Annika H. Holmbom

Visual Analytics for Behavioral and Niche Market Segmentation

Turku Centre *for* Computer Science

TUCS Dissertations
No 195, June 2015

Visual Analytics for Behavioral and Niche Market Segmentation

Annika H. Holmbom

DOCTORAL DISSERTATION

To be presented with the permission of the Faculty of Social Sciences, Business and Economics at Åbo Akademi University for public criticism in ICT house, ground floor, Auditorium Gamma, on the 8th of June, 2015 at 12 o'clock noon.

Åbo Akademi University
TUCS – Turku Centre for Computer Science
Joukahaisenkatu 3-5A, FIN-20520 Turku, Finland

2015

Supervisors

Docent Tomas Eklund
Faculty of Science and Engineering
Åbo Akademi University
Turku, Finland

Professor Barbro Back
Data Mining and Knowledge Management Laboratory
Faculty of Science and Engineering
Åbo Akademi University
Turku, Finland

Reviewers

Professor Maria Holmlund-Rytkönen
Department of Marketing
CERS-Centre for Relationship Marketing and Service Management
Hanken School of Economics
Helsinki, Finland

Professor Ari Visa
Department of Signal Processing
Tampere University of Technology
Tampere, Finland

Opponent

Professor Maria Holmlund-Rytkönen
Department of Marketing
CERS-Centre for Relationship Marketing and Service Management
Hanken School of Economics
Helsinki, Finland

ISBN 978-952-12-3204-6
ISSN 1239-1883

Abstract

Companies require information in order to gain an improved understanding of their customers. Data concerning customers, their interests and behavior are collected through different loyalty programs. The amount of data stored in company data bases has increased exponentially over the years and become difficult to handle.

This research area is the subject of much current interest, not only in academia but also in practice, as is shown by several magazines and blogs that are covering topics on how to get to know your customers, Big Data, information visualization, and data warehousing.

In this Ph.D. thesis, the Self-Organizing Map and two extensions of it – the Weighted Self-Organizing Map (WSOM) and the Self-Organizing Time Map (SOTM) – are used as data mining methods for extracting information from large amounts of customer data. The thesis focuses on how data mining methods can be used to model and analyze customer data in order to gain an overview of the customer base, as well as, for analyzing niche-markets. The thesis uses real world customer data to create models for customer profiling. Evaluation of the built models is performed by CRM experts from the retailing industry. The experts considered the information gained with help of the models to be valuable and useful for decision making and for making strategic planning for the future.

Keywords: Customer Relationship Management (CRM), Customer Portfolio Analysis (CPA), customer segmentation, visual analytics, KDD process, SOM, WSOM, SOTM, data mining of customer data, profiling of green consumers, niche-market segmentation.

Sammanfattning

Kundrelationsstyrning (eng. Customer Relationship Management, CRM) med de tillhörande slagorden Big Data, datalagerhantering, visuell analytik samt data- och textutvinning med olika metoder är ett aktuellt forskningsområde som väckt intresse både inom den akademiska världen och företagsvärlden. Då utbudet överskrider efterfrågan, är kampen om kunder hård. Genom aktiv kundrelationsstyrning och med kännedom om kundbasen kan ett företag få en konkurrensfördel i kampen om större marknadsandelar.

Företag lär känna sin kundbas med hjälp av nyttig information. Data som beskriver kunder, deras intressen och beteende samlas in genom olika förmånsprogram bl.a. genom bonuskortinköp. Mängden data som lagras i företagens databaser ökar exponentiellt med tiden och blir svår att hantera.

I min doktorsavhandling forskar jag i köpbeteendet hos kunder bl.a. på stora varuhus, där kundbasen består av flera miljoner personer. Avhandlingen fokuserar på hur datautvinningsmetoder kan användas för att skapa modeller över och analysera verklig kunddata med målsättningen att få en överblick av kundbasen som motsvarar verkligheten. Den huvudsakliga datautvinningsmetoden som används är den självorganiserande kartan (eng. the Self-Organizing Map, SOM), dvs. ett neuronnät som visualiserar komplicerad flerdimensionell data på en tvådimensionell karta.

De skapade modellerna har vidareutvecklats för att användas för analys av nischer, dvs. mindre kundgrupper som innehar ett specifikt beteende. Ett exempel på en nischmarknad är gröna kunder, som föredrar att köpa ekologiska produkter.

Nyttan av de skapade modellerna har validerats genom att intervjua specialister på varuhusen. Enligt specialisterna kan den med modellerna erhållna informationen användas som stöd vid beslutsfattande, för målinriktad marknadsföring, till att förbättra kundbetjäningen och till planering av företagets strategi.

Tiivistelmä

Asiakkuudenhallinta (engl. customer relationship management, CRM) kuten myös siihen liittyvät avainsanat massadata (engl. big data), tietovarastointi (engl. data warehousing), visuaalinen analytiikka (engl. visual analytics) sekä tiedon- ja tekstinlouhintamenetelmät ovat ajankohtaisia aiheita, jotka ovat herättäneet mielenkiintoa sekä akateemisessa että yritysmaailmassa. Tunnetusti markkinoiden kysynnän ylittäessä tarjonnan on taisto asiakkaista kova. Aktiivisella asiakkuudenhallinnalla ja asiakastuntemuksella yritys voi saavuttaa tuntuvan kilpailuedun taistellessaan markkinaosuudestaan.

Jotta yritykset oppisivat tuntemaan asiakkaansa, tarvitaan hyödyllistä tietoa. Tieto, joka kuvaa asiakaskuntaa, asiakkaiden mielenkiinnon kohteita ja heidän käyttäytymistään kerätään eri etuohjelmien, kuten bonuskorttiostosten, avulla. Yritysten tietopankkeihin varastoitavan tiedon määrä kasvaa eksponentiaalisesti ajan myötä, mikä vaikeuttaa sen työstämistä.

Väitöskirjassani tutkin suuren tavarataloketjun asiakkaiden ostokäyttäytymistä. Ketjun asiakaslukumäärä on monta miljoonaa henkilöä. Tavoitteena on luoda todellisuutta vastaava yleiskuva asiakaskunnasta käyttämällä tiedonlouhintamenetelmiä asiakastiedon mallintamiseen. Pääasiallisesti käytetty tiedonlouhintamenetelmä on itseorganisoiva kartta, nk. Self-Organizing Map, SOM. Tämä neuroverkko visualisoi monimutkaista moniulotteista tietoa kaksiulotteiselle kartalle.

Malleja on kehitetty, jotta niitä voitaisiin käyttää niche-markkinoiden (engl. niche market) analysointiin. Nichet ovat tietynlaisen käyttäytymisen omaavia pienempiä asiakasryhmiä, esimerkiksi vihreät asiakkaat eli ekologisia tuotteita suosivat asiakkaat.

Mallien hyöty on arvioitu haastattelemalla tavarataloketjun asiantuntijoita. Asiantuntijoiden mukaan malleista saatua tietoa voidaan käyttää päätöksenteon tukena, asiakaspalvelun parantamiseen, kohdemarkkinoinnin kehittämiseen ja tukena yrityksen strategian suunnittelussa.

Acknowledgements

The PhD process is often seen as a lonely and challenging journey. As becoming a doctor often means that you are growing mentally, I admit that my PhD process has been challenging at times, but never lonely. Therefore, I want to express my gratitude to those who have supported me throughout my journey.

First, I want to thank my supervisors Docent Tomas Eklund and Professor Barbro Back for introducing me to the world of Business Intelligence and Data Mining, and for their guidance and support throughout my whole PhD process. Our meetings often started with a strict agenda, but ended up in discussions way beyond the academic matters.

Then, I would like to thank the reviewers of my dissertation Professor Maria Holmlund-Rytkönen and Professor Ari Visa for their time and effort for giving suggestions and comments on the manuscript. A special thank you is due to Maria for also acting as my opponent.

Collaboration with two case companies, co-authors, colleagues, Titan-project members and other scientists have taught me a great deal and widen the perspective of my thesis. I am sincerely grateful to the two case organizations for providing data, a special thanks is due to Hannele Humaloja-Virtanen, Tero Mustonen and Karlos Kotkas from SOK; my co-authors: amongst others Dr. Peter Sarlin, Dr. Zhiyuan Yao, Samuel Rönnqvist, and Professor Hannu Vanharanta; other former and present members of the Data Mining and Knowledge Management Laboratory: Dr. Hongyan Liu, Dr. Henri Korvela, Piia Hirkman and Minna Tähtinen; other scientists I have collaborated with: Professor Jan Westerholm and Artur Signell and last but not least our after work (AW) team amongst others Henrik Nyman, Xiaolu Wang, Dr. Robin Wikström, Susanne Ramstedt and Solveig Vaherkylä. Piia Hirkman and Annika Holmbom are in addition thanked for reviewing my Finnish and Swedish abstracts.

Many thanks to the whole administrative staff at the Faculty of Science and Engineering, as well as TUCS: to name a few, Christel Engblom, Joakim Storrank, Irmeli Laine, Outi Tuohi and Tomi Suovo.

I am grateful for the financial support from Tekes (Titan, grant no. 40063/08), TUCS – Turku Centre for Computer Science, Nokia Foundations, Hans Bang stiftelse, Turun Kauppaopetussäätiö, Seniorernas Råd inom Åbo Akademis Studentkår, Ella och Georg Ehrnrooths stiftelse, Waldemar von Frenckells stiftelse and Rektor's stipendium. I also want to thank Dr. Gerhard Kranner at

Viscovery Software GmbH for providing me a full license of their SOMine software to be used for my research.

My friends have had an important role throughout my PhD process. Some of them have experienced a similar growing process struggling with writing articles and their thesis. A special thank you to our lunch group Linda Nisula, Affi Leppänen and Anders Strand, for discussions while enjoying good food. Other friends have experienced a similar never ending dilemma of fitting work with family life. Special thanks are due to my friends Tiina Heinistö and Britta Haahti for your friendship and support especially during the last stages of my PhD process.

My parents-in-law Erika and Professor Bjarne Holmbom and sister-in-law Annika Holmbom are thanked for your support, especially the help we are receiving concerning our daughter. When the hours in one day are not enough, it is priceless to have a safe place for our daughter where I know she feels at home. In addition, nice discussions during delicious family dinners are well appreciated.

I would like to thank my parents Liisa and Henrik Rosqvist, sister Nina Närvänen and brother Thomas Rosqvist with families for standing by me at all times and for supporting me in my decisions. I really enjoy our social gatherings where the cousins are able to play together.

Last but not least I would like to thank my husband Thomas for sharing the everyday life with me and our daughter Linnea. The PhD process has been stressful at times, but my family have kept me attached to reality and reminded me of what is important in life. According to our daughter, now 4 years of age, the most important outcome of Mum doing a PhD is the hat similar to Moominpappa's. She wants to have one too when she grows up.

Turku, 27th of April 2015,

Annika H. Holmbom

Contents

List of Publications

1. Yao, Z., Holmbom, A.H., Eklund, T., Back, B. (2010). Combining Unsupervised and Supervised Data Mining Techniques for Conducting Customer Portfolio Analysis, In: Petra Perner (Ed.), *Advances in data mining, applications and theoretical aspects, Lecture Notes in Computer Science 6171*, 292–307, Springer Berlin Heidelberg.

2. Holmbom, A.H., Eklund, T., Back, B. (2011). Customer Portfolio Analysis Using the SOM, *International Journal of Business Information Systems*, 8(4), 396–412.

3. Vanharanta, H., Magnusson, C., Ingman, K., Holmbom, A.H., Kantola, J. (2012). Strategic Knowledge Services, In: Jussi Kantola, Waldemar Karwowski (Eds.), *Knowledge Service Engineering Handbook*, 529-557, CRC Press, Taylor and Francis Group.

4. Holmbom, A.H., Sarlin, P., Yao, Z., Eklund, T., Back, B. (2013). Visual Data-Driven Profiling of Green Consumers, *Proceedings of the International Conference on Information Visualization*, 291–298, IEEE.

5. Holmbom, A.H., Rönnqvist, S., Sarlin, P., Eklund, T., Back, B. (2013). Green vs. Non-Green Customer Behavior. A Self-Organizing Time Map over Greenness. In: Wei Ding, Takashi Washio (Eds.), *IEEE 13th International Conference on Data Mining Workshops*, *IEEE International Conference on Data Mining*, 1–7 December, IEEE.

6. Holmbom, A.H., Eklund, T. and Back, B. (2014). A Weak-form Expert Evaluation of Customer Profiling Models, *8th European Conference on IS Management and Evaluation ECIME2014*, 11-12 September 2014, Ghent, Belgium.

List of Figures

List of Tables

List of Acronyms

ANN	Artificial Neural Network
B2B	Business to Business
B2C	Business to Customer
BI	Business Intelligence
BMU	Best Matching Unit
CART	Classification and Regression Tree
CPA	Customer Portfolio Analysis
CRM	Customer Relationship Management
DM	Data Mining
DSR	Design Science Research
DT	Decision Trees
IS	Information Systems
IT	Information Technology
KD	Knowledge Discovery
KDD	Knowledge Discovery in Data bases
MBA	Market Basket Analysis
MCIF	Marketing Customer Information File
NN	Neural Network
OLAP	Online Analytical Processing
RFM	Recency, Frequency, Monetary
SO	Sub Objectives
SOM	Self-Organizing Map
SOTM	Self-Organizing Time Map
U-matrix	Unified Distance Matrix
WSOM	Weighted Self-Organizing Map

Part I

Research summary

Chapter 1

Introduction

1.1. Background

Trade in goods has existed throughout the history of humankind, including consumption for the display of material wealth, status, and fashion. However, early consumption of other than basic sustenance goods was exclusive to kings and tribal chiefs. Modern consumption in today's sense can be traced back to the late 16th century, when nobles during the reign of Elizabeth I ignited a consumer boom while competing for the attention of the Queen (McCracken 1998). Some of the most important events that have influenced consumption have been the creation of fashion during the 16th century (McCracken 1988, Corrigan 1997), the emergence of mass consumption, developments in advertising and marketing from the 18th century onwards (Corrigan 1997; McCracken 1988; McKendrick et al. 1982), the availability of products in department stores in the 19th century (Corrigan 1997; Bowlby 1997), mass production of products in the middle of the 20th century (Leiss et al. 2000), and the change in marketing views from being product-centered to customer-centered in the late 20th century (Pope 1983).

Through the more customer-centric view of marketing, companies' interest in identifying the wants and needs of their customers has increased (Datta 1996; Heinrich 2005; Chalmeta 2006; Buttle 2004, p.4). Earlier, when shops were generally quite small and geographically limited, shop keepers could keep track of their limited customer base without using special follow-up tools. As these businesses grew and became less regionally restricted, the number of customers also increased, and other methods were needed in order to understand the customers.

One method of gathering information about customers and their shopping behavior was to ask them directly. Data about customers were collected through questionnaires, polls and queries. It was, however, quickly discovered that the problem with these self-reporting methods is that people tend to answer according to their intention, not their actual behavior (Mainieri et al. 1997; Young et al. 2010).

Instead, marketers began to turn to the behavioral data generated by the increasingly powerful IT-based systems that keep track of everyday transactions

and events. As the capacity of IT systems grew, an increasing amount of transactions were performed electronically, leading to even more data gathered and stored. The need for tools to manage these increasing amounts of data led to the emergence of the field of *Customer Relationship Management (CRM).*

Today, research concerning CRM is still a topic of interest. It is widely used in ecommerce, but also in brick and mortar shops, as these still make over 90% of retail sales (eMarketer 2014). For example, a quick search of "Customer relationship management" using Google Scholar gave over 2 million hits, with nearly 70,000 hits for 2014. An existing trend within CRM amongst companies is to gather and store as much customer data as possible, e.g., through different loyalty card programs. The aim is to achieve a competitive edge in the race for larger market shares, by using the attained data for gaining a deeper understanding of the customers.

In order to explore and utilize the information hidden in the vast amounts of stored data, an increasing demand for methods for data mining and knowledge management has arisen (Dutta et al. 2015). For example, classification methods are often used for grouping of customers into beforehand known groups (Berry and Linoff 2004). The classification (or segmentation) is usually based on a combination of demographical, behavioral and psychographic information. The problem with classification methods is that as the groups are determined beforehand, they tell more about the manager's view of the customer base, than the actual customer behavior (Berry and Linoff 2004).

In order to reveal the customers' actual behavior or abilities, data driven clustering methods are used for grouping of customers. Statistics, tables and figures are often used to describe these segments. For this task, different data mining models have been applied since the 1990's. In order for managers to fully understand and be able to use the information extracted from data, the models used for clustering should give an intuitive output using simple and intuitive principles (Sarlin 2013d).

Visual analytics is an emerging field that aims to help the average manager to interpret and use information extracted from data. With visual analytics, vast amounts of different kinds of data in different formats are analyzed in a process where human judgment, visual presentations and different kinds of interaction techniques are combined (Keim et. al. 2008; Thomas and Cook 2006). The aim of visual analytics as a part of data mining is to turn large amounts of unstructured data into useful information. Advanced computer applications are used for the information discovery process allowing decision makers to fully concentrate on the analytical process and to visualize the information useful for them (Keim et. al. 2008). Thus, the use of interactive visualization methods is integral to visual data mining.

The most promising visualization techniques are based on data and dimension reduction (Sarlin 2013d). Their aim is to represent multidimensional data in a more understandable format. In his study, Sarlin (2013d) made a comparison of data and dimension reduction techniques. He came to the conclusion that the Self-Organizing Map (SOM) has many advantages over others: e.g., trustworthy neighbors, low computational cost, flexibility for problematic data and a regularly shaped grid. The SOM is an analytical tool that has been widely applied in different business related areas (Kaski et al. 1998; Oja et al. 2003; Deboeck and Kohonen 1998), including market segmentation (D'Urso and Giovanni 2008; Lee et al. 2006). The SOM is a highly visual, non-parametric and robust method for segmentation (Kohonen 2001).

The outcome of a segmentation with the SOM is an intuitive map that visualizes different groups that the customer base consists of. Most analyses concentrate on collecting information regarding profitable and unprofitable customer segments. The tough competition for customers has created an interest in analyzing even smaller specific market segments. These so-called niches are groups of customers with distinctive interests or behavior. By understanding them the company can serve these niches according to their needs and gain profitable and loyal customers. One example of a niche-market is green consumers, where the common factor among the customers is their interest in green, or eco-friendly, products. So far, earlier studies based on mainly self-reporting methods and interviews, have not been able to identify a unanimous profile describing the green consumer (see, e.g., Roberts 1996; Tsakiridou et al. 2008; Banyte et al. 2010).

1.2. Research aim and objectives

The overall aim of this thesis is to build and evaluate segmentation models for customer relationship management (CRM), based on customer behavior. In order to fulfill the overall aim I have derived three sub objectives (SO).

SO1. To investigate how segmentation has evolved and what are the current requirements. The aim is to investigate what factors have influenced segmentation throughout time, in order to understand the requirements for segmentation today.

SO2. To build and evaluate customer segmentation models within retailing for extracting information and knowledge from large amounts of customer data using Self-Organizing Maps.

SO3. To further develop the built models in order to study current "niches."

1.3. Research Methodology

The research methodology used within this thesis follows the framework for IS research presented by Iivari in 1991. Iivari has based his framework on the work of Burrell and Morgan (1979), who in turn base their model on four major paradigmatic constituents: *ontology*, *epistemology*, *methodology* and *ethics of research*. Ontology focuses on the assumptions made about the investigated phenomena. Epistemology studies the nature of knowledge, how to acquire it and its limitations. Methodology is the study of research methods. The ethics of research dictate that the scientist is responsible for the results and the consequences of his or her research (Iivari 1991). This results in two paradigmatic extremes; the *nomothetic* (positivism) and *idiographic* (interpretivism) approaches. The nomothetic approach is mainly used in the natural sciences and focuses primarily on objective measurements obtained through direct observation using quantitative methods, such as formal-mathematical analysis, laboratory and field experiments, field studies and surveys. The idiographic approach is mainly used in the social sciences and is based upon interpretation of the world through qualitative methods such as case studies, interviews and action research. Within these two extremes, both the ontology and epistemology of my research lies closer to interpretivism, as I am describing customer behavior through the construction of a model. One of the implications of this is that the results of the research carried out in this thesis are not generalizable.

However, information systems research differs from research in the traditional natural and social sciences, and therefore, also requires a different research approach (Galliers and Land 1987). Simon (1969) pioneered the research in this area by focusing on engineering and technical development as a research approach. Iivari (1991) and March and Smith (1995) continued his line of work by formalizing the constructive approach, which has to do with creation of socio-technical artifacts, e.g., decision support systems, modeling tools, governance strategies, and methods for IS evaluation (Simon 1969; Iivari 1991; March and Smith 1995, Gregor and Hevner 2013). Simultaneously and independently, Kasanen et al. (1993) pioneered a comparable approach, also termed constructive research, within the field business administration, where models were used specifically for problem solving.

Constructive Research and *Design Science Research (DSR)* are often referred to as same type of research. In this thesis, the term DSR is used. DSR is a problem solving process, where for humans relevant and practical problems are solved with the help of technology. The aim is to build a model, i.e., to create a valuable and innovative solution to a specific problem, and to evaluate and determine a value for the solution. One difficulty with DSR is that the evaluation of the performance of a solution or an artefact is dependent on the environment in

which it is implemented. Progress is achieved when more effective technologies replace existing ones (March and Smith 1995; Hevner et al. 2004; Hevner 2007). Typically, DSR builds upon existing research and on the newest technical advances. Methods used within DSR are case studies, qualitative and quantitative methods (Kasanen et al. 1993; Järvinen 2001).

Similar to other research approaches, DSR has its own dichotomies: Research within IT can be divided into *descriptive* and *prescriptive research*. The aim in descriptive research is to understand and describe the object of the study, i.e., the nature of IT. Descriptive research is repeatable and similar to the natural sciences in its assumptions. In prescriptive research the aim is to solve problems and to improve the performance of the object of study, i.e., to use knowledge in order to improve IT performance. It is seldom repeatable, as the object of study or the conditions for doing research are rarely exactly the same. DSR belongs to prescriptive research (March and Smith 1995). This thesis follows the DSR paradigm, and is therefore, prescriptive.

The products of DSR are of four types: *constructs*, *models*, *methods* and *instantiations*, which all strive to be innovative and valuable. The constructs or concepts define the terms for the tasks. A model expresses the relationship between propositions and statements in the constructs, as it describes how things are. A method is a guideline for the steps that need to be taken in order to perform a task. Methods are based on constructs and models. An instantiation is the implementation of the artefact in its environment. It puts into action the previous research outputs, i.e., methods, models and constructs (March and Smith 1995; Hevner et al. 2004). A research framework on DSR activities vs. outputs described above is presented in Figure 1.

	Build	**Evaluate**
Constructs		
Model		
Method		
Instantiation		

Figure 1. DSR activities vs. outputs, the research framework redrawn from March and Smith (1995).

In this thesis, the research is positioned in the design science approach, in both the build and evaluate columns of the March and Smith research framework as presented in Figure 2. Models for customer segmentation and customer profiling are built and evaluated through both technical validation measures and a weak-form expert evaluation.

	Build	Evaluate
Constructs		
Model	• Customer segmentation models • Customer profiling models	• Clustering validation measures • Quality measures of the SOM • Weak-form expert evaluation of customer profiling models
Method		
Instantiation		

Figure 2. DSR activities vs. outputs, a modified research framework from March and Smith (1995) valid for this thesis.

How to conduct DSR in practice has been described by many (e.g., Kasanen et al. 1993; Hevner et al. 2004). These guidelines describe a process from finding a relevant problem, construction of a solution, demonstration and evaluation of the solution, and examining the scope of application of the solution. In this thesis, the seven DSR guidelines described by Hevner et al. (2004) are followed:

1) Design as an artifact: The produced artifact should be a construct, model, method or instantiation.

2) Problem relevance: Development of technology-based solutions to important and relevant business problems.

3) Design evaluation: Demonstrate the utility, quality and efficacy of the artifact with the help of evaluation methods.

4) Research contributions: Contribute to DSR in the areas of design artifact, design foundations and/or design methodologies.

5) Research rigor: Rigorous methods should be applied within DSR construction and evaluation of the artifact.

6) Design as a search process: Searching for an effective solution requires utilization of available resources, while satisfying laws in the problem domain.

7) Communication of research: Efficient presentation of DSR both to technology-oriented and management-oriented audiences.

In this thesis, models for customer segmentation are built using data mining methods (1). As information gained with self-reporting methods mainly reflect the customers' intention (Mainieri et al. 1997; Young et al. 2010), data driven clustering methods are used for revealing customers' actual behavior or abilities. The SOM and two adaptions of it are used for clustering in order to gain an intuitive output consisting of valuable information (Sarlin 2013d). The first analyses performed in this thesis concentrate on identifying the profitable and unprofitable customer segments. Tough competition of customers has created an interest in analyzing even smaller specific market segments called niches, and therefore, further analyses are made with the aim to understand niche markets and their distinctive interests or behavior (2). Evaluation of the solution was performed through technical measures and face validation by sales experts (3). We have built different models for conducting customer profiling on large amounts of customer data. An understanding of the topic was gained through studying the literature related to the subject. Relevant concepts were CRM, customer segmentation, knowledge discovery in databases (KDD), visual analytics, and data mining. A working solution was designed by combining the concepts and our method described above. The conclusions derived from this study and ideas for future research have been published in scientific conferences and journals, e.g., papers 1-6 and will also be summarized in Chapters 6 and 7 (4, 5 and 6). The results have been communicated through presentations at meetings, memos and reports to both technology-oriented and management-oriented audiences at the case organization in question (7).

1.3.1. Evaluation

In Guideline 3, Hevner et al (2004) highlight evaluation as an important part of DSR. Among others, Venable, Pries-Heje and Baskerville (2012) have studied DSR, with the aim of making evaluation more understandable and easier to conduct. They have constructed frameworks for selection of the most suitable evaluation strategy and method (cf. Venable et al. 2012 and Pries-Heje et al. 2008). The framework consists of four fields introduced by Pries-Heje et al. (2008), where *Ex Ante* and *Ex Post* strategies are plotted against *Naturalistic* (i.e., field setting) and *Artificial* (i.e., lab setting) paths. The Ex Ante strategy implies that evaluation is conducted straight after designing an artefact, before actually constructing it. The Ex Post strategy, in turn, implies that evaluation is conducted after designing and constructing the artefact. In another paper (Venable et al. 2012), the authors classified methods used for evaluation

according to the fields of the framework and designed a new DSR evaluation methods selection framework. This framework is illustrated in Figure 3.

	Ex Ante	**Ex Post**
Naturalistic	• Action research • Focus group	• Action research • Case study • Focus group • Participant observation • Ethnography • Phenomenology • Survey (qualitative or quantitative)
Artificial	• Mathematical or Logical proof • Criteria-Based Evaluation • Lab experiment • Computer simulation	• Mathematical or logical proof • Lab experiment • Role playing simulation • Computer simulation • Field experiment

Figure 3. The framework for selection of the most suitable evaluation strategy and method. Redrawn from Venable et al. 2012.

In research paper 6, we have built two models that serve as socio-technical products. The models are on customer shopping behavior and customer profiling and the outcome is information on customer shopping behavior. According to Venable et al. (2012), our evaluation could belong to both the Ex Ante and the Ex Post strategy in Figure 3. The Ex Ante strategy could be suitable as the product has not yet been implemented in the company and the users cannot use the model themselves. The Ex Post strategy could also be suitable, as the model is almost ready and the users are evaluating the product/outcome from their own point of view (and opinion), i.e., information gained from the model. In both of these cases, our evaluation follows the Naturalistic strategy, as we are dealing with real people, experts in their field from a real company.

According to the Ex Ante - Naturalistic path of the framework, a suitable method for our evaluation would be Action Research or Focus Group. As we did not participate in the study ourselves, Action Research was not a suitable method for our study. The Ex Post – Naturalistic path mentions qualitative or quantitative surveys, which are more suitable for our evaluation. Therefore, we have chosen to use the Ex Post-Naturalistic strategy and qualitative interviews as our evaluation method.

As guidance for the exact method used for the evaluation, Venable et al. (2012) and Pries-Heje et al. (2008) suggest following the existing literature on research methods. As an example they mention the DeLone and McLean (1992; 2003) model on IS Success. The authors systematically analyzed 180 IS success studies that they had collected, and divided these measures of IS usefulness into six categories: *system quality*, *information quality*, *information use*, *user satisfaction*, *individual impact* and *organizational impact*. In our research, we have built two models to support experts in customer profiling tasks. Therefore, a field study utilizing experts and potential end users was selected as an appropriate evaluation setting.

We have chosen to use the information quality path of the DeLone and McLean model (1992; 2003) for our evaluation, as we evaluate the information gained from the model. The questionnaire was built based on the End-User Computing Satisfaction (EUCS) framework developed by Doll & Torkzadeh (1988). The five most important factors in assessing user satisfaction with information are: content, accuracy, format, ease of use, and timeliness. The Doll and Torkzadeh framework was used in this study because it focuses more on information quality and use than many other available models, and was, therefore, found to be more suitable for this study (Doll and Torkzadeh 1988).

1.4. Contribution and Publications

This thesis consists of six research publications that have been published at peer reviewed conferences (3/6), in scientific journals (2/6) and as a book chapter (1/6). In the first publication, Yao et al. 2010, different ways for building a segmentation model were investigated. A combination of unsupervised and supervised data mining techniques was used for conducting customer portfolio analysis. Writing of the article was a joint effort. My contribution was focusing on the segmentation analysis.

The second publication, Holmbom et al. 2011, deals with customer data from a company conducting B2B. The case company contributed with customer data, which I preprocessed and used for building a segmentation model. I performed the literature review, the analysis of the data and the segmentation analysis. The writing process of this article was a joint effort of the authors.

The third publication, Vanharanta et al. 2012, is a book chapter for which all of the authors contributed with a specific part. My contribution was one of the case studies “From data to knowledge: A case study in customer segmentation with the Self-Organizing Map”. The case study is based on the department store data set. The case company contributed with customer data, which I prepared and

used for building a segmentation model. I analyzed the data, conducted the segmentation analysis, and wrote the case study.
The fourth publication, Holmbom et al. 2013a, was a joint effort between the authors. The aim was to perform a weighted segmentation of green customer data with a Weighted Self-Organizing Map (WSOM), in order to identify niche markets. My contribution was the literature review on green consumers and providing data concerning green products of interest. I also contributed to the analysis of the results. The writing process was a joint effort.

The aim of the fifth publication, Holmbom et al. 2013b, was to perform a segmentation of green customers using the Self-Organizing Time Map (SOTM), in order to analyze changes in degree "greenness" of the formed segments of green customers. My contribution was the literature review on green customers and the analysis of the formed green customer segments and changes based upon the degree of greenness. The writing process was a joint effort between the authors.

The sixth publication, Holmbom et al. 2014, was a weak form evaluation of the information extracted from transaction data with the help of Market Basket Analysis and customer segmentation analyzes. The weak form evaluation was performed in the form of qualitative interviews of experts of the department store chain. I have performed the literature reviews, constructed the surveys, conducted the interviews in May-June 2013, and analyzed the data. The writing process was a joint effort between the authors.

1.5. Overview of the thesis

In this section, an overview of the thesis is presented.

Chapter 1 includes the introduction, research aim and objectives, an introduction to the two case studies used in this thesis, my contributions to the publications and an overview of the thesis. This chapter provides a summary of the thesis. Existing problems studied in this thesis are presented.

In Chapter 2, the history and development of customer buying behavior are presented in relation to the key motivation for shopping in different eras. This chapter explains how customer segmentation has evolved over time and discusses current requirements for segmentation. The chapter essentially serves as a state of the art of the study of customer buying behavior and explains why this topic continues to be of high interest. Customer buying behavior is analyzed using data driven exploratory customer segmentation methods in papers 1-3 and 6.

Chapter 3 presents the concepts of CRM and segmentation and why they are important. This chapter provides background information on customer segmentation, which is the key method used throughout this thesis. Data driven exploratory customer segmentation is performed in papers 1-3 and 6. In addition, a study on a niche market, green consumer behavior, is presented. Green consumer behavior is a topic that has been of interest both in academia and industry since the 1970s. Earlier studies have been mainly based on self-reporting methods, giving no information on actual green consumer behavior. In papers 4 and 5, green consumer behavior is analyzed using data driven segmentation techniques instead of relying on self-reporting methods as has been done previously.

Chapter 4 continues by presenting concepts within visual analytics and how the knowledge discovery process has been adapted within this thesis. The SOM is the main technique used for data mining and visualization of the results. The preparation of data follows the KDD process, while the visualization of the results follows the Visual Analytics model. Therefore, a hybrid method where the KDD process is merged with the Visual Analytics method is proposed. The hybrid visual model for knowledge discovery has been used throughout the thesis, including all of the papers, for building the models and visualizing the results.

In Chapter 5 the data mining methods used within this thesis are described in more detail. The main method used for performing customer segmentation is the SOM, as is presented in papers 1-3 and 6. Modifications of the SOM, called the WSOM and the SOTM, were used for studying green consumer behavior through segmentation, as presented in papers 4 and 5. Other methods, for example, Decision Trees, were used in paper 1 to analyze further the created segments achieved with the SOM.

Chapter 6 explains the results from the six research papers, i.e., papers 1-6.

Chapter 7 concludes the thesis by answering the research aim through the three sub objectives. The contribution of this thesis is discussed and suggestions for future research are given.

The summary of the structure of the thesis and their connections to the publications are illustrated in Figure 4. The overall aim of this thesis is to build and evaluate segmentation models for customer relationship management (CRM), based on customer behavior. In order to fulfill the overall aim I have derived three sub objectives (SO), which are visualized in Figure 4 using different colors:

SO1. To investigate how segmentation has evolved and what are current requirements. In Chapter 2 the evolution of customer buying behavior is described from Elizabethan until present time, pointing out the different drivers for each era. The aim is to investigate what factors have influenced segmentation throughout time in order to understand the requirements for segmentation today.

SO2. To build and evaluate customer segmentation models within retailing for extracting information and knowledge from large amounts of customer data using Self-Organizing Maps. The background information needed for building segmentation models is presented in Chapters 1 and 3 to 5, while the building of the models is presented in papers 1 to 3. The evaluation process is described in Chapter 1, while the evaluation of the model built in paper 3 is presented in paper 6.

SO3. To further develop the built models in order to study current "niches." As is described in Chapter 3, since the 1970s green consumer behavior is a niche market that has raised interest in both academia and industry. New ways for studying niche markets are described in Chapter 5 and used on department store data for profiling of green consumers niche market in papers 4 and 5.

Publications vs. Chapters	1	2	3	4	5	6
Chapter 1: Introduction						
Chapter 2: Customer Buying Behavior						
Chapter 3: CRM						
Chapter 4: KD and Visual Analytics						
Chapter 5: DM						
Chapter 6: Results						
Chapter 7: Conclusion						

Figure 4. Overview of the thesis, where the sub objectives are visualized with color-coding in relation to the chapters and publications. The publications and chapters correlating to the three sub objectives are highlighted in color: SO1 in red, SO2 in blue and SO3 in green.

Chapter 2

Customer Buying Behavior

This chapter describes the evolution of customer buying behavior starting from the beginning of the 16th century. We identify the drivers behind the changes, the effect of advances in technology, changes in marketing methods, customers and products. Techniques for profiling of customers are introduced. The aim of this chapter is to investigate what factors have influenced segmentation throughout time in order to understand the requirements for segmentation today.

The shopping behavior of people from different classes were about to undergo huge changes. An extensive literature review showed that there exists a number of common drivers for shopping, describing the customer buying behavior throughout time. McCracken (1988), Corrigan (1997) and McKendrick (1982) defined three of these drivers as politics, economics and romanticism. In addition, based on the literature review, I identified six more drivers for shopping in order to describe customer buying behavior from the beginning of the 16th century until present time. The added drivers were: need, fashion, dreams, social status, image, and awareness. Below, customer buying behavior is described from the beginning of the 16th century until present time using varying drivers for shopping.

2.1. Need - Nobles purchasing long-lived products for family

Our study concentrates on Europe and begins at the beginning of the 16th century, where goods were bought according to a need with previous and coming generations in mind. Patina on the furniture spoke of a high status and bore witness that the furniture had belonged to the family for generations. The family had been rich for some time and spoke of stable care taking of the family wealth. Patina was an indicator of status (McCracken 1988, pp. 1-12; Corrigan, 1997, pp. 1-14).

2.2. Politics – creation of fashion

During the reign of Elizabeth I (1558-1603) in the late 16th century, there was a consumer boom in England as nobles, i.e., the aristocratic elite, competed for the attention of the Queen. The driver for shopping was politics. Queen Elizabeth I did not rely on second hands when it came to distributing money to the nobles. Instead, the nobles were ordered to appear in her court. They needed to spend in order to be noticed by the Queen as well as to keep their status in the social competition. It was a time of elite consumption where men did the shopping. The gap between nobles and common people was huge. The buying behavior changed from noble men shopping for family and valuing patina, to purchasing new things for themselves that could increase their status, and therefore, be favored by the Queen. Fashion was created (McCracken 1988, pp. 1-12; Corrigan 1997, pp. 1-14).

2.3. Economics – fashionable goods made available

During the 18th century, consumption boomed as fashionable goods of increasing variety were made available also for other social classes than just nobles. The driver for shopping was economics. New goods were bought by women more frequently than men and the patina of the furniture lost its importance as an indicator of status. However, the goods in one's possession were still a symbol of social status. The beginning of a consumer society with mass consumption was created. Josiah Wedgwood (1730-1795), an English potter who industrialized the manufacturing of pottery, was a key person when it came to advertising. He had noticed that the demand of goods could be manipulated. He called it the "trickle-down" effect for fashion, by which he meant that what was fashionable in court soon became fashionable amongst nobility, followed by the gentry, the middle class and finally the lower classes. Wedgwood started using advertisements in newspapers. Soon after this, other marketing devices were also introduced, such as fashion magazines, fashion plates and fashion dolls (Corrigan, 1997, pp. 1-9; McCracken 1988, pp. 6-18; McKendrick et. al. 1982, pp. 110-112).

2.4. Romanticism – self-expression through consumption

In the late 18th to 19th centuries, the driver for shopping was romanticism. The middle and working class women were doing the shopping. During romanticism, people started to pay attention to self-expression and self-development. Shopping was a way to satisfy one's wants and needs. Shopping of novels was very typical for this era (Corrigan, 1997, pp. 1-15).

The first study on consumer behavior was made in 1857 in Germany. Information was gathered on production and consumption ratios of working class families in Sachsen Germany (Engel, 1857). Engel analyzed this information and concluded in his study (1895) that an increase in a family's income, decreases the share of income spent on food, and therefore, increases the proportion spent on other goods, i.e., luxury goods. Therefore, he concluded that the share of income used on food was the best measure of the material welfare of the working class (Sellerberg 1978, pp. 31-32).

By the end of the 19th century, Veblen (1899) coined the term "conspicuous consumption" while studying the consumption of a social class that had become newly rich. By this term he meant that consumption of unpractical and unnecessary things was a sign of wealth and social status. For example, women who wore unpractical clothes, in which it was certainly impossible to do any work, were status symbols for wealthy men whose wives did not have to work and who had servants to take care of the household. The impression of higher status was given by visible consumption and by displaying leisure time, i.e., indicating not having to work for a living, but still being able to afford to consume (Corrigan, 1997, pp. 17-26).

Industrialization and the technical developments in the production processes in the middle of the 19th century made it possible for factories to produce products much faster and on a larger scale compared to earlier, when products were made by hand. The availability of products and the growth of the middle class both in size and in wealth were the drivers for the creation of the department store. Some of the most famous department stores are Harrods in London (established 1834), Le Bon Marché in Paris (est. 1838) and Selfridges in London (est. 1909).

The department stores were built massive in order to impress. They contained all products one could wish for, under the same roof, and offered means for recreational shopping. Because of the free entry to the department stores, anybody could make purchases and in this way luxury was made available to everybody. Mainly middle class women came to the department store, not just in order to buy items, but to meet other middle class women. For the first time a segment based on consumer behavior was identified. The department store gave women a public space to socialize in, employment opportunities, and created a standard-model of the female body which was visible in the media (Corrigan 1997, p. 50-51, 55-65; Bowlby, 1997, pp. 96-97).

Shopping in department stores differed from earlier ways of shopping, as now the prices were fixed and products were on display making it possible to look at products without making purchases. As customers were not obligated to make any purchases, the advertisements, the stores, products and sales persons needed

to be constructed, presented and displayed in a new way in order to sell (Corrigan 1997, pp. 50-51, 55-65).

2.5. Fashion – an expression of status

One element of fashion is that it changes rapidly with time. According to Blumer (1969), the social function of fashion is to make a clean break with the past and satisfy the need for something new. Both Lefebvre and Albinsson have studied how fashion affects the demography of products: 1) the physical lifetime of a product, i.e., until the product breaks, 2) the economical one, i.e., how long the product has some value for exchanging it, and 3) the psychological or social lifetime of a product, i.e., when you get bored of your product (Sellerberg 1978, pp. 78-80). The insight of the study was that an old product does not have to be physically broken in order to be exchanged for a new one. In many cases, a product that is not fashionable anymore, will be exchanged for a more fashionable one. This means that fashion can be used as a reason for replacing a working product with a new, more fashionable one.

In the 19th century, the upper class used fashion for expression of their social status (McCracken 1988, p. 6; Sellerberg 1978, p. 3). By following fashion, and therefore, affording to buy new clothes often, you gave the impression of being of higher status. Simmel (1904) developed Veblen's idea further and stated that consumption and fashion gives social groups an opportunity to deviate from the norm, and at the same time express resemblance with others of equal social class. People of same social status bought often similar clothes, read the same newspapers and drove similar brands of cars (Corrigan, 1997, pp. 26-32). Clothing was seen as a language of its own (Sellerberg 1978, pp. 9-10).

Between the years 1890-1925, advertising had a product-oriented approach (Leiss et. al. 2000, pp. 244-246). The main aim was to spread information about the product. The ads were mainly textual, printed in magazines and newspapers and they visualized the product and its functions. The advertising was done mainly by national agencies (Leiss et. al. 2000, p. 249). With the creation of fashion, the role of consumption evolved from informing the consumers and just satisfying a basic need (Falk, 1994, p. 151). Advertisement was used as an active strategy in 1904 in America, where John E. Kennedy, a marketing expert, was one of the first to use it. Kennedy called advertisement "a salesman in print". The target of the advertisement was called the consumer. Advertising was meant to fulfil the consumer's needs and desires. The desires and needs are realized into a definite will to buy a product (Falk, 1994, pp. 151-153).

The first methods used within marketing in the 1920s were surveys, where newspapers and magazines interviewed their readers regarding their shopping

habits. The information was given to the advertisers of the magazine. Paul Lazarsfeld made use of several marketing surveys in 1932 in order to create a profile for the proletarian (i.e., low class) consumer and to compare it with the profile of the middle class consumer (Sellerberg 1978, p. 34).

2.6. Dreams – emotional advertising

Between the years 1925-1945, advertising focused on product symbols. Advertising on the radio made it possible to influence public policy. Advertising agencies started to conduct preliminary research on media audiences, although the information retrieved was initially very limited. Marketing started to shift towards non-rational or symbolic reasoning of consumption, meaning that advertising went from being rational to emotional. The products were sold based on the wanted qualities and status of the consumer, i.e., on dreams. Sponsorships and brand images were connected to social basis of consumption (Leiss et. al., 2000, pp. 249-250).

2.7. Social status – deviation from mass

Between the years 1945-65, advertising focused on personalization. The agencies got a new media, the television, alongside radio, newspapers and magazines. Celebrities were used for marketing products in all these media. More knowledge of the consumer was collected, even if the consumer was still treated as one big mass (Leiss et. al., 2000, pp. 250-251).

Between the years 1965-85, advertising focused on market segmentation. The first person to talk about market segmentation was Wendell R. Smith in 1956 (Pope, 1983, p. 259). With market segmentation, mass marketing was replaced by specifically tailored marketing campaigns for different types of customers. Advertising campaigns were carefully created based on statistics and marketing management and with the help of computers. Consumers were divided into subgroups according to activities, media they used, consumption preferences and lifestyle attitudes (Leiss et. al., 2000, pp. 251).

As an example of early market segmentation studies, Stone (1954) was the first to analyze the sociology of shopping. In his study he tried to prove that there existed social integration in the city life and not just depersonalized relationships. Stone identified different orientations towards shopping, one of them being the economic shopper, as will be described below (Hewer & Campbell, 1997, pp. 186-187).

There is a difference between men and women in their orientations towards shopping. Men are mainly categorized as what Stone described as economic shoppers, i.e. men think of shopping as work (Campbell 1997, pp. 167-171). Men go in to a shop, purchase what they need and come out. Women, on the other hand, tend to enjoy shopping and regard it as leisure time (Hewer & Campbell, 1997, p. 187). According to Campbell (1997, pp. 170-171), women like to browse around and look at alternatives in different shops before they make a decision to purchase a product. They are more likely to shop around and do shopping trips without a specific agenda. One exception is when women do their grocery shopping, as that is seen as work and is done as convenient as possible.

Bellenger and Korgaonkar (1980) revisited Stone's way of categorizing shoppers according to their orientations towards shopping, due to changes in the environment as well as the consumer. Instead, the authors divided shoppers into groups according to the time they spend on shopping. The authors carried out a survey, achieving 324 answers, and found that there are several social and personal motives for shopping. They identified a profile for a "recreational" shopper who finds pleasure in the shopping process and for whom shopping is so much more than just the purchase of products that are needed. They found that 69% of the respondents belonged to this group. The general characteristics for a recreational shopper are active women, who while shopping appreciate a pleasant atmosphere and a large variety of quality products. In comparison to the economic shopper, the recreational shopper spends more time shopping, will not shop only for one item, makes impulse purchases and does not dwell for too long before making a purchase. The recreational shopper is more likely to shop with others and mainly in department stores. The distance to the store is not as relevant as the quality, variety and décor of the store. Other recreational activities the recreational shopper might enjoy are hiking, camping and sports. She invites guests to her home, reads women's magazines and local newspapers, i.e., she seeks information. The demographic characteristics are women who are members of white-collar families (Bellenger & Korgaonkar, 1980, pp. 83-91).

2.8. Customers and their image of lifestyle

In the late 20th century the old product-centered view of marketing dating back to the industrial revolution was no longer an efficient way of selling products. In 1993 Tom Siebel introduced *Customer Relationship Management (CRM)* with its customer centered marketing perspective, with value maximization as a goal (Datta 1996; Heinrich 2005; Chalmeta 2006; Buttle 2004, p.4). The building blocks for CRM technology date back to the 1970s, where segmentation is one of the most important methods. CRM is used to identify and group customers with similar profiles or requirements. The grouping is based on customer

information readily available through Enterprise Resource Planning (ERP), corporate data warehouses, and the Internet (Rygielski et al. 2002; Buttle 2004). CRM and segmentation are presented in more detail in Chapter 3.

As consumers were grouped according to similar profiles or requirements, new market segments were created. It was soon discovered that the profile of a consumer could change, e.g., with time. The developments in technology and the availability of customer data made it possible to make the segmentation of customers based on a smaller "niche". Examples of these "niches" could, be e.g., green consumer behavior, customers interested in football, golf or fishing. Segmentation based on "niches" and green consumer behavior is presented in more detail in Chapter 3.

2.9. Requirements of today - Awareness and expression of one's lifestyle

As stated earlier, fashion changes with time and advances in technology aids in creation of new products and services. Consumer profiles change with time and the marketing needs to keep up with the changes. The new market segments need new ways and a new language for communication.

At the end of the 20th century, advertising agencies marketed products by creating a suitable atmosphere, an image of life style. The consumer was perceived as a combination of several personas with different life styles, depending where he was, i.e., at work, at home, etc. The marketing campaign needed to reach the consumer at the right time with the right advert in order to affect both the rational and emotional appeal of the consumer (Mort, 2000, pp. 276-278; Leiss et. al., 2000, pp. 244-246).

In the beginning of the 21st century, in addition to segmentation based on "niches", the customer profiles were more definable according to the consumer's awareness and expression of lifestyle. Issues including preservation of energy and natural resources, locally produced products, transparent production processes, ethical products and origin of the product have increased in importance in the discussion of a customer whose awareness is affecting her way of shopping (Diamantopoulos et al., 2003; Banyte et al., 2010).

As has been discussed, different drivers have affected customer buying behavior in different eras. Based on the discussion in this chapter, an attempt to visualize key developments and trends in customer segmentation is made in Figure 5. The figure shows key developments in the degree of customization of products, degree of wealthy customers and segment fraction, i.e., the share of people of the whole population belonging to the segment. The figure clearly shows that the

1950s formed a turning point for customer buying behavior, as that is when mass marketing ended and the customer-centric view of marketing began.

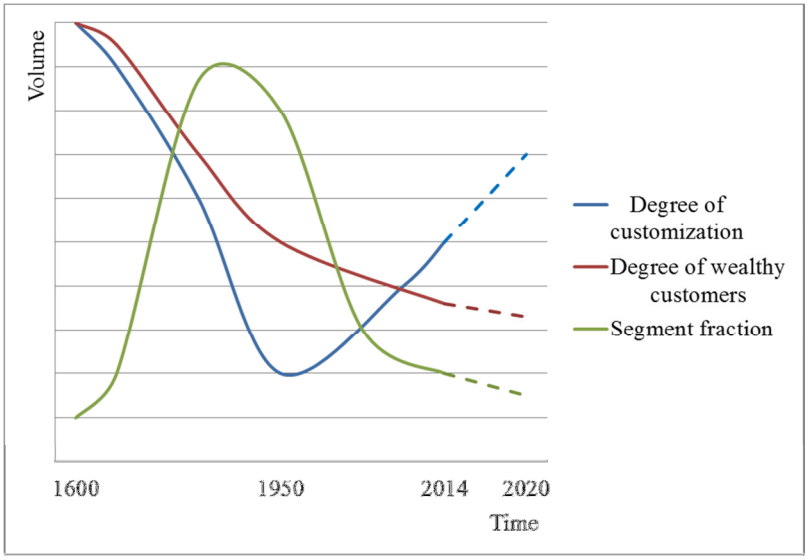


Figure 5. Key features of customer buying behavior and their changes throughout time.

As described earlier in this chapter, and visualized with red in Figure 5, in the beginning only nobles and wealthy people were able to purchase products. With mass marketing products became more affordable, and with the department stores, luxury products became available even for the middle class. In addition, in the western world today, the divide between wealthy and poor customers is smaller than in the 16th century.

The second key feature, visualized in blue in Figure 5, describes the degree of customization of products. Before the 16th century, all products sold were custom made by hand for the customer. Advances in technology and the industrial revolution led to mass production of products, which decreased the degree of customization of the products. Today, the trend is to sell the product as a package that includes service around the product and manufacturing of the custom-made product according to the customer's specifications.

