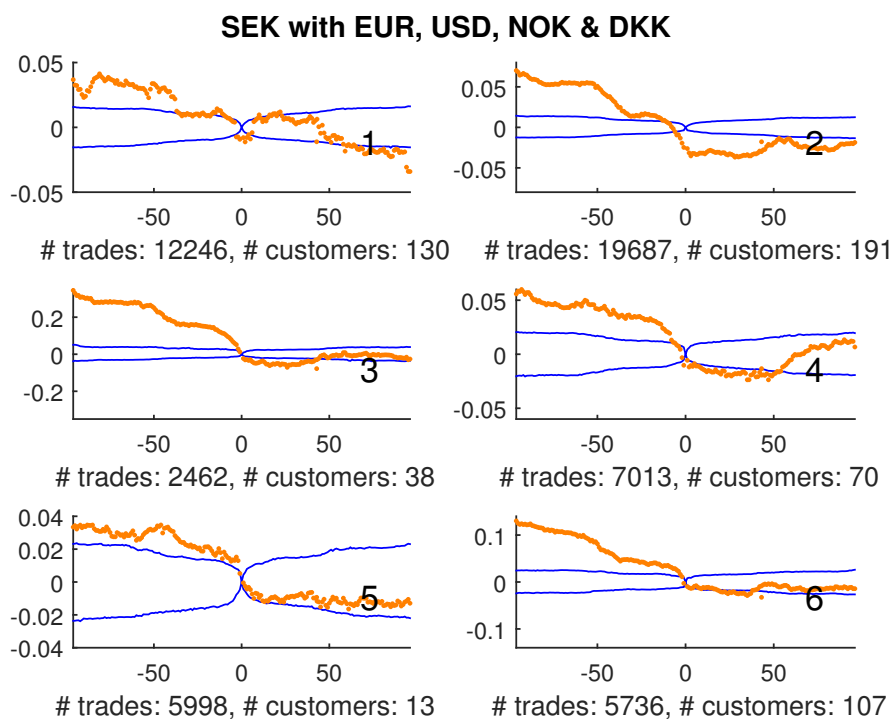


FIGURE 7.29: The orange dots are the volume-weighted indices of the clusters of the K -means clustering of the asset managers. The blue solid lines show the 99% empirical confidence bounds. All clusters are showing a contrarian behaviour but the percentage changes differ across clusters. Cluster 3 and 6 have the largest price movement before trading, they contain customers trading in high volatility or customers trading directional and in round numbers. After trading there is no significant spot impact except for cluster 2 whose customers are directional but do not trade in round numbers.

Chapter 8

Conclusions

8.1 Order Flow

From calculating indices for the four customer groups it can be concluded that there are differences in the trading styles across the groups trading with Nordea. Private clients and corporate clients are clearly showing a contrarian behaviour, while the asset managers and hedge funds have a tendency to be more trend-following. These results are confirming the conclusions made in the papers on customer order flow [Della Corte, Rime, Sarno, Tsiakis, 2011, Evans, Lyons, 2002, 2006, Marsh, O'Rourke, 2005, Menkhoff, Sarno, Schmeling, Schrimpf, 2013]. As expected the indices are more significant for the corporates group which contains the largest number of customers. The results are further clarifying that the negative relationship between non-financial customers' flow and exchange rate changes comes from their trading behaviour and not that they are constantly losing on their trades. In the same way the hedge funds are seen to have a trend-following trading style and seem to be informed.

The variation in informedness and professionalism between different customers within the same group affects the patterns that are seen. The goal of the Machine Learning techniques has been to discover other parameters than the group affiliation that affect the price impact of different customers.

The index plots of the prime brokered clients show that they are trend-following in their trading style. Their behaviour before the trade time is a lot more significant than both

the asset managers and the hedge funds. This indicates that the prime broker customers are most likely professional financial customers or large corporations.

8.2 Clustering

For NOK, trading a large volume seems to be the only feature that is connected to a trend-following behaviour and a positive price impact. Except from this the contrarian trading style is dominating all other clusters. The connection between large volumes and trend-following behaviour is not present for the SEK data.