Segmentation became a popular method within marketing in the middle of the 20th century. Before this, customers were treated as one mass. The first group of people that could have been considered to form a segment was the nobles, as they were the only ones with purchasing power. The first actual segmentation was made on geographical bases to make the distribution of products as effective as possible. The number of segments was small and the proportion of people from the whole population that belonged to each segment was small, as is

visualized in green in Figure 5. When products became available to everybody during the mass production era, the segments grew in size as a larger share of the population belonged to only a few big segments. In the late 20th century, when retailers were introduced to CRM, they began to use customer segmentation as a method for getting to know their customer base. Demographical information on the customers became an important segmentation base, followed by behavioral and psychographic information. The size of the segments became smaller containing a smaller fraction of the population, as there was more information on customers and the grouping could be made on a more detailed level. This resulted in many smaller segments. Today, the sizes and fractions of the segments have become even smaller as techniques and information are readily available for grouping customers into smaller and more focused segments, e.g., for describing niche-markets. The requirements for segmentation today are to deliver more usable precise data as an aid for:

- *Product development*, customer responses regarding new products.
- *Strategic planning*, information on market share and customer base as an aid for decision making and planning of strategy. Information on prospects and how to develop these into new customers.
- *Service development*, as an aid to understand the requirements of the customer.

2.10. Summary

This chapter is summarized in Table 1, presenting the evolution of customer buying behavior over time. The inspection of customer buying behavior started from the beginning of the 16th century, when mainly wealthy men did the shopping. Only necessities were purchased, things that brought wealth to the whole family and generations to come. During the reign of Queen Elisabeth I, nobles competed for the Queen's attention in court, and spent money on novelties in order to be noticed. During this time, fashion was created. The driver for shopping was politics.

In the 18th century, women from a variety of social classes did the shopping. The first advertising and marketing campaigns were launched. The industrial revolution contributed to the beginning of mass-consumption. The driver for shopping was economics.

During the 19th century, the first marketing study of consumers was made based on statistics. Mainly middle and working class women were shopping in the newly created department stores. The driver for shopping was romanticism with the aim of self-expression through consumption.

From the 20th century onward, advertising developed at a rapid pace. Customer profiling was performed through classifying of customers according to information received from marketing surveys. Advertising in newspapers and magazines grew in popularity. Mass production made it possible for even working class people to afford expensive necessities, e.g., a T-Ford.

During the late 20th century and beginning of the 21st century, advertising had spread to several new media in addition to newspapers and magazines, i.e., radio, television and internet. Advertisement had changed from containing symbols, using celebrities, more targeted marketing campaigns to "niches". With the increase in capacity for storing data, the methods used for profiling of customers had advanced from simple statistics to highly complicated methods and techniques run by efficient computers. The production of products had changed from mass production of only a few product models per product category, to customized products according to the purchaser's wants and needs. Service around the product gained importance.

Customer segmentation methods have been used since the introduction of CRM for gaining an understanding of the customer base. For the segmentation demographical, behavioral and psychographic information on the customers was used. With the increasing amount of customer information and more advanced segmentation methods that were able to group customers on a more detailed level, the size of the segments became smaller. Today, even niche-markets are described using segmentation. The requirements for segmentation today are to deliver more usable precise data as an aid for decision making and, e.g., product development, service development and strategic planning.

Table 1. Summary of customer buying behavior.

Time	Driver	Information	Consumer	Methods	Products
Before 16th century	Need	Word of mouth	Wealthy men	Reputation	Necessities
Late 16th century	Politics, status	Competition in Queen Elisabeth I court	Nobles, men	Reputation	Novels, show off products
18th century	Economics, status	Advertising, marketing	Upper social class women	Marketing	Novels, fashionable products
Late 18th-19th centuries	Romanticism, self expression	Marketing, studies on consumer behavior, Department stores	Middle and working class women	Statistical studies	Novels
Beginning of 20th century	Fashion as an expression of status	Written product oriented advertising in newspapers and magazines, profiling consumers	All classes	Grouping, classifying, marketing surveys	Luxury goods, unpractical and unnecessary things, mass production
1925-1954	Dreams, emotional	Radio, advertising focusing on symbols	All classes	Research on media audiences, gained limited information	Wanted qualities and status, mass production
1945-1965	Social status	TV, personalization, celebrities used for advertising	All classes	Knowledge of consumer collected. Mass marketing	Mass production
1965-85	Social status	Market segmentation, targeted marketing, computers	All lifestyle segments	Segmentation, marketing management, advanced statistical methods (e.g. Factor analysis, variance analysis) membership cards, technical methods.	No more mass production
Late 20th century	Image of lifestyle	ERP, CRM, changing consumer profiles, new market segments, profiling niches	All niches segments	Insufficient data, shortcomings of measuring techniques, self-reporting methods (polls), demographical data, psychographical data, loyalty-card programs.	No more mass production
21st century	Awareness and expression of that	Consumers have many profiles, marketers need to reach consumers with the right ad at the right time. Internet, social media.	All awareness segments	Profiling, MBA, loyalty card customer data, transaction data. Technical methods: ANNs, algorithms, software, broad loyalty card programs, visualization.	More and more customized products

This chapter has provided an overview of the development of customer buying behavior over time. In the following chapter CRM is presented with a focus on segmentation.

Chapter 3

Customer Relationship Management – CRM

In this chapter, CRM is presented in more detail starting with a definition, the development of CRM and a description of how CRM is done in practice. The most important methods used within CRM are presented, with a focus on segmentation.

3.1. Definition

There is no single definition of CRM. Different authors have stressed the concept differently. For example, CRM can be seen as a business strategy that focuses on customers. An understanding of CRM can give a leader of a company an advantage in the competition for valuable and demanding customers in a market where growth has stagnated (Grönroos 2002, p. 19; Heinrich 2005, p. 710; Datta 1996, p. 797). On the other hand, many authors have emphasized technology as the medium through which the customer is interacted with. This includes an integration of sales, marketing and customer care service in order to get one complete view of the customer (Chalmeta 2006, p. 1015; Grönroos 2002, p. 33; Parvatiyar & Sheth 2001, pp. 3-4; Dibb 2001, p. 194).

Though these two views on CRM may at first glance appear different, they both have in common the use of customer information to support better decision making. The goal of this customer-centric strategy is to create and add value both for the company and the customers (Chalmeta 2006, p. 1015). It has to be ensured that "…the right product is offered with the right service, via the right channel, using the right communication, to the right client at the right moment." (Paas & Kuijlen 2001, p. 55; Rigby & Ledingham 2004, p. 120; Soper 2002, p. 67).

Richards and Jones (2008) define seven value drivers as a reason for implementing CRM into all business processes. These value drivers are: "1) improved ability to target profitable customers; 2) integrated offerings across

channels; 3) improved sales force efficiency and effectiveness; 4) individualized marketing messages; 5) customized products and services; 6) improved customer service efficiency and effectiveness; and 7) improved pricing" (Richards and Jones 2008, p.123).

3.2. Development of CRM

Earlier, as most shops and businesses were small, every shop keeper was able to remember their customers either by knowing every customer personally, or by keeping a book of the customers. In the middle of the 19th century shops grew bigger and less geographically bounded. The industrial revolution, as discussed in Chapter 2, made it possible for new products to be invented and manufactured more cost effectively through mass production. These products were sold in department stores and larger chains of stores. When the market got saturated, i.e., there were more available products than buyers, selling products at a profit became extremely competitive. There was no point in producing even more products and selling them at zero profit. Therefore, mass production was no longer a way for cost reduction. New methods were needed.

Advances in technology in the latter part of the 20th century made it possible to add flexibility into standard processes (Paas & Kuijlen 2001, p. 53; Ehret 2004, p. 466). Within the product-oriented view, the production process was seen as design-build-sell. Now, this thinking was changing to a more customer-centric view, where the production process could be described as sell-build-redesign (Rygielski et al. 2002, p. 484). As it became apparent in the 1980s that buyer-seller relationships could create value, the management became interested in CRM (Ehret 2004, p. 466).

In 1993 Tom Siebel from Siebel Systems Inc. coined the concept of value maximization (Buttle 2004, p. 4; Boulding et al. 2005, p. 155). This new view, implying customer-centered marketing, meant that relationships with the customer became the new central issue. The companies studied their customers in order to reveal their wants and needs. Then, the products, services and communication channels were developed according to the customers' needs, in order to keep them satisfied (Paas & Kuijlen 2001, p. 53; Rygielski et al. 2002, p. 484).

Technology-enhanced CRM dates back to the 1970s, e.g., to call centers, sales force automation systems and customer information files. These technologies were connected with each other in the late 1980s. The CRM IT market was created in the early 1990s and later on an increasing amount of web technologies have been taken into use (Buttle 2004, pp. 59-62, 96).

Between the years 1990 to 2000, millions of dollars in the US alone have been invested in CRM implementation programs (Rigby & Ledingham 2004, p. 118; Paas & Kuijlen 2001, p. 59). 70% of these have failed, creating an economic loss for the company and damaging customer relationships (Lindgreen et al. 2006, p. 58; Heinrich 2005, p. 711; Wilson et al. 2007). The reason for projects failing is often the management's misconception about CRM and customer behavior (Verhoef and Langerak 2002). In several cases of these studies above, management underestimated the effort needed for implementing a customer-centric view throughout the whole company. In addition, trends in customer behavior are often misinterpreted. Instead of identifying the true reason for customer behavior, management based their decisions on their own interpretations (Verhoef and Langerak 2002). Recent research within CRM covers new techniques within CRM and their usage, e.g., Arroyo et al. (2014) who have generated instant messaging contacts for CRM systems; and Taylor et al. (2015) who have created a portal system for CRM. Other topics recent studies have focused on are customer value and customer experience management (e.g., Neslin et al. 2013), performance measurement within CRM (e.g., Ernst et al. 2011), and social or mobile CRM (e.g., Reinhold and Alt 2012). Nguyen et al. (2012) have in their article reviewed the entire CRM process pointing out successes, advances, pitfalls and futures. As key issues within CRM, they point out fairness and customer trust as well as the use of social media (Nguyen et al. (2012).

3.3. CRM in practice

By conducting effective CRM, a company should gain an understanding of its markets and customers (Rygielski et al. 2002, pp. 491-492; Paas & Kuijlen 2001, p. 53; Parvatiyar & Sheth 2001, p. 5). According to earlier studies, the 20-80 rule applies for profitable customers, i.e., only 20% of the customers contribute to 80% of the profits, while the top 20% of the unprofitable customers produces 80% of the costs. Studies have also shown that it costs less to keep an existing customer than to attract a new one (Paas & Kuijlen 2001, p. 53; Parvatiyar & Sheth 2001, p. 11; Park & Baik 2006, p. 263; Kim et al. 2006, p. 101; Abbot et al. 2001, p. 291; Lingras et al. 2005, p. 231). The main question a CRM team wants to know is: Have we been able to attract, establish, maintain and enhance relationships with the customer through loyalty, satisfaction and retention (Heinrich 2005, p. 710; Parvatiyar & Sheth 2001, pp. 3-4; Grönroos 2002, p. 51).

Implementation of CRM in a company means that business processes are re-engineered to be customer-centric (Buttle 2004, p. ix). There are usually four levels of CRM that are implemented (Buttle 2004, pp. 4-5; Paas & Kuijlen 2001, p. 52; Chalmeta 2006, p. 1021):

1) Collaborative CRM, which includes collaboration between the customer and company.
2) e-CRM, which is e-commerce driven CRM.
3) Operational (transactional) CRM, where the CRM is optimized with suitable automated information collecting business processes, e.g., marketing, sales and after-sales.
4) Analytical CRM, which includes the application of analytical tools to data.

Several different channels can be used at the customer interface including sales and service personnel, call centers, Internet websites, marketing departments, fulfilment houses as well as market and business development agents (Parvatiyar & Sheth 2001, p. 18).

The research conducted within this thesis falls into the category of analytical CRM, which aims to create value by analyzing customer data. As the amount of collected customer data is large and grows rapidly, data mining is often used (Buttle 2004, pp. 9-11; Chalmeta 2006, p. 1021).

Another way of looking at CRM, is to use the value creation process. This is called the CRM value chain. It is described through internal processes and functions and external networks and has five stages (Buttle 2004, pp. xi, 40-41):

1) *Customer intimacy* captures the information and knowledge for, from and about the customer concerning products, markets and suppliers (Buttle 2004, pp. xi, 40-41; Salomann et al. 2005, pp. 393-394).
2) *Network development*, where the relationship between customers and network are managed.
3) *Value proposition development*, where the means for satisfying the customer are created by identifying the source of value (Buttle 2004, pp. xi, 40-41).
4) *Managing the customer lifecycle* follows the different stages in the relationship between the customer and the company. In general, this lifecycle can be described in four stages: *Prospect*, i.e., not yet customers in the target market; *Responders*, i.e., aspiring customers; *Active* (or current) *customers*; and *Former customers*, i.e., those who have churned or become unprofitable customers (Rygielski et al. 2002, p. 493; Heinrich 2005, p. 711; Buttle 2004, p. 18).
5) *Customer Portfolio Analysis (CPA)*, where the different categories of customers are identified (Buttle 2004, pp. 9-11; Chalmeta 2006, p. 1021).

The research conducted in this thesis also falls within customer portfolio analysis (CPA), as it aims to identify different categories of customers. The

objective of one of the research papers is to build and evaluate a model for CPA using visual data mining methods.

CRM in practice can be described as four steps (Rygielski et al. 2002, pp. 483, 492-493; Buttle 2004, pp. 93, 99-101; Grönroos 2002, p. 43; Paas & Kuijlen 2001, p. 57):

1. Collecting information about the customers. The data can be collected from internal customer data, from the internet or purchased from external sources. The customer information collected can also be from blogs, discussion pages and from social media through text mining using netnography or other similar methods.
2. Creating a place, e.g., an enterprise data warehouse, where the collected data can easily be stored and accessed.
3. Analyzing of data with a suitable method and tool, e.g., statistical tools, OLAP or data mining. For large amounts of data, data and text mining methods and tools are used for extracting hidden information from large databases. CRM methods are discussed in more detail in the next section, i.e., Section 3.3.1.
4. Execution and tracking of, e.g., a marketing campaign with suitable software.

In this thesis, the knowledge discovery process described in Chapter 4 is used for creation of a model and analysis of data. The four steps of the CRM process listed above are indirectly incorporated in the KDD process.

3.3.1. CRM methods

The key issue within CRM is to collect data on customers in order to understand them better as was described in Section 3.1. As several customers have similar wants and needs, it is efficient to group customers according to similarities. Different segmentation methods have, therefore, been the most used methods. Examples of methods for segmentation are:

- Customer segmentation, i.e., grouping of customers according to their attributes. Customer segmentation will be presented in more detail in the next section, i.e., Section 3.4.
- Decision trees, i.e., dividing large data collections into smaller sets. Decision trees will be presented in more detail in Section 5.2.

Another approach is to use the RFM approach (Kim 2006, p. 542; Lingras et al. 2005, p. 233), where information from transaction data is used for calculation of different values for a customer:

- Recency, time passed since the last purchase
- Frequency, how often the customer makes purchases
- Monetary, the value of the purchases that one customer has made.

Other possible methods used for grouping of customers are (Chalmeta 2006, p. 1015; Paas and Kuijlen 2001, p. 53):

- Share of wallet, a survey method that indicates the percentage (share) of a customer's expenses (wallet) that goes to the business within a given time.
- Customer value, the difference between what the customer will benefit from your product, and what the customer has to give in order to get it.
- Potential lifetime value, a prediction of a net profit of the whole future relationship with the customer.
- Loyalty, customers continue their relationship with the business.

In this thesis, segmentation is used as a method for customer profiling through CPA, i.e., as described in the last section, for grouping customers according to their buying behavior in order to identify different categories of customers. The RFM approach is used to calculate more informative variables used in the segmentation. Segmentation is introduced in the next section, i.e., Section 3.4, whereas data mining methods will be presented in more detail in Chapter 5.

3.4. Customer segmentation

Customer segmentation based on customer value is the most used technique within CRM (Chalmeta 2006, p. 1015; Buttle 2004, pp. 99-101). As presented in Chapter 2, mass marketing ended when Wendell R. Smith introduced market segmentation in 1956 (Wedel & Kamakura 1999, p. XIX, 3; Wilkie & Cohen 1977, p. 1). Smith found differences, i.e., heterogeneity, between customer demands, wants and needs, and grouped similar customers to a segment. For the customers in each segment, products were produced and sold according to these customer needs. Through this market-oriented way of thinking involving identifying and fulfilling the needs of customers in a segment, a company could gain a competitive advantage.

In this thesis, customer segmentation is discussed from three perspectives: 1) segmentation bases, with an example of the difficulty of choosing segmentation bases demonstrated with a study on green consumers; 2) methods for customer segmentation; and 3) evaluation of created customer segments.

3.4.1. Segmentation bases

The bases for segmentation are chosen according to the aim of the segmentation. Earlier, as described in Chapter 2, because of difficulties with transportation of goods, market segmentation was based on geographical areas. Later, market segmentation was based on social classes as these differed severely in consumer behavior (Frank et al. 1972, pp. 4-5). As information is now readily available, segmentation is done based on a multitude of different criteria.

In general, the bases for customer segmentation are divided into two categories: 1) *general customer characteristics*, e.g., demographical, socio-economical and lifestyle measures; and 2) *product-specific customer characteristics*, e.g., customer brand attitudes, brand preferences, benefits sought, product usage and response sensitivity to different marketing campaigns (Dibb & Simkin 1996, p. 14; Wedel & Kamakura 1999, p. 7; Dickson 1982, p. 60; Frank et al. 1972, pp. 26-27, 42, 66-89, 90-111; Badgett & Stone 2005, p. 107; Tsai & Chiu 2004, pp. 265-265). Others, e.g., Kotler (2002), have divided the same information into four segmentation bases: *geographical, demographical, psychographic* and *behavioral* bases. These four bases for segmentation of consumer markets are presented in Table 2.

Table 2. Major segmentation variables for consumer markets (redrawn from Kotler 2002, p. 149)

Geographical	Region, city, density, climate
Demographical	Age, family size, family life cycle, gender, income, occupation, education, religion, race, generation, nationality, social economic ranking, life stage
Psychographic	Lifestyle, personality
Behavioral	Buying pattern, usage data, loyalty status, channel, attitude towards product, profitability

As information about customers is readily available through different channels, it is difficult to know which segmentation bases to use and which variables to choose. To illustrate difficulties in choosing between segmentation bases and variables, a summary of studies on green consumer profiling is presented in the next section.

3.4.2. Green customer segmentation

Green consumer behavior is a phenomenon and potentially exploitable niche market that has been of interest for companies for a number of years. The identification of the profile of an eco-conscious consumer, as well as information concerning the future development of the trend, has been the subject of several research studies since the 1970s, with a peak in the 1990s. At that time, as described in Section 2.6, segmentation was used within marketing with the aim to group customers according to their buying behavior. However, these studies have been challenged by a number of factors, including insufficient data, shortcomings of the measurement techniques, and differences in green consumer behavior in different countries and over time. Defining the terms of eco-product and green consumer has in itself proven to be problematic.

In many cases, a green consumer is defined as a consumer that takes into account the environment in her purchasing behavior or in processes after the purchase, e.g., buys green products, chooses products and packages made from renewable or recycled materials, or conducts recycling.

A literature review including more than 70 articles regarding green customer profiling was made, with a focus on finding out which variables were used for the profiling and what the achieved profile of a green customer was. Keywords used for the literature review were:

- Ecological products, environmentally friendly products, ecofriendly products
- Sustainable, low carbon living
- Energy saving devices
- Green products, organic products
- Green buying, Green buying behavior
- Green consumer, organic consumer.

Traditionally, demographic variables are used as the basis for market segmentations (Wedel and Kamakura 2000). However, the results concerning the usefulness of demographic data for profiling of green consumers are contradictory (see, e.g., Roberts 1996; Mainieri et al. 1997; Straughan and Roberts 1999; Diamantopoulos et al. 2003). In order to better profile green customers, psychographic and behavioral variables are used in addition to demographical variables in a number of studies (Roberts 1996; Bui 2005; Banyte et al. 2010).

Based on the literature review, we have compiled three tables. Table 3 includes studies that rely upon interviews, questionnaires, polls, and queries as data

collection instruments, in order to draw inferences on green customers. It is split into two parts, i.e., studies that use demographical variables and studies that use psychographic and behavioral variables. The table illustrates the contradictory results of previous studies in the use of demographical variables for the analysis of green consumers. It also illustrates that psychographic and behavioral variables are useful.

Table 3. Studies for and against the use of demographical, psychographic and behavioral variables for determining the profile of a green consumer.

Demographical	Useful	Davies et al. 1995; Wandel and Bugge 1997; Thompson and Kidwell 1998; Wedel and Kamakura 2000; Magnusson et al. 2001; Chinnici et al. 2002; Banyte et al. 2010
	Not useful	Frank et al. 1972; McCann 1974; Herberger 1975; Samdahl and Robertson 1989; Banerjee and McKeage 1994; Scott and Willits 1994; Stern et al. 1995; Roberts 1996; Mainieri et al. 1997; Straughan and Roberts 1999; Tsakiridou et al. 2008
Psychographic	Useful	Straughan and Roberts 1999; Chinnici et al. 2002; Banyte et al. 2010
	Not useful	
Behavioral	Useful	Chinnici et al. 2002; Banyte et al. 2010
	Not useful	

Table 4 summarizes previous research on demographical profiles of green customers and findings concerning the most important variables and the metrics used. The first column indicates studies that found a particular demographical variable (the second column) to be a significant indicator of green buying behavior. The third column further specifies what range of the variable was found to partly define the profile of a green consumer, while the fourth column specifies the studies that reached this conclusion. In general, a larger number of references in column four can be taken to indicate that the category of the variable is important. Table 4 shows that age, gender, income, education, and the existence of children in the household are the most important demographical variables that influence green customer's buying behavior (see, e.g., Chinnici et al. 2002; Padel and Foster 2005; Tsakiridou et al. 2008; Banyte et al. 2010). The main psychographic variables used in a number of studies comprise activities, lifestyle, characteristics of personality, and intention of purchasing green products. The main behavioral variables include attitude, knowledge, motives (e.g., concerned with environmental protection, health fanatics or animal lovers)

and benefit of purchasing green products (Roberts 1996; Bui 2005; Banyte et al. 2010).

Table 4. Demographical variables that influence green customer's buying behavior, as reported in previous studies.

Importance	Variable	Categories	References
Anderson and Cunningham 1972; Van Liere and Dunlap 1981; Samdahl and Robertson 1989; Roberts 1996; Bui 2005; Banyte et al. 2010	**Age**	Younger 18-30	Anderson and Cunningham 1972; Weigel 1977; Jolly 1991; Roberts 1996; Roberts and Bacon 1997; Chinnici et al. 2002; Banyte et al. 2010
		Middle aged 31-50	Roberts 1996; Chinnici et al. 2002; Bui 2005; Banyte et al. 2010
		Older 51->	Van Liere and Dunlap 1981; Samdahl and Robertson 1989; Roberts 1996; Wandel and Bugge 1997; Thompson and Kidwell 1998; Fotopoulos and Krystallis 2002; Tsakiridou et al. 2008
Van Liere and Dunlap 1981; Davies et al. 1995; Roberts 1996; Banyte et al. 2010	**Gender**	Male	
		Female	Banerjee and McKeage 1994; Davies et al. 1995; Roberts 1996; Mainieri et al. 1997; Wandel and Bugge 1997; Laroche et al. 2001; Chinnici et al. 2002; Bui 2005; Lea and Worsley 2005; Padel and Foster 2005; Banyte et al. 2010
Van Liere and Dunlap 1981; Newell and Green 1997; Banyte et al. 2010	**Income**	Low	Samdahl and Robertson 1989; Fotopoulos and Krystallis 2002
		Middle	Chinnici et al. 2002; Bui 2005
		High	Herberger 1975; Davies et al. 1995; Magnusson et al. 2001; Bui 2005; Padel and Foster 2005; Tsakiridou et al. 2008; Banyte et al. 2010
Van Liere and Dunlap 1989; Schwarts and Miller 1991; Roberts 1996; Bui 2005; Banyte et al. 2010	**Education**	Low	Samdahl and Robertson 1989
		Middle	Chinnici et al. 2002
		High	Herberger 1975; Arbuthnot 1977; Jolly 1991; Schwartz and Miller 1991; Newell and Green 1997; Wandel and Bugge 1997; Magnusson et al. 2001; Hill and Lynchehaum 2002; Bui 2005; Padel and Foster 2005; Banyte et al. 2010; Ishaswini et al. 2011
	Children in household	Yes	Davies et al. 1995; Thompson and Kidwell 1998; Laroche et al. 2001; Fotopoulos and Krystallis 2002
		No	

Table 5 summarizes previous research on psychographic and behavioral profiles of green customers. The table is interpreted in a similar manner as Table 4, but instead of demographical variables, psychographic and behavioral variables are presented in column two. Table 5 shows that the main psychographic and behavioral variables that influence green customer behavior are environmental concern, intention, price, knowledge of the products, perceived consumer effectiveness, and healthy lifestyle (see, e.g., Davies et al. 1995; Roberts 1996; Kalafatis et al. 1999; Follows and Jobber 2000; Bui 2005; Tsakiridou et al. 2008; Ishaswini et al. 2011; Jensen et al. 2011).

Table 5. Psychographic and behavioral variables that influence green customer's buying behavior, as reported in previous studies.

Importance	Variable	Categories	References
Davies et al. 1995; Kalafatis et al. 1999; Follows and Jobber 2000; Bui 2005; Tsakiridou et al. 2008; Ishaswini et al. 2011; Jensen et al. 2011	**Environmental concern**	**yes** (23 references)	Kinnear et al. 1974; Herberger 1975; Van Liere and Dunlap 1981; Antil 1984; Hines et al. 1987; Gerstman and Mayers Inc. 1989; Simmons and Widmar 1990; Mandese 1991; Roberts 1991; Shetzer et al. 1991; Shabecoff 1993; Tregear et al. 1994; Davies et al. 1995; Roberts 1995; Roberts and Bacon 1997; Schifferstein and Oude Ophuis 1998; Squires et al. 2001; Padel and Foster 2005; Bui 2005; Tsakiridou et al. 2008; Banyte et al. 2010; Ishaswini et al. 2011; Jensen et al. 2011
		no (5 ref.)	Gill et al. 1986; Kleiner 1991; Schwepker and Cornwell 1991; Schlossberg 1991; Winski 1991
Davies et al. 1995; Chinnici et al. 2002; Tsakiridou et al. 2008; Jensen et al. 2011	**Healthy lifestyle**	**yes** (12 ref.)	Tregear et al. 1994; Davies et al. 1995; Huang 1996; Schlegelmilch et al. 1996; Wandel and Bugge 1997; Magnusson et al. 2001; Squires et al. 2001; Chinnici et al. 2002; Padel and Foster 2005; Tsakiridou et al. 2008; Banyte et al. 2010; Jensen et al. 2011
		no	
Roberts 1996; Bui 2005	**Perceived consumer effectiveness, PCE**	**yes** (11 ref.)	Kinnear et al. 1974; Herberger 1975; Webster 1975; Antil 1979; Wiener and Doescher 1991; Berger and Corbin 1992; Roberts 1995; Roberts 1996; Roberts and Bacon 1997; Straughan and Roberts 1999; Bui 2005
		no	
Davies et al. 1995; Tsakiridou et al. 2008	**Limited availability**	**yes** (7 ref.)	Jolly 1991; Tregear et al. 1994; Davies et al. 1995; Roddy et al. 1996; Wandel and Bugge 1997; Lea and Worsley 2005; Jensen et al. 2011
		no	
Davies et al. 1995; Tsakiridou et al. 2008	**Higher price**	**yes** (6 ref.)	Jolly 1991; Tregear et al. 1994; Davies et al. 1995; Roddy et al. 1996; Vindigni et al. 2002; Padel and Foster 2005
		no	
	Safety, free from chemicals	**yes** (7 ref.)	Childs and Polyzees 1997; Zotos et al. 1999; Baltas 2001; Laroche et al. 2001; Squires et al. 2001; Fotopoulos and Krystallis 2002
		no	
Rao 1974; Herberger 1975; GfK 2011; Ishaswini et al. 2011; Jensen et al. 2011	**Knowledge**	**yes** (6 ref.)	Herberger 1975; Bui 2005; Padel and Foster 2005; Banyte et al. 2010; Ishaswini et al. 2011; Jensen et al. 2011
		no	
Follows and Jobber 2000; Bui 2005	**Attitudes, environmental concern**	**yes** (4 ref.)	Roberts 1996; Kilbourne and Beckmann 1998; Diamantopoulos et al. 2003; Bui 2005
		no (5 ref.)	Wicker 1969; Rotschild 1979; Gill et al. 1986; Lee and Green 1991; Tsakiridou et al. 2008
Davies et al. 1995	**Nutritious taste**	**yes** (2 ref.)	Squires et al. 2001; Chinnici et al. 2002
		no	
	Quality	**yes** (4 ref.)	Laroche et al. 2001; Vindigni et al. 2002; Krystallis and Chryssohoidis 2005; Jensen et al. 2011
		no	
Stern et al. 1993	**Altruism**	**yes** (4 ref.)	Stern et al. 1993; Straughan and Roberts 1999; Banyte et al. 2010; Griskevicius 2010
		no	
	Politically and socially active	**yes** (3 ref.)	Hine and Gifford 1991; Roberts 1996; Banyte et al. 2010
		no	
	Motives, e.g. Concern for environment	**yes** (3 ref.)	Padel and Foster 2005; Tsakiridou et al. 2008; Banyte et al. 2010
		no	
Follows and Jobber 2000	**Intention**	**yes**	
		no (3 ref.)	Wicker 1969; Laroche et al. 1996; Nakarado 1996
	Habit of buying green products	**yes** (2 ref.)	Magnusson et al. 2001; Tsakiridou et al. 2008
		no	
	Social responsibility	**yes** (1 ref.)	Herberger 1975
		no	
	Curiosity	**yes** (1 ref.)	Chinnici et al. 2002
		no	

The literature review shows that there is a large body of research into what drives green consumer behavior. There exists a lack of unanimity on a single profile of a green consumer. The poor results concerning the usefulness of demographic variables for profiling may be explained by the existence of multiple green consumer profiles. A conclusion of this, also supported in the literature, is that demographic data alone are insufficient for discriminating between green and non-green consumers. Green consumers adhere to a green purchasing behavior to a certain degree, as they purchase both environmentally friendly products and nongreen products. Therefore, profiles for different degrees of green purchasing behavior are of interest. The findings presented in Tables 3, 4 and 5 provide, therefore, a starting point for identifying how multiple green profiles may be characterized.

Most studies seeking to identify the profile of a green consumer have primarily relied upon self-reporting methods, such as questionnaires, polls, and queries. Typically, these studies have constructed the profile based upon demographic and psychographic data provided by the respondents. Several studies conclude that it is not possible to identify the profile of a green consumer based solely on demographic or psychographic data and instead base the profile on a combination of the two types of data. However, questionnaire data can be problematic for this purpose. While consumers are assumed to answer the questions according to the best of their knowledge, there is a tendency to answer according to their intentions or expected social norms, which might not reflect their actual behavior (Mainieri et al. 1997; Young et al. 2010). Instead of using questionnaire data, a preferable way to analyze green consumer behavior would be to use BI processes and methods and perform a data-driven assessment based on actual transaction data.

3.4.3. Segmentation methods

Market segmentation methods are mainly divided into four categories that are connected to each other as presented in Table 6.

Table 6. Grouping of market segmentation methods. Table redrawn from Wedel & Kamakura (1999, p. 17).

	A priori	**Post hoc**
Descriptive	Contingency tables Log-linear models	Clustering methods, e.g., k-means, unsupervised ANNs (SOM)
Predictive	Discriminant analysis, Logistic regression RFM	Decision tree algorithms, e.g., CART, CHAID Supervised ANNs

Table 6 is based on the division made by Wedel and Kamakura (2000, pp. 17, 40):

> 1) *A priori*, where the researcher has determined the type and number of segments in advance, e.g., into two types: profitable and non-profitable customers. In the predictive case, the researcher knows beforehand according to which function the predictions are made;
> 2) *Post-hoc* methods, where the outcome of the analysis determines the type and number of the segments, i.e., a data-driven method. The researcher does not know beforehand which function gives the best outcome;
> 3) *Descriptive* statistical methods, where no distinction is made between dependent and independent variables. The outcome of the methods is used to describe the customer, but nothing can be said about future customers; and
> 4) *Predictive* methods, where the association between two sets of variables are analyzed. Based on the built model, future cases can be classified according to previous results.

Methods used in the *A priori – Descriptive* category are Contingency tables and Log-linear models. Contingency tables and log-linear models are used for gaining summary information of categorical data. These methods are mainly used in database marketing (Verhoef et al. 2003).

In the *Post hoc – Descriptive* category, clustering methods are used to group customers into segments (Wedel & Kamakura 1999, pp. 17, 40). K-means and SOM are examples of clustering methods. K-means is considered an efficient and widely used clustering algorithm suitable for a large number of cases. It can deal with several attribute variables. The main weakness with k-means is that the number of segments must be determined beforehand and it is sensitive to outliers (Brimicombe 2007, p. 7; Lingras et al. 2005, p. 217; Ripley 1997, p. 113). Another widely used clustering algorithm is the unsupervised ANN called the Self-Organizing Map (SOM). The biggest difference between the SOM and k-means algorithm is their capacity in handling missing variables and outliers, when the number of segments is decided in the segmentation process and the presentation of the segmentation results. The Post hoc - Descriptive category is the main category followed in this thesis. A more detailed discussion on the requirements of a suitable clustering method is presented in Section 4.3. The clustering methods relevant for this thesis will be presented in Chapter 5.

A common thing for the methods used in the *A priori – Predictive* category is that the segments are first established and then an appropriate method is used for grouping of the independent variables (Wedel & Kamakura 1999, pp. 17, 40). Methods used within this category are Discriminant analysis, logistic regression

and RFM analysis. RFM analysis, which is used within this thesis, is mainly used in analyzes for direct marketing (Miglautsch 2000). It groups customers based on earlier purchasing behavior related to three variables: *recency*, i.e., time since last purchase; *frequency*, i.e., how many times purchases are made during a certain time period; and *monetary* value, i.e., money spent on purchases.

The last of the four categories is the *Post hoc- Predictive* category, where the methods are used for studying the relationship between the dependent and independent variables. Widely used methods are various Decision tree algorithms, because they are simple, transparent and are well suited for prediction (Rygielski et al., 2002). Another widely used class of methods are the supervised ANNs.

3.4.4. Evaluation of customer segmentation

The quality of the outcome, i.e., the segments, of customer segmentation can be evaluated using seven criteria (Wedel & Kamakura 1999, pp. 4-5; Dibb 2001, p. 196; Dibb & Simkin 1996, p. 15; Wilkie & Cohen 1977, p. 3):

- *Identifiability,* the managers should be able to identify those market segments that their customers belong to.
- *Substantiality,* in order for the marketing strategy to be profitable, the chosen segments should constitute a large enough share of the market.
- *Accessibility,* access to the customers of the targeted segments is a key issue for a successful marketing campaign.
- *Stability,* the chosen segments should be stable and not change during a marketing campaign.
- *Responsiveness,* is a measure of how well the customers in the targeted segments respond to a marketing campaign. All segments respond uniquely.
- *Actionability,* describes if the chosen segments are meaningful for the company, i.e., consistent with the goals and core competencies of the company.
- A market segment that is valid has homogeneity within and heterogeneity between the segments.

The outcome of a customer segmentation can be validated through a weak-form evaluation, consisting of interviews of experts with knowledge on customers (Waaser et al. 2004, p. 107).

In this thesis, the aim is to model customer segments that fulfill as many criteria as possible for accomplishing high quality outcome.

3.5. Summary

CRM is the key concept in this thesis. The methods used and models built are motivated from the CRM perspective. In this chapter, CRM has been discussed through the definition of CRM, the development of CRM and by describing how CRM is performed in practice. The various methods used for CRM have been discussed with an emphasis on segmentation. An introduction to segmentation bases has been provided, with a case study on an actual niche market, i.e., green consumers. Segmentation methods and evaluation of the outcome of a customer segmentation were presented.

The next chapter will give an introduction to Visual Analytics and the knowledge discovery process and how these two are used within this thesis.

Chapter 4

Visual Analytics and the Knowledge Discovery process

In this chapter we describe the concept of Visual Analytics. Two knowledge discovery processes, namely the Visual Analytics Process and the Knowledge Discovery in Data bases (KDD-process), are described in detail and combined into a hybrid visual knowledge discovery process. The hybrid process and how it has been used in this thesis is presented step by step.

As discussed in Chapter 3, within CRM customer data from different sources of various formats are collected and stored. As has been shown, since the 1990s, different more or less complex methods have been used for analyzing these multi-format customer data. Managers have found it difficult to conduct analyzes as well as interpret and use the outcomes of these (Keim et al. 2008). There is a need for new ways for analyzing and presenting information on the customers.

Visual analytics is, according to Thomas and Cook (2006), the science of analytical reasoning supported by interactive visual interfaces. It is a multidisciplinary field where vast amounts of various data in various formats are analyzed by combining human judgment, visual presentations and different interaction techniques. The information discovery process within visual analytics allows decision makers to participate in the data analysis process. Human intelligence with its capabilities for flexibility, creativity and background knowledge is combined with the enormous storage and processing capacities of computers for knowledge discovery. By interacting directly with the computer through advanced visual interfaces, decision makers are able to make well-informed decisions (Thomas and Cook 2006; Keim et al. 2008).

The Visual Analytics Process presented in Figure 6 is a knowledge discovery process. The first step labelled Data contains collection of data, including

preprocessing and transformation of the collected data. After preprocessing of data, two paths for knowledge discovery can be chosen:

- *Visual Data Exploration*, where visualization methods are used for interaction with data. The decision maker interacts directly with the computer in order to explore the data visually and to extract information useful for decision making.
- *Automated Data Analysis*, where automatic analysis methods imply data mining methods that are used for the creation of models describing the data. The created models are evaluated and tested by interacting with the data with the help of visualization methods.

According to the Visual Analytics Process, knowledge is gained through the interaction between visualization, models and the decision maker.

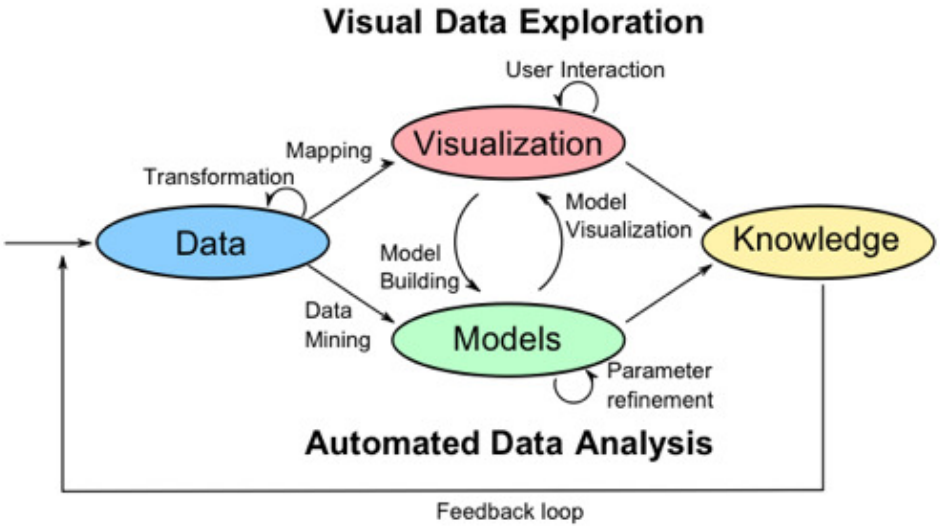


Figure 6. The Visual Analytics Process. Figure redrawn from http://www.visual-analytics.eu/faq/ (retrieved 17.11.2014).

In this thesis, both the Visual Analytics Process and the well-known KDD process have been used for knowledge discovery. The next sections present knowledge discovery starting with the KDD process and continuing with the knowledge discovery process used in this thesis.

4.1. Knowledge Discovery

The Knowledge Discovery in Data bases (KDD), is a data mining process introduced by Fayyad et al. in 1996. The aim is to identify valid, novel,

potentially useful and understandable patterns from large and complex data sets (Fayyad 1996; Maimon & Rokach 2005, pp. 1-2; Berry & Linoff 2004, p. 8, 13).

The KDD process is interactive and it comprises iterative steps with the aim to extract knowledge useful for decision making from large amounts of data. Each step is data driven, also called discovery driven, which means that the results of the mining dictates the outcome, i.e., no assumptions are made beforehand. The user needs to make many decisions while going through the numerous steps of the KDD process (Maimon & Rokach 2005, pp. 2-6; Rushmeier et al. 1997, p. 1-2; Shaw et al. 2001, p. 127; Fayyad et al. 1996). According to Fayyad et al. (1996) data mining is in practice mainly used for prediction and description. The KDD process is presented graphically in Figure 7.

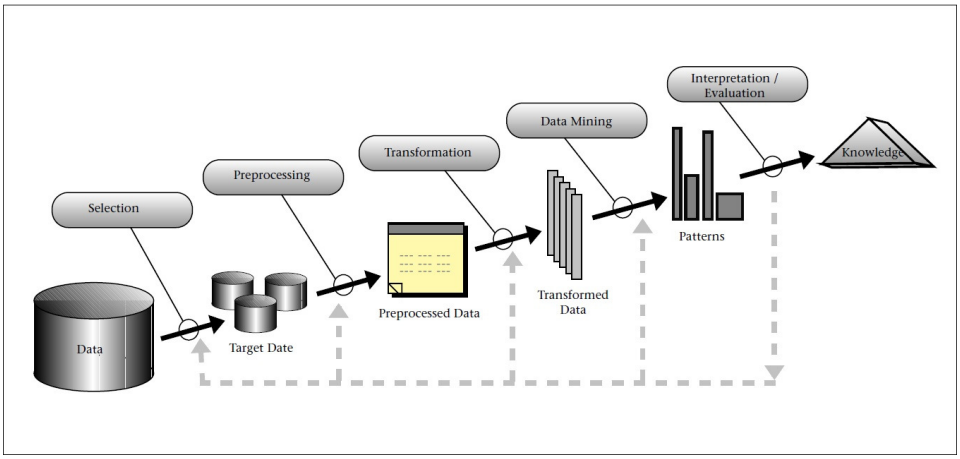


Figure 7. The KDD process, reprinted from Fayyad et al. (1996, p. 41).

As the name suggests, the KDD process starts from a data base or a data warehouse comprising *available data* (1). The goal of knowledge discovery is usually divided into two main aims according to the intended use of the system: *verification* and *discovery*. After deciding on the goal of the KDD process, an appropriate *target data* (2) set with a focus on the goal is selected. The third step is data cleaning and preprocessing, where strategies for missing data, noise and outliers are decided upon. The outcome of the *preprocessing* step is a set of *preprocessed data* (3). In the *transformation* step, data reduction and projection will take place, where useful features are found for representing the data with the goal in mind. After clearing all the data preparation issues, we have *transformed data* (4) ready for modelling with a suitable method (Fayyad et al. 1996).

The fifth step is to choose an appropriate *data mining* (5) method. Typical data mining methods are derived from machine learning, e.g., summarization, classification, regression, and clustering (Fayyad et al. 1996). Some of these

methods have been discussed in Section 3.4.3, while others will be discussed in more detail in Chapter 5. The following step includes the selection of a data mining algorithm and *selecting methods* (6) suitable for pattern searching. The seventh step is the actual *data mining* step (7), where the data is mined in search of interesting patterns. Data mining is the step that most of the work has focused on (Fayyad et al. 1996). The retrieved patterns can only be as good as the data they are based on, i.e., by performing the steps involving data selection, preprocessing and transformation well, better patterns are obtained with data mining. The next step is *interpreting* (8) the mined patterns and then taking into use or *acting on* (9) the discovered knowledge (Fayyad et al. 1996).

The KDD process model has evolved from its traditional form to more agile, making it more adaptive, flexible and human centered (Nascimento and Oliveira 2012). Furthermore, as the capability of storing data is increasing exponentially, also the information retrieved through the KDD process is much more complex. Further developed methods for visualization of multidimensional results to decision makers are needed (Keim et al. 2008).

4.1.1. Knowledge discovery tasks

Tasks performed with the KDD process are commonly called data mining tasks, and are mainly categorized into six groups: *classification, estimation, prediction, affinity grouping, clustering,* and *profiling* (Han et al. 2011, pp. 41-44; Larose 2005, p. 67; Berry and Linoff 2004, p. 8; Rygielski et al. 2002, pp. 488, 491; Hand et al. 2001, pp. 12-15; Bigus 1996, p. 12). A brief introduction to these six tasks is presented below.

1) Classification
One of the most common tasks within data mining is classification, where different objects or cases are ordered into predefined groups. The human brain uses classification in order to understand foreign objects. Within classification, the groups are predefined. The outcome of a classification is true or false. Classification is applied to tasks concerning targeted marketing, quality control and risk assessment. Methods used for classification are similar to the segmentation methods presented in Section 3.4.3, e.g., decision trees, neural networks and nearest neighbor techniques (Linoff and Berry 2011, pp. 85-86; Berry and Linoff 2004, p. 9; Bigus 1996, pp. 12, 38).

2) Estimation
Estimation is similar to classification, but in addition the outcomes are continuously valued. Estimation also allows ranking. Tasks for estimation are, for example, lifetime value of a customer, income or credit card balance. Methods used for estimation tasks are again similar to the segmentation methods

presented in Section 3.4.3, e.g., regression models, neural networks and survival analysis (Berry and Linoff 2004, pp. 9-10).

3) Prediction
Prediction is similar to both estimation and classification, with the exception that the records are used to classify something according to predicted future value or predicted behavior. The correctness of the predictions will be revealed with time. Tasks for prediction are, e.g., customer churn within a certain time interval, or customer lifetime value (Furness 2001, p. 299). Similar methods are used for prediction as for estimation and classification (Berry and Linoff 2004, pp. 10-11).

4) Association rules
Association rules, also called affinity grouping, shows which things are combined together. A typical task for association rules is market basket analysis (MBA), where the items purchased together at one time, i.e., a shopping basket, are analyzed. This task is also used for cross-selling or planning of store shelves, catalogues, and product packages (Berry and Linoff 2004, p. 11).

5) Clustering
Clustering is similar to classification, except that the clusters formed are not predefined. As presented in Table 6 in Section 3.4.3, clustering methods belong to the Post hoc - Descriptive category of market segmentation methods. Clustering is an exploratory method used for data driven grouping of data according to similarities (Berry and Linoff 2004, p. 11; Hsu and Chen 2007, p. 12; Bigus 1996, p. 38; Wedel and Kamakura 1999, p. 39). As clustering is data driven, the user needs to afterwards determine the purpose and definition of the formed clusters. Methods used for clustering tasks are, e.g., k-means and SOM. Data used for clustering are usually numeric or categorical (Hsu and Chen 2007, p. 12; Berry and Linoff 2004, p. 11; Bigus 1996, pp. 38-39; Lingras et al. 2005, p. 217). Clustering methods will be presented in more detail in Chapter 5.

6) Profiling
Within profiling, processes or customers are described based on the available data. By analyzing e.g., the purchasing behavior of a customer, it is possible to understand the customers better. Methods used for profiling tasks are decision trees, association rules and clustering (Berry and Linoff 2004, p. 12; Parr-Rud 2001, p. 184).

4.2. The knowledge discovery process in this thesis

The knowledge discovery process in this thesis follows both the Visual Analytics Process presented in Figure 6 and the KDD process presented in

Figure 7. We present a hybrid solution for knowledge discovery, where the Visual Analytics Process is superimposed on top of the KDD process as presented in Figure 8. The basic structure of the hybrid process was redrawn from the original Visual Analytics Process, i.e., containing the four stages: Data, Model, Visualization, and Knowledge. The nine different steps of the KDD process are divided into the four stages. All four stages are connected to each other and there is a feedback loop from the fourth stage (Knowledge) to the first one (Data).

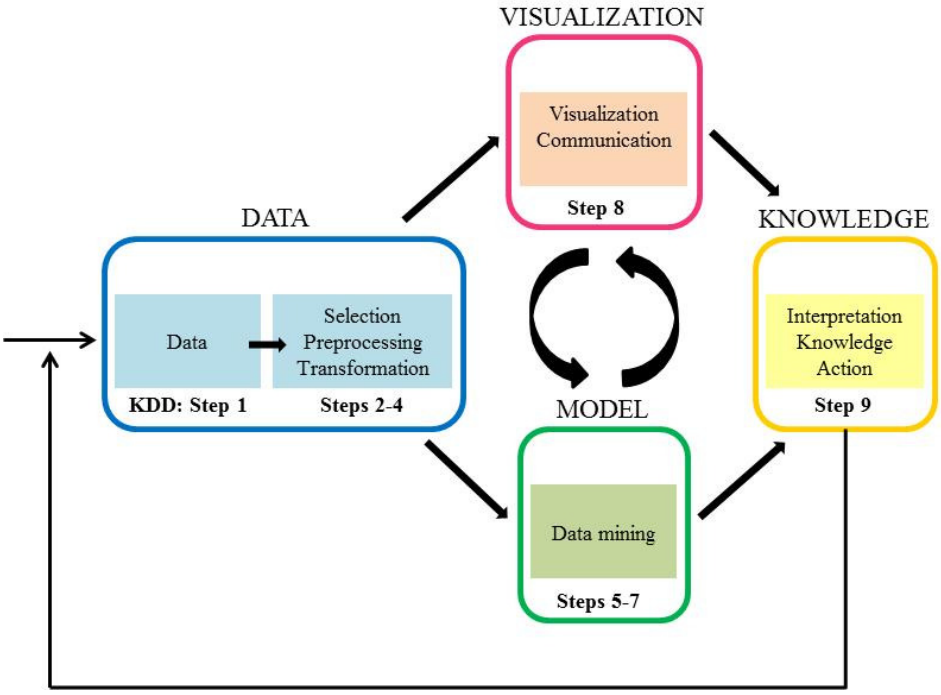


Figure 8. The hybrid visual knowledge discovery process used in this thesis is a combination of the Visual Analytics Process and the KDD process, presented in Figures 6 and 7.

In the following sub sections, the hybrid visual knowledge discovery process combining the Visual Analytics Process and the KDD process is described step by step, as presented in Figure 8, in relation to the research done in this thesis. The models built with the hybrid process are, in addition, described in the articles, i.e., research papers 1-6 and in Chapter 6, while the methods used are described in the next chapter, i.e., Chapter 5.

STAGE 1: DATA

The first stage of the hybrid visual knowledge discovery process comprises steps 1 to 4 involving available data, selection of data, preprocessing and transformation of data.

Step 1: Available data
Two different datasets are used in this thesis for construction of the models. The first dataset was from a company that sells products to other companies (B2B). The products ranged from simple periodicals to advanced consulting services. The second dataset was from a national department store chain that sells products to customers (B2C).

Dataset 1: B2B. In paper 2, Holmbom et al. 2011, data containing information on customers (i.e., businesses) and their purchasing behavior, were selected from the company data base and gathered from the customer's annual reports. The data were collected over a period of five years from 2002 to 2006 and contained information on 1,841 customers. The variables in the data set were *descriptive categories* describing the attributes of the customers, and *product categories*, containing sales information of the major products.

The descriptive categories consisted of:

- Risk factor, which is a measure of potential financial losses.
- Company age (years)
- Solvency (€), which was calculated based on the financial statement.
- Turnover (€)
- Change in turnover (%) compared to the previous year.
- Balance sheet total (€), which is a measure of company size.
- Return on equity (€), calculated based on the financial statement.

The product categories consisted of 18 different products' labeled Products A-R, where product I was the most expensive product and product L was the cheapest one. Product O stood for the overall purchases of products and product R for other products.

Dataset 2: B2C. For the remaining papers, i.e., research papers 1, 3-6, customer data from the data warehouse of a national department store chain were used. The data contained information on demographical, behavioral and product categories. The data were collected over a period of 2 years from 2007 to 2009 and contained over 1.5 million customers, i.e., about 30% of the population of Finland, and over 500,000 products.

Possible segmentation bases were presented in Section 3.4.1. The collected and calculated variables included in this study are presented in Table 7.

Table 7. The demographical, behavioral and product categories of the department store data set, i.e., Dataset 2.

	Variables
Demographical	Age, gender, child decile, estimate of income, customer tenure, mosaic group, mosaic class, service level, loyalty point level, native language
Behavioral – All the behavioral variables were calculated from transaction data	Total spending amount (€), RFM, basket size, average item value (€), total purchased items, average transaction spending (€), working time transaction (%), number of categories, purchasing frequency
Product	Products and Product categories, i.e., the departments of the department store: Leisure, Beauty, Home, Children, Sports, Men, Women and Women's shoes.

A more detailed description of the variables is as follows:

- **Age**, in years
- **Gender**, where 1 = male and 2= female
- **Estimate of income**, where a value 1, 2 or 3 is given, depending on how wealthy a household is. A larger number indicates a wealthier household.
- **Child decile**, is an estimate of the probability of children living in the same household. A value of 1-10 is assigned to a household. A larger value indicates a higher probability of children.
- **Customer tenure**; number of years the customer has been a loyalty card member.
- **Mosaic group**[1]. The whole population of Finland has been classified into 9 groups, A-I, based on a socio-demographics ranking system comprising education, lifestyle, culture and behavior. The map of Finland has been divided into a grid of 250x250 meter bits, where on average seven households per bit are assigned into one of these 9 groups.
- **Mosaic class** divides the nine Mosaic groups into 33 subclasses.

[1] Information from Experian website, visited 16th of June 2014 http://www.experian.fi/Pages/PALVELUT%20JA%20TUOTTEET/MosaicOsanaLiiketoimintaa.htm

- **Service level**, is a measure of how many other service providers from the same chain the customer has used during last year.
- **Loyalty point level**, is a measure on a scale from 1 to 4 of how many points a customer has gathered in the loyalty card program.
- **Native language** – Finnish or Swedish
- **Total spending amount** (in €), is the sum of how much a customer has purchased during the 2 year period.
- **Total purchased items**, is the number of items purchased.
- **Basket size**, is the average number of items bought per visit to the store.
- **Average item value** (in €), is the average value per item purchased.
- **Average transaction spending** (in €), is the average value per transaction.
- **Working time purchase /** Daytime shopping (%), is the share in % of purchases made during working time, i.e., Mondays to Fridays, 9 a.m. to 5 p.m.
- **Number of categories**, is the average total number of product groups per transaction.
- **Purchase frequency**, is the average number of transactions per day.
- **Total of green products** (in €), amount of green products purchased.
- **Total green products** (item), number of green products purchased.
- **Share of green products** (ratio), ration of green to non-green items purchased.

Step 2: Target data

The aim of our knowledge discovery process was to create models of customer behavior. Most data mining algorithms require a single view of the customers, i.e., all customer data comprised to one file. As the collected data were stored in several different source files, a one-dimensional flat table was created by integrating the separate files. Within analytical CRM, this one big file containing all information related to customers is called a *Marketing Customer Information File (MCIF)* (Tsiptsis and Chorianopoulos 2011, p. 148). Only the essential and most useful variables for building a model were included in the target data set. The criteria for included variables were defined specifically for each study. The excess variables were removed.

Step 3: Preprocessed data

The third step was data cleaning and preprocessing, where possible missing data, noise and outliers were handled. The data were preprocessed in order to give better data mining results.

In some of the studies, modifications to the data were made. For the B2B data, i.e., Dataset 1, a number of small and very large customers that appeared as outliers in the results of a pilot test were excluded from the data. The small customers were considered outside of the scope of this study, as they made only a few purchases per year. The large customers were already assigned their own customer service personnel, and therefore, their wants and needs were taken care of. In addition, some product categories with small and infrequent purchases were merged and called product R.

For the B2C data, i.e., Dataset 2, in paper 3 and 6, customers who made purchases of less than 100 Eur or less than two times within the 2 year period, were excluded. In papers 4 and 5, customers who did not buy any green products within the studied period, were excluded from the study.

Step 4: Transformed data
The methods used within the knowledge discovery process contained functions for data preparation and data selection. The methods were able to cope with noisy and missing data. To deal with outlier data, sigmoid (or logistic) transformation was used (Bishop, 1995). Sigmoid transformation emphasizes the center input values while reducing the influence of extreme input values (Bishop, 1995; Larose, 2005). Normalization was further used to scale the variables.

STAGE 2: MODEL
The second stage of the hybrid visual knowledge discovery process contained the steps 5 to 7, concerning building different models for customer segmentation.

Steps 5-7: Data mining
A suitable data mining method for the aim of the study was chosen in each case separately. The steps for choosing a suitable method for data mining and a suitable method for displaying the patterns in the data mining result are presented in research papers 1-5.

In this thesis, the requirements of a suitable visual data mining method for a market segmentation task are summarized as follows:

- The method should perform data-driven clustering, i.e., group customers into segments that are not determined beforehand.
- Be able to handle large amounts of customer data.
- Be able to handle missing values and outliers.
- Accept a multidimensional mix of data types.
- Perform data reduction of large data sets.

- Perform dimension reduction of multidimensional data.
- Present the analysis results visually.

Potentially applicable methods for market segmentation were presented in Section 3.4.3. For representing multidimensional data in an easily understandable format, data and dimension reduction techniques are used. The aim is to embed data into a lower dimension in order to support compression and visualization of data (Sarlin 2013d). Sarlin has in his article (2013d) assessed the suitability of data and dimension reduction methods for visual financial performance analysis. He did the assessment by comparing suitable existing methods with the help of illustrative experiments.

The choice of a suitable method is often depended on the needs for the task at hand. The task Sarlin (2013d) focused on, which is similar to the tasks in most of the articles included in this thesis, was to build a low-dimensional map from high-dimensional large data for visualization purposes. In his work Sarlin concluded that an ideal method for a visual performance analysis task combines data and dimension reduction techniques, preserves the topology of the data and has a pre-defined regular grid shape. The SOM fulfills these criteria.

In addition to presenting the analysis results visually and performing both data and dimension reduction, the SOM is a data driven method, which is able to handle large amounts of customer data of different types, as well as, missing values and outliers. The SOM and other data mining methods used within this thesis are presented in more detail in Chapter 5.

STAGE 3: VISUALIZATION
The third stage of the hybrid visual knowledge discovery process consisted of step 8, comprising the reporting, presentation and visualization of the outcome of the built models.

Step 8: Presentation and visualization of data mining results
The mined patterns were visualized in order to gain information from the knowledge discovery process. The analysis of a SOM map is presented in more detail in Chapter 5. The information was then communicated to the managers in an informative way.

STAGE 4: KNOWLEDGE
The fourth and last stage of the hybrid visual knowledge discovery process consisted of step 9, comprising creation of knowledge from the information extracted from the presented models.

Step 9: Interpretation and acting on the created knowledge
The received information was interpreted into knowledge. The knowledge was used by managers as an aid in decision making, strategy planning, product development and planning of sales campaigns.

4.3. Summary

In this chapter the concept of visual analytics was introduced. Two processes for knowledge discovery were presented, namely the Visual Analytics Process and the KDD process. A hybrid visual knowledge discovery process combining the two processes was presented, where the Visual Analytics Process was superimposed onto the KDD process.

The hybrid visual knowledge discovery process used in this thesis was presented step by step. In one of the steps, the requirements for a suitable market segmentation method were discussed.

In the next chapter, data mining methods used in this thesis are presented.

Chapter 5

Data Mining Methods

In this chapter, the methods used in the research papers 1-6 are presented. First the choice of methods used in the papers is discussed, followed by a presentation of the chosen methods.

5.1. Choice of method

As pointed out in Section 3.4.3 methods used within data driven market segmentation are located in the Post hoc - descriptive category presented in Table 6. In stage 2: Model of the hybrid visual knowledge discovery process presented in Section 4.2 containing the steps 5-7 of the KDD process, the requirements for the segmentation method used in this thesis were discussed. A discussion on the requirements for a method for clustering was made, where the outcome was that SOM is a suitable method for the customer segmentation task in this thesis. The SOM performs both data and dimension reduction, is data driven and able to handle large amount of different types of customer data, missing values and outliers, and has the capability to communicate the analyzing results visually. Therefore, the SOM was chosen as the method for customer segmentation for the research conducted in papers 1, 2 and 3. In addition, decision trees were used as a classification method in paper 1. The customer segmentation evaluated in paper 6, is based on the segmentation made in paper 3.

In papers 4 and 5, segmentation with the traditional SOM did not give detailed enough information suitable for niche-marketing. More precise information was needed for identification of green consumers. Therefore, in paper 4, an adaption of SOM called the Weighted Self-Organizing Map (WSOM) was used for giving more weight to the share of green purchases in the training process. By weighting the share of green purchases, the customers who purchased more green products were allowed to affect the map formation more. In this way, multiple profiles for green customers were identified.

In paper 5, another adaption of SOM was used. In order to find out how the profile of a green consumer changed with the amount of green products they bought, the Self-Organizing Time Map (SOTM) was used for the customer segmentation. With the SOTM, several profiles for green consumers were identified.

In the next Sections, the methods used in this thesis are introduced. First, the decision trees are presented, followed by a presentation of SOM and its two adaptions WSOM and SOTM.

5.2. Decision trees

The decision tree is a classification method used to create models for division of large data collections into smaller sets. The model is constructed top-down, starting with the whole data set and partitioning it into meaningful subsets according to one of several splitting rules proposed in the literature. The division is made through two-way and/or multi-way splits. The splitting can afterwards be translated into if-then rules and used for classification. One example of a decision tree is the *Classification and Regression tree (CART)* algorithm, which is used for building a binary decision tree. CART uses the linear combination of attributes (Murthy 1998) and the Gini-index of diversity as a splitting rule (Berry and Linoff 2004; Murthy 1998).

The aim of the decision tree method is to capture the true characteristics of the data excluding noise and randomness. Therefore, not just the right splits for the data but also the right size of the decision tree needs to be determined. Pruning is the most widely used technique for finding correctly-sized trees. After constructing a model, the quality of the decision tree needs to be determined (Murthy 1998; Safavin and Landgrebe 1991: Breiman et al. 1984).

Advantages with decision trees are (Berry and Linoff 2004; Murthy 1998; Rokach and Maimon 2008):

1) transparency, as they produce straightforward rules for classification and prediction purposes;
2) insensitive to outliers or skewed data distributions;
3) decision trees are capable of data exploration potentially usable for prediction.

Decision trees have been used for a variety of business and marketing problems. For example, Fan et al. (2006) used a decision tree to predict house prices by identifying the most significant variables. Abrahams et al. (2009) created a marketing strategy for a pet insurance company, with the help of decision trees.

Sheu et al. (2009) used a decision tree for determining what factors influence customer loyalty.

In paper 1, Yao et al. 2010, a CART algorithm was used for building a binary decision tree, in order to identify the characteristics of profitable and unprofitable customers. The relationship between customers' purchase amounts and customers' demographic and behavioral characteristics was explored with the aim to differentiate between high and low spending customers.

5.3. SOM

The Self-Organizing Map (SOM) belongs to the family of unsupervised neural networks (Kohonen 2001). The other category of neural networks, often used for classification and prediction tasks, is the supervised ANNs, also called backpropagation, where the input/output data pairs are known. The SOM performs a data driven clustering, i.e., data is grouped together into clusters according to the similarities present in the data, without almost any a priori information or assumptions concerning the input data. The most interesting properties of the SOM are that it forms a nonlinear projection of high dimensional data onto a regular low-dimensional (2D) grid, where similar inputs are self-organized and located together (Kohonen 2001). Despite the reduction in data and dimension, the neighborhood relationships are saved in the SOM.

The advantages of the SOM are (Kohonen 2001; Vesanto 1999; Vesanto and Alhoniemi 2000; Kaski 1997): 1) The SOM is a highly visual method. The results can be visually presented to decision makers, who are able to intuitively interpret the results and turn information into knowledge. 2) The SOM is a very robust non-parametric method requiring very little preprocessing of data. 3) Very little a priori knowledge is required as the SOM is an explorative tool. It is possible to uncover unexpected patterns in data. 4) The SOM has a low computational cost.

There are two versions of the basic SOM algorithm (Vesanto et al. 2000, pp. 7-9):

1) *sequential training algorithm*, where the training of the SOM is performed in sequence after introducing a new input data
2) *batch training algorithm*, where all new input data is introduced simultaneously and the algorithm is trained all at once. The advantages of this method are low computational costs and repeatability of the results.

The segmentations made in this thesis use the batch training algorithm. After introducing new input data, the network organizes itself according to recognized similarities within the input data. The result of a SOM is a topological and spatial map formed through self-organization, i.e., a visual clustering showing similarities and dissimilarities in data (Kohonen 2001, p. 105-106).

The SOM algorithm

Before starting the analysis itself, a map consisting of neurons usually in a two-dimensional array is created. The shape of the array can be either rectangular or hexagonal. As the hexagonal array is more suitable for visualization purposes (Vesanto 1999), it was chosen for the segmentation purposes in papers 1, 2 and 3. The number of nodes (neurons), i.e., the size of the map, is determined by the purpose of the study. A large map with thousands of nodes is more suitable for visualization as there will be fewer data points per node, while a small map with few hundreds of nodes is more suitable for clustering. The nodes in the input and output layer are connected to each other. A weight vector m_i is assigned to each node i, according to a chosen initialization procedure. In Viscovery SOMine, i.e., the software used for creating most of the SOM maps in this thesis, a linear procedure is used for initialization of the map. This means that the weight vectors are initialized in an orderly fashion along the linear subspace spanned by the two principal eigenvectors of the input data set. A SOM network is presented in Figure 9, redrawn from Demirhan and Gyler (2011), where the input data are projected to the output layer of the map.

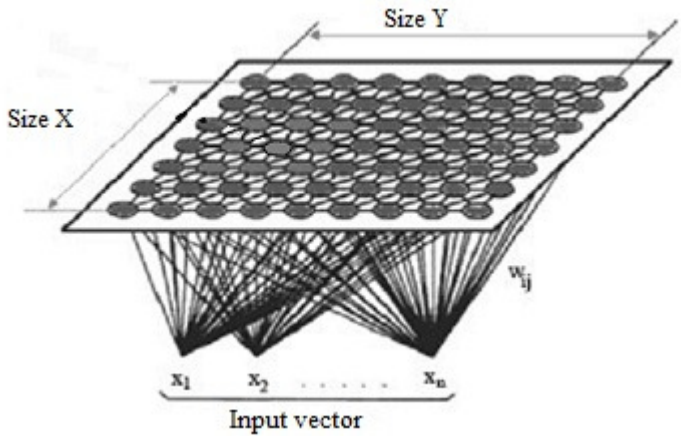


Figure 9. The structure of a typical SOM, where the input data are projected to the output layer of the map. Figure redrawn from Demirhan and Gyler (2011).

After all the input vectors x are mapped onto the hexagonal array, the linear initialization is completed and the training can begin (Kohonen 2001, p. 142). As described earlier, batch training comprises two steps involving the entire data

set: 1) finding best matching unit (BMU) for all input vectors; and 2) learning, i.e., adjusting the surrounding neighborhood nodes towards the input data vectors.

Step 1. All input data x_j (where j = 1,2,…,N) is presented to the network, the output units m_i (where i = 1,2,…,M) compete against each other in order to be declared winners m_c. The winners are the units that are most similar to the input data (BMUs), in terms of the smallest Euclidean distance, defined as $\|x - m_i\|$. m_c, i.e., the best matches can be calculated with the Formula 1 (Kohonen 2001, p. 110):

Formula 1 $$\|\boldsymbol{x_j} - \boldsymbol{m_c}\| = \mathbf{min}_i\{\|\boldsymbol{x_j} - \boldsymbol{m_i}\|\}$$

Step 2. The winning units adjust according to the batch algorithm presented in Formula 2, where *c(j)* represents the BMU of input vector j and $m_i(t)$ represents the reference vector of each input vector at time t (t = 0,1,2,…T).

Formula 2 $$\boldsymbol{m_i(t+1)} = \frac{\sum_{j=1}^{N} \boldsymbol{h_{ic(j)}(t) x_j}}{\sum_{j=1}^{N} \boldsymbol{h_{ic(j)}(t)}}$$

The connection weights for the surrounding outputs, i.e., the neighborhood $h_{ic(j)}$, are also adjusted by a decreasing factor (Bigus 1996, pp. 63-64, 71-73; Deboeck & Kohonen 1998, pp. xxxv-xxxvi, 161; Kohonen 2001, pp. 110-111). Kohonen (2001, p. 111) defines this learning process as the Gaussian equation presented in Formula 3, where r_c represents the coordinate of the best match (BMU), r_i represents the coordinate of the output unit or reference vector, and $\sigma(t)$ represents the radius of the neighborhood.

Formula 3 $$\boldsymbol{h_{ic(j)}} = \boldsymbol{\alpha(t) * exp\left(-\frac{\|r_c - r_i\|^2}{2\sigma^2(t)}\right)}$$

When similarity in the records is detected, it is used for recognizing the relationships between the segments (Rushmeier et al. 1997, p. 2).

Analysis of a SOM map

Interpretation and analysis of a SOM map is done through visualization of the map. One of the first ways of displaying the results was the so called U-matrix (unified distance matrix) (Ultsch 1993). A U-matrix is represented by the average of the distances between two neighboring reference vectors in nodes next to each other. The shape of the matrix is given by the neighborhood topology, where the landscape is visualized on the map in three dimensions by darker areas describing the peaks in the landscape, i.e., longer distances between

the reference vectors and lighter areas valleys or shorter distances. Formed clusters with lighter shade of a color in the area between the clusters are closer to each other than clusters with darker shade of color. This phenomenon, where similar units can be located on the map is called clustering via visualization (Flexer 2001). A U-matrix, is presented in Figure 10.

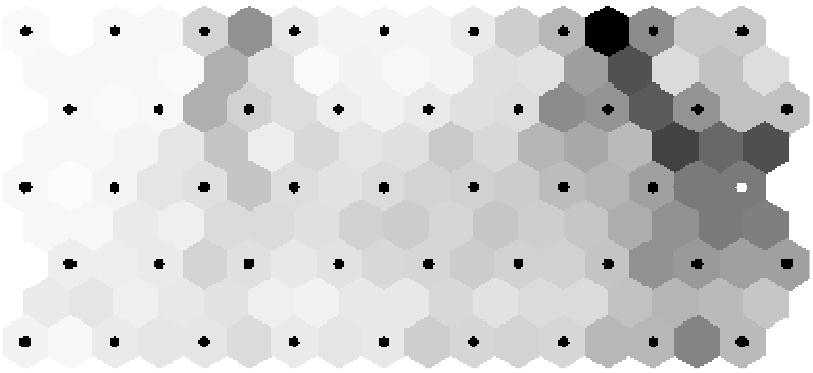


Figure 10. A U-matrix in gray-scale reprinted from Simula et al. (1999). The neurons of the network are marked as black dots. Light areas represent clusters, whereas dark areas are cluster separators. A separate cluster is formed in the upper right corner, as the clusters are separated by a dark gap.

Instead of subjective clustering via visualization, clustering of a SOM can be performed using two-stage clustering (Vesanto and Alhoniemi 2000). For example, the Viscovery SOMine software uses an adaptation of Ward's hierarchical clustering method to perform the clustering phase. The SOM-Ward's clustering method is a bottom-up method that merges the clusters with the smallest differences. The aim of this process is to minimize the total within cluster variance in order to ensure that the clusters are homogenous within the clusters and heterogeneous between the clusters. The SOM-Ward consists of two steps: 1) grouping of data with SOM into a two-dimensional display and 2) clustering of the resulting SOM using Ward's (1963) clustering. It uses the squared Euclidean distance of cluster centroids as the distance metric.

Different tools use different visualization methods. The Viscovery SOMine tool, which has been used in papers 1, 2, 3 and 6, uses SOM-Ward for optimization of the clustering. In Figure 11, a SOM-Ward map with seven clusters is presented. The map presents an overview of the customer base of the department store data presented in paper 3. The SOM-Ward map visualizes the formed 7 clusters, their location on the map and the main characteristics of the different groups of customers. In the upper left corner are the middle aged and elderly women, who are the most profitable customers of the department store. At the opposite end of the map, the right lower corner, are located the least profitable customers,

mainly men. The new customers, younger women and women with children are located in the middle of the map.

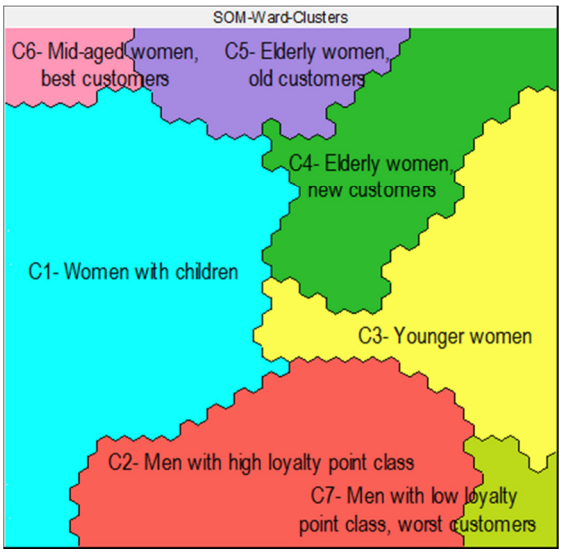


Figure 11. A SOM-Ward clustering made using Viscovery SOMine software. The formed 7 clusters are presented with different colors and named according to their attributes. This clustering is part of the study presented in paper 3.

For a closer interpretation of the formed clusters, single component-level maps called feature planes can be drawn. The feature planes visualize the distribution of individual variables in the data, and therefore, provide information on the characteristics of the clusters in the maps. In order to make the interpretation of the feature planes more intuitive, they are often displayed in color. Usually warm colors (red-yellow) indicate higher values, while cool colors (blue) indicate lower values for the variable. Feature planes from the study regarding the department store data in paper 3 are presented in Figure 12. Some conclusions regarding the customer base of the department store can be interpreted visually from the feature planes, e.g., most of the customers in clusters C2 and C7 are men (Gender), and the oldest customers are located in clusters C6, C5, C4 and C2 (Age). The full analysis of this SOM on customer data from the department store is presented in paper 3.

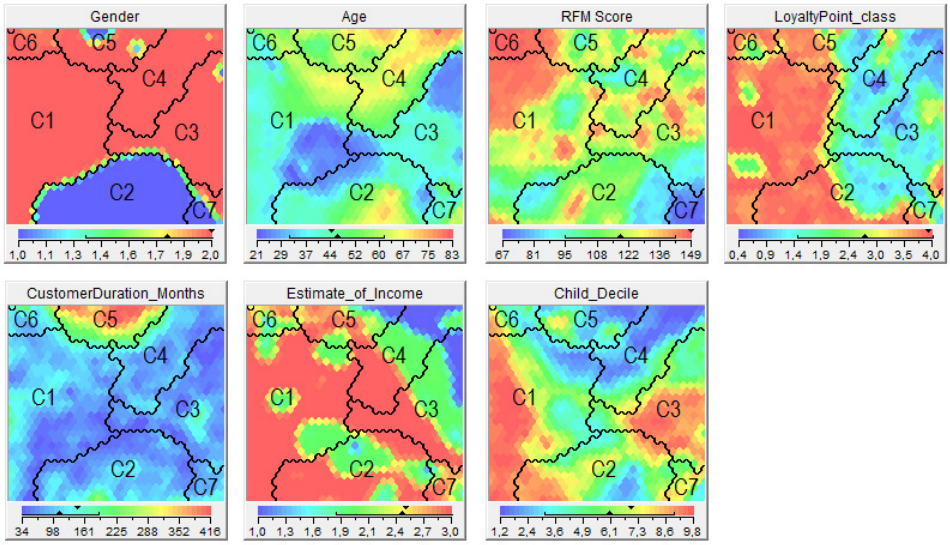


Figure 12. Feature planes for demographic variables for the department store customer data presented in paper 3.

Quality of a SOM map

The quality of a SOM map is measured by the *average quantization error* (ε_q), *distortion error* (ε_d) and the *topological error* (ε_t). The average quantization error is a measure of the average Euclidean distance, i.e., mean square deviation between the input vector x_j and its best matching reference vector $m_{c(j)}$. The aim is to minimize the quantization error. The quantization error is calculated according to Formula 4 (Desmet 2001, p. 23; Pölzlbauer 2004).

Formula 4 $$\varepsilon_q = \frac{1}{N}\sum_{j=1}^{N}\left\|x_j - m_{c(j)}\right\|$$

where N indicates the total number of samples, x_j the input vector and $m_{c(j)}$ the best matching reference vector in the output layer (Desmet 2001; Pölzlbauer 2004). This error is calculated automatically by many of the commercially available SOM software packages.

The distortion error (ε_d), presented in Formula 5, is a measure of how well the map fits the shape of the data distribution, taking into account the neighborhood radius.

Formula 5 $$\varepsilon_d = \frac{1}{N}\frac{1}{M}\sum_{j=1}^{N}\sum_{i=1}^{M} h_{ic(j)}\left\|x_j - m_{c(j)}\right\|$$

where N indicates the total number of samples, M indicates the total number of output units, $h_{ic(j)}$ the neighborhood, x_j the input vector and $m_{c(j)}$ the best matching reference vector in the output layer (Desmet 2001; Pölzlbauer 2004).

The topological error (ε_t) is a measure of the proportion of observations where the first two BMUs are adjacent. The aim is to minimize the topological error. The equation for the topological error is presented in Formula 6 (Desmet 2001, p. 23).

Formula 6 $$\varepsilon_t = \frac{1}{N}\sum_{j=1}^{N} u(x_j)$$

where N indicates the total number of samples. In the formula, $u(x_j) = 1$ if the first and second BMU's of x_j are not neighboring units. Otherwise $u(x_j) = 0$ (Desmet 2001, p. 23).

For gaining the most reliable results, the map can be trained several times with the same data for tuning the parameters and preprocessing approaches in an iterative manner. There will be variations in the quantization, distortion and topological errors between trained SOM maps. Even if there is no exact value for the best measure of map quality, these measures are useful for determining the quality of a SOM (Kohonen 2001, p. 161).

In addition, the final map can be cross validated e.g., using 10 folds *cross validation*. The map is retrained using the same parameters ten times with 90% or 9 folds of the data, each time retaining a separate 10% or one fold of the data for testing. If the differences in the error rates are small (ca 1%) between the trained maps, the final map can be considered technically validated.

In addition, a specialist in the field, in our case an expert from the department store, can validate the maps through *weak-form evaluation*. The intuition of the expert is then used for validation of the map (Waaser et al. 2004, p. 107).

5.3.1. WSOM

The Weighted Self-Organizing Map (WSOM) is an adaptation of the traditional SOM, where the focus is on a user-specified importance of data for learning. In paper 4, the WSOM was used for analyzing green consumer behavior. As discussed in Section 3.4.2, green consumer behavior has been of interest since the 1970s. Several studies have attempted to describe the green consumer with various methods, but no unanimous profile has been found.

The defining of a green consumer is not a trivial task. Issues that need to be addressed include, e.g., how many green products a consumer needs to buy in

order to be defined as a green consumer; if there is a difference between the consumer profiles of a person who buys several green products in a time period compared to a person who buys only one green product; and if there could exist several profiles for a green consumer.

In paper 4 we have come to the conclusion that there most likely exist several profiles for a green consumer, which explains why it is so difficult to define one valid profile. In segmentation using the WSOM, green consumers of the department store were given a specific weight depending on the share of green products (in number of items) that they had purchased. A larger weight meant that those consumers affected more the training of the map. Consumers that had not purchased any green products during the 2-year period, were given a weight of zero, and therefore, did not influence the segmentation. The result of the weighted segmentation was five different profiles of green consumers, instead of one.

The WSOM used in paper 4 follows the approach in Sarlin (2014). The SOM is trained using the batch training algorithm with an instance-specific weight. The general properties of the SOM are preserved while the WSOM is taking into account importance-varying instances.

The WSOM iterates in two steps, in similar manner as the traditional SOM (Kohonen 2001). As described earlier, in the first step the input vectors x are assigned to their BMUs m_c based upon the Euclidean distances. In the second step, each reference vector m_i is adjusted according to the weighted equivalent in the batch update formula, presented in Formula 7:

Formula 7 $$\mathbf{m_i(t+1)} = \frac{\sum_{j=1}^{N} \mathbf{w_j h_{ic(j)}(t) x_j}}{\sum_{j=1}^{N} \mathbf{w_j h_{ic(j)}(t)}}$$

where weight w_j represents the importance of x_j for the learning of patterns, index j indicates the input data vectors that belong to node c, and N is the number of the data vectors. The neighborhood is defined as a Gaussian function. In other words, the WSOM is similar to the standard SOM and can, therefore, be visualized using feature planes in order to indicate the distribution of individual variables.

5.3.2. SOTM

The Self-Organizing Time Map SOTM is another adaption of the SOM. Following Sarlin (2013a; 2013b) the SOTM can be described as having the same features as the traditional SOM but with means for abstractions of temporal structural changes. The SOTM illustrates how cross-sections evolve over time in

one-dimensional SOMs. Instead of time, the SOTM can be done over any variable.

In paper 5, the SOTM is used on the variable describing the greenness of a customer. The idea is to explore the changes in cluster structures over the greenness variable. More specifically, a segmentation of customers based on demographical and behavioral variables is made and the changes in the segments are monitored over the greenness variable, i.e., datasets ranging from non-green to green consumers.

Again, the SOTM over any variable *v* follows the two standard steps for creation of a SOM. First the BMUs, i.e., the units with the shortest Euclidean distances, are located and then the reference vector $m_i(v)$ is updated through a time-restricted version of the batch update formula, described in Formula 8. The neighborhood function $h_{ic(j)}(v)$ used in the batch update formula is defined as a Gaussian function, described in Formula 9. For a comprehensive description of the formulae for training of the SOTM, the reader is referred to Sarlin (2013a; 2013b; 2013c).

Formula 8 $$\mathbf{m_i(v)} = \frac{\sum_{\mathbf{j=1}}^{\mathbf{N(v)}} \mathbf{h_{ic(j)}(v) x_j(v)}}{\sum_{\mathbf{j=1}}^{\mathbf{N(v)}} \mathbf{h_{ic(j)}(v)}}$$

Formula 9 $$\mathbf{h_{ic(j)}(v)} = \mathbf{exp}\left(-\frac{\|\mathbf{r_c(v) - r_i(v)}\|^2}{\mathbf{2\sigma^2(t)}}\right)$$

where, $m_i(v)$ describes the reference vector, where $i = 1,2,\ldots,M$;
j index indicated the input data that belongs to unit c;
$h_{ic(j)}(v) \in [0,1]$ is the neighborhood function;
$\|r_c(v) - r_i(\boldsymbol{v})\|^2$ is the squared Euclidean distance between the coordinates of the reference vectors $m_c(v)$ and $m_i(v)$;
and σ is the neighborhood parameter.

The visualization of SOTM is to some extent similar as for standard SOM, where the dimensions or the variables can be visualized. The feature planes visualize the changes in variables over time (or any variable). Sammon's mapping (Sammon 1969) is used for visualizing the cluster structure of the SOTM.

5.4. Summary

In this chapter, the methods used in the research papers belonging to this thesis have been presented. A presentation of the decision trees, the traditional SOM and its two adaptions WSOM and SOTM used within this thesis have been given. In the next chapter, the results of research papers 1-6 are presented.

Chapter 6

Results

In this chapter, the research and studies in the publications of this thesis are presented. As discussed in Chapter 2, the ways for studying the customer base have evolved with time at the same pace as technology has evolved. Segmentation of customers has been done with different aims according to the available data. As described in Section 2.7, companies have an interest in gaining an overview of the customer base through segmentation. The first models constructed in this thesis concentrated on general customer segmentation, where the aim was to gain an overview of the whole customer base and to identify the most and least profitable customers. Even other data mining methods were used for enhancing the segmentation with SOM as described in Chapter 5. The information gained with the models was evaluated by experts of CRM from the retailer. Recently, segmentation of niche-markets describing the customer's awareness and expression of one's lifestyle, has gained popularity, as described in Section 2.8. Especially green consumer behavior, described in more detail in Section 3.4.2, has been a niche-market that has continued to interest companies. The more advanced models for analyzing niche-markets were built using the two adaptions of SOM, i.e., WSOM and SOTM, which are described in Sections 5.3.1 and 5.3.2.

The research published in the six papers are divided according to their themes into *general customer segmentation* (papers 1-3), *green consumer behavior* (papers 4-5), and *evaluation* (paper 6). In all of the articles, the results are visualized and analyzed with suitable methods. An overview of the papers is presented in Table 8, with regard to their purpose, techniques and dataset used, gained results, and sub objective (SO) they address. The following sections will give a deeper overview of these three themes.

Table 8. An overview of the papers and how they are connected with the sub objectives of this thesis.

Paper	Purpose	Technique	Data-set	Results	SO
1	Combination of Supervised and Unsupervised DM techniques for conducting CPA	SOM, Decision Trees	2	More detailed and accurate information on potential high-value customers from the customer base.	SO 2
2	CPA using SOM	SOM	1	Identifying profitable, average and non-profitable customers through CPA based on data driven exploratory customer segments coupled with sales data.	SO 2
3	Extracting information from data	SOM	2	Creating a visual overview of the customer base by grouping customers into segments with similar profiles or requirements.	SO 2
4	Visual data driven profiling of green consumers	WSOM	2	Identification of multiple profiles for a green consumer through a data driven exploratory analysis of a niche market.	SO 3
5	Exploring green vs. non-green consumer behavior based on visual data driven customer segmentation	SOTM	2	An analysis of how the profile for a green consumer varies with the customers' degree of greenness.	SO 3
6	A weak-form expert evaluation of customer profiling models	Evaluation	2	Determining the value based on quality of the model created in paper 3 by interviewing experts within the case company.	SO 2

6.1. Customer segmentation

Three studies on customer segmentation as a part of CPA are presented in research **papers 1, 2** and **3**. The main aim with all of these papers was to gain an overview of the customer base and to identify the most and least profitable customers. The research conducted in the papers differs regarding data sets and methods used for customer segmentation.

In **paper 1**, a two-level approach combining an unsupervised SOM-Ward clustering with the SOM and a supervised decision tree were used to identify potential high-value customers from the customer base of the case company. Dataset 2 (presented in Section 4.2) was used within this study. By using this hybrid approach more detailed information was extracted regarding the customer base. First, SOM-Ward clustering was used for customer segmentation, i.e., grouping of the customers according to similar characteristics and behavior. The characteristics of high-spending and low-spending customers were identified. Then, the decision tree algorithm was used for further exploring of the relationship between customers' spending amounts and their demographic and behavioral characteristics. Finally, by using the trained decision tree model, we were able to identify the segments representing potential profitable customers.

With similar analysis of the customer base, companies have the potential to better target their marketing efforts and make these more efficient in order to enhance their relationship with customers and increase the profitability of the entire customer base.

In **paper 2**, data-driven exploratory customer segmentation was performed based upon demographic data and coupled with product sales information. Dataset 1 (described in Section 4.2) representing purchases ranging from simple periodicals to advanced consulting services, was used. The aim was to determine which of the case company's customer relationships were profitable and worth developing, and conversely, which customers were better let go of. Overall, the strategic goal was to create a tool to be used by the sales department in the case company, in order to adjust marketing practices. These practices were used for determining suitable marketing effort levels for different, previously unknown, categories of customers. The SOM was used as the segmentation method.

A pilot test with a smaller data set was made with the aim to find the most effective segmentation parameters. During the pilot test, modifications to the data set were made in order to achieve more specific segmentation results. As described in Section 4.2, a decision was made to remove the largest customers and those customers who had only made one purchase during the time period in question. As the largest customers were served by a customer service agent of

their own, these customers were already taken good care of. Some smaller products that were rarely bought were also combined to form a single product.

The outcome of the segmentation with the SOM was ten clusters of customers displaying different demographic characteristics. The clusters were coupled with sales data, and a CPA was performed in order to identify profitable, average and non-profitable customers.

In **paper 3** the customer segmentation model is presented as a case study. The aim of the study was to gain an overview of the customer base, its demographic attributes, and customer buying behavior. More specifically, questions such as *who buys*, *how much*, *which products*, and *how recently and how often* were answered. The segmentation was based on customer data from a national chain of department stores, described as Dataset 2 in Section 4.2. The study is similar to the one presented in paper 2, but using a different dataset. Customer data were collected over a period of two years from four major department stores. The data consisted of demographical variables and eight product categories (labelled A to H), with additional information, such as the RFM variables, calculated from the transaction data as presented in Section 4.2.

The SOM was used for data-driven exploratory customer segmentation because of its advantages as discussed in Chapter 5, e.g., the visual presentation of the segmentation result, the SOM requires very little preprocessing of data and is an explorative tool that does data driven segmentation revealing unexpected patterns in data. Customers were segmented according to their demographical attributes, while purchasing behavior was associated with the demographic segmentation. The segmentation provided an overview of the customer base, consisting of seven clusters with specific characteristics for each group. Based on detailed analysis of the created segments and their feature planes, the strategic questions presented earlier could be answered.

An evaluation of the information gained from the formed segments was performed using a weak-form evaluation method. The evaluation is presented in Section 6.3.

6.2. Niche segmentation on green consumers

In research **papers 4** and **5**, green consumer profiling was performed using real customer transaction data. Instead of asking the customers through polls and queries regarding their green consumer behavior, their actual behavior was analyzed. As described in Section 3.4.2, the analysis was based on real purchases made in a department store over a two year time period. The aim of the study was to assess whether the demographics of green consumers were

distinguishable from the average consumer. Second, this enabled the comparison of multiple profiles of green consumers found in the data against the findings concerning contributing factors in previous research.

In **paper 4**, the aim was to identify green consumer profiles from customer data. The green consumers were identified through a data-driven analysis with the WSOM based on actual transaction data, i.e., Dataset 2 described in Section 4.2, including both demographic and behavioral information. The WSOM, as presented in Section 5.3.1, accounted for the 'degree' of how green a consumer was by giving a larger weight to consumers who bought more green products. The result of the weighted segmentation was visually presented in the form of five different profiles based upon which decision makers could have taken actions. The identified profiles were verified by comparison to earlier research.

In **paper 5**, patterns that emerge in the demographic and behavioral attributes of customers were identified, when studied from a degree-of-greenness perspective. The green consumer behavior was studied with the SOTM. The key idea of the SOTM, as presented in Section 5.3.2, was to enable the exploration of changes in cluster structures over not only the time dimension, but also over any other variable. Research paper 5 presents an application of the SOTM to demographical and behavioral customer data from Dataset 2 described in Section 4.2. The key focus was on assessing how customer behavior varied over customers' degree of greenness, i.e., datasets ranging from non-green to green customers. Thus, the time dimension of the SOTM was replaced with the degree of green purchases.

Among other results, the study revealed that there was no clear linear relationship between estimated income levels and the degree of greenness. While low degrees of greenness was clearly related to lower income levels, as is predicted by the literature, there was also a clear group of low income customers that displayed a high degree of greenness.

The study found that patterns across the different degrees of greenness differed significantly, and that the SOTM is a potentially useful tool for studying these patterns.

6.3. Evaluation

In **paper 6**, the customer segmentation model used for customer profiling (presented in paper 3) was evaluated. The research methodology used for the evaluation was presented in Section 1.3.1. A market basket analysis model was also evaluated in this paper, but is outside the scope of this thesis. The customer segmentation model, described in paper 3, was based on actual customer

purchasing data from a large department store for the period 2007-09, i.e., Dataset 2 described in Section 4.2. The proposed model was not yet an implemented and functioning system, which meant that the users were not able to interact with the system themselves. Therefore, the quality of the information extracted from the model for customer profiling was evaluated, instead of the technical properties of the model. The evaluation was performed using a weak-form evaluation method, consisting of interviews of experts from the department store. In this kind of evaluation, the common practice is to use a panel of experts (Huang et al. 2011; Dhillon and Torkzadeh 2006). The interviews were based on an adapted version of the five most important factors defined in the EUCS model (Doll and Torkzadeh 1988). They covered *content*, *accuracy*, *format*, and *ease of use* aspects. As we were evaluating a static model, the timeliness-aspect could not be measured. In addition, the factor "ease of use", in this case referred to the benefit and usefulness of the information.

Seven experts from the case company took part in the evaluation process. Before the actual interview, the experts were asked to fill out a questionnaire regarding their background and current access to timely sales information, i.e., how often and how useful the information was that they were receiving at the moment. Later, during an interview, the evaluator discussed the questionnaire with each respondent. Then, the experts were presented with information on their customers' buying behavior based on the results from the two models, i.e., the MBA- and the segmentation model. After each presentation, the experts were asked to respond to fifteen statements and four open-ended questions. The translated survey instruments used are presented in Appendices I and II.

In general, the information gained through the MBA- and segmentation analyzes was rated highly (4-5/max 5) by the experts. The experts considered the information gained with help of these models to be valuable and useful for decision making and for strategic planning for the future. This implies that the models could be of valuable use for managers working within CRM, e.g., planning marketing campaigns, product range planning, service development, planning of store layouts, operative and strategic planning, and for further developing loyalty programs.

6.4. Summary

This chapter summarizes the research on customer segmentation conducted in the research papers 1 to 6 in this thesis. Three different aspects of customer segmentation were covered, namely 1) different ways of conducting customer segmentation for CPA with SOM or SOM in combination with decision trees for B2B or B2C data, i.e. Datasets 1 and 2. 2) Green consumer behavior with adaption of SOM, where profiles for different green consumers were

investigated and identified either by weighting them with WSOM or analyzing them according to the degree of greenness with SOTM. Finally, 3) validating the outcome of customer segmentation with a weak-form expert evaluation of customer profiling models.

In the final chapter, key findings from this thesis and the research papers 1 to 6 will be stated, the contribution of this thesis will be presented, and implications for future research are discussed.

Chapter 7

Conclusion

In this chapter, a conclusive review of the thesis is performed with a focus on discussion regarding the aim and objectives. The contribution of the thesis is presented and, as a final issue, implications for future research are briefly discussed.

7.1 Conclusive review of the thesis regarding the aim and objectives

The overall aim of this thesis was to build and evaluate segmentation models for customer relationship management (CRM), based on customer behavior. In order to fulfill the overall aim I derived three sub objectives (SO). As the overall aim is very broad, it was assessed through the three SOs.

SO1. To investigate how segmentation has evolved and what are the current requirements. The aim is to investigate what factors have influenced segmentation throughout time, in order to understand the requirements for segmentation today.

The different eras of consumption from the beginning of the 16th century until now were presented in Chapter 2 and summarized in Table 1. The most influential developments were the creation of fashion in the court of Queen Elisabeth I as presented in Section 2.2; Products were made available for everybody as discussed in Sections 2.3-2.7, with the beginning of mass production within the Industrial revolution; The introduction of CRM in the late 20th century, as presented in Section 2.8, with an increasing amount of new methods used for collecting, modelling and analyzing of data.

In Section 2.9, the requirements of segmentation were discussed through visualization of the existing trends for customer buying behavior. The key variables and their change in time were visualized in Figure 5. The key variables are *degree of customization of products*, *degree of wealthy customers*, and

segment fraction. First, in the 16th century, custom made products were made only for the nobles, i.e., the degree of wealthy people was high. These people could be perceived as the first segment of customers, even if segmentation was not known as a method at that time. Only a fraction of people belonged to these wealthy customers. Then, the invention of new technical solutions made manufacturing of products more cost effective, and therefore, products more affordable to customers from different classes. Within mass production, the variety of products was small and the same product was offered to all customers. Therefore, the degree of customization of products decreased together with the degree of wealthy customers, while the segment fraction increased as the customer base was seen as one mass. However, saturation of the market lead to a decrease in product demand. A new trend changed the view of production and marketing. Instead of mass production, the products were custom made according to the wishes of the customers. As discussed in Chapter 3, CRM with its customer oriented view was implemented in many companies, meaning that the companies were aiming for the seven value drivers suggested by Richards and Jones (2008) presented in Section 3.1, and therefore, were able to offer interesting products and provide better service to the customers, according to their wishes.

Customer segmentation is still one of the key methods within CRM used by retailers for getting to know their customer base. The degree of wealthy customers has stabilized, while the degree of customized products has increased as the products are customized according to the customer's wishes. The segment fraction has decreased again as these contain a smaller fraction of the population. Because of the increase in available customer information and advances in technologies for analyzing large amounts of data, segmentation of customers can be made on a more detailed level, e.g., for identifying niche markets. The requirements for segmentation today are to deliver more usable precise data as an aid for:

- Decision making based on advanced data mining methods and visual analytics.
- Besides focusing on customer segmentation on a general lever, to also focus on niche markets.

SO2. To build and evaluate customer segmentation models within retailing for extracting information and knowledge from large amounts of customer data using Self-Organizing Maps.

In this thesis, models for customer segmentation were built using data mining methods. The practically relevant problem rose from the fact that there are not

many working models available for conducting customer profiling for a large customer base using visual analytics. Managers struggle with vast amounts of data, and therefore, there is a need for more methods for customer profiling.

This thesis studied how data mining methods can be used for extracting useful information from data. Patterns were identified from large amounts of real customer data using SOM. Models for data-driven segmentation were built with the aim at finding new, beforehand unknown information from data. In order to make the extracted information easier to grasp, various visualization methods were used as an aid for decision makers to intuitively understand the information and to convert it into knowledge.

In papers 1, 2 and 3, the first models for customer segmentation were built in order to gain an overview of the customer base. Customer segmentation models based on demographical, behavioral and psychographic or product information were built with the aim to group the customers into segments with different key abilities and provide information, as presented in Chapters 3, 4 and 5, e.g., on the most profitable and least profitable customers. In some of the models, other data mining methods, such as decision trees, were used for gaining more specific information on the customers.

The results of the segmentation were evaluated with technical evaluation methods presented in Section 3.4.4. The segments filled the following criteria:

- The segments were *identified* by managers of the case company.
- The segments were *substantial*, i.e., consisted of large enough share of the market to be profitable for a marketing strategy to be carried out.
- The segments were *accessible*, as the customer information was available for the case company.
- The segments were *stable*, but some related studies, e.g., Sarlin et al. 2012 and Yao et al. 2012, have suggested that some of the customers change their consumer behavior and move to another segment during marketing campaigns.
- The *responsiveness* of the segments was beyond the scope of this thesis.
- According to discussions with CRM experts in the context of evaluation of the usability of the segmentation results, the segments were *actionable*, i.e., meaningful for the case company.
- All the created segments had homogeneity within and heterogeneity between the segments.

The quality of the SOM map was verified by calculating the quantization error and topological error presented in Section 5.3.

The results of the customer segmentation model presented in paper 3 were also evaluated with a weak form evaluation method consisting of qualitative interviews of experts from the department store. The evaluation is reported in paper 6, Section 1.3.1, Section 6.3 and the translated survey instruments are presented as Appendices I and II. In general, the information gained through the segmentation analyzes was rated highly (4-5/max 5) by the experts. The experts considered the information gained with help of these models to be valuable and useful for decision making and for strategic planning for the future. This implies that the models could be of valuable use for managers working within CRM, e.g., planning marketing campaigns, product range planning, service development, planning of store layouts, operative and strategic planning, and for further developing loyalty programs.

SO3. To further develop the built models in order to study current "niches."

The average customer segmentation result gives an overview of the customer base. This is a good basis for conducting CRM, but in order to get more detailed information and knowledge about the customers, something that the company can easily use in decision making, is difficult. Current trends and "niches" are difficult to identify with normal customer segmentation methods.

In this thesis, two recently developed methods, i.e., WSOM and SOTM, were tested in order to attain new information on green consumer behavior. Green consumer behavior is a "niche" that has earlier mainly been investigated through self-reporting methods such as polls and surveys, as described in Section 3.4.2. The problem with this is that people tend to answer according to their intention, instead of their actual behavior. Therefore, an interest arose in studying green consumer behavior based on transaction data in order to analyze the actual behavior.
Two adaptions of the SOM were applied on transaction data from the department store, namely WSOM and SOTM. The WSOM was presented in more detail in Section 5.3.1 and the SOTM in Section 5.3.2. The niche segmentation on green consumer behavior was presented in Section 6.2. The built models and analyzes are visually presented in the papers 4 and 5.

7.2 Contribution claims

According to the seven guidelines described by Hevner et al. (2004) introduced in Section 1.3, design science research (DSR) should contribute both to research and practice. Therefore, the contribution of this thesis is studied from two different perspectives:

- with regard to the research community, and
- with regard to real-world practice.

7.2.1. Contribution claims with regard to the research community

With regard to the research community, following the DSR paradigm, this thesis has built and evaluated models for customer segmentation. The models were based upon real world data and domain experts from the case company were used to evaluate the built models. The models were found useful for segmentation of customers according to their buying behavior.

The thesis also provides a survey of key attributes describing customer buying behavior in different eras, as presented in Chapter 2. The first three stages of customer buying behavior were defined in the current literature, but the following stages have been derived based on an extensive literature review.

Another contribution is the positioning of the research in the overall visual analytics framework by proposing a hybrid visual knowledge discovery process combining the Visual Analytics process and the KDD process. The process was successfully used throughout the thesis in the construction of the models. This hybrid model is, to the best of my knowledge, a new contribution to the literature.

Finally, a new approach for analyzing a green consumer niche market was introduced. An extensive review of the literature shows that there are no demographic characteristics that clearly determine a green consumer, and that self-reporting studies fail to predict actual green behavior. Therefore, instead of using self-reporting methods, as most previous studies do, data driven exploratory methods were used for extracting detailed information on green consumers. The study thus contributes to the literature by clearly showing that there is no single universal profile of green consumers, and that it is instead possible to identify multiple profiles using a data-driven behavioral approach.

Thus, in accordance with the overall DSR paradigm, this research contributes to the research community with additional knowledge of how visual analytics, and

specifically the SOM, can be deployed to conduct behavioral segmentation of customers, both general and niche-driven segmentation.

7.2.2. Contribution claims with regard to real-world practice

With regard to real-world practice, following the DSR paradigm, this thesis has built and evaluated models for customer segmentation. As presented in papers 1-3 and 6 the models were based on real world data and evaluated by experts from the case company, thus strongly anchoring the research in practice. The information gained with the models was found useful for, e.g., decision making, improvement of customer service, and strategic planning. This thesis thus contributes by providing practice with concrete examples of how this approach could be fielded, as well as support for its usefulness for practice.

As was mentioned in section 7.2.1, the thesis also provides a survey of key attributes describing customer buying behavior in different eras. From the extensive literature review, the current requirements of customer segmentation were derived, again with a link to real world practice.

The positioning of the research in the hybrid visual knowledge discovery process is also a contribution for practice, by providing concrete guidelines for how the approach can be fielded in practice. This hybrid model has proven to be useful for performing visual knowledge discovery, as was verified by the experts from the case company in their evaluation.

Finally, the proposed new approach for analyzing a green consumer niche market is a clear contribution to practice as well. It is clearly shown through the literature review that self-reporting studies fail to predict actual green behavior, and that a data driven exploratory method can be successfully employed for extracting detailed information on green consumers. The techniques used for analyzing of niche markets as presented in papers 4 and 5 were useful for gaining more detailed information on green consumers.