For both currencies directional and contradicting trading styles can be associated with each other. Trading a large share of the volume in periods with high volatility is connected to a contrarian behaviour for the customers for both NOK and SEK.

In Section 7.1.2 it can be seen that the customers trading through a prime brokered agreement are showing a trend-following behaviour when the volume-weighted index is plotted. It is then concluded that most of the prime brokered clients are well-informed financial customers or large corporations. The trend-following pattern is not as clear for the customer segments that the PB clients end up in as a result of the clustering techniques. The clustering does not add extra information about the trading styles of the PB customers.

The clusters are never well-separated. Even though the silhouette values are predominantly positive, the cluster plots show that observations from different clusters lie very close to each other. The indices for different clusters also indicate that the clusters are rather similar in the way it is measured in this thesis. The conclusion is that it is difficult to distinguish any significant behaviours from the chosen parameters.

8.3 Summary

It is clear that different customers have different price impact when trading currencies, and the initially chosen customer grouping is showing some of this. The conclusions come down to that the new grouping performed by different Machine Learning Techniques does not provide better information about the trading style and price impact of different

customers than the initial one. The intuitive grouping tells more about the trading behaviour of a customer, although there are outliers breaking the normal patterns as can be seen in Section 7.2.3. The clients trading through a prime brokerage relationship are trend-followers and are most likely financial customers or large corporates.

The flow of Nordea is dominated by customers who trade according to a contrarian style. The result is that the non-contrarian customers are "drowned" by the large contrarian flow and the clustering techniques performed on the customers with the chosen features are not able to distinguish between them.

Chapter 9

Discussion

The order flow results turned out fairly well aligned with the expected ones. The expectations on the Machine Learning part were not well-defined since it is a totally new approach to this topic. There is a number of levels of uncertainty in the study and a few of them will be discussed in this chapter.

First of all anyone who has worked with raw data can sign that the data is not perfect from the beginning, and will not be perfect after manipulations either. There will still be outliers and random errors in different parameters. As an example it was noticed rather far into the study that the trades from 6th of October were not included in the table, obviously something went wrong when these trades were loaded into the table from the database. Another source of errors could be the initial customer classification which might be incorrect. It would be preferable if all customers were assigned to a customer type, but this requires a lot of time and effort so only the largest customers are classified.

The question why some customers affect the spot rate more than others does not have a simple answer. It can depend on parameters that are not visible to Nordea or not directly measurable. A popular explanation in the literature is the grade of informedness of a customer, which is difficult to observe from the outside. The initial grouping of customers according to the type of firm or institution is a rather good partitioning but still there are large differences within the groups. For instance there are corporates who trade more like an average asset manager and asset managers who trade more like most corporates. One suggestion for improving the grouping is to divide all corporations into two or more

groups for instance according to their size and to split the asset managers according to their main businesses (pension funds, insurance companies etc.).

When studying order flow it should be kept in mind that nobody in the FX market has the full picture. All market participants only see those parts of the market that are observable to them. Nordea has a significant part of the market share in SEK and NOK but not in other currencies. Therefore the flow in the Scandinavian currencies is the only order flow that could be expected to contain any significant information at all for Nordea.

The spot currency prices used are mid prices, which is an approximation. It would be more accurate to use ask prices for buys and bid prices for sells. However the results in this thesis would most likely not change considerably.

The clustering algorithms have been studied to achieve a basic understanding of several existing methods. One could also have chosen one single algorithm and put the effort on optimizing it. The Calinski-Harabasz criterion is used as a benchmark of the number of clusters. The output of `evalclusters` could have been more carefully studied. Other criteria were tested as well but are not included in this report. The result of the CH criterion gave the most satisfying results and was therefore chosen as quality measure. The Hierarchical clustering algorithm did not work out as hoped and the reason is a little unclear still. More investigations need to be made to get it to work properly. Another clustering algorithm that was tested but did not converge is the Gaussian Mixtures. Clustering techniques are sensitive to data, which means that if the data is inaccurately prepared it affects the result. There are various ways to choose which features to consider for a proper clustering of this particular data set. The features chosen in this thesis can contain errors and are not necessarily the most optimal choices.