7.3 Limitations and future work

Limitations

In this thesis, customer segmentation is performed on customer transaction data. The customers belong either to a company conducting B2B or a national department store chain conducting B2C. The models built for customer segmentation are only valid for the given customer base in Finland, for the

specific time period they were collected on. The author acknowledges that the models are not verified as overall models for customer segmentation.

Using the DSR method brings some limitations to the studies. Advances in technology might outrun the DSR results before implementation of the created model to the organization or even before the study have been published (Hevner et al. 2004).

Future work

In papers 4 and 5, demographical, behavioral and product usage variables derived from questionnaires and polls were compared to similar variables derived from actual transaction data. There were differences in the data. Therefore, it would be interesting to do data analysis and interviews on consumer behavior in parallel on same customer base and during same time period, so that the difference in data could be detected. Also, the analysis result based on data from several different sources collected in different format might bring more value to the analysis and give better insights of the customer base.

In Chapter 4, a hybrid visual knowledge discovery process combining the Visual Analytics process and the KDD process was presented. The process was successfully used throughout the thesis in the construction of the models. However, before further use of the process, the hybrid visual knowledge discovery process should be validated for visual analytics. In addition, a more extensive evaluation of the approach should be performed.

Ecommerce is growing every year, creating new markets and shopping opportunities within retail industry. The share of ecommerce is expected to grow rapidly. This gives more demand for web analytics, as consumer behavior can be investigated through text analysis and text mining via the social media. In addition, ecommerce gives an opening for creation of new customer segmentation methods and techniques (eMarketer 2014). Examples of newer methods are online analytical processing (OLAP) on transaction and customer data and Big Data. All of these combined give a vision of real time analyzes of customer and transaction data with the help of visual analytical methods, where the managers are able to analyze and make intuitive decisions based on information provided and communicate the knowledge created so that these are immediately put into action.

References

Abbott, J., Stone, M. and Buttle F. (2001). Integrating customer data into customer relationship management strategy: An empirical study, *Journal of Database Marketing*, 8(4), 289-300. Henry Stewart Publications.

Abrahams, A.S., Becker, A.B., Sabido, D., D'Souza, R., Makriyiannis, G., Krasnodebski, M. (2009). Inducing a Marketing Strategy for a New Pet Insurance Company using Decision Trees, *Expert Systems Applications*, 36, 1914–1921

Anderson, T. Jr. and Cunningham, W.H. (1972). The socially conscious consumer, *Journal of Marketing*, 36 (7), 23-31.

Antil, J.H. and Bennet, P.D. (1979). Construction and validation of a scale to measure socially responsible consumption behavior, in Henion, K.E. and Kinnear, T.C. (Eds.), *The conserver Society*, American Marketing Association, Chicago, IL, 51-68.

Antil, J. H. (1984). Conceptualization and operationalization of involvement. *Advances in consumer research*, 11(1), 203-209.

Arbuthnot, J. (1977). The roles of attitudinal and personality variables in the prediction of environmental behavior and knowledge, *Environment and Behavior*, 9, 217-232.

Arroyo, J.A, Kruger, S.P., O'sullivan, P.J., Da Silva, L.M.P (2014). Generating instant messaging contacts for customer relationship management system, *United States Patent Application Publication*, Pub. No.: US 2014/0351345 A1

Badgett, M. and Stone, M. (2005). Multidimensional segmentation at work: Driving an operational model that integrates customer segmentation with customer management, *Journal of Targeting, Measurement and Analysis for Marketing*, 13(2), 103-121. Henry Stewart Publications.

Baltas, G. (2001). Nutrition labeling: issues and policies, *European Journal of Marketing*, 35(5), 708-21.

Banerjee, B. and McKeage, K. (1994). How green is my value: exploring the relationship between environmentalism and materialism, in Allen C.T. and John, D.R. (Eds) *Advances in Consumer Research*, Association for Consumer Research, Provo, UT, 21, 147-52.

Banytė, J., Brazionienė, L., Gadeikienė, A. (2010). Investigation of green consumer profile: a case of Lithuanian market of ecofriendly food products, *Economics and Management*, 15, 374-383, ISSN 1822-6515.Bellenger, D. N. & Korgaonkar, P. K. (1980). Profiling the Recreational Shopper, *Journal of Retailing*, 56(3), 77-92. Elsevier.

Berger, I.E. and Corbin, R.M. (1992). Perceived consumer effectiveness and faith in others as moderators of environmentally responsible behaviors, *Journal of Public Policy & Marketing*, 11(2), 79-88.

Berry, M.J.A. and Linoff, G.S. (2004). *Data Mining Techniques: For marketing, Sales, and Customer Relationship Management*, 2nd ed., Indianapolis, Indiana: Wiley Publishing Inc.

Bigus, J.P. (1996). *Data mining with Neural Networks: Solving Business Problems from Application Development to Decision Support*, New Youk, NY: The McGraw-Hill Companies Inc.

Bishop, C.M. (1995). *Neural Networks for Pattern Recognition*, Avon: Oxford University Press.

Blumer, H. (1969). Fashion: From Class Differentiation to Collective Selection, *The Sociological Quarterly*, 10(3), 275–291.

Boulding, W., Staelin, R., Ehret, M. and Johnston, W.J. (2005). A customer relationship management roadmap: What is known, potential pitfalls, and where to go, *Journal of Marketing*, 69, 155-166. American Marketing Association.

Bowlby, R. (1997). Supermarket Futures, in Falk, P. & Campbell, C. (Eds.) *The Shopping Experience*, London, UK: SAGE Publications Ltd.

Breiman, L. (1984). *Classification and regression trees*, New York: Chapman and Hall/CRC.

Brimicombe, A.J. (2007). A dual approach to cluster discovery in point event data sets, *Computers, Environments and Urban Systems*, 31(1), 4-18. Elsevier Ltd.

Bui, My H. (2005). Environmental Marketing: A model of consumer behavior, In Lapidus, R.S. and Chapman, K.J. (Eds,) *Proceedings of the Annual Meeting of the Association of Collegiate*, La Jolla, California, April 2005.

Burrell, G. and Morgan, G. (1979). *Sociological paradigms and organizational analysis: elements of the sociology of corporate life*, London: Heinemann.

Buttle, F. (2004). *Customer Relationship Management Concepts and Tools*, Oxford: Butterworth-Heinemann.

Campbell, C. (1997). Shopping, pleasure and the sex war, in Falk, P. & Campbell, C. (Eds.) *The Shopping Experience*, London, UK: SAGE Publications Ltd.

Chalmeta, R. (2006). Methodology for customer relationship management, *The Journal of Systems and Software*, 79(7), 1015–1024.

Childs, N. and Polyzees, G.H. (1997). Foods that help prevent disease: consumer attitudes and public policy implications, *Journal of Consumer Marketing*, 14(6), 433-47.

Chinnici, G., D'Amico, M., Pecorino, B. (2002). A multivariate statistical analysis on the consumers of organic products, *British Food Journal*, 104 (3/4/5), 187-199.

Corrigan, P. (1997). *The Sociology of Consumption*, London, UK: SAGE Publications Ltd.

Datta, Y. (1996). Market segmentation: an integrated framework, *Long Range Planning*, 29(6), 797–811.

Davies, A., Titterington, A.J. and Cochrane, A. (1995). Who buys organic food? A profile of the purchasers of organic food in N. Ireland, *British Food Journal*, 97(10), 17-23.

Deboeck, G.J. and Kohonen, T. (1998). *Visual Explorations in Finance with Self-Organizing Maps*, Berlin: Springer-Verlag.

DeLone, W.H. and McLean, E.R. (1992). Information Systems Success: The quest for the dependent variable, *Information systems research*, 3(1), 60-95.

DeLone, W. H. (2003). The DeLone and McLean model of information systems success: a ten-year update. *Journal of management information systems*, 19(4), 9-30.

Demirhan, A. and Gyler, I. (2011). Combining stationary wavelet transform and self-organizing maps for brain MR image segmentation, *Engineering Applications of Artificial Intelligence*, 24(2), 358-367.

Desmet, P. (2001). Buying behavior study with basket analysis: pre-clustering with a Kohonen map, *European Journal of Economic and Social Systems,* 15(2), 17–30.

Dhillon, G. and Torkzadeh, G. (2006). Value-focused assessment of information system security in organizations, *Information Systems Journal*, 16(1), 293-314.

Diamantopoulos, A., Schlegelmilch, B.B., Sinkovics, R.R., Bohlen, G.M. (2003). Can socio-demographics still play a role in profiling green consumers? A review of the evidence and an empirical investigation, *Journal of Business Research*, 56, 465-480.

Dibb, S. (2001). New millennium, new segments: moving towards the segment of one?, *Journal of Strategic Marketing*, 9(3), 193–213.

Dibb, S. and Simkin, L. (1996). *The market segmentation workbook: Target marketing for marketing managers*, London, Great Britain: Routledge.

Dickson, P.R. (1982). Person-Situation: Segmentation's missing link, *Journal of Marketing*, 46(Fall), 56-64. EBSCO Host.

Doll, W.J. and Torkzadeh, G. (1988). The measurement of End-User Computing Satisfaction, *MIS Quarterly*, 12(2), 259-274.

D'Urso, P., Giovanni, L.D. (2008). Temporal Self-Organizing Maps for Telecommunications Market Segmentation. *Neurocomputing* 71, 2880–2892.

Dutta, S., Bhattacharya, S., Guin, K.K. (2015). Data Mining in Market Segmentation: A Literature Review and Suggestions, pp. 87-98. In Das, K.N., Deep, K., Pant, M., Bansal, J.C., and Nagar, A. (Eds.) Advances in Intelligent Systems and Computing. SocProS 2014, Volume 1: *Proceedings of Fourth International Conference on Soft Computing for Problem Solving*, Vol 335, Springer India.

Ehret, M. (2004). Managing the trade-off between relationships and value networks. Towards a value-based approach of customer relationship managements in business-to-business markets, *Industrial Marketing Management*, 33, 465-473. Elsevier Inc.

eMarketer (December 23rd, 2014) *Retail sales worldwide will top 22 trillion this year* (Online) eMarketer. Available: http://www.emarketer.com/Article/Retail-Sales-Worldwide-Will-Top-22-Trillion-This-Year/1011765 [Accessed 1st of April 2015]

Engel, E. (1857). Die Productions- und Consumtionsverhältnisse des Königreichs Sachsen, *Zeitschrift des statistischen Bureaus des Königlich Sächsischen Ministerium des Inneren* 8–9: 28–29.

Engel, E. (1895). *Die Lebenkosten Belgischer Arbeiter-Familien fruher und jetzt*, Dresden.Falk, P. and Campbell, C. (Eds.) (1997). *The Shopping Experience*, London, UK: SAGE Publications Ltd.

Ernst, H., Hoyer, W. D., Krafft, M., & Krieger, K. (2011). Customer relationship management and company performance-the mediating role of new product performance. *Journal Of The Academy Of Marketing Science*, 39(2), 290-306.

Falk, P. (1994). *The Consuming Body*, London, UK: SAGE Publications Ltd.

Fan, G.Z., Ong, S.E., Koh, H.C. (2006). Determinants of House Price: A Decision Tree Approach, *Urban Studies*, 43, 2301–2315.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery in databases, *Artificial Intelligence magazine*, 17(3), 37-54.

Flexer, A. (2001). On the use of self-organizing maps for clustering and visualization, *Intelligent Data Analysis*, 5(5), 373-384.

Follows, S.B. and Jobber, D. (2000). Environmentally responsible purchase behavior: a test of a consumer model, *European Journal of Marketing*, 34(5/6), 723-46.

Fotopoulos, C. and Krystallis, A. (2002). Purchasing motives and profile of the Greek organic consumer: a countrywide survey, *British Food Journal*, 104 (9), 730-765.

Frank, R.E., Massy, W.F. and Wind, Y. (1972). *Market Segmentation*, Englewood Cliffs, New Jersey: Prentice-hall Inc.

Furness, P. (2001). Techniques for customer modelling in CRM, *Journal of Financial Services Marketing*, 5(4), 293-307, Henry Stewart Publications.

Galliers, R.D. and Land, F.F. (1987). Viewpoint: choosing appropriate information systems research methodologies, *Communications of the ACM*, 30(11), 901-902.

Gerstman and Meyers Inc. (1989). *Consumer Solid Waste Management: Awareness, Attitude and Behavior Study*, New York, NY: Gerstman and Meyers Inc.

GfK 2011 survey (2011). The Environment: Public Attitudes and Individual Behavior — A Twenty-Year Evolution, *SC Johnson's GfK Roper Consulting Green Gauge® US survey*.

Gill, J.D., Crosby, L.A. and Taylor, J.R. (1986). Ecological concern, attitudes, and social norms in voting behavior, *Public Opinion Quarterly*, 50, 537-54.

Gregor, S., Hevner, A.R. (2013). Positioning and Presenting Design Science Research for Maximum Impact, *MIS Quarterly*, 37(2), 337-355.

Griskevicius, V., Tybur, J.M., and Van den Bergh, B. (2010). Going Green to Be Seen: Status, Reputation, and Conspicuous Conservation, *Journal of Personality and Social Psychology*, 98(3), 392–404

Grönroos, C. (2002). *Service management och marknadsföring – en CRM ansats*, Kristianstad, Sverige: Kristianstads Boktryckeri AB.

Han, J., Kamber, M., and Pei, J. (2011). *Data Mining: Concepts and Techniques*. Waltham, MA: Elsevier Science.

Hand, D.J., Mannila, H. and Smyth, P. (2001). *Principles of Data Mining*, Boston, USA: MIT Press.

Heinrich, B. (2005). Transforming strategic goals of CRM into process goals and activities, *Business Process Management Journal*, 11(6),709–723.

Herberger, R.A. Jr. (1975, reprinted in 2002). The Ecological Product Buying Motive: A Challenge for Consumer Education, *The Journal of Consumer Affairs*, 187-195.

Hevner, A.R., March, S.T., Park, J., and Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, 28(1), 75-105.

Hevner, A. (2007). A three-cycle view of design science research, *Scandinavian Journal of Information Systems*, 19(2), 87–92.

Hewer, P. & Campbell, C. (1997). Research on shopping – a brief history and selected literature, in Falk, P. & Campbell, C. (Eds.) *The Shopping Experience*, London, UK: SAGE Publications Ltd.

Hill, H. and Lynchehaum, F. (2002). Organic milk: attitudes and consumption patterns, *British Food Journal*, 104(7), 526-542.

Hine, D.W. and Gifford, R. (1991). Fear appeals, individual differences, and environmental concern, *The Journal of Environmental Education*, 23(1), 36-41.

Hines, J.M., Hungerford, H.R., Tomera, A.N. (1987). Analysis and Synthesis of Research on Responsible Environmental Behavior: A Meta-Analysis, *The Journal of Environmental Education*, 18(2).

Hsu, C-C. and Chen, Y-C. (2007). Mining of mixed data with application to catalog marketing, *Expert Systems with Applications*, 32, 12-23. Elsevier Ltd.

Huang, C.L. (1996). Consumer preferences and attitudes towards organically grown produce, *European Review of Agricultural Economics*, 23, 331-42.

Huang, S.M., Hung, W.H., Yen, D.C., Chang, I., Jiang, D. (2011). Building the evaluation model of the IT general control for CPAs under enterprise risk management, *Decision Support Systems*, 50(4), 692-701.

Iivari, J. (1991). A paradigmatic analysis of contemporary schools of IS development, *European Journal of Information Systems*, 1(4), 249-272.

Ishaswini, S. and Datta, K. (2011). Pro-environmental Concern Influencing Green Buying: A Study on Indian Consumers, *International Journal of Business and Management*, 6(6), 124-133.

Järvinen, P. (2001). *On research methods*, Tampere: Opinpajan kirja.
Jensen, K.O., Denver, S., Zanoli, R. (2011). Actual and potential development of consumer demand on the organic food market in Europe, *NJAS - Wageningen Journal of Life Sciences*, 58, 79–84.
Jolly, D. (1991). Differences between buyers and nonbuyers of organic produce and willingness to pay organic price premiums, *Journal of Agribusiness*, 9(1), 97-111.
Kalafatis, S.P., Pollard, M., East, R. and Tsogas, M.H. (1999). Green marketing and Ajzen's theory of planned behaviour: a cross-market examination, *Journal of Consumer Marketing*, 16(5), 441-460, MCB University Press.
Kasanen, E., Lukka, K. and Siitonen, A. (1993). The constructive approach in management accounting research, *Journal of management accounting research*, 5,June 1991, 243-264.
Kaski, S. (1997). *Data exploration using self-organizing maps*, Helsinki University of Technology, Helsinki.
Kaski, S., Kangas, J., Kohonen, T. (1998). Bibliography of Self-Organizing Map (SOM) Papers 1981-1997. *Neural Computing Surveys* 1, 102–350.
Keim, D.A., Mansmann, F., Schneidewind, J., Thomas, J. and Ziegler, H. (2008). Visual Analytics: Scope and Challenges, in: Simoff, S.J. et al. (Eds.): *Visual Data Mining*, LNCS 4404, 76-90. Berlin, Heidelberg: Springer Verlag.
Kilbourne, W.E., and Beckmann, S.C. (1998). Review and Critical Assessment of Research on Marketing and the Environment, *Journal of Marketing Management*, 14(6), 513-532.
Kim, S., Jung, T., Suh, E. and Hwang, H. (2006). Customer segmentation and strategy development based on customer lifetime value: a case study, *Expert Systems with Applications*, 31(1), 101–107.
Kinnear, T.C., Taylor, J.R. and Ahmed, S.A. (1974). Ecologically concerned consumers: who are they?, *Journal of Marketing*, 38, 20-24.
Kleiner, A. (1991). What does it mean to be green?, *Harvard Business Review*, 69, 38-47.
Kohonen, T. (2001). *Self-Organizing Maps*, Berlin: Springer-Verlag.
Kotler, P. (2002). *Marketing management*, Millennium edition, Pearson Custom Publishing.
Krystallis, A., and Chryssohoidis, G. (2005). Consumers' willingness to pay for organic food: factors that affect it and variation per organic product type. *British Food Journal*, 107(5), 320-343.
Laroche, M., Bergeron, J. and Barbaro-Forleo, G. (2001). Targeting consumers who are willing to pay more for environmentally friendly products, *Journal of Consumer Marketing*, 18(6), 503-20.
Larose, D.T. (2005). *Discovering Knowledge in Data. An Introduction to Data Mining*, Hoboken, NJ: John Wiley & Sons Inc.
Lea, E. and Worsley, T. (2005). Australians' organic food beliefs, demographics and values, *British Food Journal*, 107(11), 855-69.

Lee, C. and Green, R.T. (1991). Cross-cultural examination of the Fishbein behavioural intentions model, *Journal of International Business Studies*, 22(2), 289-305.

Lee, S.C., Gu, J.C., Suh, Y.H. (2006). A Comparative Analysis of Clustering Methodology and Application for Market Segmentation: K-Means, SOM and a Two-Level SOM. In: Zhong, N., Ras, Z.W., Tsumoto, S., Suzuki, E. (eds.) ISMIS 2006. *LNCS (LNAI)*, vol. 4203, pp. 435–444. Springer, Heidelberg.

Leiss, W., Kline, S. & Jhally, S. (2000). The Bonding of Media and Advertising, in Lee, M. J. (Ed.) (2000). *The Consumer Society Reader*, Oxford, UK: Blackwell Publishers Ltd.

Lindgreen, A., Palmer, R., Vanhamme, J. and Wouters, J. (2006). A relationship-management assessment tool: Questioning, identifying, and prioritizing critical aspects of customer relationships, *Industrial Marketing Management*, 35, 57-71. Elsevier Inc.

Lingras, P., Hogo, M., Snorek, M. and West, C. (2005). Temporal analysis of clusters of supermarket customers: conventional versus interval set approach, *Information Sciences*, 172(1-2), 215–240.

Linoff, G.S. and Berry, M. J. (2011). *Data Mining Techniques: For marketing, sales, and customer relationship management*, (3rd Edition). Indianapolis, Indiana: Wiley Computer Publishing.

Magnusson, M., Arvola, A., Koivisto Hursti, U., Aberg, L. and Sjoden, P. (2001). Attitudes towards organic foods among Swedish consumers, *British Food Journal*, 103(3), 209-26.

Maimon, O. and Rokach, L. (2005). *The data mining and knowledge discovery handbook*, NY, USA: Springer Science+Business Media Inc.

Mainieri, T., Barnett, E.G., Valdero, T.R., Unipan, J.B. and Oskamp, S. (1997). Green Buying: The Influence of Environmental Concern on Consumer Behavior, *The Journal of Social Psychology*, 137(2), 189-204.

Mandese, J. (1991). New study finds green confusion, *Advertising Age*, October 21.

March, S.T. and Smith, G.F. (1995). Design and natural science research on information technology, *Decision Support Systems*, 15(4), 251-266.

McCann, J.M. (1974). Market segment response to the marketing decision variables, *Journal of Market Responding*, 11(4), 399-415.

McCracken, G (1988). *Culture and consumption: new approaches to the symbolic character of consumer goods and activities*, Bloomington (Ind.): Indiana University Press.

McKendrick, N., Brewer, J., Plumb, J.H. (1982). *The birth of a consumer society: the commercialization of eighteenth-century England*, London: Europa Publications.

Miglautsch, J.R. (2000). Thoughts on RFM scoring, *The journal of database marketing*, 8(1), 67-72.

Mort, F. (2000). The Politics of Consumption, in Lee, M. J. (Ed.) (2000). *The Consumer Society Reader*, Oxford, UK: Blackwell Publishers Ltd.

Murthy, S.K. (1998). Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey, *Data Mining and Knowledge Discovery*, 2, 345–389.

Nascimento, G.S. and Oliveira, A.A. (2012). An Agile Knowledge Discovery in Databases Software Process, in Xiang, Y. et al. (Eds.): *ICDKE 2012*, LNCS 7696, 56–64, Berlin, Heidelberg: Springer-Verlag.

Nakarado, G.L. (1996). A marketing orientation is the key to a sustainable energy future, *Energy Policy*, 24(2), 187–193.

Neslin, S., Taylor, G., Grantham, K., & McNeil, K. (2013). Overcoming the 'recency trap' in customer relationship management. *Journal Of The Academy Of Marketing Science*, 41(3), 320-337.

Newell, S.J. and Green, C.L. (1997). Racial differences in consumer environmental concern, The Journal of Consumer Affairs, 31(1), 53-69.

Ngai, E.W.T. (2005). Customer relationship management research (1992-2002): An academic literature review and classification, *Marketing Intelligence & Planning*, 23(6), 582 – 605.

Nguyen, B., and Mutum, D.S, (2012). A review of customer relationship management: successes, advances, pitfalls and futures, *Business Process Management Journal*, Vol. 18 Iss 3 pp. 400 - 419

Oja, M., Kaski, S., Kohonen, T.(2003). Bibliography of Self-Organizing Map (SOM) Papers: 1998-2001 Addendum. *Neural Computing Surveys* 3, 1–156 (2003)

Paas, L. and Kuijlen, T. (2001). Towards a general definition of customer relationship management, *Journal of Database Marketing*, 9(1), 51–60.

Padel, S. and Foster, C. (2005). Exploring the gap between attitudes and behavior. Understanding why consumers buy or do not buy organic food, *British Food Journal*, 107(8), 606-625, Emerald Group Publishing Limited.

Park, H-S. and Baik, D-K. (2006). A study for control of client value using cluster analysis, *Journal of Network and Computer Applications*, 29, 262-276. Elsevier Ltd.

Parr-Rud, O. (2001). *Data Mining Cookbook, Modeling Data for Marketing, Risk, and Customer Relationship Management*, NY, USA: John Wiley & Sons Inc.

Parvatiyar, A. and Sheth, J.N. (2001). Customer relationship management: emerging practice, process and discipline, *Journal of Economic and Social Research*, 3(2), 1–34.

Pope, D. (1983). *The making of modern advertising*, New York, USA: Basic Books.

Pries-Heje, J., Baskerville, R., Venable, J.R. (2008). Strategies for Design Science Research Evaluation, In Proceedings of the *16th European Conference on Information Systems* (ECIS 2008), Galway, Ireland.

Pölzlbauer G (2004). Survey and comparison of quality measures for SOM. In Paralic J, Pölzlbauer G, Rauber A (Eds.) *Proceedings of the 5th Workshop on data analysis*, Vysoké Tatry (Slovakia) 24-27 June 2004. Elfa Academic Press, pp. 67-82.Rao, C.P. (1974). Consumer ecological concern and adaptive behavior, *Journal of the Academy of Marketing Science*, 2(1), 262-277.

Reinhold, O., and Alt, R. (2012). Social Customer Relationship Management: State of the Art and Learnings from Current Projects, *25th Bled eConference eDependability: Reliable and Trustworthy eStructures, eProcesses, eOperations and eServices for the Future*, June 17-20, 2012; Bled, Slovenia.

Richards, K.A. and Jones, E. (2008). Customer relationship management: Finding value drivers, *Industrial Marketing Management*, 37, 120–130.

Rigby, D.K. and Ledingham, D. (2004). CRM done right, *Harvard Business Review*, 82(11), 118-129.

Ripley, B.D. (1997). Classification, In Kotz, S., Read, C.B. and Banks, D.L. (Eds.) *Encyclopedia of Statistical Sciences*. Update 1, 110-116, NJ, USA: Wiley-Interscience.

Roberts, J.A. (1991). *The development of a profile of the socially responsible consumer for the 1990s and its marketing management and public policy implications*, Doctoral Thesis, Marketing Department, University of Nebraska, Lincoln, NE.

Roberts, J.A. (1995). Profiling levels of socially responsible consumer behavior: a cluster analytic approach and its implications for marketing, *Journal of Marketing Theory and Practice*, Fall, 97-117.

Roberts, J.A. (1996). Green Consumers in the 1990s: Profile and Implications for Advertising, *Journal of Business Research*, 36, 217-231. Elsevier Science Inc.

Roberts, J.A. and Bacon, R. (1997). Exploring the subtle relationships between environmental concern and the ecologically conscious consumer behavior, *Journal of Busine*ss Research, 40, 79-89.

Roddy, G., Cowan, C.A. and Hutchinson, G. (1996). Consumer attitudes and behaviour to organic foods in Ireland, *Journal of International Consumer Marketing*, 9(2), 41-63.

Rokach, L., Maimon, O. (2008). Data Mining with Decision Trees: Theory and Applications, Singapore: World Scientific Publishing.

Rothschild, M.L. (1979). Marketing communications in non-business situations or why it's so hard to sell brotherhood like soap, *Journal of Marketing*, 43(2), 11–20.

Rushmeier, H., Lawrence, R. and Almasi, G. (1997). Case study: visualizing customer segmentations produced by self organizing maps, *Eighth IEEE Visualization 1997* (VIS'97), 463–466.

Rygielski, C., Wang, J. and Yen, D.C. (2002). Data mining techniques for customer relationship management, *Technology in Society*, 24(4), 483–502.

Safavin, S.R., Landgrebe, D. (1991). A survey of decision tree classifier methodology, *IEEE Transactions on Systems, Man, and Cybernetics*, 21(3), 660–674.

Salomann, H., Dous, M., Kolbe, L. and Brenner, W. (2005). Rejuvenating customer management: How to make knowledge for, from and about customers work, *European Management Journal*, 23(4), 392-403. Elsevier Ltd.

Samdahl, D.M., Robertson, R. (1989). Social determinants of environmental concern: specification and test of the model, *Environmental Behavior*, 21(1), 57–81.

Sammon, J.W. (1969). A nonlinear mapping for data structure analysis, IEEE *Transactions on Computers*, 18, 401-409.

Sarlin, P. (2013a). Self-Organizing Time Map: An Abstraction of Temporal Multivariate Patterns, *Neurocomputing*, 99(1), 496-508.

Sarlin, P. (2013b). Replacing the time dimension: A Self-Organizing Time Map over any variable, in *Proceedings of the Workshop on New Challenges in Neural Computation* (NC^2), Saarbrucken, Germany.

Sarlin, P. (2013c). *Mapping the financial stability*, TUCS Dissertations 159. Åbo Akademi University.

Sarlin, P. (2013d). Data dimension reduction for visual financial performance analysis, *Information Visualization* 0(0), 1-20.

Sarlin, P. (2014). A Weighted SOM for classifying data with instance-varying importance. *International Journal of Machine Learning and Cybernetics*, 5, 101–110.

Sarlin, P., Yao, Z., Eklund, T. (2012). A framework for state transitions on the Self-Organizing Map: Some Temporal Financial Applications, *Intelligent systems in accounting, finance and management*, 19(3), 189-203.

Schifferstein, H.N.J. and Oude Ophuis, P.A.M. (1998). Health-related determinants of organic food consumption in The Netherlands, *Food Quality and Preference*, 9(3), 119-33.

Schlegelmilch, B.B., Bohlen, G.M. and Diamantopoulos, A. (1996). The link between green purchasing decisions and measures of environmental consciousness, *European Journal of Marketing*, 30(5), 35-55.

Schlossberg, H. (1991). Green marketing has been planted – now watch it grow, *Marketing News*, 4, 26-30.

Schwartz, J. and Miller, T. (1991). The earth's best friends, *American Demographics*, 13, February, 26-33.

Schwepker, C.H. and Cornwell, T.B. (1991). An examination of ecologically concerned consumers and their intention to purchase ecologically packaged products, *Journal of Public Policy & Marketing*, 10(2), 77-101.

Scott, D., Willits, F.K. (1994). Environmental attitudes and behavior: a Pennsylvania survey, *Environmental Behavior*, 26 (2), 239-60.

Sellerberg, Ann-Mari (1978). *Konsumtionens sociologi*, Stockholm: Esselte Studium, Scandinavian University Books.

Shabecoff, P. (1993). *A Fierce Green Fire: The American Environmental Movement*, New York, NY: Hill and Wang Publishers.

Shaw, M.J., Subramaniam, C., Tan, G.W. and Welge, M.E. (2001). Knowledge management and data mining for marketing, *Decision Support Systems*, 31(1), 127–137.

Shetzer, L., Stackman, R.W. and Moore, L.F. (1991). Business environment attitudes and the new environmental paradigm, *Journal of Environmental Education*, 22, 14-21.

Sheu, J.J., Su, Y.H. and Chu, K.T. (2009). Segmenting Online Game Customers - the Perspective of Experiential Marketing. *Expert Systems Applications*, 36, 8487–8495.

Simmel, G. (1904). Fashion, *International Quarterly*, 10, 130–155.

Simmons, D.H. and Widmar, R. (1990). Motivations and barriers to recycling: Toward a strategy for public education, *The journal of Environmental Education*, 22, 13-18.

Simon, H.A. (1969). The sciences of the artificial, *MIT*, Cambridge, MA.

Simula, O., Vesanto, J., Alhoniemi, E. and Hollmén, J. (1999). Neuro-Fuzzy Techniques for Intelligent Information Systems, chapter Analysis and Modeling of Complex Systems Using the Self-Organizing Map, *Physica Verlag* (Springer-Verlag), 3-22.

Smith, K. and Gupta, J. (2002). *Neural Networks in Business*, Hershey, PA: IDEA Group Publishing.

Soper, S. (2002). The evolution of segmentation methods in financial services: Where next?, *Journal of Financial Services Marketing*, 7(1), 67-74. Henry Stewart Publications.

Squires, L., Juric, B. and Cornwell, T.B. (2001). Level of market development and intensity of organic food consumption: cross-cultural study of Danish and New Zeland consumers, *Journal of Consumer Marketing*, 18(5), 392-409.

Stern, P.C., Dietz T. and Guagnano G.A. (1995). The new ecological paradigm in social–psychological context, *Environmental Behavior*, 27(6), 723-743.

Stern, P.C., Dietz, T. and Kalof, L. (1993). Value orientations, gender, and environmental concern, *Environment and Behavior*, 25(3), 322-348.

Stone, G.P. (1954). City Shoppers and Urban Identification: Observations on the Social Psychology of City Life, *American Journal of Sociology*, 60(1), 36-45.

Straughan, R.D. and Roberts, J.A. (1999). Environmental segmentation alternatives: a look at green consumer behavior in the new millennium, *Journal of Consumer Marketing*, 16(6), 558-575, MCB University Press.

Taylor, J., Roberts, J., Oram, C., Itani, M., and Pan, W. (2015) "Customer relationship management portal system and method" United States Patent Application Publication, Pub. No.: US 2015/0007168 A1.

Thomas, J. and Cook, K. (2006). Visualization viewpoints, *IEEE Computer Graphics and Applications*, Jan/Feb 2006, 10-13.

Thompson, G.D. and Kidwell, J. (1998). Explaining the choice of organic produce: cosmetic defects, prices and consumer preferences, *American Journal of Agricultural Economics*, 80(2), 277-287.

Tregear, A., Dent, J.B., McGregor, M.J. (1994). The Demand for Organically Grown Produce, *British Food Journal*, 96(4), 21–25.

Tsai, C. and Chiu, C. (2004). A purchase-based market segmentation methodology, *Expert Systems with Applications*, 27(2), 265–276.

Tsakiridou, E., Boutsouki, C., Zotos, Y., and Mattas, K. (2008). Attitudes and behavior towards organic products: an exploratory study, *International Journal of Retail & Distribution Management*, 36(2), 158-175. Emerald Group Publishing Limited.

Tsiptsis, K. and Chorianopoulos, A. (2011). *Data mining techniques in CRM: inside customer segmentation*, Wiley.

Ultsch, A. (1993). Self organized feature planes for monitoring and knowledge acquisition of a chemical process, In S. Gielen and B. Kappen (Eds.): *The International Conference on Artificial Neural Networks* (ICANN93), 864–867, London: Springer-Verlag.

Van Liere, K. and Dunlap, R. (1981). The social bases of environmental concern: a review of hypotheses, explanations, and empirical evidence, *Public Opinion Quarterly*, 44(2), 181-97.

Veblen, T. (1899). *The Theory of the Leisure Class: An Economic Study of Institutions*, USA: McMillan.

Venable, J., Pries-Heje, J., Baskerville, R. (2012). A Comprehensive Framework for Evaluation in Design Science Research, p. 423-438, in Peffers, K., Rothenberger, M., Kuechler, B. (Eds.) *Design Science Research in Information Systems Advances in Theory and Practice, Proceedings of the 7th International Conference*, DESRIST 2012 Las Vegas, NV, USA, May 14-15, 2012.

Verhoef, P.C., Langerak, F. (2002). Eleven misconceptions about customer relationship management, *Business Strategy Review*, 13(4), 70-76.

Verhoef, P.C., Spring, P.N., Hoekstra, J.C. and Leeflang, P.S.H. (2003). The commercial use of segmentation and predictive modelling techniques for database marketing in the Netherlands, *Decision Support Systems*, 34(4), 471-481.

Vesanto, J. (1999). SOM-based data visualization methods, *Intelligent data analysis*, 3(2), 111-126.

Vesanto, J. and Alhoniemi, E. (2000). Clustering of the Self-organizing Map, *IEEE Transaction on Neural Networks*, 11(3), 586-600.

Vesanto, J., Himberg, J., Alhoniemi, E. and Parhankangas, J. (2000). *SOM Toolbox for Matlab 5; Report A57*, Espoo, Finland: Libella Oy.

Vindigni, G., Janssen, M.A., Jager, W. (2002). Organic food consumption: A multi-theoretical framework of consumer decision making, *British Food Journal*, 104(8), 624–642.

Waaser, E., Dahneke, M., Pekkarinen, M. and Weissel, M. (2004). How you slice it: smarter segmentation for your sales force, *Harvard Business Review*, 82(3), 105-111.

Wandel, M. and Bugge, A. (1997). Environmental concern in consumer evaluation of food quality, *Food Quality and Preference*, 8(1), 19-26.

Ward, J.H. Jr. (1963). Hierarchical Grouping to Optimize an Objective Function, *Journal of the American Statistical Association*, 58, 236-244.

Webster, F.E. Jr (1975). Determining the characteristics of the socially conscious consume, *Journal of Consumer Research*, 2, 188-96.

Wedel, M. and Kamakura, W. (1999). *Market segmentation conceptual and methodological foundations*, Massachusetts, USA: Kluwer Academic Publishers.

Wedel, M. and Kamakura, W.A. (2000). *Market segmentation conceptual and methodological foundations*, 2nd ed., Boston: Kluwer.

Weigel, R. H. (1977). Ideological and demographic correlates of proecological behavior, *The Journal of Social Psychology*, 103, 39-47.

Wiener, J.L. and Doescher, T.A. (1991). A framework for promoting cooperation, *Journal of Marketing*, 55, 38-47.

Wicker, A.W. (1969). Attitudes versus actions: The relationship of verbal and overt behavioral responses to attitude objects, *Journal of Social Issues*, 25(4), 41-78.

Wilkie, W.L. and Cohen, J.B. (1977). An overview of market segmentation: Behavioral concepts and research approaches, *Working paper, report No. 77-105*, Cambridge, Massachusetts: Marketing Science Institute.

Wilson, H., Clark, M. and Smith, B. (2007). Justifying CRM projects in a business-to-business context: the potential of the Benefits Dependency Network, *Industrial Marketing Management*, 36(6), 770–783, Elsevier Inc.

Winski, J.M. (1991). Big prizes, but no easy answers, *Advertising Age*, October 26, GR-3.

Young, W., Hwang, K., McDonald, S., Oates, C.J. (2010). Sustainable Consumption: Green Consumer Behaviour when Purchasing Products. *Sustainable Development*, 18, pp. 20-31.

Yao, Z., Sarlin, P., Eklund, T. and Back, B. (2012). Temporal customer segmentation using Self-Organizing Time Map, In Proceedings of the *16th International Conference on Information Visualization* (IV 2012), Montpellier, France, 10-13 July, 234-240.

Zotos, Y., Ziamou, P. and Tsakiridou, E. (1999). Marketing organically produced food products in Greece: challenges and opportunities, *Greener Management International*, 25, 91-104.

Appendices

Appendix I: Survey instrument for collecting background information on the respondents (translated from Finnish)

Appendix II: Survey instrument for weak-form validation of information retrieved from a customer segmentation model (translated from Finnish)

Appendix I – Survey instrument for collecting background information

Part I – Background information

1. Background information on the respondent

1.1 Name:

1.2 Age class (in years): 18-29 30-39 40-49 50-64 65 +

1.3 Occupation:

1.4 Title:

1.5 Responsibilities / duties at work:

1.6 How many years have you been working in this or a similar position?

1.7 How many years have you been employed by Sokos?

1.8 Do you have earlier experience on IT-tools?

1.9 What IT-tools and methods are you using for your work?

1.10 Have you been employed by Sokos or been working at the women's department at Sokos during the years 2007-2009?

2. Distribution of information

2.1 Is information on customers and their shopping behavior distributed to you?

2.2 How is the information distributed (orally, e.g. at meetings, reports in writing, tables)?

2.3 How often is the information distributed (daily, weekly, monthly, or yearly)?

2.4 In your work, do you use information on customers retrieved from the retailer's database or your own expertise?

2.5 According to your experience, do you get enough information in your line of work?

2.6 Would you like to receive information on your customers more often?

2.7 What kind of information on your customers would you need in your line of work?

Part II – The present situation

3. Description of the customers

3.1 Please describe in your own words a typical customer that visits your department store daily. You can freely describe several different kind of typical customers. (E.g., Age, gender, families with children, something else?)

3.2 Does the same customers make purchases at your store often?

4. Customer shopping behavior

The number of products at one shopping visit:

4.1 How big a share of the customers purchases only one product at a shopping visit?

4.2 How big a share of the customers purchases two products at a shopping visit?

4.3 How big a share of the customers purchases three or more products at a shopping visit?

Which products do the customers combine at one shopping visit:

4.4 What is a typical product combination, i.e., which products are usually bought together? You can give many examples.

4.5 Are the products bought together often similar, e.g., two pairs of socks?

4.6 Are the products that are purchased together usually from the same area of the women's department, e.g., socks, underwear, shirts, trousers?

4.7 Are the products that are purchased together usually of the same brand, e.g., Esprit, Norlyn?

4.8 What product brands does the customer combine, i.e., which brands are purchased together?

From which departments are products purchased together at one shopping visit:

4.9 According to your experience, do the customers purchase products from other departments at one shopping visit?

4.10 From which departments do customers usually make purchases at one shopping visit, e.g., from women's, men's, or children's departments?

Appendix II – Survey instrument for weak-form validation of information retrieved from an MBA model and customer segmentation model

Part III – Presentation of the analyzes

Presentation of the analyzes using power point slide shows:
a) Market Basket Analysis (MBA) and b) Segmentation analysis.

Part IV – Evaluation of the usefulness of the analyzes

A) Market Basket Analysis

Please answer the statements by marking the suitable answer: 1 (strongly disagree) – 5 (strongly agree).

	1	2	3	4	5	
Statement	Strongly disagree	Disagree	Neutral	Agree	Strongly agree	Do not know

5. Content

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree	Do not know
5.1 The analysis gives important information.	1	2	3	4	5	x
5.2 The results of the analysis respond to my needs.	1	2	3	4	5	x
5.3 The analysis gives useful information.	1	2	3	4	5	x
5.4 The analysis gives new information.	1	2	3	4	5	x
5.5 The information extracted from the analysis is sufficient.	1	2	3	4	5	x

6. Accuracy

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree	Do not know
6.1 The results of the analysis are correct.	1	2	3	4	5	x
6.2 The results of the analysis are reliable.	1	2	3	4	5	x
6.3 I am satisfied with the accuracy of the analysis.	1	2	3	4	5	x

7. Format

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree	Do not know
7.1 The results of the analysis were visually clearly presented.	1	2	3	4	5	x
7.2 The results of the analysis are easily read.	1	2	3	4	5	x
7.3 The results of the analysis are easily understood.	1	2	3	4	5	x
7.4 Overall, I am satisfied with the format of the analysis.	1	2	3	4	5	x

8. Benefit and usefulness

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree	Do not know
8.1 The results of the analysis correlate well with my own understanding regarding the customers of the department store.	1	2	3	4	5	x
8.2 The results of the analysis were useful.	1	2	3	4	5	x
8.3 I can benefit from this kind of analysis in my work.	1	2	3	4	5	x

9. Future – open-ended questions

9.1 How would you see that the presented models could impact your work?

9.2 What information about customers and their shopping behavior do you feel is missing from the MBA model, i.e., how should the MBA model be improved in order to be even more useful?

9.3 How often would you like the MBA analysis to be updated?

9.4 In what form would you like the results to be presented to you?

B) Segmentation analysis

Please answer the statements by marking the suitable answer: 1 (strongly disagree) – 5 (strongly agree).

	1	2	3	4	5	
Statement	Strongly disagree	Disagree	Neutral	Agree	Strongly agree	Do not know

10. Content

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree	Do not know
10.1 The analysis gives important information.	1	2	3	4	5	x
10.2 The results of the analysis respond to my needs.	1	2	3	4	5	x
10.3 The analysis gives useful information.	1	2	3	4	5	x
10.4 The analysis gives new information.	1	2	3	4	5	x
10.5 The information extracted from the analysis is sufficient.	1	2	3	4	5	x

11. Accuracy

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree	Do not know
11.1 The results of the analysis are correct.	1	2	3	4	5	x
11.2 The results of the analysis are reliable.	1	2	3	4	5	x
11.3 I am satisfied with the accuracy of the analysis.	1	2	3	4	5	X

12. Format

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree	Do not know
12.1 The results of the analysis were visually clearly presented.	1	2	3	4	5	x
12.2 The results of the analysis are easily read.	1	2	3	4	5	x
12.3 The results of the analysis are easily understood.	1	2	3	4	5	x
12.4 Overall, I am satisfied with the format of the analysis.	1	2	3	4	5	x

13. Benefit and usefulness

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree	Do not know
13.1 The results of the analysis correlate well with my own understanding regarding the customers of the department store.	1	2	3	4	5	x
13.2 The results of the analysis were useful.	1	2	3	4	5	x
13.3 I can benefit from this kind of analysis in my work.	1	2	3	4	5	x

14. Future – open-ended questions

14.1 How would you see that the presented models could impact your work?

14.2 What information about customers and their shopping behavior do you feel is missing from the segmentation model, i.e., how should the segmentation model be improved in order to be even more useful?

14.3 How often would you like the segmentation analysis to be updated?

14.4 In what form would you like the results to be presented to you?

Part II

Original research papers

1

Publication 1

Yao, Z., Holmbom, A.H., Eklund, T., Back, B. (2010). Combining Unsupervised and Supervised Data Mining Techniques for Conducting Customer Portfolio Analysis, In: Petra Perner (Ed.), *Advances in data mining, applications and theoretical aspects, Lecture Notes in Computer Science 6171*, 292–307, Springer Berlin Heidelberg.

Publication 2

2

Holmbom, A.H., Eklund, T. and Back, B. (2011). Customer portfolio analysis using the SOM, *Int. J. Business Information Systems*, (8:4), pp.396–412.

Customer portfolio analysis using the SOM

Annika H. Holmbom*, Tomas Eklund and Barbro Back

Turku Centre for Computer Science (TUCS),
Department of Information Technologies,
Åbo Akademi University Turku, 20520, Finland
Fax: +358-2-2155557
E-mail: Annika.H.Holmbom@abo.fi
E-mail: Tomas.Eklund@abo.fi
E-mail: Barbro.Back@abo.fi
*Corresponding author

Abstract: In order to compete for profitable customers, companies are looking to add value using customer relationship management (CRM). One subset of CRM is customer segmentation, which is the process of dividing customers into groups based upon common features or needs. Segmentation methods can be used for customer portfolio analysis (CPA), the process of analysing the profitability of customers. The purpose of this paper is to illustrate how the self-organising map (SOM) can be used for CPA. We segment, identify and analyse a case organisation's profitable and unprofitable customers in order to gain knowledge for the organisation to develop its marketing strategies. The results are validated through cross and face validation. The SOM is able to segment the data in an innovative and reliable way and to provide new insights for decision makers.

Keywords: customer relationship management; CRM; customer portfolio analysis; CPA; data-driven market segmentation; self-organising map; SOM.

Reference to this paper should be made as follows: Holmbom, A.H., Eklund, T. and Back, B. (2011) 'Customer portfolio analysis using the SOM', *Int. J. Business Information Systems*, Vol. 8, No. 4, pp.396–412.

Biographical notes: Annika H. Holmbom is a PhD student in Information Systems at the Department of Information Technologies at Åbo Akademi University. Her main research interest is in customer segmentation using data mining methods.

Tomas Eklund is an Academy of Finland Postdoctoral Researcher and Docent at Åbo Akademi University. He received his PhD in Information Systems from Åbo Akademi University in 2004. His main research interests are in data and text mining for business intelligence and competitor analysis.

Barbro Back is a Professor in Accounting Information Systems at the Department of Information Technologies at Åbo Akademi University. Her main research interests are in the areas of data and text mining for business intelligence, ERP-systems and medical information systems.

1 Introduction

Customer relationship management (CRM) is an important topic of management today. The objective of CRM is to integrate sales, marketing and customer care service in order to add value for both the company and its customers (Heinrich, 2005; Datta, 1996; Chalmeta, 2006). CRM first emerged in 1993 and has developed rapidly in recent years thanks to advances in information technology (Buttle, 2004; Rygielski et al., 2002), to become the important function that it is in today's companies. One of the most important tasks within CRM is customer segmentation, the process of identifying and grouping customers with similar profiles or requirements (Lingras et al., 2005).

The key element in CRM and customer segmentation is overall information about customers. Today, data about customers are readily available through ERPs, corporate data warehouses, and the Internet. Data can also be purchased from external providers, which has become a category of business of its own (Buttle, 2004; Rygielski et al., 2002).

The problem is that the amount of information available for segmentation is huge and can be very challenging to deal with because of issues such as missing data, non-uniform distributions, errors, etc. The extraction of information from large databases is, therefore, often performed using data mining methods (Rygielski et al., 2002; Berry and Linoff, 2004; Berson et al., 2000; Famili et al., 1997; Shaw et al., 2001).

Customer portfolio analysis (CPA) is the process of dividing the customer base into profitable and unprofitable customers (Buttle, 2004). CPA is related to segmentation and can be seen as a subset of it. However, the purpose is different and many of the methods used are unique (Terho and Halinen, 2007). CPA can beneficially be applied to analyse segments identified using segmentation methods. There are a large variety of different methods available for CPA. However, although there is a great wealth of theoretical literature surrounding CPA available, very little literature appears to show how CPA is actually being used by companies today (Terho and Halinen, 2007). Much of the literature concerns mathematical optimisation models, such as portfolio theory (e.g., Turnbull, 1990) and customer lifetime value (e.g., Kim et al., 2006). These methods generally view the customer base in the same way as a portfolio of investments, to be managed using the same methods.

In this study, a data-driven exploratory CPA, based upon demographic data and coupled with product sales information, will be performed. The segmentation tool we will use is an unsupervised artificial neural network (ANN), the self-organising map (SOM).

This paper[1] is organised as follows. In Section 2, related literature will be presented. In Section 3, the theoretical background of CRM and CPA will be presented. The data used will also be explained, as well as the theory behind the SOM. In the following section the construction process will be detailed, and the results will be analysed and validated. The final section concludes the paper and presents implications and future research potential.

2 Related literature

As for CPA, there are a large variety of methods available for segmentation. Many commonly used segmentation methods belong to the family of clustering approaches. Most of the methods in this area are statistical tools, such as k-means clustering and

hierarchical clustering methods. When considering data mining approaches, ANNs represent the most commonly used approaches in the academic literature (Ngai et al., 2009). Other approaches, such as regression, sequence analysis, association analysis, and genetic algorithms are also employed. Other commonly used approaches are decision tree-related approaches (e.g., CHAID) and fuzzy clustering approaches (e.g., fuzzy C-means clustering, FCM).

Ngai et al. (2009) propose a classification framework for data mining applications within CRM applications. Their classification is based upon four categories depending upon the purpose of the application; *customer identification, customer attraction, customer retention, and customer development*. Customer identification is concerned with identifying new potential customers, as well as grouping existing customers into manageable segments (*segmentation*) and identifying customer value and potential (*target customer analysis*). Customer attraction concerns attracting new customers, including direct marketing approaches. Customer retention deals with efforts to maintain customers' loyalty, including loyalty programs and churn analysis. Customer development includes lifetime value analysis and other approaches aiming to maximise the value of a customer. The authors classify 87 different articles (appearing in top journals in the field) published between 2000 and 2006 based upon their framework, noting that most (57) dealt with customer retention.

There are many recent applications representative of the different classes proposed by Ngai et al. (2009) in the literature, indicating a still strong interest in data mining for CRM. For example, in customer identification, Romdhane et al. (2010) used FCM and back propagation ANNs to create customer segments usable for targeted marketing. Hu et al. (2008) used a back propagation ANN to segment long range communications customers based upon behavioural patterns. In customer attraction, Crone et al. (2006) apply a number of different techniques, including ANNs, decision trees and support vector machines (SVMs) to a direct marketing database, in order to test for the effect of pre-processing of data. McCarty and Hastak (2007) used RFM (recency, monetary and frequency) analysis, CHAID trees and logistic regression for direct marketing segmentation. In the area of customer retention, Hosseini et al. (2010) used K-means and RFM analysis for customer loyalty modelling. Wang et al. (2009) used decision trees for churn analysis of wireless customers.

Many studies are difficult to classify into just one group. For example, Chan (2008) used RFM analysis and the customer life time value (LTV) model, combined with genetic algorithms, to model and identify potential customers for marketing campaigns in the automobile industry. This study would represent an example of the combination of customer identification, customer attraction, and customer development.

In this study, we are interested in assessing the potential of existing customers, making it an example of customer identification, more specifically target customer analysis. While CPA is not included in the classification proposed by Ngai et al., for the purposes of this study it can be considered to be the same as target customer analysis.

The SOM has been widely applied in finance, economics, and marketing (Kohonen, 1998; Kaski et al., 1998; Oja et al., 2003). For example, the SOM has been used for financial benchmarking (Back et al., 1998; Eklund et al., 2003), macro-economic analysis (Kaski and Kohonen, 1996; Länsiluoto, 2007), and bankruptcy prediction (Martín-del-Brío and Serrano-Cinca, 1993; Kiviluoto, 1998; Back et al., 1995).

However, regardless of its obvious benefits, the SOM has not been very widely applied in customer segmentation tasks, specifically to CPA. Examples include

Rushmeier et al. (1997), who used the SOM to visualise demographic customer segments for marketing purposes, Vellido et al. (1999a), who used the SOM for demographic segmentation of online customers, Lee et al. (2004, 2005), who used the SOM for demographic segmentation of online gamers, and Lingras et al. (2005), who used the SOM for temporal analysis of supermarket customers during a period of 24 hours. Mo et al. (2010) use a two-stage SOM model for multi-region segmentation, based upon the principle that each region is specific in its segmentation basis, and thus generalised segmentation models must be generated from region-specific models. Yao et al. (2010) use a two-level approach to CPA by combining SOM-Ward clustering and decision trees to analyse customer profitability.

This study differs from the previous ones in that the SOM was here used for CPA, based upon demographic information as well as product sales information, and for multiple years of data. The work builds upon the research initiated in Holmbom (2007) and continued in Holmbom et al. (2008), in which a model for segmentation of customer data was built. The model was based upon customer data provided by a case company, and the SOM was used to construct the model.

3 Methodology

3.1 Customer relationship management

CRM, also called relationship marketing (Parvatiyar and Sheth, 2001), is a technology-enhanced way for a marketer to interact with customers in order to create cooperative and collaborative relationships, in other words partnerships, between the company and its customers (Parvatiyar and Sheth, 2001; Dibb 2001). In order to successfully lead a company in a market where growth has stagnated and there is a competition for valuable but demanding customers, an understanding of CRM is necessary (Heinrich, 2005; Datta, 1996).

The process of applying analytical tools to customer transaction data is called Analytical CRM (Chalmeta, 2006; Paas and Kuijlen, 2001). Typically, data-mining tools are used in this application (Buttle, 2004). One of the most common applications of analytical CRM is customer segmentation, i.e., the use of analytical tools to study customer data (Paas and Kuijlen, 2001). Customer segmentation is the process of grouping customers into subgroups (segments) with similar behaviour or needs, in order to better serve or target the customers (Buttle, 2004; Lingras et al., 2005). The identified segments can then be more effectively targeted with suitable marketing strategies (Frank et al., 1972; Wedel and Kamakura, 1999). Customer segmentation is also used to identify profitable and unprofitable customers in the customer base, as well as customer relationships with development potential.

There are two main bases for segmentation, i.e., demographic data, such as socioeconomic and lifestyle measures, and product-specific measures, such as product usage, customer brand attitudes, brand preferences, benefits sought and response sensitivity to different marketing campaigns. Demographic data are the most commonly used base for segmentation (Frank et al., 1972; Wedel and Kamakura, 1999; Tsai and Chiu, 2004). Segmentation can also be divided into two major groups based upon the approach used: market-driven and data-driven segmentation. Market-driven segmentation uses data to divide customers into segments. These segments are beforehand set

according to characteristics that describe a specified customer profile, e.g., one that has been determined to be profitable. Data-driven segmentation is performed on actual customer data, e.g., the shopping behaviour of a customer (Berson et al., 2000).

3.2 The data

The data were provided by a case company that sells products ranging from simple periodicals to advanced consulting services, to other companies (B2B). The company wanted to perform a CPA in order to determine which of its customers were profitable and worth developing its relationship with, and conversely, which customers were better let go of. In addition, the company wanted to determine which groups of customers purchased which products. Overall, the strategic goal was to create a tool to be used by the sales department in order to adjust company marketing practices, i.e., to determine suitable marketing effort levels for different, previously unknown, categories of customers.

The data were extracted from the case company's data warehouse and consist of data about customers and their purchasing behaviour. The customers are companies from different lines of business, e.g., service, construction, industrial, wholesale, and retail. The data originate from the customers' annual reports and the case company's own data warehouse. They contain descriptive categories, describing the attributes of the customers, as well as sales information concerning the major products. The variables were selected in collaboration with the case company, based upon which data:

1 the case company considered important demographic indicators

2 were available (and considered reliable and complete) through the case company's data warehouse

3 were suitable for analysis with the SOM (quantitative variables).

Based upon a pilot test of the data and a review in cooperation with the case company, a number of small and very large customers (appearing as outliers in the results) were removed, and some product categories with small and infrequent purchases were merged. The motivation for doing this was that the largest customers are already individually served by an own sales representative, and the smallest customers were usually one-time purchasers.

The descriptive categories consisted of:

- risk factor, which is an internally calculated measure of potential financial losses
- company age in years
- solvency, which was calculated from the financial statement
- turnover
- change in turnover (%), compared to the previous year
- balance sheet total, which serves as a measure of company size
- return on equity (ROE), which was calculated from the financial statement.

The product categories consisted of 18 different products, labelled Products A–R. The products are generally speaking information service products, where Product I

(a significant consulting service) is the most expensive product and Product L (a simple filtered data product) is the most inexpensive one. Product O stands for overall purchases of products and Product R for other products (one-time analyses and other products difficult to categorise). The data collected were for the period of 2002 to 2006.

The data set contained 1,841 customers, i.e., 9,205 rows of data. 12.8% of the customers had incomplete descriptive data, i.e., 3.6% of the data values were missing. The missing data were not considered a problem, as the SOM is able to deal with small amounts of missing data (Bigus, 1996).

3.3 The SOM

ANNs have been widely applied to various business problems (Vellido et al., 1999b; Smith and Gupta, 2002). ANNs are commonly divided into two main categories: supervised and unsupervised learning approaches (Haykin, 1999). Supervised networks learn patterns by using target outcomes, and are thus most often used for classification tasks, i.e., where classes are predetermined. Market-driven segmentation would be performed using supervised learning ANNs.

Unsupervised learning is used for exploratory analysis, clustering, and visualisation (Kohonen, 1998). Kohonen's SOM is the most commonly used unsupervised ANN. The SOM is a two-layer feed forward network, in which each neuron learns to recognise a specific input pattern (Kohonen, 2001). Each neuron is represented by a prototype vector, i.e., an n-dimensional weight vector. The algorithm is basically a two-step process; in the first step, the best matching neuron [best matching unit (BMU)] for an input data row is located on the map, and secondly, it and its surrounding neurons within a certain neighbourhood radius are tuned to better match (i.e., learn from) the input data, based upon a learning rate factor. The process is repeated until a certain stopping criterion is reached, for example, the training length. The result of the training process is a visual clustering that shows similarities and dissimilarities in the data (Kohonen, 2001).

Essentially, the SOM is a nonlinear projection technique that displays high-dimensional data on a two-dimensional grid, by preserving the relationships (or topology) in the data but not the actual distances (Deboeck and Kohonen, 1998). Commonly, the SOM is visualised using the U-matrix (unified distance matrix) of the map, which displays the Euclidean distances between neurons in shades of colour (Ultsch, 1993).

4 Constructing the model

Viscovery SOMine 4.0 (http://www.viscovery.net/) was used to train the maps in this study. SOMine is based upon the batch-SOM training algorithm (Kohonen, 2001) and also uses a stepwise increasing map size during the training process, which makes it a very efficient implementation of the SOM algorithm (Deboeck, 1998). In addition, SOMine is very user friendly and includes a number of advanced data pre-processing and analysis tools, such as automated clustering of the map based upon Ward's hierarchical clustering method.

The demographic data of the companies (risk factor, age, solvency, turnover, and change in turnover percent, balance sheet total, and ROE) were first used to create one

map. The results of the demographic segmentation were then matched with the sales information for each product.

Even though the SOM is fairly tolerant towards noisy or missing data (Bigus, 1996; Smith and Gupta, 2002), data pre-processing is an important part of the data-mining task. Pre-processing refers to the task of dealing with data quality issues such as missing, erroneous, or outlier data (Berson et al., 2000; Famili et al., 1997; Pyle, 1999; Hand et al., 2001). In this application, sigmoid (or logistic) transformation (Bishop, 1995) was used to deal with outlier data. The sigmoid transformation was used because it emphasises the centre input values while reducing the influence of extreme input values (Bishop, 1995; Larose, 2005). Normalisation was further used to scale the variables.

Generally speaking, the size of the map is dependent upon the purpose of the application. A large hexagonal map is good for visualisation (more accurate on the individual record level), whereas a small map is more suitable for clustering (squeezes data into a smaller number of groups) (Kohonen, 2001; Desmet, 2001). In this case, a map size of 700 nodes was selected as a balance between clustering and visualisation since the groups were not expected to be very homogeneous and we wanted to be able to accurately judge the intra-cluster differences. As the software uses the batch SOM algorithm, the learning rate does not need to be specified (Deboeck, 1998), and the only other parameter required is the tension. The tension is essentially a value for the neighbourhood radius in the final training stage, where a small tension results in high local detail (accuracy), while a high tension has an averaging (smoothing) effect on the map. In this case, the default value 0.5 (average) was used. The neighbourhood function used in SOMine is always Gaussian.

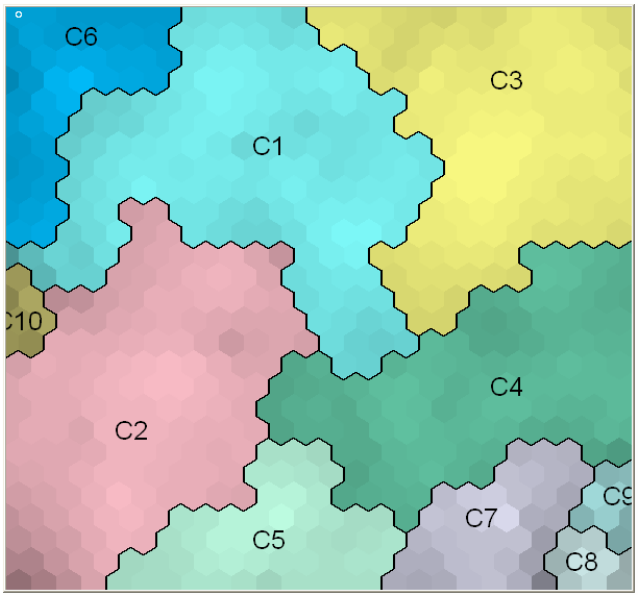
During the course of the experiment, hundreds of different maps were trained, tuning the parameters and data pre-processing approaches in an iterative manner. The final map was validated using 10 folds cross validation, i.e., the map was retrained ten times using 90% (9 folds) of the data using the same parameters, each time retaining a different 10% (one fold) of the data for testing. The differences in average quantisation error and normalised distortion[2] were very small between maps trained using the different folds, circa 1%. Therefore, the final map can be considered technically validated.

Although the U-matrix of the SOM can be manually interpreted to identify the clusters, two-stage clustering (Vesanto and Alhoniemi, 2000) is an easier and more objective method of identifying the clusters on the map. In two-stage clustering, the neurons on the map are clustered based upon their Euclidean distances, using a suitable clustering algorithm. In this case, Ward's hierarchical clustering method, included in the SOMine software, was used to identify the clusters on the final map. The final map is displayed in Figure 1. The clustering of the map resulted in ten clusters of various sizes, labelled C1–C10. The shade of the cluster only signifies cluster membership, and does not imply any value.

In order to interpret the map, and in particular the characteristics of each cluster, the component planes (displayed in Figure 2) of the map are used. The component planes show the distribution of values across the map, according to one variable at a time. The values according to one variable are displayed by the colour of the neuron, where 'warm' colours (red, orange, and yellow) illustrate high values and 'cool' colours (blue) illustrate low values[3]. The approximate values are indicated by the scale under each component plane. The map is interpreted by reading the component planes for each cluster. For example, Cluster 6 (upper left corner) displays medium to high values in solvency and ROE, and low values in age, turnover, and balance sheet total. Cluster 6 also shows

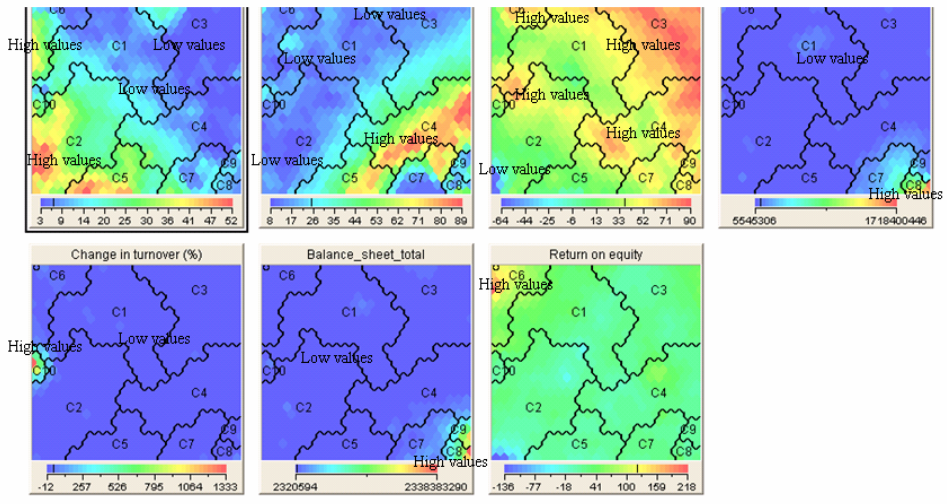
varying risk factors, from low to high. We can conclude that these are fairly young and small companies, although very profitable. We can also see that the risk factor is extremely large in segments C2 and C5, which means that these segments contain less reliable companies. The oldest companies are found in segments C4, C5 and C7, and to some extent, in C3.

Figure 1 A U-matrix of the segmentation results, using the customer data from years 2002–2006 (see online version for colours)



Note: The map was created according to the variables from the descriptive categories.

Figure 2 The component planes of the map, showing the values according to one component at a time (see online version for colours)



5 Analysis of the results

The results of the customer segmentation are summarised in Table 1, which shows the sizes and distinguishing features of each of the clusters. Clusters C1, C2 and C3 are the largest ones.

Table 1 Division of customers into clusters according to the segmentation presented in Figure 1

Clusters	*Customers*	*%*	*Distinguishing feature(s)*	*Purchased products*
Cluster 1	1,659	20.68	No specific attribute	A–R, except M
Cluster 2	1,689	21.05	High risk factor	A–R, except M
Cluster 3	1,699	21.17	Highest solvency	A–R, except M
Cluster 4	1,139	14.19	Oldest companies, high solvency	A–R, except M
Cluster 5	640	7.98	Large companies, high solvency	A–R, except M and Q
Cluster 6	565	7.04	High risk factor, good solvency, very high profitability	A–R, except M
Cluster 7	398	4.96	Both old and young companies, good turnover	A–R, except M
Cluster 8	91	1.13	Largest turnover, large balance sheet total	A–R, except M and Q
Cluster 9	80	1.00	Large balance sheet total	A–R, except L and M
Cluster 10	64	0.80	Largest change in turnover (%)	A–R, except M and Q

The clusters are identified as follows:

- Cluster 1: an average group with no specifically identifying characteristics. Risk factor, age, turnover, and balance sheet total are low, and solvency is medium to high. Return on equity is good on average. One of the three largest groups in terms of number of customers.
- Cluster 2: exhibits a considerably higher risk factor, lower solvency, and lower return on equity than in Cluster 1. This group also contains the customers with the lowest return on equity, as well as the companies with the lowest solvency. One of the three largest groups in terms of number of customers.
- Cluster 3: similar to Cluster 1 except for a considerably higher solvency. The average company age is somewhat higher, although risk factor seems to be similar to that of Cluster 1. One of the three largest groups in terms of number of customers.

- Cluster 4: contains the oldest companies in the dataset, and generally exhibits a high solvency and good profitability.
- Cluster 5: is a mid-size cluster containing larger than average companies. Solvency is good, and the companies are fairly new. Profitability is average.
- Cluster 6: a mid-size cluster that contains the most profitable companies in the dataset. In general small, growing companies. Nearly half of the cluster displays a very high risk factor, but some companies are also very solvent.
- Cluster 7: a mid-size cluster that contains fairly large companies in terms of turnover and total assets. Solvency is good to average, and the cluster contains a mix of old and new companies.
- Cluster 8: is one of the three small clusters and contains the largest companies in terms of assets and turnover. The companies are solvent and fairly profitable, and their risk factor is very low. Company age is above average.
- Cluster 9: is another small group of large companies. This cluster differs from Cluster 8 in that the companies are newer and turnover is lower.
- Cluster 10: is the smallest and final cluster identified. It contains rapidly growing companies that are fairly profitable and solvent, and have a fairly low risk factor.

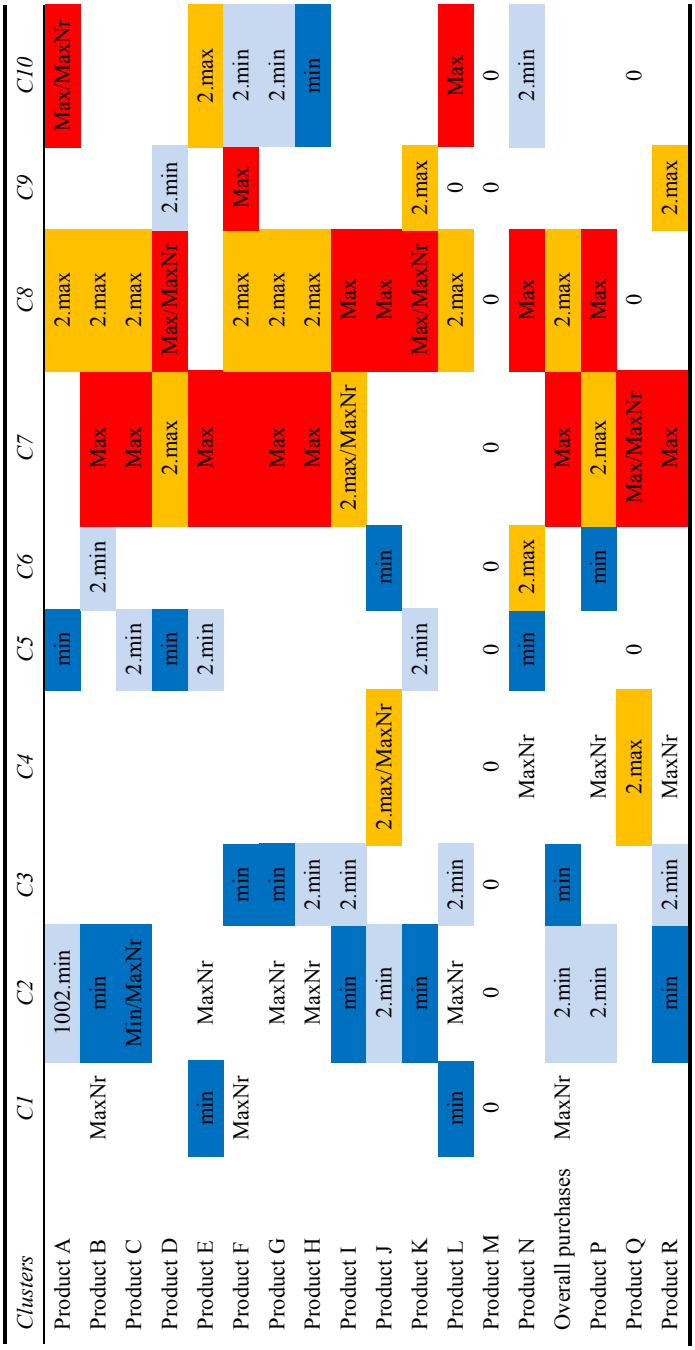
After the clusters were identified, the next step was to compare the sales information for each product category to the created segments. The full table can be found in Table 2. Table 2 visualises how the clusters can be divided into groups consisting of the largest, average, and smallest customers according to the sales information. For the experiment, we did not have information on what the profit margin is for every product. Therefore, by profitable customers we mean customers who have the highest total purchase amounts. For our case company, these customers might not be the same as the most profitable ones.

In Table 2, the two clusters with the highest average purchases (Max, 2.max) of a particular product, as well as the two clusters with the smallest average purchases (Min, 2.min) are marked with shaded boxes[4]. The comparison was made according to average cluster sales. Also, the cluster in which the company that purchased the most (measured in Euros) is located was marked for each of the products separately (MaxNr).

The Product M was purchased only a few times. Therefore, no statistical values could be calculated and no comparison to the other products could be made. Product L was not purchased by any of the customers in Cluster C9 and Product Q was not purchased by any of the customers in Clusters C5, C8 and C10. These have been marked with the value of zero.

The analysis presented in Table 2 tells us that customers located in Clusters C4, C7–C10 purchase a lot of products, while the customers located in segments C1–C3 and C5–C6 do not. If the amount of sales work expended for each of the segments is the same, the division of the customer segments can be extended to describe profitable, average, and non-profitable customers.

Table 2 Comparison of the sales of the product categories against the created clusters (see online version for colours)



Clusters	*C1*	*C2*	*C3*	*C4*	*C5*	*C6*	*C7*	*C8*	*C9*	*C10*
Product A		1002.min			min			2.max		Max/MaxNr
Product B	MaxNr	min				2.min	Max	2.max		
Product C		Min/MaxNr			2.min		Max	2.max		
Product D					min		2.max	Max/MaxNr	2.min	
Product E	min	MaxNr			2.min		Max			2.max
Product F	MaxNr		min					2.max	Max	2.min
Product G		MaxNr	min				Max	2.max		2.min
Product H		MaxNr	2.min				Max	2.max		min
Product I		min	2.min				2.max/MaxNr	Max		
Product J		2.min		2.max/MaxNr		min		Max		
Product K		min			2.min			Max/MaxNr	2.max	
Product L	min	MaxNr	2.min					2.max	0	Max
Product M	0	0	0	0	0	0	0	0	0	0
Product N				MaxNr	min	2.max		Max		2.min
Overall purchases	MaxNr	2.min	min				Max	2.max		
Product P		2.min		MaxNr		min	2.max	Max		
Product Q				2.max	0		Max/MaxNr	0	0	
Product R		min	2.min	MaxNr			Max		2.max	

Notes: The table indicates the cluster that is the largest (Max), second largest (2.Max), smallest (min) and second smallest (2.min) purchaser of a particular product, in terms of average cluster sales. MaxNr indicates the cluster in which the largest single purchaser of the product can be found.

5.1 Largest customers in terms of sales

The largest customers are located in Clusters C7 and C8. 6.1% of the companies in the customer base belong to this group. These companies have the largest average sales figures for nearly every product. They stand for 22.4% of the total purchases. In the SOM map presented in Figure 1, these clusters are located in the lower right corner of the map. There are both young and old companies that possess a large turnover in Cluster C7. Some of the younger companies in this segment have a slightly higher risk factor, and some of the customers have a high solvency. The large balance sheet total for the customers in Cluster C8 is a sign that these companies are large in size. These companies have the largest turnover and a high solvency.

5.2 Average customers in terms of sales

According to the segmentation model, the average customers in terms of sales are located in Clusters C1, C4, C6, C9 and C10. They stand for 42.5% of the total purchases. They constitute 43.7% of the total customer base. The companies in Cluster C1 do not have a dominating descriptive component. A small part of the customers in this cluster display an increased risk factor, another part is slightly older, and a third part has a good solvency. The oldest customers are located in Cluster C4. These have a high solvency, and some of them have an increased risk factor. The companies in Cluster C6 have a high risk factor, but they possess a high solvency and the largest return on equity. The companies in Cluster C9 are large in size, as they have a large balance sheet total. The customers in Cluster C10 are growing companies. Also, the specific company that makes the highest overall purchases belongs to this cluster.

The order of priority for these average clusters, according to the information gained from this segmentation, would be as follows: Clusters C10 and C9 (made most purchases), Cluster C4 (average) and Clusters C1 and C6 (made least purchases). According to the model, Cluster C9 is very similar to the clusters with the companies who conducted most purchases, i.e., Clusters C7 and C8. This would indicate the possibility that future good customers could be found in Cluster C9. Similarly, the poorest customers in this group are located in Cluster C6, which is very different from the best performing clusters.

5.3 Poorest customers in terms of sales

The companies in Clusters C2, C3 and C5 are the poorest customers, i.e., they purchase the least amount of products. They stand for 35.1% of the total purchases. Their share of the customer base is 50.3% and these customers, therefore, constitute the largest group. According to the model, the companies in Cluster C2 have a high risk factor. However, many of the single companies that have purchased the largest amounts of a specific product are located in this cluster. The customers in Cluster C3 have the highest solvency. A small share of these companies has a slightly increased risk factor, and another share is slightly older. The companies in Cluster C5 are older with a high risk factor. A common factor amongst the poorest customers is the high risk factor, indicating that dealing with them implies a higher risk of default than with most other customers.

6 Validation of results

The resulting model was face validated by two experts from the sales department of the case organisation. The two experts each have several years' experience, one a head of sales and the other as development manager. Therefore, the two managers have a strong knowledge of the customer base of the company. During the discussions with these experts, several findings appearing in the results of the analysis provided evidence to support intuitive ideas that the managers had thought of but never been able to show conclusively. New information was also presented to the experts, primarily regarding where to look for new profitable customers.

From the perspective of segmentation-based CPA, the SOM has several advantages. Compared to mathematical optimisation methods and most statistical approaches, the main advantage of the SOM is that it is a highly visual method. This makes it simple to present and explain results to business decision makers. Also, judging the results is more intuitive for a non-mathematically inclined audience. The SOM is also very robust, requiring very little pre-processing of the data, and unlike most statistical approaches, is non-parametric. The SOM is an explorative tool, meaning that very little *a priori* knowledge is required, and it is possible to uncover unexpected patterns in data. Decision trees are simple to use and highly visual approaches, but correctly deciding the split lines is imperative (Pyle, 1999), and they are unsuitable for exploratory analyses where no predefined classes exist. Regression approaches and classification-based neural networks are also unable to deal with data when predefined classes are not available.

7 Conclusions and future research

In this paper, the use of the SOM for CPA has been illustrated. A customer segmentation based upon demographic data was performed using the SOM, identifying ten clusters of customers displaying different demographic characteristics. The resulting clusters were then coupled with sales data, and a CPA was performed in order to identify profitable and unprofitable customers. The resulting model was face validated by experts from the sales department of the case organisation.

7.1 Contribution and implications

This study contributes to the literature and practise in both segmentation and the SOM by providing a real world example of how the SOM can be used to perform CPA. The example demonstrates that the SOM is a visual and easy to understand exploratory tool for the purpose. In terms of implications, the face validation showed that the case company's sales department could potentially develop its marketing strategies based on the results of this work, based upon the new information about the customer base that the model provides.

7.2 Limitations of the research

There are a number of limitations with this research. Firstly, we were restricted to the data available in the case company's data warehouse, i.e., no external data were used. Of course, any small errors inherent in the data warehouse data would also have affected the

result of this analysis, although effort was made to identify erroneous or outlier data. Secondly, we were forced to leave out the largest and smallest customers when training the model, as these were very different from the customer base in general. However, the largest customers are usually individually served by their own customer service contact and the smallest customers' usually only make one-time purchases. Thus, discluding these customers from the segmentation model can be motivated. Thirdly, as a method, the SOM performs dimensional reduction, squeezing several dimensions onto a two dimensional plane. This invariably leads to a certain loss of accuracy. Finally, the model was trained on the data available for a specific period of time, i.e., no longitudinal comparison has been made.

7.3 Future research directions

There are several interesting topics of research that should be pursued in the future. Firstly, the model cannot identify a universal demographic feature or set of features that can predict customer profitability, although customer size gives an indication of purchase potential. Further research should be conducted to see if the addition of other demographic features could increase the predictive performance of the model. Secondly, predicting purchase potential is potentially a valuable addition to the model. This could be done using statistical models. Thirdly, predicting the level of effort required to push a customer to a more profitable level of relationship should be researched, e.g., using Markov chain analysis. Finally, market basket analysis was preliminarily performed in Holmbom (2007), and should be further developed.

Acknowledgements

The authors would like to thank the case organisation for its participation in the study. The authors also gratefully acknowledge the financial support of the National Agency of Technology (Titan, grant no. 40063/08) and the Academy of Finland (grant no. 127656). The constructive and helpful comments of the anonymous reviewers of this paper are also gratefully acknowledged. Finally, we are thankful for the helpful comments of Prof. Dr. Christer Carlsson concerning the early stages of this article.

References

Back, B., Oosterom, G., Sere, K. and Van Wezel, M. (1995) 'Intelligent information systems within business: bankruptcy predictions using neural networks', in G. Doukidis, R.D. Galliers, T. Jelassi, H. Kremer and F.F. Land (Eds.): *The 3rd European Conference on Information Systems (ECIS'95)*, pp.99–111, June 1–3.

Back, B., Sere, K. and Vanharanta, H. (1998) 'Managing complexity in large data bases using self-organizing maps', *Accounting Management and Information Technologies*, Vol. 8, No. 4, pp.191–210.

Berry, M.J.A. and Linoff, G.S. (2004) *Data Mining Techniques: For marketing, Sales, and Customer Relationship Management,* 2nd ed., Wiley Publishing Inc., Indianapolis, Indiana.

Berson, A., Smith, S. and Thearling, K. (2000) *Building Data Mining Applications for CRM,* McGraw-Hill Companies Inc., USA.

Bigus, J.P. (1996) *Data mining with Neural Networks: Solving Business Problems from Application Development to Decision Support,* The McGraw-Hill Companies Inc., New York, NY.

Bishop, C.M. (1995) *Neural Networks for Pattern Recognition*, Oxford University Press, Avon.

Buttle, F. (2004) *Customer Relationship Management Concepts and Tools*, Butterworth-Heinemann, Oxford.

Chalmeta, R. (2006) 'Methodology for customer relationship management', *The Journal of Systems and Software*, Vol. 79, No. 7, pp.1015–1024.

Chan, C.C.H. (2008) 'Intelligent value-based customer segmentation method for campaign management: a case study of automobile retailer', *Expert Systems with Applications*, Vol. 34, No. 4, pp.2754–2762.

Crone, S.F., Lessmann, S. and Stahlbock, R. (2006) 'The impact of preprocessing on data mining: an evaluation of classifier sensitivity in direct marketing', *European Journal of Operational Research*, Vol. 173, No. 3, pp.781–800.

Datta, Y. (1996) 'Market segmentation: an integrated framework', *Long Range Planning*, Vol. 29, No. 6, pp.797–811.

Deboeck, G.J. (1998) 'Software tools for self-organizing maps', in G.J. Deboeck and T. Kohonen (Eds.): *Visual Explorations in Finance Using Self-Organizing Maps*, pp.179-194, Springer-Verlag, Berlin.

Deboeck, G.J. and Kohonen, T. (1998) *Visual Explorations in Finance with Self-Organizing Maps*, Springer-Verlag, Berlin.

Desmet, P. (2001) 'Buying behavior study with basket analysis: pre-clustering with a Kohonen map', *European Journal of Economic and Social Systems*, Vol. 15, No. 2, pp.17–30.

Dibb, S. (2001) 'New millennium, new segments: moving towards the segment of one?', *Journal of Strategic Marketing*, Vol. 9, No. 3, pp.193–213.

Eklund, T., Back, B., Vanharanta, H. and Visa, A. (2003) 'Using the self-organizing map as a visualization tool in financial benchmarking', *Information Visualization*, Vol. 2, No. 3, pp.171–181.

Famili, A., Shen, W., Weber, R. and Simoudis, E. (1997) 'Data preprocessing and intelligent data analysis', *Intelligent Data Analysis*, Vol. 1, No. 1, pp.3–23.

Frank, R.E., Massy, W.F. and Wind, Y. (1972) *Market Segmentation*, Prentice-hall Inc., Englewood Cliffs, New Jersey.

Hand, D.J., Mannila, H. and Smyth, P. (2001) *Principles of Data Mining. Boston*, MIT Press, USA.

Haykin, S. (1999) *Neural Networks – A Comprehensive Foundation*, Prentice Hall International, Inc., Upper Saddle River, NJ.

Heinrich, B. (2005) 'Transforming strategic goals of CRM into process goals and activities', *Business Process Management Journal*, Vol. 11, No. 6, pp.709–723.

Holmbom, A.H. (2007) 'Identifying customer segments using the self-organizing map', Unpublished Master's Thesis, Abo Akademi University, Turku, Finland.

Holmbom, A.H., Eklund, T. and Back, B. (2008) 'Customer portfolio analysis using the SOM', in *Proceedings of 19th Australasian Conference on Information Systems (ACIS 2008)*, December 3–5, pp.412–422.

Hosseini, S.M.S., Malekia, A. and Gholamiana, M.R. (2010) 'Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty', *Expert Systems with Applications*, Vol. 37, No. 7, pp.5259–5264.

Hu, M.Y., Shanker, M., Zhang, G.P. and Hung, M.S. (2008) 'Modeling consumer situational choice of long distance communication with neural networks', *Decision Support Systems*, Vol. 44, No. 4, pp.899–908.

Kaski, S. and Kohonen, T. (1996) 'Exploratory data analysis by the self-organizing map: structures of welfare and poverty in the world', in P.N.R. Apostolos, Y. Abu-Mostafa, J. Moody and A. Weigend (Eds.): *Neural Networks in Financial Engineering*, pp.498–507, World Scientific, Singapore.

Kaski, S., Kangas, J. and Kohonen, T. (1998) 'Bibliography of self-organizing map (SOM) papers 1981–1997', *Neural Computing Surveys*, Vol. 1, pp.102–350.

Kim, S., Jung, T., Suh, E. and Hwang, H. (2006) 'Customer segmentation and strategy development based on customer lifetime value: a case study', *Expert Systems with Applications*, Vol. 31, No. 1, pp.101–107.

Kiviluoto, K. (1998) 'Predicting bankruptcies with the self-organizing map', *Neurocomputing*, Vol. 21, Nos. 1–3, pp.191–201.

Kohonen, T. (1998) 'The SOM methodology', in G.J. Deboeck and T. Kohonen (Eds.): *Visual Explorations in Finance: with Self-Organizing Maps*, pp.159–167, Springer Verlag, Berlin.

Kohonen, T. (2001) *Self-Organizing Maps*, Springer-Verlag, Berlin.

Länsiluoto, A. (2007) 'Suitability of self-organising maps for analysing a macro-environment – an empirical field survey', *International Journal of Business Information Systems*, Vol. 2, No. 2, pp.149–161.

Larose, D.T. (2005) *Discovering Knowledge in Data. An Introduction to Data Mining*, John Wiley & Sons Inc., Hoboken, NJ.

Lee, S.C., Suh, Y.H., Kim, J.K. and Lee, K.J. (2004) 'A cross-national market segmentation of online game industry using SOM', *Expert Systems with Applications*, Vol. 27, No. 4, pp.559–570.

Lee, S.C., Xiang, J.Y. and Jing, L.B. (2005) 'Who are the target customers in Chinese online game market? Segmentation with a two-step approach', in *Proceedings of the 2005 IEEE International Conference on e-Business Engineering*, IEEE, pp.736–743.

Lingras, P., Hogo, M., Snorek, M. and West, C. (2005) 'Temporal analysis of clusters of supermarket customers: conventional versus interval set approach', *Information Sciences*, Vol. 172, Nos. 1–2, pp.215–240.

Martín-del-Brío, B. and Serrano-Cinca, C. (1993) 'Self-organizing neural networks for the analysis and representation of data: some financial cases', *Neural Computing and Applications*, Vol. 1, No. 2, pp.193–206.

McCarty, J.A. and Hastak, M. (2007) 'Segmentation approaches in data-mining: a comparison of RFM, CHAID, and logistic regression', *Journal of Business Research*, Vol. 60, No. 6, pp.656–662.

Mo, J., Kiang, M.Y., Zou, P. and Li, Y. (2010) 'A two-stage clustering approach for multi-region segmentation', *Expert Systems with Applications*, Vol. 37, No. 10, pp.7120–7131.

Ngai, E.W.T., Xiub, L. and Chaua, D.C.K. (2009) 'Application of data mining techniques in customer relationship management: a literature review and classification', *Expert Systems with Applications*, Vol. 36, No. 2, pp.2592–2602.

Oja, M., Kaski, S. and Kohonen, T. (2003) 'Bibliography of self-organizing map (SOM) papers: 1998–2001 addendum', *Neural Computing Surveys*, Vol. 3, pp.1–156.

Paas, L. and Kuijlen, T. (2001) 'Towards a general definition of customer relationship management', *Journal of Database Marketing*, Vol. 9, No. 1, pp.51–60.

Parvatiyar, A. and Sheth, J.N. (2001) 'Customer relationship management: emerging practice, process and discipline', *Journal of Economic and Social Research*, Vol. 3, No. 2, pp.1–34.

Pyle, D. (1999) *Data Preparation for Data Mining*, Academic Press, San Diego, CA.

Romdhane, L.B., Fadhel, N. and Ayeb, B. (2010) 'An efficient approach for building customer profiles from business data', *Expert Systems with Applications*, Vol. 37, No. 2, pp.1573–1585.

Rushmeier, H., Lawrence, R. and Almasi, G. (1997) 'Case study: visualizing customer segmentations produced by self organizing maps', Eighth IEEE Visualization 1997 (VIS'97), pp.463–466.

Rygielski, C., Wang, J. and Yen, D.C. (2002) 'Data mining techniques for customer relationship management', *Technology in Society*, Vol. 24, No. 4, pp.483–502.

Shaw, M.J., Subramaniam, C., Tan, G.W. and Welge, M.E. (2001) 'Knowledge management and data mining for marketing', *Decision Support Systems*, Vol. 31, No. 1, pp.127–137.

Smith, K. and Gupta, J. (2002) *Neural Networks in Business*, IDEA Group Publishing, Hershey, PA.

Terho, H. and Halinen, A. (2007) 'Customer portfolio analysis practices in different exchange contexts', *Journal of Business Research*, Vol. 60, No. 7, pp.720–730.

Tsai, C. and Chiu, C. (2004) 'A purchase-based market segmentation methodology', *Expert Systems with Applications*, Vol. 27, No. 2, pp.265–276.

Turnbull, P.W. (1990) 'A review of portfolio planning models for industrial marketing and purchasing management', *European Journal of Marketing*, Vol. 24, No. 3, pp.7–22.

Ultsch, A. (1993) 'Self organized feature planes for monitoring and knowledge acquisition of a chemical process', S. Gielen and B. Kappen (Eds.): *The International Conference on Artificial Neural Networks (ICANN93)*, pp.864–867, Springer-Verlag, London.

Vellido, A., Lisboa, P.J.G. and Meehan, K. (1999a) 'Segmentation of the on-line shopping market using neural networks', *Expert Systems with Applications*, Vol. 17, No. 4, pp.303–314.

Vellido, A., Lisboa, P.J.G. and Vaughan, J. (1999b) 'Neural networks in business: a survey of applications (1992–1998)', *Expert Systems with Applications*, Vol. 17, No. 1, pp.51–70.

Vesanto, J. and Alhoniemi, E. (2000) 'Clustering of the self-organizing map', *IEEE Transactions on Neural Networks*, Vol. 11, No. 3, pp.586–600.

Wang, Y., Chiang, D., Hsu, M., Lin, C. and Lin, I. (2009) 'A recommender system to avoid customer churn: a case study', *Expert Systems with Applications*, Vol. 36, No. 4, pp.8071–8075.

Wedel, M. and Kamakura, W. (1999) 'Market segmentation conceptual and methodological foundations', Kluwer Academic Publishers, Massachusetts, USA.

Yao, Z., Holmbom, A.H., Eklund, T. and Back, B. (2010) 'Combining unsupervised and supervised data mining techniques for conducting customer portfolio analysis', *Advances in Data Mining. Applications and Theoretical Aspects*, Springer, Heidelberg, pp.292–307.

Notes

1 This paper is an extended version of a paper presented at the 19th Australasian Conference on Information Systems (ACIS, 2008).

2 The average quantisation error is a measure of the fit of the nodes to the data, while the normalised distortion is a measure of how well the map has retained the topological properties of the data. See Kohonen (2001) for details.

3 Due to reproduction reasons, the figures are displayed in greyscale. Approximate values are indicated by labels in the figures.

4 Due to reproduction reasons, colours have again been omitted. The clusters making the largest average purchases of a particular product were originally marked in shades of red, while the opposites were marked in shades of blue.

Publication 3

Vanharanta, H., Magnusson, C., Ingman, K., Holmbom, A.H., and Kantola, J. (2012). Strategic Knowledge Services. In Kantola, J. and Karwowski, W. (Eds.) *Knowledge Service Engineering Handbook*. CRC Press, Taylor and Francis Group.

21 Strategic Knowledge Services

Hannu Vanharanta, Camilla Magnusson, Kari Ingman, Annika H. Holmbom, and Jussi I. Kantola

CONTENTS

21.1 INTRODUCTION

With the development of the knowledge society, we have entered the postindustrial era of the service economy. Knowledge is strategic in nature and can lead to significant competitive advantage and added value if it is understood and used correctly in business situations. Explicit knowledge comes from data and information. Applicable explicit and tacit knowledge, in turn, is created through people and novel tools. The question is how we can manage all these system inputs—including data warehouses, information flows, knowledge creation and perception, and situation-specific life and business information—and whether it is even possible to create more knowledge using the capabilities of the organization and modern technology. The service economy uses technology to meet these challenges and requirements, as well as to reach clients, organizations, and groups to offer them value-added services. Service, in broad terms, refers to the action of helping or doing work for someone else (cf. Oxford English Dictionary 2011); the concept is apt for the knowledge society and integral in knowledge services.

In this chapter, we examine the strategic knowledge services that the company can use to understand the recommended response and action in different business situations in order to continuously create added value for its customers. We first introduce the continuous strategy ontology, which shows the main dimensions, constructs, and variables of strategic knowledge services in different organizations. We then present three case studies: the first is a case study in customer segmentation with the self-organizing map (SOM) (from data to knowledge service), the second case is a collocation network-based knowledge service (from information to knowledge service), and the third case is a collective human knowledge-based service using an ontology (from knowledge to knowledge service).

21.2 CONTEXT OF STRATEGY MAKING AND STRATEGIC KNOWLEDGE SERVICES

Strategy makers try to predict and influence the future state of the organization. They often work together with their executives and the decision makers, as the work involves linking the organization's goals with the external environment. Success in strategy-making activity requires that all the people involved arrive at a shared vision as the basis for progression. Often though, the process of strategy making is fraught with misunderstanding and conflict, especially at its outset. Participants may

lack a shared understanding of the future path of the organization; they may not have a common view of the operating environment or may lack a shared appreciation of the overlying world structure. On top of this, different participants perceive managerial issues and organizational characteristics in particular in different ways, and, thus, the strategy-making process becomes even more complex and problematic. This is true of any organization, public or private, because frequently the decision makers themselves do not understand the organization's structure and variables or information and knowledge well enough.

Faced with these problems, strategy makers often express the need for comprehensive, reliable, and commonly assimilable information that they can use to analyze and synthesize the current performance of their organization and to estimate its future potential. The software industry has tried to offer knowledge and information storage systems for this purpose for many years, yet the computer-based executive support systems (ESS) and decision support systems (DSS) developed so far have only provided partial solutions. Such systems support either specific activities or specific processes, but they do not provide executives with the kind of support that would enable a collective, holistic understanding of the organizational situation and also of the relationships and interrelationships that must be mastered in strategic management, and especially in sustainable development nowadays.

Prior-art ESS and DSS lack a statement of direction that would serve as a general framework for strategic management and a content-specific construct as a generic ontology for support systems. They also lack a basic strategic construct for guiding the organization toward purposeful progress and integrating its internal and external worlds. However, rapid advances in research and technology have propelled us into a new era where it is possible to incorporate also "soft" and "unstructured" abstract concepts, like those encountered in strategic management generally, into computer-based working environments. We can therefore demand more now that strategy making and planning can benefit from the new generation of computerized applications. Of course, strategy making this way requires new ways of thinking, as well as novel technological software and approaches.

These new ontology and content-oriented approaches to support systems development will take pride of place over prior-art activity and process-oriented approaches. New theories and methodologies, as well as technology, will facilitate the integration of the organization's internal world with its external world. This will definitively and radically influence ESS and DSS design and thinking to the benefit of strategic management and leadership. Computers will then become integrative tools of communication within members of the organization, and they will become more collective in nature than specific. This, in turn, will advance the use of computer applications in strategic decision-making work; those applications require more from management. The tools will enhance the use of data acquisition and knowledge creation in strategic management in a two-way or three-way fashion, that is, top-down, bottom-up, and middle-top-down, as previously referred to in strategic management literature.

To reach these goals and objectives in practice, software technology must be developed so that the executive support and decision support structures are coherent with the real structure of the world. The system architecture must emulate the reality

of the organizational and individual behavior and allow the strategy makers to gain a holistic view of the organization's activities, its operating environment, and its external structure. Metaphors, as tools, can assist us in achieving these aims, and ontologies help us to understand more through different perceptions of the same content.

For several years, we have worked to develop ESS and DSS that will enhance actual decision making through visual perception, as well as through textual meanings and meta-knowledge formation. These new support systems are based on the ontologies of various organizational constructs and on the overall conceptual framework the continuous strategy. The basic principles of the continuous strategy ontology are described in the following chapter. With this ontology, we have good opportunities to fix the case research and applications to the dimensions, contents, and concepts of it. The resulting data can help executives in their daily tasks and turn their activities into strategic knowledge services. Through case studies, we want to show how important and meaningful it is for decision makers to process data, information, and knowledge from many viewpoints. Presentation of the data, constructs, and variables in a visual form offers new possibilities to improve decision making in strategic management.

21.3 CONTINUOUS STRATEGY ONTOLOGY

The continuous strategy ontology is a framework for strategic planning processes. The framework is built on metaphorical insights into the "company"—an "organism" seen as part of the living system. The continuous strategy ontology is supported by a chain of construction metaphors: the external world metaphor, the business world metaphor, the company world metaphor, the product world metaphor, and the buyer world metaphor. These represent conceptual models, and they are used to construct a coherent picture of the real world that exists in and around the company.

The metaphors described here are designed to facilitate the holistic understanding of management issues, business interrelationships, and company characteristics, so that they may be better recognized and understood in strategic management. The metaphors enhance the visual perception of decision makers and help them to arrive at a shared strategic vision for the company. As a chain, the metaphors depict a world structure that is coherent with the basic goals of the company, and they describe this structure in terms of three basic components, or cornerstones, of the business world and the companies within it. These cornerstones are capital, work, and people, and they are arranged in each metaphor on three perpendicular axes to form a multidimensional space. The projections formed by the axes vary according to the scale and the content of the metaphor (cf. Vanharanta 1995).

21.3.1 External World Metaphor

In the largest scale metaphor, which depicts the external environment as an overlying world structure, the dimensions are environment (capital), organized world (work), and people (their organizational role). The projections of these dimensions are ecosystem, infrastructure, and organized knowledge. See Figure 21.1.

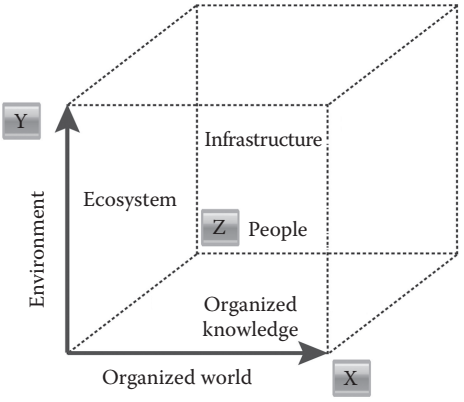


FIGURE 21.1 The external world metaphor.

21.3.2 Business World Metaphor

Similar metaphors are used to depict the general framework of the business world. The business world metaphor represents the company's external commerciopolitical environment to which it is exposed. It has the same components, that is, capital, work, and people, as the external world metaphor. See Figure 21.2.

All business activity originates from the combination and interplay of capital, work, and people. These components are very close to those described by Archibald (1973) (resources, labor force, and technical knowledge), Galbraith (1963) (capital, production, and technostructure), and Ohmae's Japanese view (1982) (kane, mono, and hito). The axes form three projections: business assets, business structure, and business knowledge.

21.3.3 Company World Metaphor

The company world metaphor determines the company's characteristics in the same three dimensions and can be extended to cover the corresponding activities, that is,

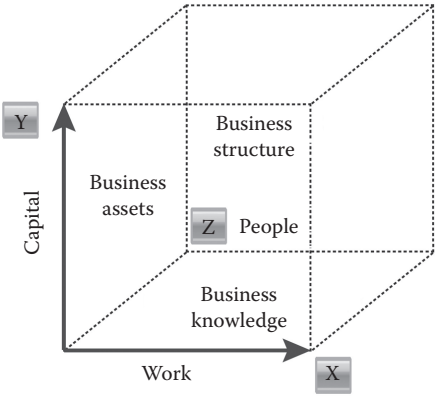


FIGURE 21.2 The business world metaphor.

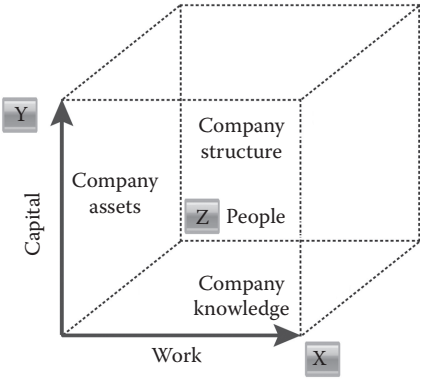


FIGURE 21.3 The company world metaphor.

financing, operations, and management. The three axes form the projections company assets, company structure, and company knowledge. The combined vector of these projections is akin to company performance. The metaphor holistically depicts the living company, its main characteristics, and its performance. See Figure 21.3.

The company world metaphor can be used to create a wide variety of active computer templates for incorporation into human–computer interfaces of company-specific ESS and DSS.

21.3.4 Product World Metaphor

The fourth business metaphor is the product world metaphor. The metaphor is designed according to the concept of supply and depicts the framework of supply as a subclimate within the company environment. Supply is crucial to company survival. The product world metaphor is constructed according to the same principles and the same sequential assembly process as the company world metaphor; see Figure 21.4.

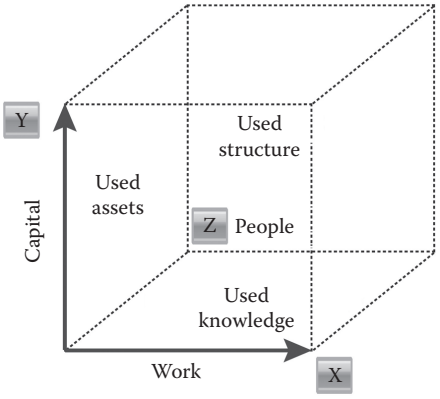


FIGURE 21.4 The product world metaphor.