The way of measuring the result of the different clusters in this thesis is not the only way but is convenient since the indices provide information about trading patterns. The volume-weighted indices give an indication about the price impact of the different clusters on average. Taking the average is not very sensitive to outliers and the median would be another option to see if there are customers that draw the result in a certain direction. Some robustness tests for the indices were made by checking the median as well and using different confidence bounds. Other possibilities could be to check if the

results are consistent if using only a part of the data. One could also use Bootstrapping to re-sample the data and see if the behaviour differs from randomness.

Some interesting steps further in this study would be to investigate other features of the customers than the ones chosen in this project. Moreover much work could be done to optimize the clustering techniques by for instance specifying initial values¹. Another option is to try other software or clustering techniques. When studying the underlying customers of the prime brokered deals one alternative method is that instead of including the prime brokered clients in the clustering process, supervised learning methods could be used. Then the data is trained on the pre-labelled customers and so the classes of the PB customers can be predicted using classification techniques.

The results of this thesis do not provide any new insights about how customers can be grouped in a better way than before. It tells that the FX market is very efficient and complex. Nordea's composition of customers is not preferable since corporates are not usually informed traders. If there would have been more hedge funds and professional asset managers the results might have been different and more significant. Single market participants might also change their trading styles as they receive new information, new systems or new personnel that impact their way of trading. In this way a single customer could have changed their trading behaviour during the studied time period. The investigations made have still contributed with a broader understanding of the customers of Nordea. Knowing how different customers trade in the market can be exploited for instance when the sales team attracting and supporting customers.

¹One suggestion is to perform clustering with one method and then use the resulting cluster centres as initial values in another algorithm.

Appendix A

Correlations and Regressions

The studies presented in this Appendix are more similar to the ones done in the papers about order flow and are included for a better comparison with earlier results [Della Corte, Rime, Sarno, Tsiaklis, 2011, Evans, Lyons, 2002, 2006, Marsh, O’Rourke, 2005, Menkhoff, Sarno, Schmeling, Schrimpf, 2013]. Two methods will be presented in the next section and the results in the following section.

A.1 Methods

For this analysis log spot changes are used as a measure of the price change in exchange rates, defined as in Equation 3.2. The spot prices are obtained at 17:00 on the actual day. The order flow aggregated over day t is denoted x_t . Only customer flow is used here, it means that all the interbank trades and prime brokered trades are removed.

A.1.1 Correlation Coefficients

As a measure of the dependence between flows and log spot changes Pearson’s correlation coefficient is used. For two different sets X and Y of n observations the correlation coefficient is defined as

$$\hat{r} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}}, \quad (\text{A.1})$$

where \bar{X} and \bar{Y} are the sample means. \hat{r} takes values in the range $[-1, 1]$, where 1 represents total dependence, 0 means independence and -1 stands for total but reversed dependence. The calculations are made with Matlab's function `corr` which also gives the P-values of the tests of the null hypothesis of no correlation against the hypothesis of nonzero correlation. A P-value that is less than a significance level, say 0.01, then the correlation is significantly different from zero.

A.1.2 Regressions

The second method for measuring dependence uses techniques called *simple linear regression*¹, which are performed with Matlab's `regress`. The regression model can be written

$$\Delta s_t = \beta_0 + \beta_1 x_t + \varepsilon_t \quad (\text{A.2})$$

and is solved with least squares. If there is a positive dependence then $\beta_1 > 0$ and if it is negative then $\beta_1 < 0$.

A.2 Results

This section presents results of the calculations described in previous section. All trades done in the interbank market have been excluded in this analysis. Pearson's correlation coefficient is calculated between the customer order flow aggregated over different horizons and the log spot change over the same horizon. The coefficients in the regression are estimated for the same horizons and the results are presented in table [A.1](#).

For the daily horizon the spot rates are taken at 17 : 00 (GMT+1) at the actual trading day. The change is therefore not in the exact same period as the trades (since they are aggregated with midnight as separation between days). The difference is considered to be insignificant since not many trades occur during night time anyway.

The analysis is also performed for the order flow of asset managers and corporations separately to see how the trading of groups are related to currency movements. Then x_t is exchanged to x_t^{AM} and x_t^{CO} respectively in Equation [A.2](#). The resulting estimated coefficients are presented in table [A.2](#).

¹This is a special case of multiple linear regression, where there can be more explanatory variables.

TABLE A.1: Estimates of correlation and regression coefficients for the daily SEK customer order flow and log spot changes.

horizon (days)	-10	-5	-2	-1	1	2	5	10
\hat{r}	-0.121	-0.186	-0.157	-0.139	0.003	-0.005	0.060	0.045
$\hat{\beta}_0$	-0.034	-0.022	0.000	-0.002	0.002	-0.005	-0.035	-0.071
$\hat{\beta}_1$ (10^{-3})	-2.29	-2.70	-1.50	-0.97	0.02	-0.05	0.88	0.85
P-value	0.002	0.000	0.000	0.000	0.948	0.893	0.122	0.250

Notes: Included currency pairs are SEK with EUR, USD, NOK and DKK. The negative correlations at horizons prior to trade are statistically significant with low P-values, while the results in future times are not.

TABLE A.2: Estimates of correlation and regression coefficients for the daily SEK customer order flow and log spot changes for the customer groups AM and CO.

Asset Managers								
horizon (days)	-10	-5	-2	-1	1	2	5	10
\hat{r}	0.002	0.025	0.042	0.091	0.024	-0.013	-0.005	-0.005
$\hat{\beta}_0$	-0.037	-0.023	-0.001	-0.001	0.003	-0.005	-0.035	-0.069
$\hat{\beta}_1$ (10^{-3})	0.049	0.545	0.589	0.094	0.246	0.180	0.099	0.150
P-value	0.964	0.514	0.282	0.019	0.536	0.744	0.907	0.892
Corporations								
horizon (days)	-10	-5	-2	-1	1	2	5	10
\hat{r}	-0.125	-0.193	-0.171	-0.188	-0.036	-0.008	0.051	0.043
$\hat{\beta}_0$	-0.034	-0.022	0.000	-0.002	0.002	-0.005	-0.034	-0.071
$\hat{\beta}_1$ (10^{-3})	-3.47	-4.09	-2.39	-1.91	-0.361	-0.115	1.11	1.21
P-value	0.001	0.000	0.000	0.000	0.355	0.832	0.184	0.263

Notes: Included currency pairs are SEK with EUR, USD, NOK and DKK. There is no really significant correlation found for the asset managers at any horizon, except for at the contemporaneous horizon (-1) which shows a small positive correlation. The corporates flow is negatively correlated with log spot changes at horizons before the trading day, but uncorrelated afterwards.

Nordeas aggregated daily customer order flow turns out to be negatively correlated with log spot changes back in time according to Table A.1. The P-values indicate that the coefficients are significant for all horizons prior to trade time. There is no significant correlation between the order flow and future log spot changes.

The results for the asset managers flow and the corporates flow are rather different from each other as can be seen in Table A.2. There is no significant correlation at any horizon

for the asset managers order flow except for a very small positive correlation one day before. In contrast to the asset managers, the corporates' flow is negatively correlated with previous log spot changes. The result for the corporations is very similar to the aggregated flow. Since the corporations account for about 56% of the total flow in SEK this result is not surprising. The results also align with the results of the indices from Section [7.1.1](#).

Appendix B

Features for Initial Customer Groups

Figures [B.1](#) and [B.2](#) show the scatter plots of the customer data for NOK and SEK respectively, where the objects are divided into the initial customer groups. It can be seen that there is a large diversity in the features within the pre-labelled customer groups.

It can be seen that the customers of all groups are spread out in all features. There are no clear patterns except that of course all XX clients trading the lowest volumes. It can also be noticed that the PB clients are non-directional.