The three planes of projection characterize the creation of a product or service, a process that consumes a finite quantity of assets (assets used), requires specific activities inside and outside the company (structure used), and is created using a finite amount of knowledge (knowledge used). The metaphor defines the characteristics and performance of a product or service in the market. It is also possible to think through the added-value concept of each dimension. The metaphor is directly convertible into a graphical template for incorporation into human–computer interface of ESS and DSS.

21.3.5 Buyer World Metaphor

The fifth metaphor is the buyer world metaphor. It is a metaphor for the concept of demand, showing it as a subclimate in the company environment. Demand is just as crucial for company survival as supply. The buyer world metaphor, shown in Figure 21.5, defines buyer characteristics. Again, there are three planes of projection, this time characterizing buyers. The metaphor emphasizes the relationships and interrelationships between the producer, product, and buyer.

Buyers have finite assets and related financial potential (buyer assets), they consume specific products and belong to various buyer groups (buyer structure), and they use a finite amount of knowledge to make their purchasing decisions (buyer knowledge). A template can be developed for incorporation into human–computer interfaces of ESS and DSS from this metaphor.

21.3.6 Company Continuum Metaphor

These five metaphors can be combined through a sixth metaphor, that is, the company continuum metaphor. In this metaphor, each of the previous metaphors has a specific place and a specific content, which gives us a dynamic picture of the living company and its relationships and interrelationships. The company continuum metaphor is both a static and dynamic representation of a continuous company strategy, in which the company is part of the living system and is represented by capital,

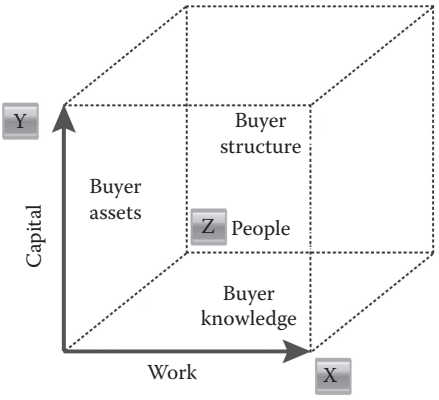


FIGURE 21.5 The buyer world metaphor.

work, and people. Company activities and company characteristics are formed in the company world, in terms of financing (capital), operations (work), and management (people). The formation of the supply concept in the product world is through assets used (capital), structure used (work), and knowledge used (people). The formation of the demand concept in the buyer world is through buyer assets (capital), buyer structure (work), and buyer knowledge (people). The formation of the main components of the business world is through business assets (capital), business structure (work), and business knowledge (people). In the external world, projections are formed through the ecosystem (capital), infrastructure (work), and organized knowledge (people).

The visual and cognitive perceptions, both static and dynamic, show the mutual interdependence of these five metaphors, but they are not easily perceivable holistically and do not show the dynamic dimension of the metaphorical content. Thus, it was necessary to create the company continuum metaphor as a template. See Figure 21.6.

The company continuum template is an integrated working method for use in management. It is a concept map for navigating and combining the dynamic content of the metaphors in the chain, that is, data, information, and knowledge. The template illustrates the interdependence of the five metaphors from macrocosm to microcosm. It shows that all company activities start and are maintained within the freedom and constraints of the supporting frameworks provided by the five-metaphor chain, that is, by the continuous strategy ontology. The template itself is self-explanatory and guides users onward, for example, in the computer context toward other templates and applications.

Managers are responsible for, and should be committed to, development of the company. If they are active users of the metaphor chain and its computer applications, their responsibility and commitment also become an integral part of the supporting framework. In other words, responsibility and commitment are conditional on the freedom and constraints of the metaphor framework.

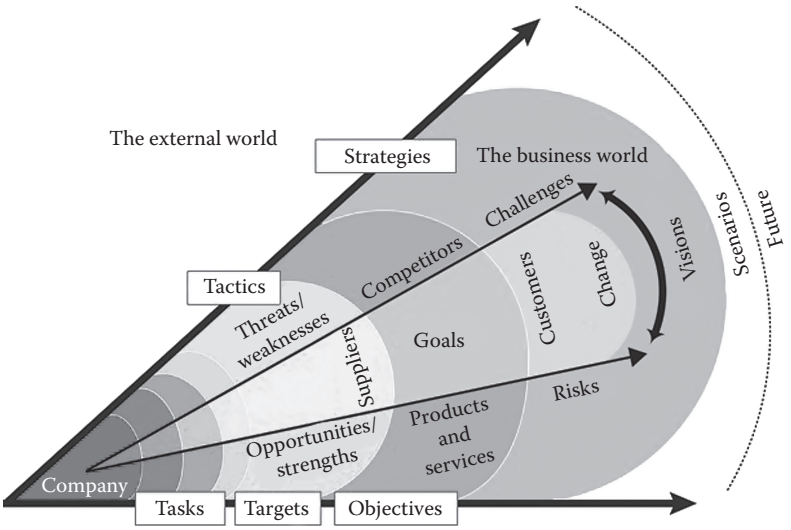


FIGURE 21.6 The company continuum metaphor.

Integration of the top-down overview with the bottom-up market-driven view results in a holistic perception of the company, that is, its business issues, interrelationships, characteristics, its present position, and its progress in the competitive environment. If this visual perception can, on a continuous basis, be instilled into decision makers as a shared vision, it will give them improved competence in strategic management and will diminish the misunderstanding, isolation, and conflict that may otherwise arise. This will result in better strategies and business activities for the survival, progress, and continuity of the company. The following case examples show how knowledge services help managers to create possibilities within strategic decision making.

21.4 FROM DATA TO KNOWLEDGE: A CASE STUDY IN CUSTOMER SEGMENTATION WITH THE SELF-ORGANIZING MAP

In this case study, customer segmentation (Lingras et al. 2005), one of the most important tasks in customer relationship management (CRM) (Datta 1996; Heinrich 2005; Chalmeta 2006), is used to identify and group customers with similar profiles or requirements. The grouping is based on customer information readily available through ERPs, corporate data warehouses, and the Internet (Rygielski et al. 2002; Buttle 2004). Problems arising from the vast amounts and varying quality of the data are solved using data mining methods (Famili et al. 1997; Berson et al. 2000; Shaw et al. 2001; Rygielski et al. 2002; Berry and Linoff 2004).

The aim is to get an overview of the customer base, its demographic attributes, and consumer behavior, with focus on converting large amounts of customer data into actual knowledge of the customer base (cf. Product World Metaphor and Buyer World Metaphor). More specifically, questions such as who buys, how much, what products, how often, and how recently need to be answered. The data are analyzed with the SOM (Kohonen 2001), and a data-driven exploratory customer segmentation (Berson et al. 2000; Holmbom et al. 2011) is carried out. Customers are grouped according to their shopping behavior, and an analysis is conducted based on the demographic information (Frank et al. 1972; Wedel and Kamakura 1999; Tsai and Chiu 2004).

21.4.1 Data

The study focused on the customers of four major department stores and spanned a period of 2 years, from August 2007 to July 2009. The data consisted of the following:

- Demographic data, such as customer loyalty point class, service level, customer duration, gender, child decile, estimate of income, and age
- Product data, consisting of eight product categories, labeled A to H
- RFM (recency, frequency, monetary, and RFM score) information, calculated with IBM SPSS Modeler

The data were preprocessed in order to give better data mining results. Some limitations were made, for example, customers that made purchases for less than

EUR 100.00 (monetary > EUR 100) or less than two times (frequency > 2) during the 2-year period were excluded. This eliminated about 30% of the customers of the data set.

21.4.2 Training the Model

The data were obtained in the form of text files consisting of transaction data and customer demographic data. IBM SPSS Modeler (http://www.ibm.com/) was used to preprocess and translate the data into the right format, and Viscovery SOMine 5.0 (http://www.viscovery.net/) was used to create a map based on the demographic data. The results of the demographic segmentation were then matched with the sales information for each product category.

The resulting data contained significant outliers, and even though the SOM is fairly tolerant toward noisy or missing data (Bigus 1996; Smith and Gupta 2002), sigmoid (or logistic) transformation of the data (Bishop 1995) was used. Sigmoid transformation reduces the influence of extreme input values and emphasizes center input values (Bishop 1995; Larose 2005), while variance scaling makes the variables comparable. Other input parameters used when training the map were a map size of 1000 nodes, a tension of 0.5, and Gaussian neighborhood function. The clusters on the final map, displayed in Figure 21.7, were identified with the help of Ward's hierarchical clustering method and resulted in seven clusters of various sizes, labeled C1–C7. The color of the cluster in the figure signifies cluster membership and does not imply value.

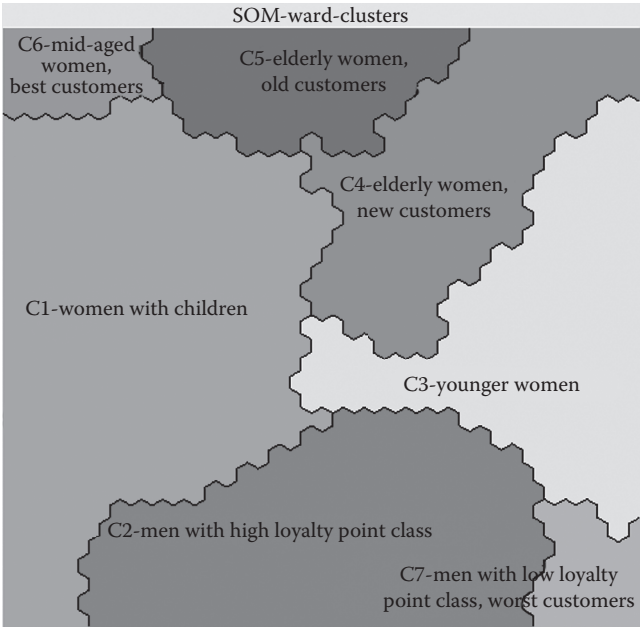


FIGURE 21.7 A clustered SOM model of the segmentation. The segments are labeled according to their characteristic features.

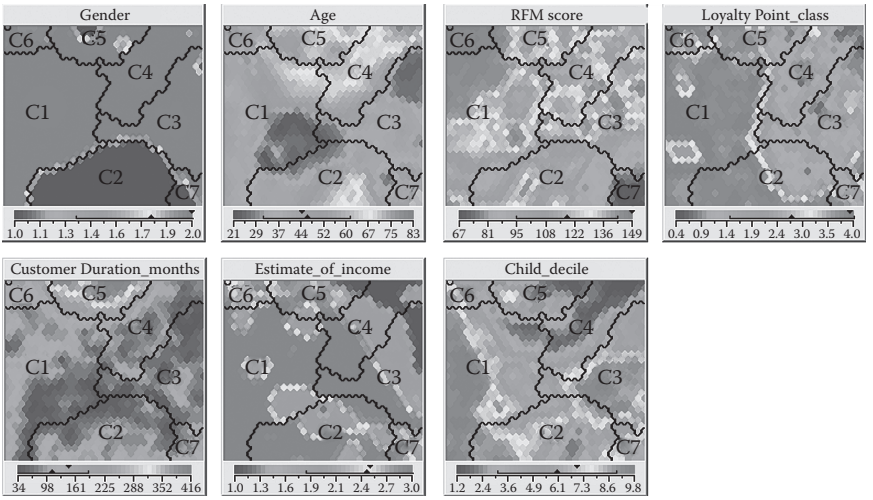


FIGURE 21.8 The feature planes of the main variables (gender, age, RFM score, loyalty point class, customer duration in months, estimate of income, and child decile) that contributed to the map, showing the values by each component.

The feature planes, displayed in Figure 21.8, were used to interpret the characteristics of each cluster. They show the distribution of values across the map according to one variable at a time. The values for each variable are displayed by the color of the neuron, where "warm" colors (red, orange, and yellow) illustrate high values and "cool" colors (blue) illustrate low values. The approximate values are indicated by the scale under each component plane. The map is interpreted by reading the feature planes for each cluster. For example, Cluster C5 displays medium to high values in recency, customer duration, gender, and age and low values for all the product categories. We can conclude that these customers were mainly women of a higher age, who have been customers for a longer time and have visited the department store recently, but that they do not buy much in terms of Euros.

21.4.3 Demographic Segmentation

Demographic segmentation aimed to find out who buys, what products, how much, and how often, based on customers' demographic information (cf. External World Metaphor). Therefore, the demographic and RFM variables were given equal weight in the training process, but the variables describing the product categories, along with the RFM score, were not given any weight. Thus, the model is entirely based on the demographic attributes of the customers, and purchasing behavior is associated with the demographic segmentation.

Similar results were found for all four department stores. As Figures 21.7 and 21.8 show, segmentation was mainly based on gender, age, and the purchase amount. Figure 21.9 shows closer analysis of the clusters.

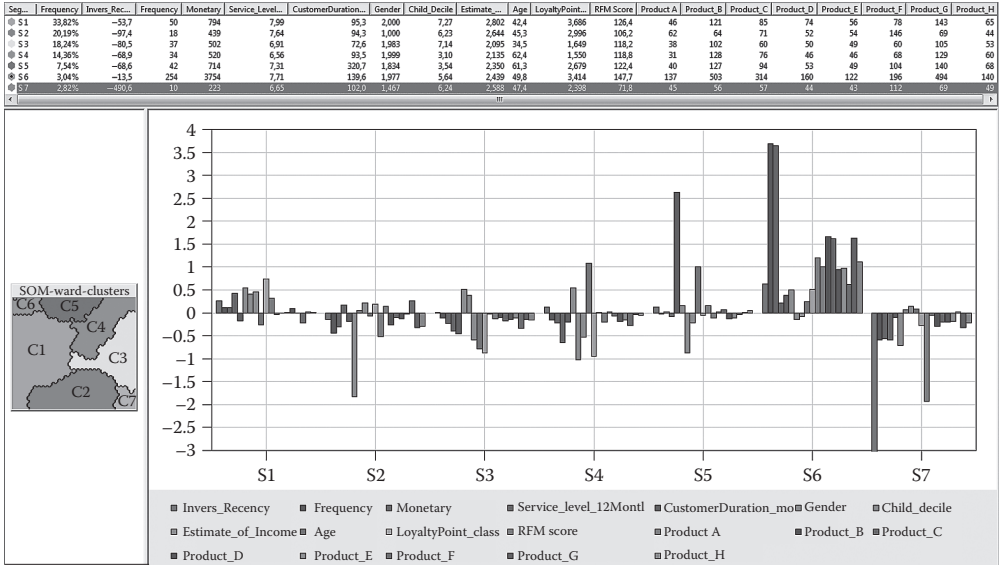


FIGURE 21.9 The statistical analysis of the created clusters. The clusters are displayed separately at the top of the figure, including their size (%) and an average of each of the variables. The bars in the diagram show the deviation from the average (in standard deviation) for each of the variables. The explanation for each bar (from left to right) is found underneath the diagram, showing first the RFM variables, then the demographic variables, and last the products.

With the help of the clustering, the feature planes, and statistical analysis of the created clusters (Figures 21.7 through 21.9), we obtain an overview of the customer base and are able to answer the strategic questions:

- *Who buys?* The purchases are mainly made by customers from two of the clusters: elderly women (average 60 years) in Cluster 5 (7.54%), with a low loyalty point class, service level, and income and who have been customers for a longer time, and middle-aged women (average 45–50 years) in Cluster 6 (3.04%), with a high loyalty point class, service level, and income. They visit the store often and make a significant amount of the total purchases.
- *How much?* The middle-aged women in Cluster 6 buy the most. They spent an average EUR 3750 within the 2 year period. Men in Cluster 7 (2.82%), who do not visit the store often and have a low loyalty point class and service level, are the ones who buy the least—they spent an average EUR 220. Customers from the other clusters purchased between EUR 440 and EUR 800 in the period.
- *Which products?* Women with children in Cluster 1 (33.8%), who have a high loyalty point class, service level, and estimate of income, buy products mainly in Category D. Men with a high loyalty point class, service level, and income in Cluster 2 (20.1%) buy products mainly in Categories A and F. Elderly women with a low loyalty point class, service level, and income in Cluster 4 (14.3%) have not been customers for long. They largely purchase products in Category B. The elderly women in Cluster 5 buy products mainly in Categories B, C, and H. The middle-aged women in Cluster 6 buy from every product category. The younger women (average 30 years) in Cluster 3 (18.2%), with a low income, loyalty point class, and service level, and men in Cluster 7 do not buy from any specific product category.
- *How recently and how often?* The middle-aged women in Cluster 6 have visited the store recently, and they visit the store often. The men with a low loyalty point class in Cluster 7 do not visit the store often and have not done so recently. Customers from the other clusters fall in between these two extremes.

21.4.4 Case Discussion

Customer data from a chain of department stores were successfully analyzed with the SOM to obtain a demographic and behavioral overview of the customer base. Customers were grouped according to their shopping behavior through data-driven exploratory customer segmentation and analyzed based on the demographic information, resulting in a map consisting of seven clusters.

21.5 FROM INFORMATION TO KNOWLEDGE: A CASE STUDY OF COLLOCATIONAL NETWORKS

This case study demonstrates how DSS can assist in creating strategic knowledge from information. Telecommunications companies' annual reports were visualized using the text visualization method collocational topic networks and then shown to

interviewees who follow the industry closely. The interviews show that the visualizations were able to capture what the interviewees considered to be actual developments in the telecommunications industry. Furthermore, the visualizations were found to provide new insights into this market.

21.5.1 Background on Competitive Intelligence

Competitive intelligence, sometimes also known as business intelligence or market intelligence, has been defined by the Society of Competitive Intelligence Professionals (SCIP) as "a systematic and ethical programme for gathering, analyzing, and managing any combination of data, information, and knowledge concerning the business environment in which a company operates that, when acted upon, will confer a significant competitive advantage or enable sound decisions to be made" (SCIP, 2011). Competitive intelligence practice owes much to Porter's (1980) work on competitive strategy, which introduces five forces that shape a company's business environment: current competitors, the threat of substitute products, the threat of possible new entrants to the market, the bargaining power of suppliers, and the bargaining power of customers. It is essential that companies know these forces and how they are changing as best as they can; however, some issues make this very difficult.

First, there is information overload. Companies are overwhelmed by the amount of data that is publicly available about their competitors, customers, and suppliers. They need tools to be able to distinguish the relevant from the irrelevant.

Second, competitive analysis in strategic management often relies heavily on quantitative financial data. Fleisher and Bensoussan (2003) call this phenomenon "ratio blinders." According to them, many organizations make the mistake of relying too much on the financial data of their business environment. This can lead to a situation where companies see a financial gap between their organization and a competitor but cannot see the reasons behind it, and thus do not have the means to close it. This issue relates to a commonly made mistake—the confusion of operational data with strategic data, as mentioned by Zahra and Chaples (1993). To deepen analysis based on operational financial data, there is a demand for methods that allow competitive intelligence practitioners to systematically include qualitative industry data in their analysis. Such data are available, for example, in public texts produced by competitors or other companies in the industry.

In order to incorporate qualitative data effectively into the analysis, a systematic methodology for analyzing large quantities of text is needed. In today's complex business environment, with a high number of companies that need to be tracked and a constantly increasing amount of texts being produced, a manual scan is ineffective. Text mining and text visualization are possible solutions to this issue, and they will be discussed in more detail next.

21.5.2 Text Mining and Text Visualization

Text mining or text data mining, defined as the discovery of trends and patterns within textual data (Hearst, 1999), promises to alleviate the task of processing texts produced in a company's business environment.

Fan et al. (2006) provide a brief overview of text mining methods intended for business users that are available as commercial software packages. However, there is still a long way for these methods to become everyday tools in companies. Krier and Zacca (2002) point out that although various text mining methods have been available for some time within the competitive intelligence community, their use by intelligence practitioners has been limited because they are perceived as "black box" methods, that is, the workings of these methods have been difficult to understand by users who do not have a background in linguistics.

This case study addresses these issues by presenting a text mining method that produces a visual network consisting of words as its outcome. Text visualization is something that allows users to interpret the results in a more intuitive way. However, when working with any intelligence tools, the final interpretation of the results and their significance is always a qualitative process that cannot be automated. Qualitative interviews were therefore conducted in order to evaluate the collocational topic networks as a text visualization method.

21.5.3 Evaluation of Collocational Topic Networks as a Text Visualization Method

Temporal text mining is a term that covers the methods used to discover temporal patterns in texts collected over time (Mei and Zhai, 2005). Temporal text mining suits the purposes of strategic competitive analysis well, because an essential part of competitive intelligence analysis lies in understanding not only competitors' or partners' current strategic decisions, but also their development over time. This creates a starting point for predicting their future decisions.

In this case study, various public companies' annual reports were used as visualization material. Annual reports are ideal material for temporal competitive analysis, as they reflect changes in a company's strategy and competitive situation over time. They are also particularly suitable for the task from a text visualization point of view, as the document structure is usually similar (passages, length) from 1 year to the next.

For the purposes of this case study, six visualizations from 2003 to 2008 were created from the annual reports of seven large telecommunications companies (see Table 21.1).

TABLE 21.1
Texts Included in the Visualization

Company	Country of Origin	Type of Document	Years Included
AT&T	United States	Annual reports	2003–2008
BT	United Kingdom	Annual reports	2004–2009
France Télécom	France	Annual reports	2003–2008
Telecom Italia	Italy	Annual reports	2003–2008
Telenor	Norway	Annual reports	2003–2008
TeliaSonera	Sweden	Annual reports	2003–2008
Verizon	United States	Annual reports	2003–2008

A text file was created for each year included in the visualization by merging the seven annual reports published by the seven companies that year. A similar visualization could be made out of the reports of a single company, but it was decided that for the purposes of this evaluation, a general view of the industry (cf. Business World Metaphor) would be produced based on all company reports, as this would be something that all the interviewees would have an opinion on and would be easier to discuss than developments at a single company.

21.5.3.1 Visualization Method

The visualization method applied in this study, *collocational topic networks*, has its roots in lexicographical research (Williams, 1998) and, more generally, text linguistics (Phillips, 1985). A collocational topic network is a network containing a user-defined topic word as its central node and a user-defined number of links to words that occur in a statistically significant way in the vicinity of the topic word in the text that is visualized. The method is based on the assumption that changes in the networks will reflect changes in how the topic is discussed in the underlying texts.

A collocation is defined as "the occurrence of two or more words within a short space of each other in a text" by Sinclair (1991). Significant collocation takes place when two or more words occur together more frequently than would be expected by coincidence. The significance of collocation is measured using the mutual information or MI score, which compares the frequency of two words with the frequency of their occurrence independent of each other; it is widely used in linguistics when dealing with text masses.

The initial stage of producing the networks consisted of calculating the MI score for all words occurring within a span of four words. A maximum span of this size has been recommended by Sinclair (1991) for studying collocations in English. So-called stop words with little semantic content were left out, such as prepositions, articles, conjunctions, and words referring to figures and currency. Because the case company provided value-added services to telecommunications companies, the word "service" was selected to be the central node in the networks as the context in which the companies discussed the concept of service that was deemed to be of interest to the interviewees. This process produced the six topic networks depicted in Figure 21.10.

Theoretically, the size of the networks is limited only by the contents of the text. It would be possible to make an extensive network where each appearing word is linked further to its own most significant collocations. However, for the purposes of this case study, simple networks containing only the six main collocates of the word "service" were produced. No further links from these six words were drawn. The networks were then shown to interviewees. The eight interviewees represented the company's management team, product management team, and the sales department. All of the interviewees had been working within the technology industry for over 10 years. The interviews were semistructured, allowing for a conversation between interviewer and interviewee and for the introduction of new topics by the interviewee.

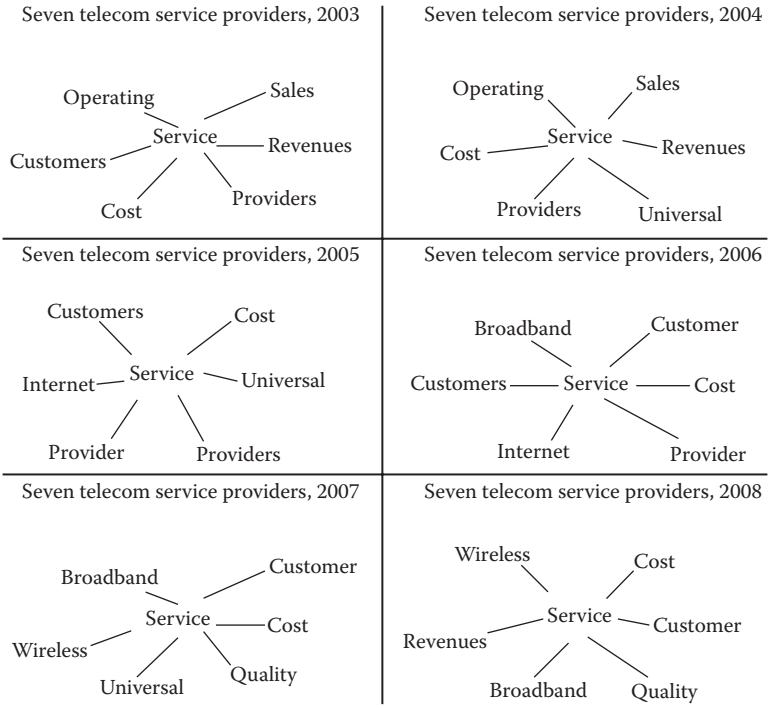


FIGURE 21.10 The topic networks for the word "service" in seven telecommunications companies' annual reports 2003–2008.

21.5.3.2 Interpretation of Visualizations

The interviewees were shown the list of companies included in the study (Table 21.1) and the collocational topic networks in Figure 21.10. The basic workings of the visualization were explained to them in the following way: "These networks are based on annual reports by all seven telecommunications service providers, year by year. The six words around the word 'service' occur in these reports in a statistically significant way together with 'service.'" The interviewees where then asked to look at the networks and discuss any thoughts that emerged. Further clarification of the method was provided if requested.

After studying the networks, most interviewees mentioned at least two of the following observations:

- The word "cost" appears in connection with service in every network. Interviewees believed this reflected telecommunications companies' struggles with cost reduction throughout the period.
- The word "revenues" appears in the first 2 years then disappears and reappears in 2008. Several interviewees noticed this, but did not quite know how to interpret it. During the period, revenues from new services had been rising, while traditional service revenues had declined.

- The earliest networks do not mention any specific service types, but in 2005, "Internet" occurs for the first time. In 2006, the more specific "broadband" appears, and in 2007, "wireless" appears. The interviewees found this interesting and believed it showed that the companies have become more elaborate about the different types of service they provide and now discuss the services in more detail.
- The word "universal" is present in 2004, 2005, and 2007. The interviewees believed this was because most of the included companies are incumbents in their home country and so they may be required by law to provide a universal service, that is, a service that reaches all residents of that country.
- The word "quality" appears in 2007 and remains in the network in 2008. The interviewees saw this as an indication that fierce price competition has driven telecommunications companies to emphasize service quality in their external communication.

Generally, the interviewees considered the networks to reflect actual developments in the telecommunications service industry during 2003–2008. It should be noted that this method allows users to go to the original source texts, if needed, to look for explanations for the occurrence of unusual, ambiguous, or otherwise particularly interesting words in the networks. In that sense, this is not a "black box" method, and user interpretations of changes in the networks can usually find support (or be disconfirmed) through the original texts.

21.5.3.3 Interview Themes on Text Visualization

During the interviews, three recurring themes on the subject of text visualization began to emerge from the conversations. The themes were as follows:

1. Visualization of qualitative versus quantitative data. All interviewees said that they work with both qualitative and quantitative data. They considered qualitative data to include both written and verbal nonnumeric data. The interviewees said that they were familiar with, and frequently worked with, quantitative data models such as pie charts or bar charts. In contrast to this, the use of text visualization and text mining methods in a corporate context is still rare. None of the interviewees had encountered such methods during their working career.
2. Uses of text visualization. The interviewees were asked to spontaneously present situations where they would consider using such visualizations. All interviewees quickly responded with various scenarios where they could see the topic networks to be useful. Such scenarios include an analysis of competitors' marketing materials, including differences between competitors and changes within the industry over time.
3. Visualization for analysis versus visualization for presentation. The interviewees' responses to the networks and their suggestions for usage scenarios show that they see the visualizations as accurate representations of the underlying texts. Furthermore, they see them as tools that could be used to make users aware of issues that would otherwise go undetected. In some cases, they could also be used in presentations as a basis for discussion.

21.5.4 Case Discussion

In conclusion, all interviewees considered the collocational topic networks that they were shown to reflect actual developments in the telecommunications service industry during 2003–2008. Furthermore, all interviewees found the visualizations to provide interesting insights and thought of several possible uses for the method within the domain of competitive intelligence.

The collocational topic network method presented here could easily be used as a tool for assisting competitive intelligence analysis in situations where textual data from an organization's competitive environment are needed to balance financial data. The tool could produce networks out of several user-specified topics and highlight changes according to the user's wishes. Building on Porter's (1980) theory of five basic competitive forces, the tool could be used not only for competitor analysis but also to keep an eye on strategic changes in potential entrants to the market or vendors of substitute products. It would also be useful for following developments at suppliers and customers, particularly in a business-to-business context.

Topic networks could, in some cases, also be shown to an audience, internal or external, but the need for making this type of presentation has to be carefully considered; the unfamiliarity of the visualization method to most viewers means that the visualization itself might get more attention than the argument that the presenter is trying to make using it.

21.6 FROM KNOWLEDGE TO KNOWLEDGE: FROM INNER PERCEPTIONS TO STRONG SALES CULTURE USING ONTOLOGY

The medical technology company in this case study has experienced a strong expansion of its sales force outside the domestic markets. Sales personnel sell the same products in different market areas, but the cultural background of the customer base varies greatly. Single salespersons have remote locations in different continents without close connection to the company's technical staff, and, therefore, multicultural understanding and cooperation is needed at the headquarters. Cultural understanding is also required to understand customers' needs in the correct way based on the globally incoming information.

The objectives in this case were, first, to define the culture in the industrial environment and, second, to build an ontology of sales culture and its relation to organizational culture in order to understand underlying processes within sales culture. The third objective was to use the new ontology to study and analyze the export sales force in the company to find out which areas of the sales culture needed to be developed. The fourth objective was to see how well the results match the literature on the most important organizational competencies in creating a strong sales culture. Finally, after bridging the results and the literature, a prioritized list of actions was suggested for strategic decision making based on the individuals' collective inner perceptions.

21.6.1 Sales Culture Ontology

As sales culture is one aspect of the company culture, each culture affects the other. The company culture comprises the company's vision, company values, and company strategy. Though the sales culture has its own strategy, it naturally must be in concordance with the company strategy. Company culture is affected by the environment, stakeholders, and customer feedback. Sales culture adapts the relevant elements of these factors via vision, values, and strategy from the company culture. Sales culture consists of five main components: sales culture support, sales force, sales system, sales process, and sales culture assumptions.

Sales culture support consists of three concepts. The first contains suitable elements of company strategy, vision, and values. The second is the ethics of sales, and this probably has the most important role in the company because sales processes vary so much depending on the customers and market areas. Also, the opportunity to act in an unethical way is biggest during sales and marketing activities, for example, selling unsuitable products or selling for personal benefit. The third and the most important subconcept is leadership, which is to build, change, and strengthen the culture. Leadership can further be divided into leadership, management, and time management, but to keep this ontology lean, the term represents all three aspects in this study.

The *sales force* concept establishes the paths through which a single sales activity takes form. These paths are represented by dotted lines in Figure 21.11. Some of the elements can be regarded as tools of leadership and are affected by leaders, but some will be developed over a longer time period and cannot be controlled directly. Elements along a path constitute the logical chain that is behind a single action, for example, training and development improves salespersons' skills and thus improves their understanding of selected actions. Following the elements on a path, leaders and managers can focus their development efforts on the correct issues.

Sales system defines the objects of sales activities and the market areas in which these activities are performed. The general sales strategy is set within the sales system subconcept and defines the company's strategic position in the marketplace. When leaders/top managers communicate the company strategy to the whole organization, the sales system is the detailed projection of the strategy from the sales point of view. Sales system also includes the prescription of functions and resources outside the sales organization that are needed to support sales processes. For example, the more technically advanced the products the company is selling, the more technical expertise is needed from other functions. The attitude toward sales and how well sales are integrated in the other functions are essential to the organization's success. If sales remain an isolated function, the company may lose many opportunities to improve its customer service; in customer meetings, it might be hard to convince a customer without the presence of experts, but, on the other hand, if salespersons are ignored, much valuable information from the marketplace might not reach the technical staff at all.

The *sales process* concept comprises the theoretical context of the sales procedure and describes all the processes that are needed to realize sales efforts in the company. The quantity of processes varies depending on the company's business.

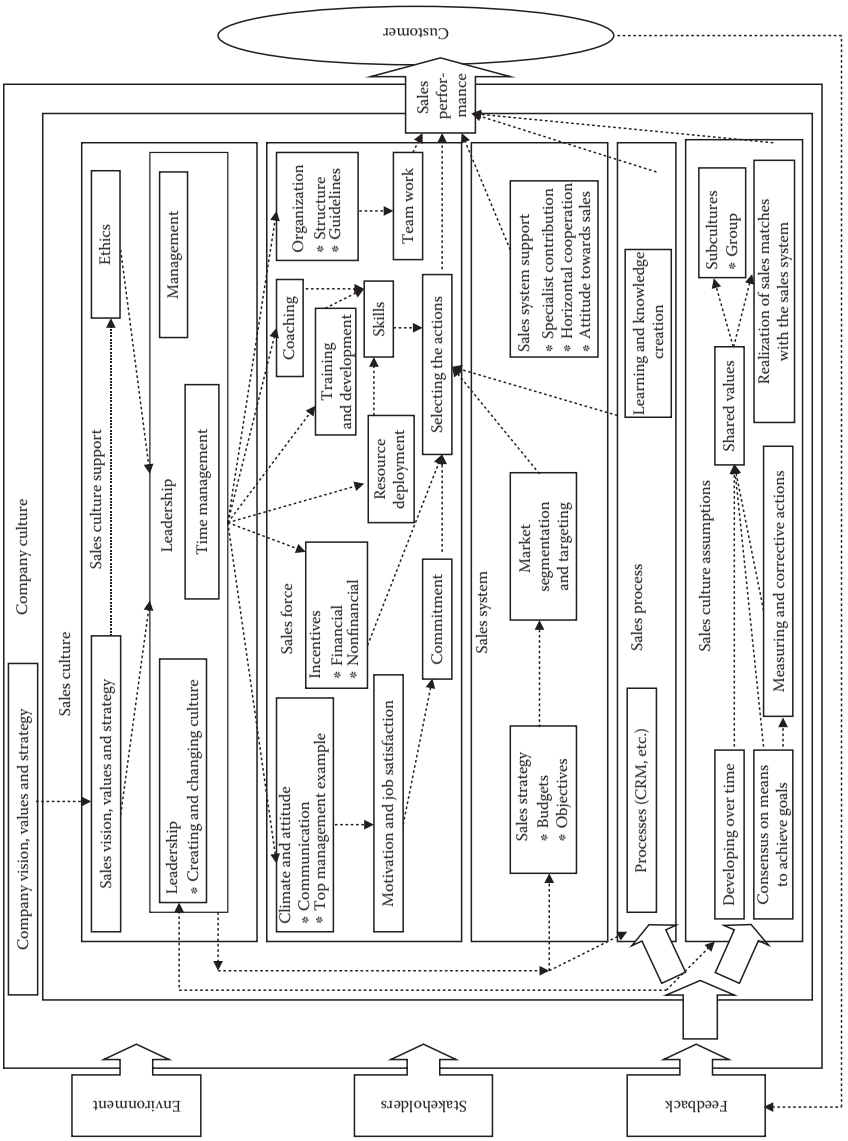


FIGURE 21.11 The SCO.

The simplest sales process is for agencies, which have only a few major customers in the domestic market. The most complicated processes are for multinational companies, which have their own sales offices and agencies serving customers worldwide. A vital part of the sales process is continuous learning and knowledge creation. If training—which is included in the sales force concept—is more focused on individual salespersons, improving their personal sales competencies, learning, and knowledge creation includes processing sales-related innovations and processes, where tacit knowledge is converted to explicit knowledge. Continuous learning and knowledge creation should be a controlled process to ensure proper results.

Sales culture assumptions reflect the underlying assumptions that guide salespersons' professional decisions and daily actions. Most of these assumptions are not conscious, but are behind decisions. These assumptions develop over time, and if no assumptions can be found, it could be that there is no culture or the culture is weak. Furthermore, in the case of young companies, certain elements of the culture may have not yet evolved. Leaders have a very strong effect on the progression of the sales culture. Typically, the organization follows the example of the leader (often the leader is also the founder): what the leader pays attention to, how the leader reacts to certain situations, etc. By repeating certain patterns of behavior, leaders build the underlying assumptions for the organization. When the organization as a whole can agree on the objectives and the applicable methods to reach those objectives, the culture starts to form and strengthen. One sign of a more advanced culture is the consensus on measuring results and especially on the actions needed if the objectives are not reached. Though the leaders can develop and change the culture, the culture exists among the individuals in the organization. Therefore, the culture affects how well the sales strategy is realized in operations, the sales systems, and sales processes on a daily basis. The difference between defined and actual behavior reflects the level of culture. It is critical that the group shares values that are important to it, and from the company point of view, those values should be in conjunction with the company values. When the organization grows larger, it may have several groups with different value bases. This is acceptable as long as the most important values are the same for each group and a single group's values do not contradict those of the company. The sales culture can be seen through customers' eyes as the performance of sales force. A theoretical "best way" of sales processes exists, but how the underlying assumptions are affecting in practice, that is, whether the written method is actually followed or not, is at the core of the sales culture. It is easy to have an ideal way of handling a sales case and other customer management actions on paper, but in reality, how well each salesperson operates according to this way is crucial. As shown in Figure 21.11, several elements influence sales performance, as illustrated by the dotted lines leading up to sales performance. The customer's positive or negative feedback on the sales performance is mainly handled in the sales process concept. Customer feedback also affects sales culture assumptions because there is a two-way relationship between sales culture assumptions and leadership. Customer feedback is also handled by leaders because they are able to change the culture, and feedback-based changes can often be tracked to leadership. In the next section, we explain how the sales culture ontology (SCO) was used to analyze the case company's sales force.

21.6.2 Evolute System

The evolute system supports the use of fuzzy logic (Zadeh 1965, 1973) applications on the Internet (Kantola et al. 2006; Kantola 2009). The evolute system "hosts" domain ontologies and presents the ontologies online to target groups through semantic entities, such as statements. Different classes (concepts) in the SCO are described in familiar language in the statements. A few indicative statements relate to each class in the SCO. One statement can relate to more than one concept, and each statement has a weight. Statements are compared to linguistic labels on a fuzzy scale, which means that the meaning of the statement in the individual's mind can be captured and a conversion from the meaning to the crisp numerical value during the evaluation process is eliminated. When the statements are evaluated, inputs are converted into fuzzy sets (fuzzification). The inference engine evaluates the fuzzified inputs using rules in the rule base. This results in one fuzzy set for each SCO class (inferencing). Fuzzy sets are then converted into crisp class values and, furthermore, into visual reports. The evolute system works as a generic fuzzy rule base system, as described in Figure 21.12.

Participants evaluate the current reality and the future vision of their surroundings according to the statements in the evolute system. The difference between the self-perceived current reality and the vision is called creative tension (Senge 1994). In the case of the SCO (external object), the difference between the perceived current reality and future vision can be called the proactive vision. Evolute provides SCO-based "answers" or instances (cf. Kantola 2009). The collection of instances forms the instance matrix (cf. Kantola 2009):

$$\text{ONTOLOGY}_{\text{SCO}}\ (\text{individuals}_{1-n},\ \text{instance})$$

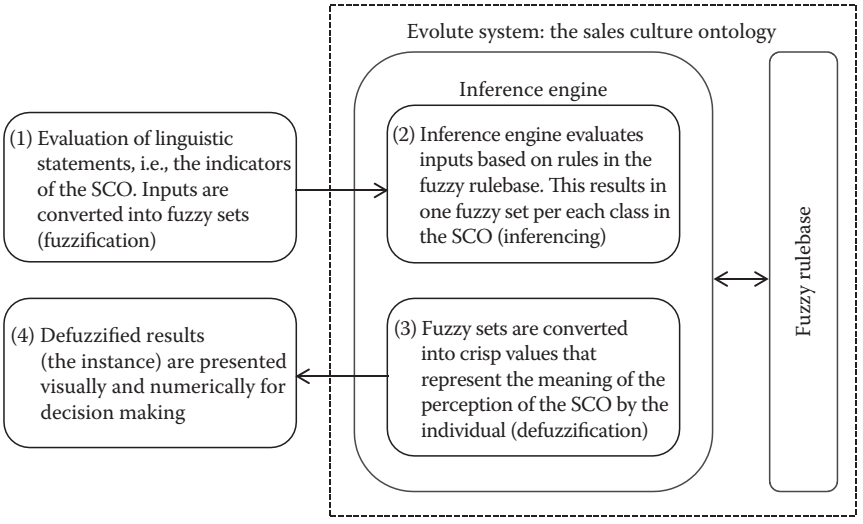


FIGURE 21.12 The SCO is evaluated through indirect statements.

The instance matrix describes the perceived state of the object under scrutiny in the organization. The instance matrix, as a function of time, can be stated as follows:

$$\text{ONTOLOGY}_{\text{SCO}}\ (\text{individuals}_{1-n},\ \text{instance}_{1-k})$$

The instance matrix represents the collective perception/collective mind of stakeholders regarding the SCO in the company. In this case, the goal is to capture a true bottom-up view of the current reality and envisioned future of the features and practices of the company's sales culture.

The SCO and its propositions in the evolute knowledge base can be fine-tuned by adjusting the fuzzy sets and fuzzy rules as more is learned about the sales culture domain. The content of the SCO will develop over time as the domain naturally evolves and as researchers learn more about it (Gomez-Perez 2004).

21.6.3 Dataset

From total 19 inquiries, a completed evaluation was received from 14 participants, so the answering ratio was 74%. After all the employees completed the inquiry, the results were summarized so that no individual's answers could be seen. The results, therefore, show the collective status in the whole company. The results are shown in graphical format.

21.6.4 Results

In Figure 21.13, the results are sorted top-down according to the biggest proactive vision (vision–current). The bigger the proactive vision, the more employees want to improve the concept. Development efforts focusing on the concepts with the biggest proactive vision can be expected to give the best results due to the high internal demand of employees. The biggest proactive vision in this case company was for "selecting the actions," the second was "development over time," and the third "strategy." At the other end, there was practically no proactive vision for "ethics," which means that employees do not see any need/place for improvements in ethics-related issues. Therefore, efforts to improve ethics in the company would probably not be successful. We concentrate on the three concepts with the biggest proactive vision and more closely examine the indicative statements behind the results.

21.6.4.1 Selecting the Actions

There are two statements behind this concept: (1) *In our company, the rewarding of certain behavior is logical*, and (2) *there are enough sales meetings to give guidelines for sales operations*. Employees answered on average that there is the biggest improvement potential according to these two indicators, in other words that rewarding behavior is not logical and there are not enough sales meetings.

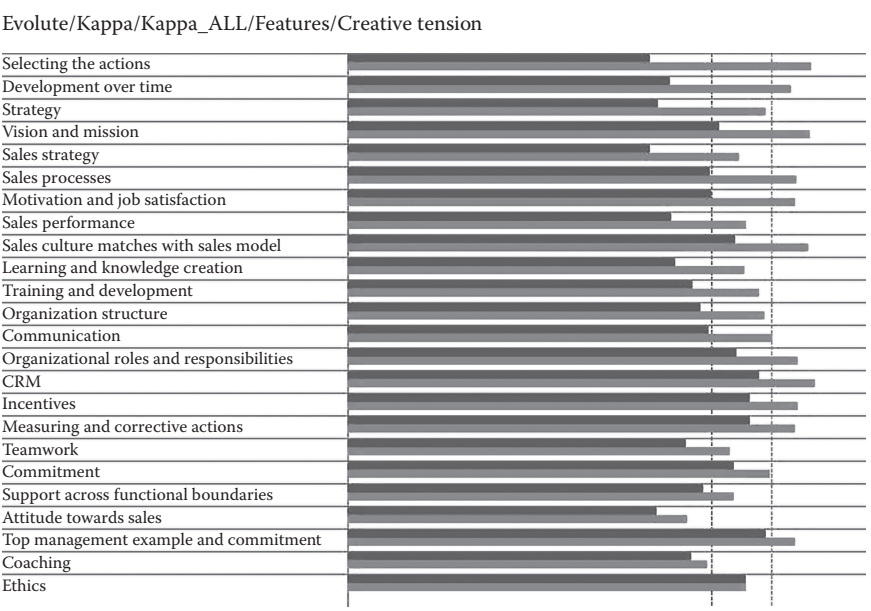


FIGURE 21.13 The concepts in the SCO sorted by biggest proactive vision.

21.6.4.2 Development over Time

The statements behind this concept are as follows: (1) *new members in sales team adapt to the team habits*, (2) *I know how my colleagues/team members will act when the customer gives negative feedback*. Statements in this class describe the level of culture in the purest form, that is, how well a group has learned how to react in challenging situations. If there is large proactive vision in this concept, it means that the culture has not been developed. There are mainly two reasons for this: either the organization is so young that the culture has not yet been developed or the leaders have not been able to establish the culture, for example, through own example and being logical in operations.

21.6.4.3 Strategy

In this concept, the indicative statements are the following: (1) *our company offerings are based on global market demand*, (2) *our company offerings are based on local market demand*, (3) our company is customer/product oriented, (4) our company is service /technology oriented, (5) *our company recognizes changes in the environment*, and (6) *the relationship between the marketing strategy and sales strategy is defined in the company strategy*. Interpretation of these results shows how well the case company has been able to develop its offerings to meet customers' requirements. According to sales personnel, there is clearly room for improvement. Also, the sales personnel consider the attitude to be a typical engineering attitude, that is, technology is the driver instead of the customer and service. However, this is an area that will probably never be—from the sales personnel's point of view—at a satisfactory level.

The company's capability to recognize changes in the environment is vital in order to adapt to changes in the strategy and meet customer requirements in the future. If no changes are recognized, the company keeps its old methods and offerings, and inevitable disaster follows. The last statement refers to the roles of sales and marketing and especially to the relationship between them, which is often poorly defined. The literature frequently shows sales under marketing, even though the sales personnel can be significantly larger. However, the distinction in roles between these two functions should be clear in order to avoid overlapping tasks or even neglecting them.

To find out the level of the case company's sales culture, one option is to compare the results to the literature regarding the creation of a strong culture. The literature gives the following as the main components when creating a strong culture: (1) the leader's role: leaders show their own example and operate logically (Oedewald and Reiman 2003; Benson and Rutigliano (2004)); (2) the organization acts as defined in process descriptions (Oedewald and Reiman 2003; Holland and Laine 2008); (3) personnel is committed to work and company (Oedewald and Reiman 2003); (4) the culture develops over time, that is, groups have had enough common experiences (Schein 1992; Leppänen 2006); and (5) the sales strategy is clear to everyone (Oedewald and Reiman 2003). Elements found from literature are not exactly matching with the concepts in the SCO, but still there are clear similarities.

21.6.5 Recommended Actions

For culture development, it is important to have repeated patterns of behavior for certain situations. Typically, these patterns evolve over time through the selection of methods that will solve problems. If top management does not support the evolution taking place within a group by operating logically, it weakens the culture.

The consensus on illogical and even unfair rewarding is obvious among the sales force. Recommended action for the management is to clarify what has been the basis for rewarding, analyze it, and set new guidelines for rewarding. The final step is to inform the whole sales force of these guidelines and have top management commit to them. The literature supports this; the leader's role is very important. Rewarding must be seen on a wide base; it does not only include incentives but also promotions, getting certain sales territories, getting customers or customer groups, being allowed to operate with larger authorizations, or public rewarding. For example, in a big sales case, a top management takeover would lead to an unsatisfactory situation from the salesperson's point of view—the top management's role should be more supportive in order to improve the salesperson's knowledge.

Culture development takes place in groups and therefore needs forums and face-to-face meetings. For this purpose, sales meetings are a good and natural option. Also information can be distributed in this kind of meetings. Although it would be easy to propose more sales meetings for the whole sales force, there would be limitations due to expenses and time usage. Rather the way to improve this issue is to develop the quality of the meetings, use modern negotiation media (net meetings) instead of traveling, and increase the amount of local meetings. Support from the literature can be found here, too: culture develops over time, that is, when groups have had enough common experiences.

Culturally, sales meetings foster the development of the subculture. When the company grows and operations become global, meetings that the whole company can join are impossible to arrange. Subcultures are acceptable as long as they share the main elements of the company culture.

21.7 DISCUSSION AND CONCLUSIONS

In strategic management, a lot of data, information, and knowledge are needed to make a sound strategy for the organization. This strategy work seems to require knowledge services, but the supply side still tries to find out new ways and better tools to serve the demanding requirements coming from turbulent business environments. Strategic management also has its problems, because many strategy frameworks presented in the literature are close to each other, but the contents inside these frameworks vary and are not simple enough for daily management.

In this chapter, we show that by using an ontology-based strategic management framework, we can reach a model that can easily be connected to the modern requirements of strategic management. Novel tools and methods can offer new types of knowledge to help managers make strategic choices when they implement their overall company strategies in turbulent business markets. We have shown in detailed case examples how it is possible to perform customer segmentation based on business data and carry out in-depth analysis with feature planes concerning the main variables of the segmentation, as well as how we can create new, relevant knowledge of competitors with text visualization methods. In the third case example, the ontology-based method and fuzzy logic tools were used to visualize current demand to improve the sales culture and show this kind of internal, bottom-up strategic information to top management.

All the knowledge from the case studies can easily be adapted to the general strategy framework presented at the beginning of this chapter. The connection to the continuous strategy ontology gives managers a strong basis to demand more situation-dependent knowledge to help their strategic decision making. We have shown that visual images contain a lot of information and are easy to understand and perceive. We believe these kinds of new knowledge services are strongly needed by strategic management and will change future ideas of strategy making and implementation.

ACKNOWLEDGMENTS

The author thanks the case organization for its participation in the study. The author also gratefully acknowledges the financial support of the National Agency of Technology (Titan, grant. no 33/31/08).

REFERENCES

Archibald, G. C. (1973). *The Theory of the Firm*. Great Britain, U.K.: Penguin Education; Richard Clay Ltd., pp. 9–10.

Benson, S. and Rutigliano, T. (2004). Discover your sales strengths, *Gallup Series*, Random House, 256p.

Berry, M. J. A. and Linoff, G. S. (2004). *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*. Indianapolis, IN: Wiley Publishing Inc.

Berson, A., Smith, S., and Thearling, K. (2000). *Building Data Mining Applications for CRM*. New York: McGraw-Hill Companies Inc.

Bigus, J. P. (1996). *Data Mining with Neural Networks: Solving Business Problems from Application Development to Decision Support*. New York: The McGraw-Hill Companies Inc.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Avon, Oxford, U.K.: Oxford University Press.

Buttle, F. (2004). *Customer Relationship Management Concepts and Tools*. Oxford, U.K.: Butterworth-Heinemann.

Chalmeta, R. (2006). Methodology for customer relationship management, *The Journal of Systems and Software* (79:7), 1015–1024.

Datta, Y. (1996). Market segmentation: An integrated framework, *Long Range Planning* (29:6), 797–811.

Famili, A., Shen, W.-M., Weber, R., and Simoudis, E. (1997). Data preprocessing and intelligent data analysis, *Intelligent Data Analysis* (1:1), 3–23.

Fan, W., Wallace, L., Rich, S., and Zhang, Z. (2006). Tapping the power of text mining, *Communications of the ACM* (49:9), 76–82.

Fleisher, C. S. and Bensoussan, B. (2003). Why is analysis performed so poorly and what can be done to improve it? In *Controversies in Competitive Intelligence: The Enduring Issues* (Fleisher, C.S. and Blenkhorn, D.L., Eds.). Westport, CT: Praeger Publishers, pp. 110–122.

Frank, R. E., Massy, W. F., and Wind, Y. (1972). *Market Segmentation*. Englewood Cliffs, NJ: Prentice-hall Inc.

Galbraith, J. K. (1963). *The Industrial State*. New York: The American Library Inc., 418pp.

Gomez-Perez, A. (2004). Ontology evaluation. In *Handbook on Ontologies* (S. Staab and R. Studer, Eds.). Springer, Berlin, pp. 251–273.

Hearst, M. (1999). Untangling text data mining, In *Proceedings of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics*, University of Maryland, June 20–26, 1999 (invited paper).

Heinrich, B. (2005). Transforming strategic goals of CRM into process goals and activities, *Business Process Management Journal* (11:6), 709–723.

Holland, J. and Laine, P. (2008). *Improvisointi Kuriin Myynnin Johtaja* (Stop Improvising, Sales Manager), Fakta. Talentum, Helsinki, Finland, December 2008.

Holmbom, A. H., Eklund, T., and Back, B. (2011). Customer portfolio analysis using the SOM, *International Journal of Business Information Systems* (8:4), 396–412.

Kantola, J. (2009). Ontology-based resource management, *Human Factors and Ergonomics in Manufacturing & Service Industries* (19:6), 515–527.

Kantola, J., Vanharanta, H., and Karwowski, W. (2006).The evolute system: A co-evolutionary human resource development methodology, *International Encyclopedia of Ergonomics and Human Factors*, 2nd edn. Boca Raton, FL: CRC.

Kohonen, T. (2001). *Self-Organizing Maps*. Berlin, Germany: Springer-Verlag.

Krier, M. and Zacca, F. (2002). Automatic categorisation applications at the European patent office. *World Patent Information* (24:3), 187–196.

Larose, D. T. (2005). *Discovering Knowledge in Data. An Introduction to Data Mining*. Hoboken, NJ: John Wiley & Sons Inc.

Leppänen, J. (2006). *Yritysturvallisuus Käytännössä.* (Company Security in Practice). Talentum, Helsinki, Finland, 403p.

Lingras, P., Hogo, M., Snorek, M., and West, C. (2005). Temporal analysis of clusters of supermarket customers: Conventional versus interval set approach, *Information Sciences* (172:1–2), 215–240.

Mei, Q. and Zhai, C. (2005). Discovering evolutionary theme patterns from text: An exploration of temporal text mining. In *Proceedings of the eleventh ACM SIGKDD international conference on knowledge discovery in data mining*. ACM, p. 198.
Oedewald, P. and Reiman, T. (2003). Core task modeling in cultural assessment: A case study in nuclear power plant maintenance. *Cognition, Technology and Work* (5:4), 283–293.
Ohmae, K. (1982). *The Mind of the Strategist*. New York: McGraw-Hill, Inc., 283p.
Oxford English Dictionary. (2011). www.oed.com, Oxford University Press.
Phillips, M. (1985). *Aspects of Text Structure: An Investigation of the Lexical Organisation of Text*. Amsterdam, the Netherlands: North-Holland.
Porter, M. E. (1980). *Competitive Strategy: Techniques for Analyzing Industries and Competitors*. New York: The Free Press.
Rygielski, C., Wang, J.-C., and Yen, D. C. (2002). Data mining techniques for customer relationship management, *Technology in Society* (24:4), 483–502.
Schein, E. H. (1992). *Organizational Culture and Leadership*, 2nd edn. San Francisco, CA: Jossey-Bass management series.
SCIP (2011). Glossary of terms used in competitive intelligence and knowledge management. http://scip.cms-plus.com/files/Resources/Prior%20Intelligence%20Glossary%2009Oct.pdf (accessed 11/08/2011).
Senge, P., The Fifth Discipline: The Art & Practice of the Learning Organization, Doubleday Business; 1st edition (October 1, 1994), ISBN-10: 0385260954, ISBN-13: 978-0385260954, 424p.
Shaw, M. J., Subramaniam, C., Tan, G. W., and Welge, M. E. (2001). Knowledge management and data mining for marketing, *Decision Support Systems* (31:1), 127–137.
Sinclair, J. (1991). *Corpus Concordance and Collocation*. Oxford, U.K.: Oxford University Press.
Smith, K. and Gupta, J. (2002). *Neural Networks in Business*. Hershey, PA: IDEA Group Publishing.
Tsai, C.-Y. and Chiu, C.-C. (2004). A purchase-based market segmentation methodology, *Expert Systems with Applications* (27:2), August, 265–276.
Vanharanta, H. (1995). Hyperknowledge and continuous strategy in executive support systems. *Acta Academiae Aboensis*. PhD thesis, Ser. B, Vol. 55, No. 1. Turku, Finland.
Wedel, M. and Kamakura, W. (1999). *Market Segmentation Conceptual and Methodological Foundations*. Boston, MA: Kluwer Academic Publishers.
Williams, G. (1998). Collocational networks: Interlocking patterns of lexis in a corpus of plant biology research articles. *International Journal of Corpus Linguistics* (3:1), 151–171.
Zadeh, L. (1965). Fuzzy sets, *Information and Control* (8:3), 338–353.
Zadeh, L. (1973). Outline of a new approach to the analysis of complex systems and decision processes, *IEEE Transactions on Systems, Man, and Cybernetics* (1:1), 28–44.
Zahra, S. A. and Chaples, S. S. (1993). Blind spots in competitive analysis. *The Academy of Management Executive* (7:2), 7–28.

Publication 4

Holmbom, A.H., Sarlin, P., Yao, Z., Eklund, T., Back, B. (2013). Visual Data-Driven Profiling of Green Consumers, *Proceedings of the 17th Conference on Information Visualisation* (IV2013), London, UK, July 15-18, 2013.

Visual Data-Driven Profiling of Green Consumers

Annika H. Holmbom, Peter Sarlin, Zhiyuan Yao, Tomas Eklund, Barbro Back
TUCS – Turku Centre for Computer Science,
Department of Information Technologies, Åbo Akademi University
Turku, Finland
{annika.h.holmbom, peter.sarlin, zhiyuan.yao, tomas.eklund, barbro.back}@abo.fi

***Abstract*—There is an increasing interest in green consumer behavior. These consumers are ecologically conscious and interested in buying environmentally friendly products. Earlier efforts at identifying these consumers have relied upon questionnaires based on demographic and psychographic data. Most of the studies have concluded that it is not possible to identify a unanimous profile for a green consumer, because: (1) there might be several profiles for green consumers, and (2) in questionnaires, consumers tend to answer according to their intentions, not according to actual behavior.**

We apply a new method, the Weighted Self-Organizing Map (WSOM) for visual customer segmentation in order to profile green consumers. The consumers are identified through a data-driven analysis based on actual transaction data, including both demographic and behavioral information. The WSOM accounts for the 'degree' of how green a consumer is by giving a larger weight to consumers who buy more green products. The identified profiles are verified by comparison to earlier research.

Keywords---Visual customer segmentation; Data-driven profiling; Green consumer behavior; Weighted Self-Organizing Map (WSOM)

I. INTRODUCTION

Green consumer behavior is a phenomenon that has been of interest for companies for a number of years. The identification of the profile of an eco-conscious consumer, as well as information concerning the future development of the trend, has been the subject of several research studies since the 1970s, with a peak in the 1990s. However, these studies have been challenged by a number of factors, including insufficient data, shortcomings of the measurement techniques, and differences in green consumer behavior in different countries and over time. Defining the terms eco-product and green consumer has in itself proven to be problematic.

Most studies seeking to identify the profile of a green consumer have primarily relied upon self-reporting methods, such as questionnaires, polls, and queries. Typically, these studies have constructed the profile based upon demographic and psychographic data provided by the respondents. Several studies conclude that it is not possible to identify the profile of a green consumer based solely on demographic or psychographic data and instead base the profile on a combination of the two types of data.

However, questionnaire data can be problematic for this purpose. While consumers are assumed to answer the questions according to the best of their knowledge, they tend to answer according to their intentions or expected social norms, which might not reflect their actual behavior. For example, Mainieri et al. [1] found that while 75% of the respondents were willing to pay a higher price for environmentally friendly products, only 14% acted upon it. Young et al. [2] found this "attitude-behavior gap" to be 30%. Green consumers adhere to a green purchasing behavior to a certain degree, as they purchase both environmentally friendly products and non-green products [3]. Therefore, profiles for different degrees of green purchasing behavior are of interest.

Business Intelligence (BI) is an umbrella term for a set of theories, methodologies, processes, architectures, and technologies that transform large amounts of raw data into useful information. BI can be used for visualizing, identifying and developing new opportunities, in order to achieve a competitive market advantage [4, 5].

Hence, instead of using questionnaire data, a preferable way to analyze green consumer behavior would be to use BI processes and methods and perform a data-driven assessment based on actual transaction data. The Self-Organizing Map (SOM) [6] has been shown to hold promise for visual customer segmentation (see, e.g., [7-14]). Unlike most other clustering algorithms, the SOM possesses simultaneous clustering and projection capabilities, which makes it a suitable technique for data and dimensionality reduction. Reasons for choosing SOM over other similar methods are: the pre-defined grid structure for linking visualizations, flexibility for missing data and computational efficiency (see [14, 15]). Moreover, the SOM allows the user to explore the cluster structure on the two-dimensional SOM grid. Therefore, the results of the clustering can be presented in a user-friendly format easy to understand and interpret [8-14].

In this study, we apply a Weighted SOM (WSOM) [16] for weighted customer segmentation of department store data. The weighting scheme enables specifying the importance of customers based upon how much of their purchases consist of green products. We analyze retail consumer data collected through the loyalty-card program of a department store. The data consist of demographic, geographic, and actual behavioral data in the form of transaction data. To our knowledge, this approach – using BI processes and methods, has not been previously used for profiling of green consumers.

The reminder of the paper is structured as follows. In section two, we provide background information on green consumer behavior. We compile findings from earlier studies into two tables, to be used in the analysis of our results. In section three, we present the data from the department store and describe the basics of the WSOM. Section four describes how we built the model, the

1550-6037/13 $26.00 © 2013 IEEE
DOI 10.1109/IV.2013.37

results, and our analysis of green consumer behavior. In section five, we compare our results to the results in the previous literature. Section six concludes the paper.

II. GREEN CONSUMERS

This section reviews the literature on green consumer profiling. First, we provide definitions of key terms related to green consumer behavior and green products, as per their use in this study. Second, we review the literature on profiling green consumers. The aim of the review is to illustrate previous findings concerning the most relevant variables for explaining green consumer behavior, and to create a basis for comparison to this study.

A. Definitions

A green consumer is defined as a consumer that takes into account the environment in her purchasing behavior or in processes after the purchase, e.g., buys green products, chooses products and packages made from renewable or recycled materials, or conducts recycling (see e.g., [1, 17, 18]). In our context a green consumer is a person who buys green products from the department store in question.

A green product is most often defined as an organic product. Several previous studies have focused on organic products in grocery stores, e.g., fruits, vegetables, milk, cereals, and meat (see e.g., [19-22]). Our study does not include organic food, but focuses on green products in the department store in question. A description of these is given in Section 3.1.

B. Green consumer profiling

There is a broad literature on green consumer profiling, with the oldest studies being from the 1970s. Traditionally, demographic variables are used as the basis for market segmentations [23]. However, the results concerning the usefulness of demographic data for profiling of green consumers are contradictory (see, e.g., [24, 1, 25, 26]). In order to better profile green consumers, psychographic variables, such as attitude, activities, characteristics of personality, and intention of purchasing green products, are used in a number of studies (e.g., [24, 27, 28]). The literature shows that the main psychographic variables that influence green consumer behavior are environmental concern, intention, price, knowledge of the products, perceived consumer effectiveness, and healthy lifestyle (e.g., [27, 29-32]).

Based on an extensive literature review, we have compiled two tables. **Table 1** includes studies that rely upon interviews, questionnaires, polls, and queries as data collection instruments, in order to draw inferences on green consumers. It is split into two parts: studies that use demographic variables and studies that use psychographic variables. The table illustrates the contradictory results concerning the usefulness of demographic variables for the analysis of green consumers in previous studies. It also illustrates that psychographic variables are useful.

As the key focus of the present paper is on a data-driven segmentation of green consumers using loyalty-card data, we do not have access to psychographic information concerning the consumers. Therefore, a more thorough review focuses on the demographic variables and their importance in profiling a green consumer.

Table 2 summarizes findings in previous research concerning the most important variables determining the demographic profile of a green consumer. The references in different categories relate to characteristics of the green profiles and references in the importance column relate to the significance of the findings. E.g. no earlier research has shown that the variable "Children in household" would have a strong influence, or that having no children in the household would influence the green consumer behavior. The results show that the average green profile consists of young and old female consumers with a high education and income and children in the household, and that the most important variables determining a profile are age, gender, income and education (e.g., [22, 30, 33, 34]).

Our literature review shows that there is a large body of research into what drives green consumer behavior. Yet, Tables 1 and 2 also illustrate the lack of unanimity on a single profile of a green consumer. The poor results concerning the usefulness of demographic variables for profiling may be explained by the existence of multiple green consumer profiles. A conclusion of this is that demographic data are insufficient for discriminating between one profile of green and non-green consumers. The findings in Table 2 provide a starting point for identifying how multiple green profiles may be characterized.

TABLE 1. THE USEFULNESS OF DEMO- AND PSYCHOGRAPHIC VARIABLES FOR PROFILING GREEN CONSUMERS.

Demographics	Useful	Davies et al. 1995 [35]; Wandel and Bugge 1997 [36]; Thompson and Kidwell 1998 [19]; Wedel and Kamakura 2000 [23]; Magnusson et al. 2001 [20]; Chinnici et al. 2002 [33]; Banyte et al. 2010 [43]
	Not useful	Frank et al. 1972 [37]; McCann 1974 [38]; Herberger 1975 [39]; Samdahl and Robertson 1989 [40]; Banerjee and McKeage 1994 [41]; Scott and Willits 1994 [42]; Stern et al. 1995 [43]; Roberts 1996 [24]; Mainieri et al. 1997 [1]; Straughan and Roberts 1999 [25]; Tsakiridou et al. 2008 [30]
Psychographics	Useful	Straughan and Roberts 1999 [25]; Chinnici et al. 2002 [33]; Banyte et al. 2010 [34]
	Not useful	

TABLE 2. A REVIEW OF DEMOGRAPHIC VARIABLES THAT INFLUENCE GREEN CONSUMER BEHAVIOR. THE REFERENCES IN THE COLUMN "IMPORTANCE" IMPLY THAT THIS VARIABLE HAS A STRONG INFLUENCE ON GREEN CONSUMER BEHAVIOR.

Importance	Variable	Categories	References
Anderson and Cunningham 1972 [44]; Van Liere and Dunlap 1981 [45]; Samdahl and Robertson 1989 [40]; Roberts 1996 [24]; Bui 2005 [27]; Banyte et al. 2010 [34]	*Age*	Younger 18-30	Anderson and Cunningham 1972 [44]; Weigel 1977 [46]; Jolly 1991 [47]; Roberts 1996 [24]; Roberts and Bacon 1997 [48]; Chinnici et al. 2002 [33]; Banyte et al. 2010 [34]
		Middle aged 31-50	Roberts 1996 [24]; Chinnici et al. 2002 [33]; Bui 2005 [27]; Banyte et al. 2010 [34]
		Older 51->	Van Liere and Dunlap 1981 [45]; Samdahl and Robertson 1989 [40]; Roberts 1996 [24]; Wandel and Bugge 1997 [36]; Thompson and Kidwell 1998 [19]; Fotopoulos and Krystallis 2002 [49]; Tsakiridou et al. 2008 [30]
Van Liere and Dunlap 1981 [45]; Davies et al. 1995 [35]; Roberts 1996 [24]; Banyte et al. 2010 [34]	*Gender*	Male	
		Female	Banerjee and McKeage 1994 [41]; Davies et al. 1995 [35]; Roberts 1996[24]; Mainieri et al. 1997 [1]; Wandel and Bugge 1997 [36]; Laroche et al. 2001 [18]; Chinnici et al. 2002 [33]; Bui 2005 [27]; Lea and Worsley 2005 [21]; Padel and Foster 2005 [22]; Banyte et al. 2010 [34]
Van Liere and Dunlap 1981 [45]; Newell and Green 1997 [52]; Banyte et al. 2010 [34]	*Income*	Low	Samdahl and Robertson 1989 [40]; Fotopoulos and Krystallis 2002 [49]
		Middle	Chinnici et al. 2002 [33]; Bui 2005 [27]
		High	Herberger 1975 [39]; Davies et al. 1995 [35]; Magnusson et al. 2001 [20]; Bui 2005 [27]; Padel and Foster 2005 [22]; Tsakiridou et al. 2008 [30]; Banyte et al. 2010 [34]
Van Liere and Dunlap 1989 [45]; Schwarts and Miller 1991 [51]; Roberts 1996 [24]; Bui 2005 [27]; Banyte et al. 2010 [34]	*Education*	Low	Samdahl and Robertson 1989 [40]
		Middle	Chinnici et al. 2002 [33]
		High	Herberger 1975 [39]; Arbuthnot 1977 [50]; Jolly 1991 [47]; Schwartz and Miller 1991 [51]; Newell and Green 1997 [52]; Wandel and Bugge 1997 [36]; Magnusson et al. 2001 [20]; Hill and Lynchehaum 2002 [53]; Bui 2005 [27]; Padel and Foster 2005 [22]; Banyte et al. 2010 [34]; Ishaswini et al. 2011 [32]
	Children in household	Yes	Davies et al. 1995 [35]; Thompson and Kidwell 1998 [19]; Laroche et al. 2001 [18]; Fotopoulos and Krystallis 2002 [49]
		No	

III. METHODOLOGY

In this section, we first present the department store data used in this study. Then, we introduce the basics of the Weighted WSOM used in the study.

A. Data

The data, spanning the period 2007-2009, are from a department store belonging to a national retailer. Customers' demographic information is obtained through the retailer's loyalty card system, and their purchasing behavior is summarized from the transaction database.

In this study, a customer is considered to be green to a degree if he or she has purchased any green products within the two year period. The green products are located in different departments of the department store. Beauty products include soaps, shampoos, natural supplements (e.g., blueberry extract) and cosmetics and beauty items (e.g., sponges) produced according to organic requirements including heavy restrictions on the use of pesticides or other chemicals. In the departments for men's and women's clothing, women's shoes, sports and leisure, the products consist mainly of clothes, socks, underwear and shoes made from recycled or organic materials, e.g., organic cotton or wool, or shoes made of organic materials instead of leather. In the children's department, green products include both clothes made from organic materials as well as toys and children's products made according to green specifications In the home department, there are products and packaging made from recycled, renewable or organic materials.

The degree of green consumer behavior demonstrated by the customer is defined according to the percentage of purchases of green products. This sets the basis for weighting customers as per their eco-consciousness. However, to increase the reliability of the results, only customers with a total number of 24 or more purchased products are included. The training dataset, containing over 1.5 million customers (almost 30% of the whole population in Finland), consists of six demographic variables, eight product categories and nine purchasing behavior variables.

The demographic variables are gender, child decile (an estimated probability of children living in the same household), estimate of income, age and native language (Finnish or Swedish). The behavioral variables are average transaction spending (Eur), total spending (Eur), basket size, purchase frequency, total number of items, total spending amount in green products (Eur), total number of green products (no of items), shopping occurrences during working time (%), average item value (Eur), per-department spending at the department store: Leisure, Beauty, Home, Children, Sports, Men, Women, and Women's shoes (spending in % /department).

B. Weighted Self-Organizing Map

The SOM [54] is a method for clustering and projection. Weighting of the SOM is in itself not novel. In his early works, Kohonen [55] introduced one type of weighting of data for eliminating border effects to the edges of a SOM grid. Likewise, Kim and Ra [56] and Kangas [57] weighted data blocks based upon their statistical properties in image compression. Similar instance weighting may also be applied to balance imbalanced samples (e.g., [58]), and to vary the influence of variables on distance calculations in the matching phase of SOM learning (e.g., [58]). The weighting in the WSOM [16] differs by focusing on a user-specified

importance of data for learning, in the unsupervised case for forming clusters. The weighted version of the SOM used herein thus follows the approach in Sarlin [16], by augmenting the batch version of the SOM training algorithm with an instance-specific weight. The functioning of the WSOM is a simple and intuitive approach for dealing with importance-varying instances, while still preserving the general properties of the SOM. Motivations for using the batch rather than the sequential algorithm are the reduction of computational cost and reproducible results (given similar initializations). The WSOM is initialized by setting the reference vectors to the direction of the two principal components of the input data. This type of initialization not only has the advantage of decreased computational cost and reproducibility, but has also been shown to be important for convergence when using the batch SOM [59].

Following Kohonen [6], the WSOM iterates in two steps through 1,2,…,t. The first step follows the standard SOM matching formula by assigning each input data vector x to its best-matching unit (BMU) mc based upon the Euclidean distance, i.e., $\|x - m_c(t)\| = \min_i \|x - m_i(t)\|$. In the second step, each reference vector m_i (where i=1,2,…,M) is adjusted using the following weighted counterpart of the batch update formula:

$$m_i(t+1) = \frac{\sum_{j=1}^{N} w_j h_{ic(j)}(t) x_j}{\sum_{j=1}^{N} w_j h_{ic(j)}(t)}, \quad (1)$$

where weight w_j represents the importance of x_j for the learning of patterns, index j indicates the input data vectors that belong to node c, and N is the number of the data vectors. The neighborhood $h_{ic(j)} \in (0,1]$ is defined as a Gaussian function that increases in the distance between the coordinates of the reference vectors m_c and m_i on the two-dimensional grid, i.e., $\|r_c - r_i\|^2$, and where the radius of the neighborhood $\sigma(t)$ is a monotonically decreasing function of time t. The information products of the WSOM can, as those of standard SOMs, be linked to the grid of units. Hence, the high-dimensional space of the WSOM may be described using feature planes by visualizing the spread of individual variables (or other linked information) on the same WSOM grid structure.

C. Clustering of the WSOM

A second-level clustering of the SOM has in previous studies been shown to be an effective approach for clustering and visualization (see, e.g., [60]). This involves clustering data into units of the SOM and the SOM into homogeneous groups of units. In particular, Li [61] shows the superiority of the combinatorial approach of the SOM and Ward's [62] hierarchical clustering over other classical clustering algorithms. Hence, this study applies Ward's method for clustering the WSOM. Ward's clustering starts with each unit being treated as a separate cluster. Then, the two clusters (or units) with the minimum distance are merged in each step until there is only one cluster left on the map. Then, a suitable cut-off (number of clusters) is chosen for analysis. In this paper, the number of clusters K is determined by the Silhouette index [63]. It is a widely used cluster validity measure that takes into account both cluster compactness and cluster separation. The Silhouette index for a clustering solution is simply the average of the Silhouette coefficients of all observations (i.e., units here), which are based upon the average distance between unit i and all other units in the same cluster. The higher the value, the better the observations are clustered. Similar to the feature plane visualizations, the clustering may be shown on the WSOM grid through color coding.

IV. Visual Data-Driven Profiling of Green Consumers

This section presents the green consumer data, explains the segmentation model, and visually presents the profiles for green consumers.

A. Data on green consumers

When computing the degree of green consumption of a customer, we have relied upon total purchases of green products during a two-year period. That is, the degree of green behavior of a customer is defined to be the share of green products (in items) in her basket during a two-year period.

Of the 1.5 million customers of the department store, 20% have purchased at least one green product during the two-year period. These customers account for 44.7% of total purchases, while 7.7% of their total purchases consist of green products. Figure 1 presents how green consumers differ from the general customer base of the department store.

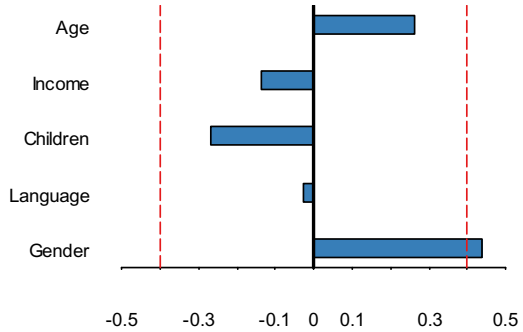


Figure 1. Distances of the average eco customer to the general customer base.

Figure 1 shows the deviation of the green consumers' mean from that of all customers. The mean for the green customers is a weighted average as per the degree of greenness. The only statistically significant difference at level .05 is that the green consumers are mainly female. Other differences include that the green consumer is older than the average consumer, has a lower estimated income and a low probability of having children living in the same household. The lower probability of having children

is likely explained by the higher age of the consumers, whose children are no longer living in the same household. According to the earlier studies presented in Table 2, gender and age are the most significant demographic variables for the profile of a green consumer. As the earlier studies also pointed out, there is no unanimous profile for a green consumer. Instead, several profiles for green consumers exist with a varying degree of green consumer behavior.

B. The segmentation model

Antil and Bennett [3] argued that there are different green consumer profiles depending on the degree of green consumer behavior that they exhibit. Therefore, in order to find out how 'green' green consumers are, the profiles were derived using demographic data combined with actual transaction data. The result is a segmentation of green consumers with the WSOM, where the consumers are given a certain weight depending on the share of green products (in number of items) that they purchased. Accordingly, those who have not purchased any green products, are given a weight of zero, and therefore, are excluded from this study. The WSOM was applied to the demographic variables, whereas behavioral data were only associated to the model. Association refers to computing averages of additional variables for each unit of the WSOM, and does thus not affect training. As shown in Figure 2 with the Silhouette index, the optimal number of clusters on the WSOM was determined to be five.

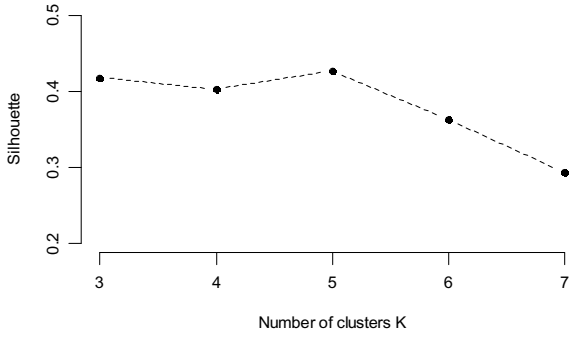
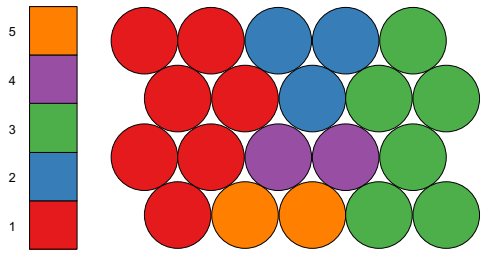


Figure 2. Validation of the Ward's clustering for K=3,4,…,7 with the Silhouette index.

The five clusters (or segments) are shown in Figure 3. The feature planes of the demographic variables are presented in Figure 4. Here, darker colors indicate higher values and lighter lower values, e.g. the variable age indicates that younger consumers are located in segments 1 and 5, while older consumers are located in segment 3. Figure 5 presents the associated behavioral and product mix variables, in addition to the degree of green products. E.g. consumers who purchased most of the green products are located in segments 3 and 4.



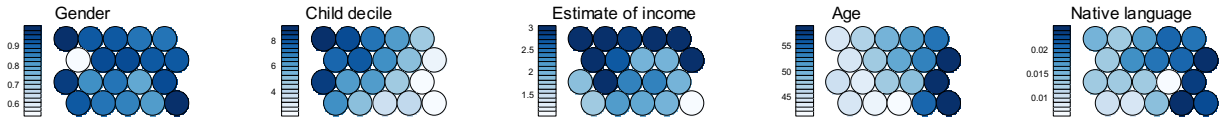
Notes: The coloring uses ColorBrewer's qualitative scale [64], such that segments are differentiated in hue contrast with nearly constant saturation and lightness.

Figure 3. The WSOM with five segments. Each color indicates a segment.

C. Profiles of green consumers

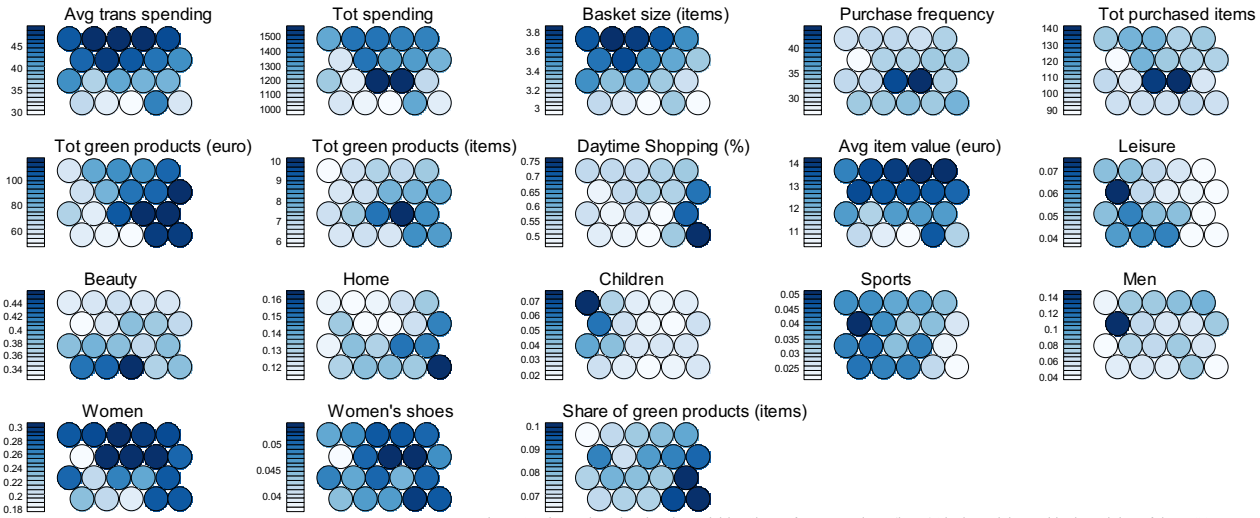
The second level clustering results in five different profiles of green consumers, as illustrated in Table 3. The results show that there are two groups displaying a high degree of green purchases: Segments 3 and 4. Segment 3 contains typically elderly, high spending customers, who purchase a high degree of green products, mainly during daytime. The majority of the customers are female, and their product mix contains products from the Home, Women and Women's shoes departments. Segment 4, on the other hand, consists of middle-aged, frequent and high-spending shoppers, who purchase a high share of green products and shop from a broader category of departments than customers in Segment 3.

There are also three groups exhibiting lesser degrees of green behavior. Segment 2 includes high spending middle-aged consumers, who also spend a fair amount on green products. They primarily purchase items from the women's section of the department store. Segment 1 represents a group of high spending consumers, who visit the department store seldom, but often purchase many items at a time and from several different departments. These consumers purchase very little green products. They are younger than the average green consumer, have lower than average estimated income levels, and have a higher likelihood of children in household. Segment 5 is a group of low-spending young customers, who mainly purchase items from the beauty section and purchase very little green products.



Notes: The feature planes represent the spread of values of individual variables on the WSOM grid. The coloring uses a scale from ColorBrewer [64], in which variation of a blue hue occurs in luminance and light to dark represent low to high values according to a feature plane-specific scale.

Figure 4. WSOM feature planes for the demographic variables.



Notes: See notes to Figure 4. It is worth noting that the variable "Share of green products (items)" is the weight used in the training of the WSOM.

Figure 5. WSOM feature planes for the behavioral and product variables.

TABLE 3. THE FIVE IDENTIFIED GREEN CONSUMER PROFILES.

Segment	Size (%)	Demographics	Product Mix	Behavioral Profile
Segment 1: High spending families, with low degree of green purchases	38.2	Have children Younger to middle-aged High est. income	Wide variety of departments	High avg transaction spending High tot spending Low purchase frequency Large basket size High item value Low purchases of green products
Segment 2: Elderly high spending consumers, fair degree of green purchases	9.7	Avg likelihood of children Middle-aged High est. income	Women Women's shoes	High avg transaction spending Large basket size High item value Low purchase frequency
Segment 3: Elderly, highly green consumers	41.6	Older	Home Women Women's shoes	High tot nr of eco products High avg transaction spending High item value Shop during working hours
Segment 4: Middle-aged frequent shoppers, high degree of green purchases	0.4	Middle-aged	Wide variety of departments	High tot nr of eco products High share of eco items High tot spending High frequency High tot sales items
Segment 5: Younger beauty item shoppers, low degree of green purchases	10.1	Younger	Beauty	Low tot spending Low avg item value Avg frequency

V. DISCUSSION AND COMPARISON OF RESULTS

The inconclusive results of earlier studies presented in Table 2 verify that there exists no unanimous profile for a green consumer. Often, the variables age and gender are cited as most important, as are income and education. According to the literature, the existence of children in the same household does not seem to be a distinguishing variable regarding green consumer behavior.

According to our study, the customers with the highest degree of green consumer behavior are located in Segments 3 and 4. They are middle-aged or older customers with a lower likelihood of children living in the same household. This description of green consumers has been provided by many studies since the 1970s. Most of the studies claim that a higher income is important, but we could not find any support for this.

Customers that have a slightly lower degree of green consumer behavior are located in Segments 2 and 1. These customers are profiled as younger consumers with higher income and with children living in the same household. As was mentioned, earlier studies have shown that the existence of children in the same household is not a significant variable of the profile for a green consumer. We do find it as a determinant of green consumers, but only of those with a low degree of green consumer behavior, which supports the existence of multiple profiles.

Customers in Segment 5 possess the lowest degree of green consumer behavior. The profile for the customers in this segment is younger beauty item shoppers. Several

studies have concluded that even if young people are often interested in environmentally friendly issues, they do not have the purchasing power to shop only green products. Even though earlier studies have shown that younger people often have a lower income compared to other age groups, our data are not suitable for assessing this effect as loyalty-card members are seldom young enough.

VI. Conclusion

While green consumer behavior has been the focus of both research in academia and the private sector for many years, the literature is inconclusive concerning the importance of different demographic and psychographic variables for determining the profile of a green consumer. In fact, previous research has indicated that there likely exist several different profiles for green consumers, which could be identified based on green consumer behavior. Previous research has also questioned how reliable analyses based on questionnaires are when it comes to green consumers. In this study, we have used Business Intelligence and visualization processes and methods to analyze green consumer behavior, particularly a data-driven method in the form of the Weighted Self-Organizing Maps (WSOMs). We have shown how these analyses and processes with real-world data from a department store can be performed. The result of such a process and the data-driven method were visually presented as in the form of five different profiles based upon which decision makers could have taken actions.

We have identified five different profiles of green consumers; two with high purchases, two with low purchases, and one in between. The most clearly discriminating variables were age and children in the household. Green consumers were likely to be older than average, and less likely to have children in the household, while younger consumers and customers with children in the household exhibited less green behavior. Income level, on the other hand, did not prove to be a strong indicator of propensity towards green consumer behavior, which is somewhat in contradiction with expectations based upon the literature.

While there is a broad body of literature on green consumers, the focus of those studies has mainly been on propensity to buy organic food. In our case, data are from a department store that does not carry organic food products. Instead, green products have been defined as products that have been produced according to green specifications, e.g., beauty products and children's products produced according to organic requirements, shoes manufactured of other, more green materials than leather, clothes made of organic cotton or organic wool and products made of recycled materials. Also, consumer behavior and attitudes toward green products varies over time and across countries. We acknowledge that a study is only applicable for the time and area from where the consumers come from. In our case, our study is limited to green consumer behavior at this specific department store in Finland, between the years 2007 to 2009. Nevertheless, we provide an approach which is suitable for describing the real behavior of consumers independent of time and location.

Acknowledgment

We thank the case organization for fruitful cooperation and providing data. The financial support from the Academy of Finland (grants Nos. 127656 and 127592) and the Foundation of Nokia Corporation is gratefully acknowledged. We also thank anonymous reviewers, who helped us to improve this paper.

References

[1] Mainieri, T., Barnett, E.G., Valdero, T.R., Unipan, J.B. and Oskamp, S. (1997). Green Buying: The Influence of Environmental Concern on Consumer Behavior, The Journal of Social Psychology, 137 (2), 189-204.

[2] Young, W., Hwang, K., McDonald, S., Oates, C.J. (2010). Sustainable Consumption: Green Consumer Behaviour when Purchasing Products. Sustainable Development, 18, pp. 20-31.

[3] Antil, J.H. and Bennet, P.D. (1979). Construction and validation of a scale to measure socially responsible consumption behaviour, in Henion, K.E. and Kinnear, T.C. (Eds.), The conserver Society, American Marketing Association, Chicago, IL, 51-68.

[4] Rud, O. (2009). Business Intelligence Success Factors: Tools for Aligning Your Business in the Global Economy. Hoboken, N.J.: Wiley & Sons.

[5] Chung, W., Chen, H., Nunamaker Jr., J.F. (2005). A Visual Framework for Knowledge Discovery on the Web: An Empirical Study of Business Intelligence Exploration. Journal of Management Information Systems, Vol. 21, No.4, pp. 57-84.

[6] Kohonen, T. (2001). Self-Organizing Maps, 3rd edition, Berlin: Springer-Verlag.

[7] M.C. Ferreira de Oliveira and H. Levkowitz, From visual data exploration to visual data mining: A survey, IEEE Transactions on Visualization and Computer Graphics, vol. 9, 378-394, 2003.

[8] Yao, Z., Sarlin, P., Eklund, T., Back, B. (2012). Combining Visual Customer Segmentation and Response Modeling. Proceedings of the 20th European Conference on Information Systems (ECIS'12), Barcelona, Spain, June 11--13, AISeL.

[9] Holmbom, A.H., Eklund, T. and Back, B. (2011). Customer portfolio analysis using the SOM, International Journal of Business Information Systems, vol. 8, 396-412.

[10] Yao Z, Holmbom A, Eklund T, Back B (2010). Combining unsupervised and supervised data mining techniques for conducting customer portfolio analysis. In: Proceedings of the 10th Industrial conf on data mining, Springer, Germany, Berlin.

[11] Lingras P, Hogo M, Snorek M, West C (2005) Temporal analysis of clusters of supermarket customers: conventional versus interval set approach. Information Sciences 172 (1–2):215-240.

[12] Lee SC, Suh YH, Kim JK, Lee KJ (2004) A cross-national market segmentation of online game industry using SOM. Expert systems with applications 27 (4):559-570.

[13] Vellido A, Lisboa P, Meehan K (1999) Segmentation of the on-line shopping market using neural networks. Expert Systems with Applications 17 (4):303-314

[14] Sarlin P (1999) Data and Dimension Reduction for Visual Financial Performance Analysis. TUCS Technical Reports. TUCS

[15] Vesanto J (1999) SOM-based data visualization methods. Intelligent data analysis 3 (2):111-126.

[16] Sarlin, P. (2013). A Weighted SOM for classifying data with instance-varying importance. International Journal of Machine Learning and Cybernetics, In press.

[17] Follows, S.B. and Jobber, D. (2000). Environmentally responsible purchase behavior: a test of a consumer model. European Journal of Marketing, Vol. 34, No. 5/6, pp. 723-46.

[18] Laroche, M., Bergeron, J. and Barbaro-Forleo, G. (2001). Targeting consumers who are willing to pay more for environmentally friendly products, J of Consumer Marketing, 18 (6), 503-20.

[19] Thompson, G.D. and Kidwell, J. (1998). Explaining the choice of organic produce: cosmetic defects, prices and consumer preferences, American J of Agricultural Econ, 80 (2), 277-87.

[20] Magnusson, M., Arvola, A., Koivisto Hursti, U., Aberg, L. and Sjoden, P. (2001). Attitudes towards organic foods among Swedish consumers, British Food Journal, 103 (3), 209-26.

[21] Lea, E. and Worsley, T. (2005). Australians' organic food beliefs, demographics and values, British Food Journal, 107 (11), 855-69.

[22] Padel, S., Foster, C. (2005). Exploring the gap between attitudes and behavior. Understanding why consumers buy or do not buy organic food. British Food Journal, 107 (8), 606-625. Emerald Group Publishing Limited.

[23] Wedel M., Kamakura W.A. (2000). Market segmentation-conceptual and methodological foundations, 2nd ed. Boston: Kluwer.

[24] Roberts, J.A. (1996). Green Consumers in the 1990s: Profile and Implications for Advertising, Journal of Business Research 36, 217-231. Elsevier Science Inc.

[25] Straughan, R.D., Roberts, J.A. (1999). Environmental segmentation alternatives: a look at green consumer behavior in the new millennium. Journal of Consumer Marketing, 16 (6), 558-575, MCB University Press

[26] Diamantopoulos, A., Schlegelmilch, B.B., Sinkovics, R.R., Bohlen, G.M. (2003). Can socio-demographics still play a role in profiling green consumers? A review of the evidence and an empirical investigation, J of Business Research 56, 465-480.

[27] Bui, My H. (2005). Environmental Marketing: A model of consumer behavior. In Proceedings of the Annual Meeting of the Association of Collegiate Marketing Educators, Dallas, Texas March 1-5, 2005, 20-28.

[28] Paco, A., and Raposo, M. (2010). Green consumer market segmentation: empirical findings from Portugal, International Journal of Consumer Studies, Vol 34, pp. 429-436.

[29] Kalafatis, S.P., Pollard, M., East, R. and Tsogas, M.H. (1999). Green marketing and Ajzen's theory of planned behaviour: a cross-market examination, Journal of Consumer Marketing, 16 (5), 441-460. MCB University Press.

[30] Tsakiridou, E., Boutsouki, C., Zotos, Y., and Mattas, K. (2008). Attitudes and behavior towards organic products: an exploratory study, International Journal of Retail & Distribution Management. 36 (2), 158-175. Emerald Group Publishing Limited.

[31] Coad, A., de Haan, P., Woersdorfer, J.S. (2009). Consumer support for environmental policies: An application to purchases of green cars. Ecological Economics, 68, pp. 2078-86.

[32] Ishaswini, S. and Datta, K. (2011). Pro-environmental Concern Influencing Green Buying: A Study on Indian Consumers. Int. Journal of Business and Management, 6 (6), 124-133.

[33] Chinnici, G., D'Amico, M., Pecorino, B. (2002). A multivariate statistical analysis on the consumers of organic products. British Food Journal, 104 (3/4/5), 187-199.

[34] Banytė, J., Brazionienė, L., Gadeikienė, A. (2010). Investigation of green consumer profile: a case of lithuanian market of eco-friendly food products. Economics and Management: 2010. 15. ISSN 1822-6515, 374-383.

[35] Davies, A., Titterington, A.J. and Cochrane, A. (1995). Who buys organic food? A profile of the purchasers of organic food in N. Ireland, British Food Journal, 97 (10), 17-23.

[36] Wandel, M. and Bugge, A. (1997). Environmental concern in consumer evaluation of food quality, Food Quality and Preference, 8 (1), 19-26.

[37] Frank, R.E., Massy, W.F., Wind, Y.(1972). Market segmentation, Englewood Cliffs: Prentice-Hall.

[38] McCann, J.M. (1974). Market segment response to the marketing decision variables, J of Market Responding, 11 (4), 399-415.

[39] Herberger, R.A. Jr. (1975, reprinted in 2002). The Ecological Product Buying Motive: A Challenge for Consumer Education, The Journal of Consumer Affairs, 187-195. EBSCO Publishing

[40] Samdahl D.M., Robertson R. (1989). Social determinants of environmental concern: specification and test of the model. Environmental Behavior, 21 (1), 57–81.

[41] Banerjee, B. and McKeage, K. (1994). How green is my value: exploring the relationship between environmentalism and materialism, in Allen C.T. and John, D.R. (Eds) Advances in Consumer Research, Association for Consumer Research, Provo, UT, 21, 147-52.

[42] Scott, D., Willits, F.K. (1994). Environmental attitudes and behavior: a Pennsylvania survey, Environmental Behavior, 26 (2), 239-60.

[43] Stern, P.C., Dietz T., Guagnano G.A. (1995). The new ecological paradigm in social–psychological context. Environmental Behavior, 27 (6), 723-43.

[44] Anderson, T. Jr. and Cunningham, W.H. (1972). The socially conscious consumer, Journal of Marketing, 36 (7), 23-31.

[45] Van Liere, K. and Dunlap, R. (1981). The social bases of environmental concern: a review of hypotheses, explanations, and empirical evidence, Public Opinion Quarterly, 44 (2), 181-97.

[46] Weigel, R. H. (1977). Ideological and demographic correlates of proecological behavior, The J of Social Psychology, 103, 39-47.

[47] Jolly, D. (1991). Differences between buyers and nonbuyers of organic produce and willingness to pay organic price premiums, Journal of Agribusiness, 9 (1), 97-111.

[48] Roberts, J.A. and Bacon, R. (1997). Exploring the subtle relationships between environmental concern and the ecologically conscious consumer behaviour, J of Bus. Research, 40, 79-89.

[49] Fotopoulos, C. and Krystallis, A. (2002). Purchasing motives and profile of the Greek organic consumer: a countrywide survey, British Food Journal, 104 (9), 730-765.

[50] Arbuthnot, J. (1977). The roles of attitudinal and personality variables in the prediction of environmental behavior and knowledge, Environment and Behavior, 9, 217-232.

[51] Schwartz, J. and Miller, T. (1991). The earth's best friends, American Demographics, 13, February, 26-33.

[51] Newell, S. J. and Green, C.L. (1997). Racial differences in consumer environmental concern, The Journal of Consumer Affairs, 31 (1), 53-69.

[53] Hill, H. and Lynchehaum, F. (2002). Organic milk: attitudes and consumption patterns, British Food Journal, 104 (7), 526-542.

[54] Kohonen, T. (1982). Self-organized formation of topologically correct feature maps, Biological Cybernetics 66, 59–69.

[55] Kohonen, T. (1993). Things you haven't heard about the Self-Organizing Map, Proceedings of the International Conference on Neural Networks (ICNN 93), 1147-1156.

[56] Kim, K.Y., Ra, J.B. (1993). Edge preserving vector quantization using self-organizing map based on adaptive learning, Proceedings of the International Joint Conference on Neural Networks (IJCNN 93), 11, 1219-1222. IEEE Press

[57] Kangas, J. (1995). Sample weighting when training self-organizing maps for image compression, Proceedings of the 1995 IEEE Workshop on NNs for Signal Processing, 343-350.

[58] Vesanto, J., Himberg, J., Alhoniemi, E., Parhankangas, J. (1999). Self-Organizing Map in Matlab: the SOM Toolbox, Proceedings of the Matlab DSP Conference, 1999, 35-40.

[59] Forte, J.C., Letrémy, P., Cottrell, M., 2002. Advantages and drawbacks of the Batch Kohonen algorithm, in: Verleysen, M., (Ed.), Proceedings of the 10th European Symposium on Neural Networks, Springer-Verlag, Berlin, pp. 223–230.

[60] Vesanto, J. and Alhoniemi, E. (2000). Clustering of the self-organizing map. IEEE Transactions on NNs, 11 (3), 586-600.

[61] Li, H. (2005). Data visualization of asymmetric data using Sammon mapping and applications of self-organizing maps, University of Maryland (College Park, Md.)

[62] Ward Jr J.H. (1963). Hierarchical grouping to optimize an objective function, Journal of the American statistical association 58 (301), 236-244.

[63] Rousseeuw, P.J., Kaufman, L. (1990). Finding groups in data, JohnWiley & Sons, Inc, New York.

[64] Harrower, M. and Brewer, C. (2003). Colorbrewer.org: an online tool for selecting colour schemes for maps," The Cartographic Journal, vol. 40(1), pp. 27–37.

Publication 5

Holmbom, A.H., Rönnqvist, S., Sarlin, P., Eklund, T., Back, B. (2013). Green vs. Non-Green Customer Behavior: A Self-Organizing Time Map Over Greenness, In Wei Ding, Takashi Washio (Eds.), *IEEE 13th International Conference on Data Mining Workshops*, IEEE International Conference on Data Mining, 1–7, IEEE.

2013 IEEE 13th International Conference on Data Mining Workshops

Green *vs*. non-green customer behavior

A Self-Organizing Time Map over greenness

Annika H. Holmbom[1], Samuel Rönnqvist[1], Peter Sarlin[1,2], Tomas Eklund[1] and Barbro Back[1]
[1]TUCS – Turku Centre for Computer Science, Dept. of IT, Åbo Akademi University, Turku, Finland
{annika.h.holmbom, samuel.ronnqvist, peter.sarlin, tomas.eklund, barbro.back}@abo.fi
[2]Arcada University of Applied Sciences, Dept. of Business, IT and Media, Helsinki, Finland

***Abstract*—Companies have traditionally used segmentation approaches to study and learn more about their customer base. One area that has attracted considerable amounts of research in recent years is that of green customer behavior. However, the approaches used have often been static clustering approaches and have focused on identifying green vs. non-green customers. In fact, results have been non-unanimous and not seldom contradictory. An alternative approach is to study customers according to degrees of green purchases. Recently, a Self-Organizing Time Map (SOTM) over any variable of cardinal, ordinal or higher level of measurement has been proposed. The key idea is to enable the exploration of changes in cluster structures over not only the time dimension, but also any other variable. This paper presents an application of the SOTM to demographic and behavioral customer data, in which the key focus is on assessing how customer behavior varies over customers' degree of greenness.**

Keywords—Self-Organizing Time Map (SOTM); Green customer behavior; Clustering; Customer Relationship Management (CRM), Visual Analytics

I. INTRODUCTION

With increasing competition in the international retail and wholesale industries, companies have long been applying techniques from the field of Customer Relationship Management (CRM) to gain a competitive advantage through greater knowledge about their customer base. A commonly used approach for understanding the customer base and its needs is through the use of customer segmentation (see, e.g., [1,2]). The focus of customer segmentation is on dividing the customer base into distinct and separable segments of similar customers, with similar needs and behavior. Customer segmentation is not a trivial task; it is difficult to obtain both rigorous and at the same time relevant (actionable) results.

One area of considerable interest in both industry and academia is green customer behavior, an area which has long been predicted to become an important market segment ([3-5]). However, while it has been the topic of both research and market studies, there is little consensus about the actual profile of a green customer, and therefore, very little actionable knowledge for decision makers. Research conducted in the area of green customer behavior can be roughly divided into three streams: (*i*) research based solely on demographic data, (*ii*) research based on demographic and psychographic data provided by respondents, and (*iii*) research based on demographic and purchasing data. A common conclusion is that one cannot identify the profile of a green customer based solely on demographic or psychographic data, and most studies instead base the profile on a combination of the two types of data (see, e.g., [3,4]). In particular, studies profiling green customers have primarily relied upon self-reporting methods, such as questionnaires, polls, and queries (see, e.g., [6]). However, questionnaire data can be problematic for this purpose. While customers are assumed to answer the questions according to the best of their knowledge, they tend to answer according to their intentions or expected social norms, which might not reflect their actual behavior. For instance, Mainieri et al. [4] found that 75% of the respondents were willing to pay a higher price for environmentally friendly products, while only 14% actually acted upon it. Defra [7] reported that this "attitude-behavior gap" was 30% in UK in 2006. Green customers adhere to a green purchasing behavior to a certain degree, as they purchase both environmentally friendly products and non-green products [5, 8]. Hughner et al. [9] also found evidence of this gap. They showed that despite generally favorable attitudes customers hold for organic food (between 46-67% of the population), actual purchase behavior forms only 4-10% of different product ranges.