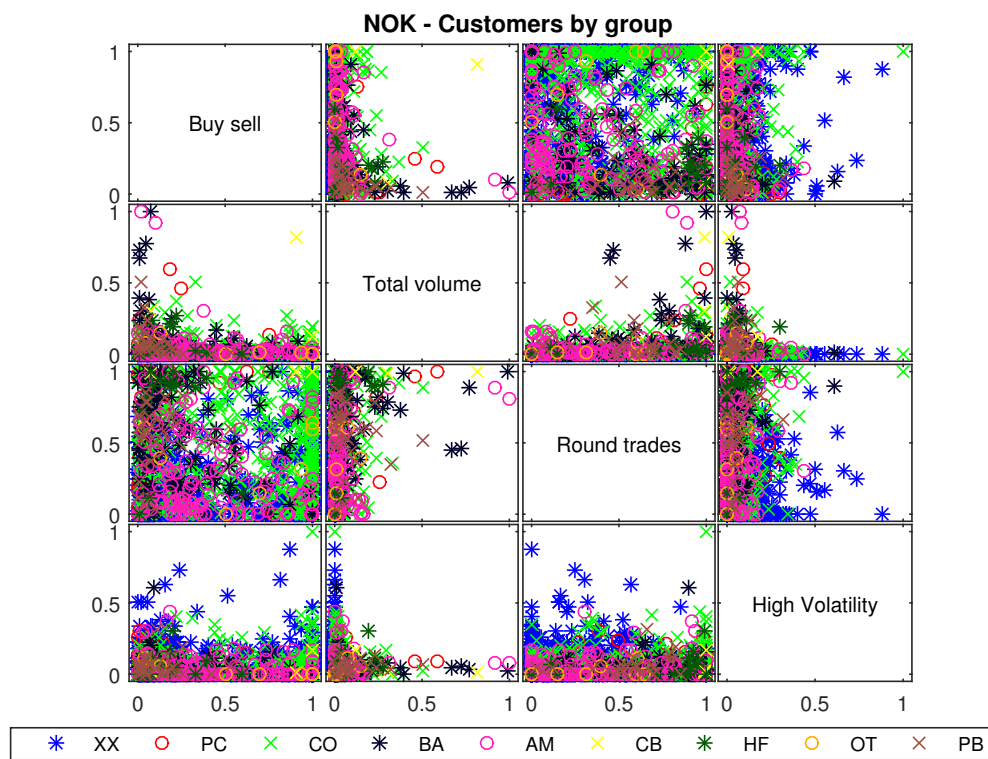


FIGURE B.1: This scatter plot matrix shows the normalised values of the four features of the customers according to the initial grouping for NOK.

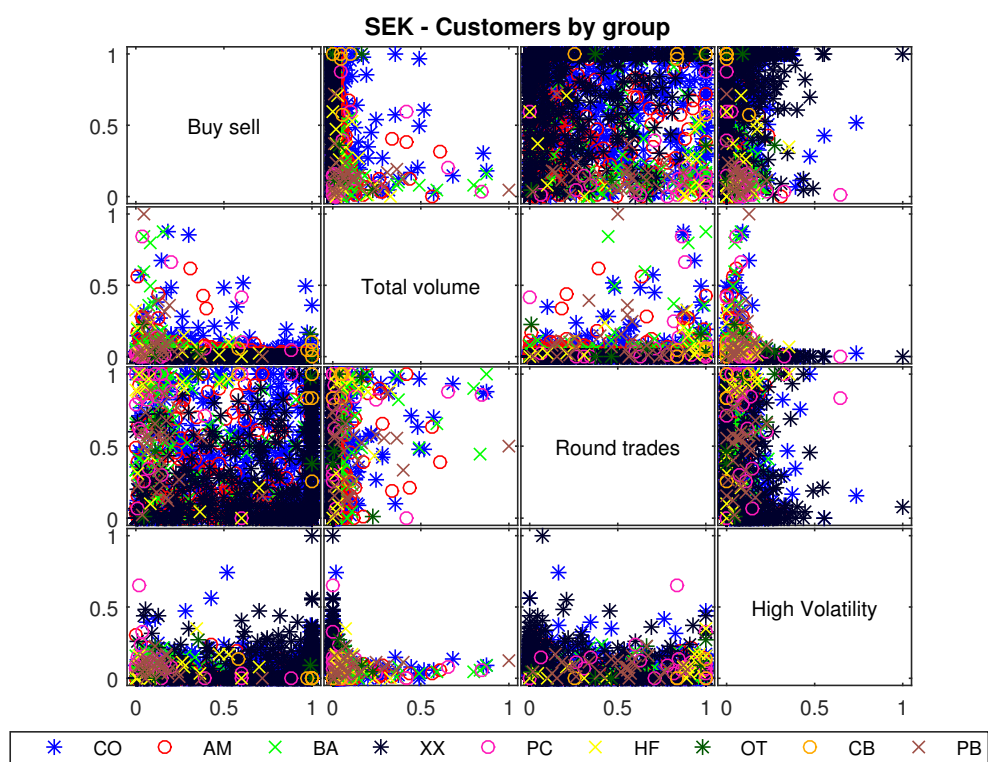


FIGURE B.2: This scatter plot matrix shows the normalised values of the four features of the customers according to the initial grouping for SEK.

Bibliography

- Bank for International Settlements.** September 2013. *Trennial Central Bank Survey*.
www.bis.org
- Bishop, C. M.** New York: Springer 2006. *Pattern Recognition and Machine Learning*.
- Bjønnnes, G., Osler, C., Rime, D.** *Assymetric Information in the Interbank Foreign Exchange Market*, Emerging Markets Group Working Paper Series 2009. http://www.cass.city.ac.uk/_data/assets/pdf_file/0019/29008/WP-EMG-16-2009.pdf
- Della Corte, P., Rime, D., Sarno, L., Tsiakas, I.** 2011. *(Why) Does Order Flow Forecast Exchange Rates?* http://www.gla.ac.uk/media/media_231269_en.pdf
- Evans, M. D. D., Lyons, R. K.** *Order Flow and Exchange Rate Dynamics*, Journal of Political Economy, pp. 110, 170-180, 2002.
- Evans, M. D. D., Lyons, R. K.** *Understanding Order Flow*, International Journal of Finance and Economics 11, pp. 2-23, 2006.
- Fan, M., Lyons, R. K.** *Customer trades and extreme events in foreign exchange*, Paul Mizen (ed.) Monetary History, Exchange Rates and Financial Markets: Essay in Honour of Charles Goodhart, pp. 160-179, Edward Elgar: Northampton, MA, USA, 2003.
- Foster, F. D., Rosov, S.** *Measuring the Information Content of Customer Foreign Exchange Orders*, Australian Journal of Management Vol 39(2), pp. 247-264, 2014.
- Hastie, T., Tibshirani, R., Friedman, J.** New York: Springer 2009. *The Elements of Statistical Learning*. Vol. 2. No. 1.
- Lindström, E., Madsen, H., Nygaard Nielsen, J.**, Chapman and Hall/CRC, 2015. *Statistict for Finance*.

- Marsh, I., O'Rourke, C.** *Customer Order Flow and Exchange Rate Movements: Is There Really Information Content?* https://www.cass.city.ac.uk/__data/assets/pdf_file/0008/29069/marsh_customerorder.pdf.
- Meese, R. A., Rogoff, K.** *Empirical Exchange Rate Models of the Seventies: Do They Fit Out of Sample?*, *Journal of International Economics* 14, pp. 3-24, 1983.
- Menkhoff, L., Sarno, L., Schmeling, M., Schrimpf, A.** *Information Flows in Dark Markets: Dissecting Customer Currency Trades*. BIS Working Papers No 405, 2013.
- Matteucci, M.** *A Tutorial on Clustering Algorithms*. Politecnico de Milano. http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/
- Mooi, E., Sarstedt, M.** Springer Texts in Business and Economics, 2014. *A Concise Guide to Market Research - The Process, Data, and Methods Using IBM SPSS Statistics Series*, 2nd ed.
- Rousseeuw, P. J.** *Silhouettes: A graphical aid to the interpretation and validation of cluster analysis*, *Journal of Computational and Applied Mathematics*, Vol. 20, pp. 53-65, 1987.