The third, newer line of research in the area focuses on data-driven assessment based upon actual transaction data. Thus, the clustering or segmentation can be performed based upon transaction data, with which also the degree of green behavior demonstrated by a customer can be objectively defined by the share of purchased green products. The Self-Organizing Map (SOM) [10] has shown promising results for visual customer segmentation (see, e.g., [11-17]). Unlike most other clustering algorithms, the SOM possesses simultaneous clustering and projection capabilities, which makes it a suitable technique for data and dimensionality reduction. Reasons for choosing the SOM over other similar methods are: the pre-defined grid structure that supports comparison and linking of visualizations, flexibility for missing data, as well as computational efficiency (see [18, 19]).

A common problem with the segmentation literature is that it is mainly based on static, general clusters, where much of the variance within the clusters is potentially disregarded. For instance, in a segmentation identifying green customers, one would likely miss non-linear relationships between variables within the cluster. For example, the influence of a variable describing income level, while often cited as an important contributing factor in green behavior, might vary within the segment of green customers. In a recent study, Holmbom et al.

978-0-7695-5109-8/13 $31.00 © 2013 IEEE
DOI 10.1109/ICDMW.2013.103

[6] used a further developed SOM, namely the Weighted SOM (WSOM), to analyze green customer behavior. The WSOM accounts for the 'degree' of how green a customer is by giving customers who buy more green products a higher weight, thereby better extracting the profiles of customers actually exhibiting green purchasing behavior.

While [6] provides a solution to a truly green segmentation, it does not aid in comparing clustering results on two different datasets, such as customers with different degrees of actual green purchases (e.g., 10% vs. 35% of shopping basket content or even green vs. non-green customers). In this paper, we present an approach to green customer analysis that uses another adaptation of the standard SOM, namely the Self-Organizing Time Map (SOTM) [20]. A recent extension to the standard SOTM presents a solution over any variable of cardinal, ordinal or higher level of measurement [21]. The key idea is to enable the exploration of changes in cluster structures not only over the time dimension, but over any variable. This enables comparing cluster structures in a number of different datasets. In this paper, we present an application of the SOTM to demographic and behavioral customer data, in which the key focus is on assessing how customer behavior varies over customers' degree of greenness, i.e., datasets ranging from non-green to green customers. Thus, we replace the time dimension of the SOTM with the degree of green purchases.

This paper is organized as follows. In section II, we present the SOTM and the data used in this study. In section III, we present the experiments carried out, and in particular, analyze the results. Finally, in section IV, we present the conclusions of the study.

II. Methods and Data

This section presents the SOTM over any variable, its visualization and the used data.

A. Self-Organizing Time Map over any variable

The SOTM uses the capabilities of the SOM for abstraction of patterns in data. In the form that the SOTM was presented in [20], it provides means for abstractions of temporal structural changes by illustrating how cross-sections evolve over time in one-dimensional SOMs [10]. The standard SOTM focuses on visual dynamic clustering. This section presents the SOTM over any variable [21], which replaces the x-dimension of time t with any variable v. The properties of variable v (where v=1,2,…,V) follow those of time t in the original SOTM by being a discretized ordinal or cardinal variable (or higher level of measurement), having arbitrary frequency and being related to all entities in Ω.

To observe the cross-sectional structures of the dataset over variable v, the SOTM performs a mapping from the input space $\Omega(v)$, with a probability density function $p(x,v)$, onto a one-dimensional array $A(v)$ of output units $m_i(v)$ (where i=1,2,…,M) and preservers the orientation between consecutive patterns through short-term memory that initializes reference vectors of $A(v\text{-}1)$ with those of $A(v)$. In essence, the SOTM over variables follows the two standard steps from the SOM paradigm. First, it locates best-matching units (BMUs) by a matching of each data to the unit with shortest Euclidean distance, i.e., $\min\|x(v) - m_i(v)\|$, and then it updates each reference vector $m_i(t)$ through a time-restricted version of the batch update formula.

$$m_i(v) = \frac{\sum_{j=1}^{N(v)} h_{ic(j)}(v) x_j(v)}{\sum_{j=1}^{N(v)} h_{ic(j)}(v)}, \quad (2)$$

where index j indicates the input data that belong to unit c and the neighborhood function $h_{ic(j)}(v) \in [0,1]$ is defined as a Gaussian function

$$h_{ic(j)}(v) = \exp\left(-\frac{\|r_c(v) - r_i(v)\|^2}{2\sigma^2}\right), \quad (3)$$

where $\|r_c(v) - r_i(v)\|^2$ is the squared Euclidean distance between the coordinates of the reference vectors $m_c(v)$ and $m_i(v)$ on the one-dimensional array, and σ is the user-specified neighborhood parameter.

B. Visualizing the SOTM

The visualization of the SOTM makes use of approaches common in the SOM literature (see [20] for further information). The multidimensionality can be disentangled with feature planes. They show the distribution of values for each variable on the SOTM grid, which provides means for assessing how individual variables evolve over time. The feature planes are produced using a blue color scale (light represents low and dark high values). Sammon's mapping [22], a multidimensional scaling technique focusing on local pairwise distances, is used for visualizing the multidimensional structural properties of SOTMs. In brief, it enables examining structural properties for each cross-section v (vertically) and changes in structures (horizontally) by matching pairwise distances of the SOTM units with their original distance in the high-dimensional space. A Sammon's mapping of the SOTM plots all SOTM units ($m_i(v)$ where v=1,2,…,V) to one dimension and assigns the value of Sammon's dimension to each unit. Then, the Sammon's dimension is used as an input to a coloring method for visualizing the cluster structure of the SOTM. The coloring uses the uniform color space CIELab, where perceptual differences of colors represent distances in data. However, as also the individual SOMs of the SOTMs are one-dimensional, we only use one color dimension (blue to yellow) to represent differences between units. Fig. 1 illustrates this visualization procedure.

C. The Data

This paper uses data, spanning the period 2007-2009, from a department store belonging to a national retailer. The demographic information of customers is obtained through the loyalty card system of the retailer, and their transaction database is used for summarizing purchasing behavior for the customers.

The training dataset, containing over 1.5 million customers (almost 30% of the whole population in Finland), consists of six demographic variables, eight product categories and nine purchasing behavior variables.

The demographic variables are gender, child decile (an estimated probability of children living in the same household), estimate of income, age and native language (Finnish or Swedish). The behavioral variables are average transaction spending (EUR), total spending in all stores belonging to the loyalty card program (EUR), basket size, purchase frequency, total number of items, total spending amount in green products (EUR), total number of green products (no. of items), shopping visits during working hours (%), average item value (EUR), per-department spending at the department store: Leisure, Beauty, Home, Children, Sports, Men, Women, and Women's shoes (spending in % / department).

In this study, a customer is considered to be green to a degree if he or she has purchased any green products within the two year period. The green products are located in different departments of the department store. Beauty products include soaps, shampoos, natural supplements (e.g., blueberry extract) and cosmetics and beauty items (e.g., sponges) produced according to organic requirements including heavy restrictions on the use of pesticides or other chemicals. In the departments for men's and women's clothing, women's shoes, sports and leisure, the products consist mainly of clothes, socks, underwear and shoes made from recycled or organic materials, e.g., organic cotton or wool, or shoes made of organic materials instead of leather. In the children's department, green products include both clothes made from organic materials as well as toys and children's products made according to green specifications. In the home department, there are products and packaging made from recycled, renewable or organic materials.

The degree of green behavior demonstrated by the customer is defined by the percentage of purchased products being green. When computing the degree of green consumption of a customer, we have relied upon total purchases of green products during a two-year period. That is, the degree of green behavior of a customer is defined to be the share of green products (in items) in her basket during a two-year period. To increase reliability of the results, only customers with a total number of 24 or more purchased products are included.

Of the 1.5 million customers of the department store, 20% have purchased at least one green product during the two-year period. These customers account for 44.7% of total purchases, while 7.7% of their total purchases consist of green products.

III. Experiments

This section presents the experiments on customer data. First, we discuss the architecture and parameterization of the SOTM as well as model training. Then, we present the model and results related to green customer profiles, as well as other potential purposes of use within the problem domain.

A. Model Training

We apply the SOTM to the customer dataset presented in the previous section. The SOTM uses the demographic variables as input, and only visualizes the values of behavioral variables associated to the SOTM model. This is motivated by the interest in relating eco-consciousness and green consumption to demographics, rather than to purchasing behavior. In practice, the association means that we compute averages of additional variables for each unit of the SOTM, and thus, associating variables does not affect training. The model architecture is set to 6x10 units, where 10 unit columns represent the degree of greenness and 6 unit rows represent the cross-sectional structures. The number of units on the greenness dimension is set to span the degree of greenness in tenths, whereas the number of units for each greenness degree is determined based upon its descriptive value. It is worth noting that the SOTM, like the SOM, is not restricted to treat each unit as an individual cluster; only the dense locations in the data attract units due to the property of approximating probability density functions $p(x,v)$. When choosing the final specification of the SOTM, we use the following quality measures introduced in [20]: quantization error and topographic error. For a SOTM with 6x10 units, we chose a neighborhood radius $\sigma = 2.4$, as it has the highest quantization accuracies given no topographic errors. We stress topographic error as it is relatively more important for the interpretation of a SOTM.

B. Results and Analysis

Fig. 1 illustrates the SOTM, where the timeline below the figure represents the customers' degree of greenness. As described in the previous section, the coloring of the SOTM illustrates the proximity of units based upon a Sammon's mapping. Differences between units along the horizontal axis show differences in the degree of greenness, while differences between the units along the vertical axis show differences in cross-sections for one specific degree of greenness. Except for low and high greenness, the SOTM in Fig. 1 illustrates minor differences between customers of different degrees of greenness. For a more detailed view, we can, however, turn to patterns in the individual variables.

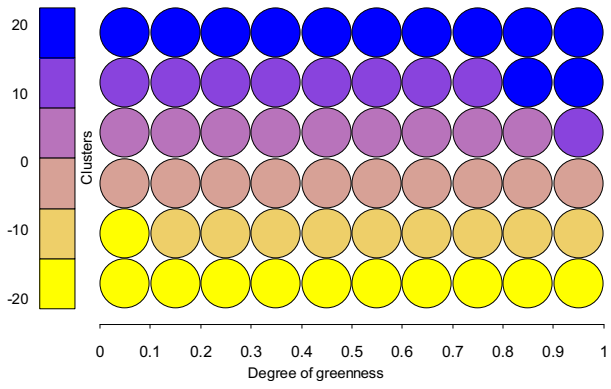


Fig. 1. Visualization of the trained SOTM. The x-axis represents degree of greenness and the y-axis clusters per cross-section.

The feature planes in Fig. 2 (layers of Fig. 1) show the spread of variable values for individual inputs using a blue scale (with individual scales on the left of each plane). They are particularly useful for gaining an understanding of how individual variables evolve over the degree of greenness. By assessing the feature planes, one can visually both discover the spread of values in the cross-section, and their variation over the degree of greenness. We first discuss the demographic and then the behavioral variables. Rather than focusing on detailed analysis per node (i.e., cluster or segment), the analysis herein discusses differences in the cross-section as a whole over the degree of greenness, starting from non-green customers. The demographic variables show the following patterns (the lettering refers to that of the feature planes in Fig. 2):

a. Whereas most customers are female, the non-green customers include a small share of male customers as well. Yet, as the degree of green purchases increases, the share of male customers gradually decreases to zero.

b. The likelihood of having children in the same household is average for non-green customers, whereas a higher degree of greenness leads to a division of customers to those with clearly lower (upper part of the map) and higher likelihoods (lower part of the map).

c. Similar to the pattern of having children, the income of customers is average for non-green customers, whereas more green customers are gradually divided into low (upper part) and high income customers (lower part).

d. Rather than having large differences within cross-sections, the share of older customers seems to gradually decrease over the degree of greenness.

e. Whereas the customers are mainly Finnish speaking, one can clearly observe that the share of Swedish speakers decreases with an increasing degree of greenness. This occurs gradually and evenly for the entire cross-section.

The above patterns relate to the demographic characteristics of customers over their degree of greenness. While demographics are used for training the model, we can also assess how additional variables are distributed on the SOTM by associating them to the model. Fig. 3 represents behavioral variables on the SOTM, and how they change over the degree of greenness. As parts of the behavioral variables are relatively stable regardless of the degree of greenness, we group in the following those that change with degree of greenness:

f. Average spending per transaction is high for all non-green customers, but decreases with the degree of greenness.

g. Total spending stands for the total amount of purchases made by the customer in all the different shops within the loyalty card chain the department store belongs to. The customers can be roughly divided into high and low spenders in the chain, but the values remain close to constant over degree of greenness.

h-j. Likewise, basket size (h) varies between small and large baskets, but does not exhibit significant differences over degree of greenness. Purchase frequency (i) also shows cross-sectional differences, but no changes over greenness. The total number of items purchased by customers (j) is relatively stable over greenness, except for higher values in the upper part for customers of average greenness.

k-l. The total amount of Euros spent on green products (k) increases with the increase in greenness (degree of greenness is based upon number of items). However, the total number of green products (l) decreases over greenness, indicating that the purchased items are more expensive.

m. We can observe that there is a group of customers that often visit the department store during daytime, who exhibit a high degree of greenness, whereas customers with lower levels of greenness rarely shop during daytime hours. Customers who have the possibility to shop during daytime are mainly mothers who stay at home, students, and pensioners.

n. Customers who buy expensive products on average possess a low degree of greenness. The variable decreases gradually with an increase in the degree of greenness.

Further, we can view which types of customers shop in the various departments of the store. This is presented in Fig. 4. In the following, we again present differences over the degree of greenness:

o. Customers with a low degree of greenness make most of their purchases from the department leisure. The trend is decreasing with the degree of greenness, except for a slight increase in parts of the cross-section for high degrees of greenness.

p. Customers who exhibit a higher degree of greenness make most of their purchases in the beauty department, whereas non-green customers purchase only little. This result is probably influenced by the large number of relatively inexpensive green items (e.g., cosmetics, nutritional supplements, etc.) in this section.

q. Purchases from the home department increase with an increase in the degree of greenness.

r. Purchases from the children's department increase slightly with the degree of greenness, especially in the upper part of the map. It is worth noting that the customers with a high likelihood of having children in the same household do not purchase more products from the children's department. This is partially in contradiction with many of the green customer profiles found in the literature (see e.g., [23]).

s. Percentages of purchases from the sports department decrease slightly with the degree of greenness. This might be related to the limited number of green products in the sports department.

t. Likewise, purchases from the men's department decrease with an increase in the degree of greenness. This variable follows the distribution of the variable gender, which points to men mainly shopping in the men's department.

u. Purchases from the women's department increase with the degree of greenness.

v. In contrast, percentages of purchases from the department of women's shoes decrease with the degree of greenness.

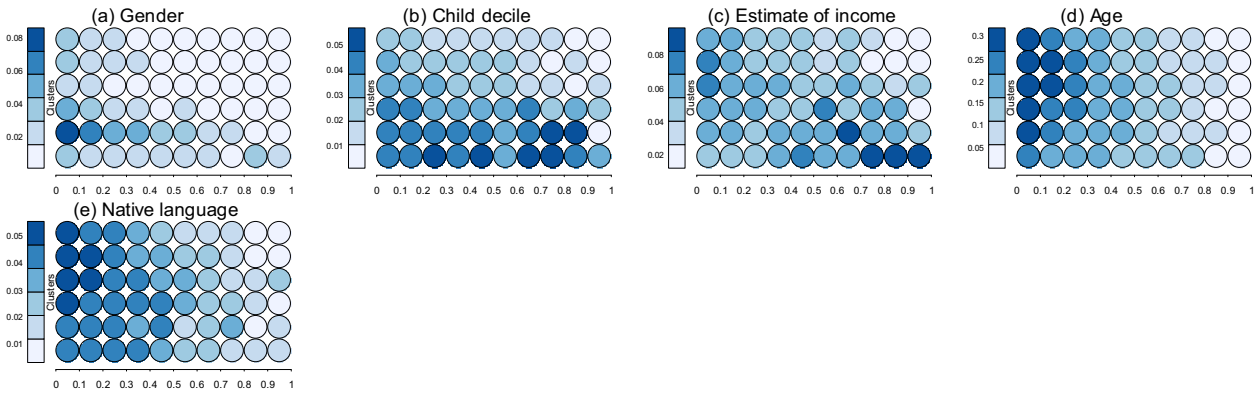


Fig. 2. Demographic variables illustrated on individual feature planes (a-e), where the y-axis refers to clusters and and the x-axis to degree of greenness.

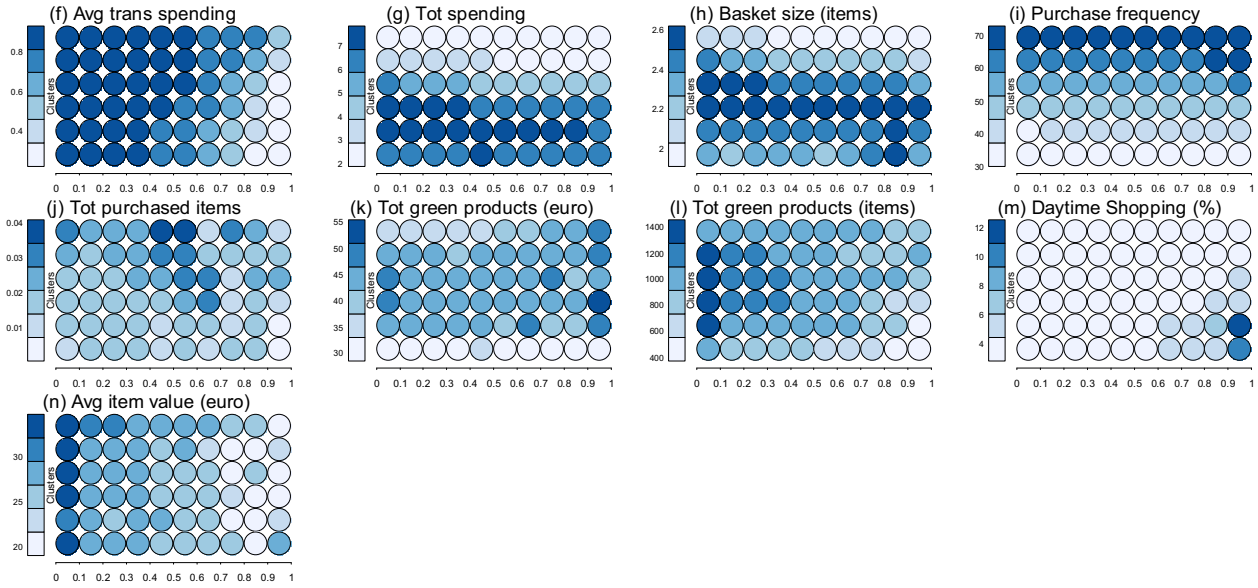


Fig. 3. Purchasing behavioral variables illustrated on feature planes (f-n), where the y-axis refers to clusters and and the x-axis to degree of greenness.

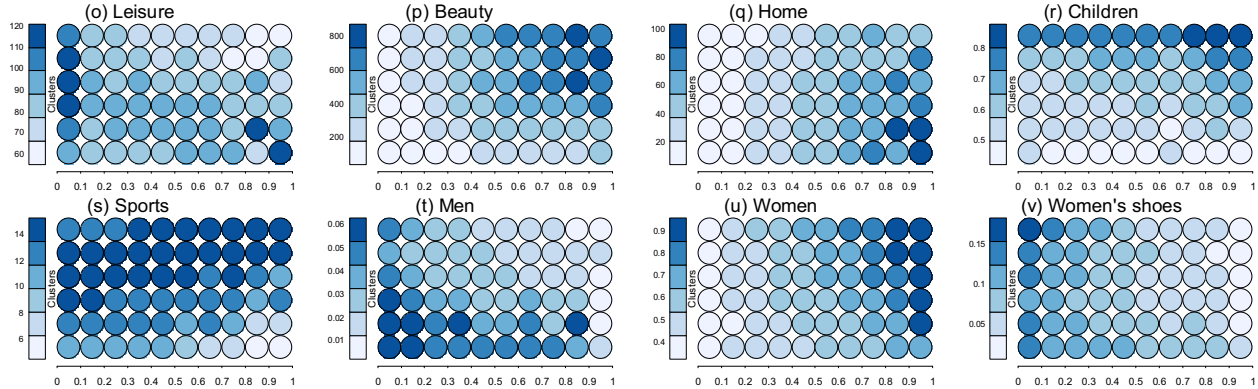


Fig. 4. Per department spendning variables illustrated on feature planes (o-v), where the y-axis refers to clusters and and the x-axis to degree of greenness.

Further insights in customer behavior can be gained by assessing relationships among the variables. For instance, one might be interested in how some behavioral variables follow patterns in demographic characteristics of customers. Below, we present a couple of connections between different variables:

- Customers that buy from the children's department are not those with children. This could point to green products being bought as presents, and somewhat surprisingly, the results point to young, low-income customers.
- While the share of purchases in the women's department positively correlates with the share of female customers, the purchases of women's shoes does not correlate, particularly not over the degree of greenness. This indicates that a large share of green customers are female, but particularly the ones who do not buy accessories like shoes, bags, belts, etc.

IV. CONCLUSIONS

In this paper, we have presented the application of the Self-Organizing Time Map (SOTM) to the analysis of department store transaction data, in order to identify underlying characteristics of green customer behavior. Instead of using a general segmentation approach and trying to separate green vs. non-green customers, as is typically done in the literature, we have used an approach based upon customers' differing degrees of greenness. Firstly, we have split the demographic and behavioral data based upon customers' degrees of green purchases. Then, we have trained a SOTM based upon the customers' demographic attributes, which illustrates differences across the degree of greenness. We studied the patterns that emerge in the demographic and behavioral attributes of customers, when studied from a degree-of-greenness perspective. Among other results, we find that there is not a clear linear relationship between estimated income levels and degree of greenness. While low degrees of greenness is clearly related to lower income levels, as is predicted by the literature, there is also a clear group of low income customers that display a high degree of greenness.

The study finds that patterns across the different degrees of greenness differ significantly, and that the SOTM is a potentially useful tool for studying these patterns. The SOTM enables examining the multidimensional structural properties of each cross-section v (vertically) and changes in structures (horizontally) by a Sammon's mapping-based coloring. By assessing the feature planes, we have been able to visually both discover the spread of values in the cross-section, and their variation over the degree of greenness. Future use of such a procedure is not, however, restricted to the greenness dimension, but could as well highlight differences in cluster structures over other variables of interest, such as distance to a department store, age of customers or total spending.

The generalized SOTM holds promise as a practical visual data mining tool. Analysis of the high-dimensional input data is made easier by being viewed from the perspective of a single variable of primary interest, while the benefits of the regular Self-Organizing Map are being exploited as well. However, future work should focus on evaluation of the SOTM's knowledge discovery capabilities, in order to assess the quality of patterns that users apprehend and how it compares to other methods.

ACKNOWLEDGMENT

We thank the case organization for fruitful cooperation and for providing data. We also thank Zhiyuan Yao for help with data preparation. The financial support from the Foundation for Economic Education, Finland and the Foundation of Nokia Corporation is gratefully acknowledged.

REFERENCES

[1] C. Rygielski, JC. Wang, DC. Yen, "Data mining techniques for customer relationship management," Technology in Society 24(4), 2002, pp. 483-502.

[2] M. Wedel., WA. Kamakura, "Market segmentation- conceptual and methodological foundations," 2nd ed. Boston: Kluwer, 2000.

[3] R.D. Straughan, JA. Roberts, "Environmental segmentation alternatives: a look at green consumer behavior in the new millennium," Journal of Consumer Marketing 16 (6), 1999, pp. 558-575.

[4] T. Mainieri, E.G. Barnett, T.R. Valdero, J.B. Unipan, and S. Oskamp, "Green Buying: The Influence of Environmental Concern on Consumer Behavior," The Journal of Social Psychology 137 (2), 1997, pp. 189-204.

[5] W. Young, K. Hwang, S. McDonald, C.J. Oates, "Sustainable Consumption: Green Consumer Behaviour when Purchasing Products," Sustainable Development 18, 2010, pp. 20-31.

[6] A.H. Holmbom, P. Sarlin, T. Eklund, Z. Yao, and B. Back, "Visual Data-Driven Profiling of Green Consumers," in Proceedings of the International Conference on Information Visualisation (iV2013), London, UK, 2013.

[7] Defra, "Sustainable Consumption and Production: Encouraging sustainable consumption," Department for Environment, Food and Rural Affairs, UK, 2006. Available at: http://web.archive.org/web/20070208050732/http://www.sustainable-development.gov.uk/what/priority/consumption-production/consumption.html

[8] J.H. Antil, and P.D. Bennet, "Construction and validation of a scale to measure socially responsible consumption behaviour," in K.E. Henion, and T.C. Kinnear (Eds.), The conserver Society, American Marketing Association, Chicago, IL, 1979, pp. 51-68.

[9] R.S. Hughner, P. McDonagh, A. Prothero, C.J. Shultz II, J. Stanton, "Who are organic food consumers? A compilation and review of why people purchase organic food," Journal of Consumer Behaviour 6, 2007, pp. 1-17.

[10] T. Kohonen, "Self-Organizing Maps, 3rd edition, Berlin: Springer, 2001.

[11] M.C. Ferreira de Oliveira and H. Levkowitz, "From visual data exploration to visual data mining: A survey," IEEE Transactions on Visualization and Computer Graphics 9, 2003, pp. 378-394.

[12] Z. Yao, P. Sarlin, T. Eklund, B. Back, "Combining Visual Customer Segmentation and Response Modeling," in Proceedings of the 20th European Conference on Information Systems (ECIS12), Barcelona, Spain, 2012. AISeL.

[13] A.H. Holmbom, T. Eklund, and B. Back, "Customer portfolio analysis using the SOM," International Journal of Business Information Systems 8, 2011, pp. 396-412.

[14] Z. Yao, A.H. Holmbom, T. Eklund, B. Back, "Combining unsupervised and supervised data mining techniques for conducting customer portfolio analysis," in Proceedings of the 10th Industrial Conference on Data Mining, Berlin: Springer, 2010.

[15] P. Lingras, M. Hogo, M. Snorek, C. West, "Temporal analysis of clusters of supermarket customers: conventional versus interval set approach," Information Sciences 172 (1-2), 2005, pp. 215-240.

[16] S.C. Lee, Y.H. Suh, J.K. Kim, K.J. Lee, “A cross-national market segmentation of online game industry using SOM,” Expert systems with applications 27 (4), 2004, pp. 559-570.

[17] A. Vellido, P. Lisboa, K. Meehan, “Segmentation of the on-line shopping market using neural networks,” Expert Systems with Applications 17 (4), 1999, pp. 303-314.

[18] P. Sarlin, “Data and Dimension Reduction for Visual Financial Performance Analysis,” Information Visualization, 2013, in press. DOI: 10.1177/1473871613504102.

[19] J. Vesanto, “SOM-based data visualization methods,” Intelligent data analysis 3 (2), 1999, pp. 111-126.

[20] P. Sarlin, “Self-Organizing Time Map: An Abstraction of Temporal Multivariate Patterns,” Neurocomputing 99(1), 2013, pp. 496-508.

[21] P. Sarlin, “Replacing the time dimension: A Self-Organizing Time Map over any variable,” in Proceedings of the Workshop on New Challenges in Neural Computation (NC^2), Saarbrucken, Germany, 2013.

[22] J.W. Sammon, “A nonlinear mapping for data structure analysis,” IEEE Transactions on Computers 18, 1969, pp. 401-409.

[23] M. Laroche, J. Bergeron, and G. Barbaro-Forleo, “Targeting consumers who are willing to pay more for environmentally friendly products,” Journal of Consumer Marketing 18(6), 2001, pp. 503-20.

Publication 6

Holmbom, A.H., Eklund, T. and Back, B. (2014). A Weak-form Expert Evaluation of Customer Profiling Models, In Devos, J. and De Haes, S. (Eds.) *The proceedings of the 8th European Conference on IS Management and Evaluation - ECIME2014*, 11-12 September 2014, Ghent, Belgium.

A Weak-form Expert Evaluation of Customer Profiling Models

Annika H. Holmbom, Tomas Eklund, Barbro Back
Department of Information Technologies, Åbo Akademi University and Turku Centre for Computer Science (TUCS), Turku, Finland
Annika.h.holmbom@abo.fi
Tomas.eklund@abo.fi
Barbro.back@abo.fi

Abstract: In this paper, we evaluate two models that were used for customer profiling: 1) a market basket analysis (MBA) model and 2) a customer segmentation model. The models were based on actual customer purchasing data from a large department store for the period 2007-09. A weak-form evaluation method, consisting of qualitative interviews of experts from the department store, was used. The questions focused on information quality according to the DeLone and McLean framework and were derived from the Doll and Torkzadeh model, i.e., they covered content, accuracy, format, and ease of use aspects. Seven experts from the case company took part in the evaluation process. Before the actual interview, the experts were asked to fill out a questionnaire regarding their background and current access to timely sales information, i.e., how often and how useful the information was that they were receiving at the moment. Later, during an interview, the evaluator discussed the questionnaire with each respondent. Then, the experts were presented with information on their customers' buying behavior based on the results from the two models, i.e., the MBA- and the segmentation model. After each presentation, the experts were asked to respond to fifteen statements and four open-ended questions. All in all, the information gained through the MBA- and segmentation analyses was rated highly (4-5/max 5) by the experts. The experts considered the information gained with help of these models to be valuable and useful for decision making and for making strategic planning for the future. This implies that the models could be of valuable use for managers working within CRM, e.g., planning marketing campaigns, product range planning, service development, planning of store layouts, operative and strategic planning, and for further developing loyalty programs.

Keywords: Weak-form evaluation, qualitative interviews, data mining, market basket analysis, customer segmentation

1. Background

Department stores today are facing increasing competition from many sources, including other department stores, specialty stores, and e-commerce. In this increasingly competitive market it has become even more important for retailers to understand the needs of their customers. Through widespread implementation of customer relationship management (CRM), companies have been striving to become more customer-centric. The aim of CRM is to create added value for both the companies and their customers, by integrating data from sales, marketing and customer support to obtain a better understanding of customers' needs (Heinrich 2005; Datta 1996; Chalmeta 2006).

The key element within CRM is thus customer information (Buttle 2004; Rygielski et al. 2002). Today, production of data, as well as the capacity to store the produced data, is growing rapidly, constantly outpacing companies' abilities to analyze them. While, different data mining methods have been applied since the 1990's, many approaches have been difficult for the average manager to interpret and use. Visual analytics is an emerging field that aims to bridge this gap. As a multidisciplinary field, within visual analytics vast amounts of different kinds of data in different formats are analyzed in a process where human judgment, visual presentations and different kinds of interaction techniques are combined (Keim et. al. 2008; Thomas and Cook 2006). The aim of visual analytics as a part of data mining is to turn large amounts of unstructured data into useful information. Advanced computer applications are used for the information discovery process allowing decision makers to fully concentrate on the analytical process and to visualize the information useful for them (Keim et. al. 2008). Thus, the use of interactive visualization methods is integral to visual data mining.

Within CRM, the area of analytical CRM relates to advanced modeling and profiling of customers based upon demographic and transaction data. One important application within this area is market basket analysis (MBA). MBA typically uses association rule mining to analyze transaction data and identify products that are purchased together, so-called baskets. Results from this type of analysis are used, e.g., for planning store layouts, catalogue design, upselling, and designing marketing campaigns (Olson and Delen 2008). Another important application within analytical CRM is segmentation, where

customers with similar profiles or requirements are identified and grouped together (Lingras et al. 2005). Segmentation is a suitable approach for gaining an overview of the customer base. Through segmentation, the manager gains a description of the main types of customers and is able to identify key customers and their needs (Buttle 2004; Lingras et al. 2005).

In this paper, we evaluate two models that were used for customer profiling: 1) an MBA-model and 2) a customer segmentation model. The models were based on actual customer purchasing data from a large department store for the period 2007-09. The MBA-model was built using the Apriori algorithm and the segmentation model was built using the self-organizing map (SOM). The objective of this study is to evaluate and determine the added value of the two models for decision makers at the department store.

The rest of the paper is structured as follows: In Section two, we describe the methodology used in this paper. The data and the two models that are evaluated in this paper are described in Section three. The evaluation of the models is discussed in Section four. We present the results in Section five, and in Section six we conclude the paper and discuss future work.

2. Methodology

The overall research process that this evaluation pertains to, i.e., the creation of the MBA and segmentation models, follows the design science research (DSR) paradigm. The goal of DSR is to find innovative solutions to relevant and practical problems with the help of primarily technological artifacts. DSR is an iterative process based upon two ongoing cycles; the build and evaluate cycles. Both cycles are integral and equally important activities in the DSR paradigm, as the actual functioning of the artifact must be proven (March and Smith 1995; Hevner 2007). One difficulty with design science is that evaluation of performance of a solution or an artefact is dependent on the environment it is working in. Progress is achieved when more effective technologies replace the existing ones (March and Smith 1995). In our research, we have built two models to support experts in customer profiling tasks. Therefore, a field study utilizing experts and potential end users was selected as an appropriate evaluation setting.

In order to evaluate the usefulness of the created models, an information systems evaluation approach was taken. Different ways to evaluate the usefulness of information systems has been extensively discussed in the literature. For example, over 100 different measures used to evaluate IS usefulness were identified by DeLone and McLean (1992). The authors systematically analyzed 180 IS success studies that they had collected, and divided these measures of IS usefulness into six categories: *system quality, information quality, information use, user satisfaction, individual impact* and *organizational impact*. The model is causal, indicating that system quality and information quality influence user satisfaction and information use, in turn influencing impact. Based upon a multitude of studies in the field, the authors have updated and enhanced the framework in DeLone and McLean (2003). The authors add the factor service quality, remove information use and individual and organizational impact, and replace them with intention to use/use and net benefits. The authors note that the choice of which variable to use is, of course, dependent on the objective of the study. The model has been widely applied and validated in the literature, and clearly shows the causal importance of the factor of information quality in terms of IS success.

In the setting of this study, the focus is on measures of information quality and information use of models that the users cannot directly interact with at this stage. Therefore, one potentially applicable model is the End-User Computing Satisfaction (EUCS) framework, developed by Doll and Torkzadeh (1988). According to the authors, the five most important factors in assessing user satisfaction with information are: *content, accuracy, format, ease of use,* and *timeliness*. The Doll and Torkzadeh framework was used in this study because it focuses more on information quality and use than many other available models, and was, therefore, found to be more suitable for this study (Doll and Torkzadeh 1988).

3. Two models for customer profiling

The models presented in this study were based on actual customer purchasing data from a large department store for the period 2007-09. Data mining methods were used to identify patterns in the customers' buying behavior. The extracted patterns describe the customer's behavior and purchasing habits, in this case, mainly connected to the women's department in two major department stores of the national chain.

3.2 The Market Basket Analysis (MBA) model

An important task for CRM managers is to determine which products are purchased together, i.e., to perform MBA. The most well-known method within MBA is the Apriori algorithm (Agrawal, Srikant 1994; Olson, Delen 2008; Rajaraman, Ullman 2011). The Apriori algorithm tries to identify frequent item sets, i.e., in an MBA setting, products that appear together significantly more often than would be statistically expected beforehand. For more detailed information on the Apriori algorithm, see e.g., the work of Agrawal and Srikant (1994).

The aim of the MBA model was to gain information on customers' purchasing behavior. The analysis seeks to answer questions such as:

- How many products are purchased together in one shopping transaction (basket)?
- Which products are bought together?
- From which departments are products in a shopping basket combined?
- Which clothing brands do the customers combine?

For the MBA model, we used a large market basket data set in the form of a sparse matrix, with transaction data from a two year period. The data consists of 16 million transactions, (i.e., baskets) and in total 39 million product purchases (i.e., individual line items), with an average of 2.43 purchases per transaction. The total number of products is 557,000 and the total number of customers is 1.5 million. Almost 50% of the transactions contained only one product, i.e., almost every other customer bought only one product per visit to the department store.

An implementation of the Apriori algorithm based upon smart preprocessing of data was used for the analysis. The outcome of the analysis was tables expressing the number of products in a shopping basket, the connection between products according to existing relations in the transaction data, dependency diagrams that show the relationships between different departments and the connections between different product brands.

3.3 The customer segmentation model

Another important task for CRM managers is to have an overview of their customer base, i.e., to perform customer segmentation. A well-known unsupervised artificial neural network (ANN), the Self-Organizing Map (SOM) (Kohonen 2001), was used for the customer segmentation task. The SOM is a widely used unsupervised data mining method for data-driven clustering. With the SOM it is possible to explore relationships in multidimensional input data by projecting them onto a two-dimensional topological map. The topological properties of the SOM mean that similar data are located close to each other on the grid, preserving relationships but not actual distances (as opposed to, e.g., multidimensional scaling). The SOM is a highly visual, non-parametric and very robust method that requires very little preprocessing of data (Kohonen 2001).

The purpose of the customer segmentation analysis was to group customers according to their behavior and demographical abilities. The analysis seeks to answer the questions:

- Who buys?
- Which products?
- For how much (€)?
- How often?

The data used for the segmentation model consist of two parts; 1) demographic background information, and 2) purchasing transaction data describing the purchasing behavior of the customers. The demographic data are obtained through the loyalty card program of the retailer, and transaction data are taken from the retailer's transaction systems. The training dataset contains over 1.5 million customers (almost 30% of the population in Finland), i.e., 1.5 million rows of data.

The variables included are the following:
Demographic: Age, gender, child decile (an estimated probability of children living in the same household), estimate of income, native language (Finnish or Swedish), customer tenure, and loyalty point level.
Purchasing behavior: Recency, Frequency, and Monetary (RFM), and per-department spending at the department store: Leisure, Beauty, Home, Children, Sports, Men, Women, and Women's shoes.

First, the data were transformed into a suitable format using SPSS Modeler. The SOM model and analyses were performed using Viscovery SOMine. SOMine provides means for both data preprocessing and visualization of the results. The outcome of the analysis was a segmentation of the whole customer base. The formed segments were visually presented using topological maps and information, and statistics for each segment were presented. The outcome of the customer segmentation has been published in the Knowledge Service Engineering Handbook (Vanharanta et. al. 2012).

4. Evaluation of the models

In this study, the proposed models are not yet implemented and functioning systems, and therefore, the users will not be interacting with the systems themselves. Therefore, we have chosen to evaluate the quality of the information extracted from the two models for customer profiling, instead of the technical properties of the models. The evaluation is performed using a weak-form evaluation method, consisting of interviews of experts from the department store. The interviews are based on an adapted version of the five most important factors defined in the EUCS model (Doll and Torkzadeh 1988). They covered *content, accuracy, format,* and *ease of use* aspects. As we in this case are evaluating a static model, the timeliness-aspect could not be measured. In addition, the factor "ease of use", in this case refers to the benefit and usefulness of the information.

For the interviews, two questionnaires and two PowerPoint presentations were created. The first questionnaire was sent to the respondents beforehand. Its purpose was to collect background information on the respondents and to map the current situation regarding available information on customer profiling. During the interviews, the PowerPoint presentations were used for communication of the results of the analyses gained with the two models, i.e., the MBA and customer segmentation models. The second questionnaire was used for the evaluation of the potential usefulness of the analyses. The questionnaire was administered in Finnish, which is the official company language.

Seven specialists, of eight originally contacted, were interviewed between May and July 2013. The interviews took from 45 minutes to one hour and consisted of 1) discussions of the initial background questionnaire, 2) presentation of the results from the two analyses, and 3) the evaluation of the potential usefulness of the results gained from the analyses with the two models. The interviews were recorded.

The experts were asked to evaluate the potential use of the outcomes of the 1) MBA and 2) segmentation analyses by answering fifteen statements (on a Likert-scale of 1 to 5/ max 5, or *strongly disagree, disagree, neutral, agree,* and *strongly agree*) and four open-ended questions.

5. Results

The results of the evaluation consist of background information on the respondents and evaluation of the information quality of the two models on customer profiling.

5.1 Demographic information on the respondents

The first questionnaire was used for collection of background information on the respondents. The seven respondents were between 30 to 64 years old (see Table 1). They had differing backgrounds ranging from economics, sales, marketing, and management. Their titles at the company were head of sales, store manager, product range manager, head of division, and concept manager. Their areas of responsibility were all linked with women's clothing, management of the department store, and product sourcing. The respondents had been working at the department store chain between 8 and 26 years, of which three persons less than 2 years and one person for 25 years. All of the respondents were familiar with IT tools and different reporting software, but only few of them had any experience with advanced tools used for analyzing data.

Table 1. Age distribution of the respondents.

Age class (years)	**18-29**	**30-39**	**40-49**	**50-64**	**65+**
Respondents (N=7)	0	2	3	2	0

According to the respondents, information on customers and their shopping behavior is distributed to some extent, but several of the respondents think that the information is on a too general level to be useful in their actual work. In particular, the experts are in need of information in support of management tasks. The respondents receive information on sales daily, while other information is

updated on a monthly or annual basis. In their work, the respondents use both their own expertise and information on customers retrieved from the retailer's database.

When the respondents were asked what kind of information they would need in their line of work, they responded that they would need information such as brand loyalty studies, studies concerning customer segment purchasing behavior in and out of campaigns, and information on how to reach customers through marketing. Information supporting product portfolio selection and services was also called for. In general, there was a clear need for more precise information and information, on a deeper and more specific level.

5.2 Evaluation of the Market Basket Analysis (MBA) model

The results of the MBA evaluation are presented in Table 2. For each statement, the table shows the number of the respondents that responded with a certain rating. The MBA model was rated according to statements assessing four different factors: *content (5 statements), accuracy (3 statements), format (4 statements),* and *benefit/ usefulness (3 statements)*. Overall, the MBA analysis received good ratings from the experts measured in terms of both average and median. In particular, the accuracy and format of the MBA received high ratings (median values between 4 and 5). The content and usefulness of the analysis was also generally highly rated (median 4-5), but in particular statement 2.4 (The analysis provides new information) received a lower rating, likely because the used data was from 2007-09. Overall, the respondents were clearly pleased with the information quality of the MBA model.

Table 2. The results of the MBA evaluation.

	Average	Median	Strongly disagree	Disagree	Neutral	Agree	Strongly agree	Do not know	N
2. Content									
2.1 The analysis gives important information.	4,1	4	0	1	0	3	3	0	7
2.2 The results of the analysis respond to my needs.	3,6	4	0	2	0	4	1	0	7
2.3 The analysis gives useful information.	3,7	4	0	2	0	3	2	0	7
2.4 The analysis gives new information	2,9	2	0	4	1	1	1	0	7
2.5 The information extracted from the analysis is sufficient.	3,7	4	0	2	0	3	2	0	7
3. Accuracy									
3.1 The results of the analysis are correct.	4,6	5	0	0	1	1	5	0	7
3.2 The results of the analysis are reliable.	4,7	5	0	0	0	2	5	0	7
3.3 I am satisfied with the accuracy of the analysis.	4,3	4	0	0	1	3	3	0	7
4. Format									
4.1 The results of the analysis were visually clearly presented.	4,7	5	0	0	0	2	5	0	7
4.2 The results of the analysis are easily read.	4,7	5	0	0	0	2	5	0	7
4.3 The results of the analysis are easily understood.	4,7	5	0	0	0	2	5	0	7
4.4 Overall, I am satisfied with the format of the analysis.	4,3	5	0	1	0	2	4	0	7
5. Benefit and usefulness									
5.1 The results of the analysis correlate well with my own understanding regarding the customers of the department store.	4,7	5	0	0	0	2	5	0	7
5.2 The results of the analysis were useful.	3,3	4	0	3	0	3	1	0	7
5.3 I can benefit from this kind of analysis in my work.	4,3	4	0	0	0	5	2	0	7

In addition to the structured, Likert-scale statements above, the experts were asked to respond to four open-ended questions in order to assess how they would like to see the models used and developed in the future.

Firstly, the respondents were asked how they would see that the presented models could impact upon their work. Several respondents (3/7) emphasized that the models would provide support for product display planning, and would also provide support for tracking changes in sales based upon display changes and marketing efforts. Product range planning was also mentioned by several of the managers (3/7). Two managers mentioned decision support in general, focusing on new perspectives and support for "gut-feeling" decisions. One respondent also mentioned that the MBA model would directly support cross-selling efforts.

Next, respondents were asked what information about customers and their shopping behavior they feel is missing from the MBA model, i.e. how should the MBA model be improved in order to be even more useful. The respondents specifically mentioned brand loyalty information, product group level comparisons, and department level analyses. One respondent raised the interesting question of how many customers paid at several different cash registers during the same shopping visit, something that has not at all been addressed previously in our models.

The respondents were also asked how often they would like to see the MBA analysis updated. Most respondents (5/7) responded with either once or twice a year, but two respondents would have liked to see the results updated considerably more often. One of these respondents specifically mentioned marketing cycles as the motivation for needing the updates 5 times in a year. This indicates differences related to the job descriptions of the respondents; the respondents involved in marketing campaign planning and product procurement are likely more interested in timely MBA data than floor level managers.

Finally, the respondents were asked in what form they would like to see the results presented. All respondents emphasized the use of graphical displays and brief reports, e.g., PowerPoint presentations. One of the respondents mentioned also the visualization of how the gained information changed with time, which is an important aspect to think about when implementing a system that gives updated reports on timely information. It is obvious that the respondents do not want to spend a lot of time going through long daily reports, instead requiring visually intuitive and simple presentations.

5.3 Evaluation of the customer segmentation model

The results of the segmentation analysis evaluation are presented in Table 3. The same statements as for rating the MBA model were used for rating the segmentation model. Overall, the experts rated the segmentation analysis highly. The accuracy and format of the segmentation analysis received the highest ratings (median values between 4 and 5). The content and usefulness of the analysis was also generally highly rated (median 4-5). Overall, the respondents were clearly pleased with the information quality of the segmentation analysis model.

As for the MBA model evaluation, the experts were asked to respond to four open-ended questions concerning the use of the model.

Firstly, the respondents were asked how they would see that the presented models could impact upon their work. Some of the respondents (2/7) emphasized that the models would provide support for product range planning, marketing and service. One respondent also mentioned that the segmentation model would directly support short term actions, and in the long run, strategic planning. One manager mentioned decision support in general. This can be seen as a strong support for the visual analytics capabilities of the model.

Next, respondents were asked what information about customers and their shopping behavior they feel is missing from the segmentation analysis model, i.e., how should the segmentation model be improved. The respondents specifically mentioned brand loyalty information and information concerning average frequency of visits. In addition, more information on profitable and potential customers with different shopping behavior was of interest for the respondents. One respondent raised the interesting question of how information has changed with time since the presented analysis, when there have been made changes in the product range and developments in services.

The respondents were also asked how often they would like the segmentation analysis to be updated. Most respondents (5/7) responded with either once or twice a year, while two respondents would have liked to see the results updated once a month.

Finally, the respondents were asked in what form they would like to see the results presented. All respondents emphasized the use of graphical displays and brief reports, e.g., PowerPoint presentations and maps. Visualization of how the gained information changed with time was also mentioned by the respondents. Again, the respondents required visually intuitive and simple presentations, instead of long written reports.

Table 3. The results of the segmentation analysis evaluation.

	Average	Median	Strongly disagree	Disagree	Neutral	Agree	Strongly agree	Do not know	N
6. Content									
6.1 The analysis gives important information.	4,6	5	0	0	0	3	4	0	7
6.2 The results of the analysis respond to my needs.	3,7	4	0	2	0	3	2	0	7
6.3 The analysis gives useful information	3,6	4	0	2	0	4	1	0	7
6.4 The analysis gives new information	3,1	4	0	3	0	4	0	0	7
6.5 The information extracted from the analysis is sufficient.	3,4	4	0	3	0	2	2	0	7
7. Accuracy									
7.1 The results of the analysis are correct.	4,6	5	0	0	0	3	4	0	7
7.2 The results of the analysis are reliable.	4,7	5	0	0	0	2	5	0	7
7.3 I am satisfied with the accuracy of the analysis.	3,7	4	0	2	1	1	3	0	7
8. Format									
8.1 The results of the analysis were visually clearly presented.	4,3	5	0	1	0	2	4	0	7
8.2 The results of the analysis are easily read.	4,4	5	0	1	0	1	5	0	7
8.3 The results of the analysis are easily understood.	4,7	5	0	0	0	2	5	0	7
8.4 Overall, I am satisfied with the format of the analysis.	4,4	5	0	1	0	1	5	0	7
9. Benefit and usefulness									
9.1 The results of the analysis correlate well with my own understanding regarding the customers of the department store.	4,4	5	0	1	0	1	5	0	7
9.2 The results of the analysis were useful.	4	4	0	1	0	4	2	0	7
9.3 I can benefit from this kind of analysis in my work.	4,3	4	0	0	0	5	2	0	7

6. Conclusions and future work

All in all, the information gained through the MBA- and segmentation analyses was rated highly (4-5/max 5) as to content, accuracy, format, and ease of use aspects by the experts. Most of the respondents would like the analyses to be updated once or twice a year and they preferred to receive the information as brief reports with graphical displays including details on occurred changes in time. The experts considered the information gained with help of these models to be valuable and useful for decision making and for strategic planning. This can be seen as a strong support for the visual analytics capabilities of the model, and therefore, be of valuable use for managers working within CRM.

The respondents were very interested in gaining deeper and more specific information regarding both MBA and segmentation to support their daily work. Also, changes in time were of high interest. Based upon the expert evaluation, it is possible to develop the analyses further according to their needs and

in this way extract more valuable and useful information. As the results of the weak-form evaluation were positive, there is support for further developing the models into system implementations.

Acknowledgements
We thank the case organization for fruitful cooperation and for providing data. Mr. Artur Signell is thanked for performing the executions of the Apriori algorithm. The financial support of The Foundation of Economic Education is gratefully acknowledged.

References
Agrawal, R. and Srikant, R. (1994) "Fast algorithms for mining association rules", Proc 20th Int Conf VeryLarge Data Bases VLDB, Citeseer, pp 487.
Buttle, F. (2004) *Customer Relationship Management Concepts and Tools*, Butterworth-Heinemann, Oxford.
Chalmeta, R. (2006) “Methodology for customer relationship management”, *The Journal of Systems and Software*, Vol. 79, No. 7, pp 1015–1024.
Datta, Y. (1996) “Market segmentation: an integrated framework”, *Long Range Planning*, Vol. 29, No. 6, pp 797–811.
DeLone, W.H. and McLean, E.R. (1992) “Information Systems Success: The quest for the dependent variable”, *Information systems research*, Vol. 3, No. 1, pp 60-95.
DeLone, W.H., and McLean, E.R. (2003) "The DeLone and McLean Model of Information Systems Success: A Ten-Year Update", *Journal of Management Information Systems,* Vol. 19, No. 4, pp 9-30.
Doll, W.J. and Torkzadeh, G. (1988) “The measurement of End-User Computing Satisfaction”, *MIS Quarterly*, 12(2), pp 259-274.
Heinrich, B. (2005) “Transforming strategic goals of CRM into process goals and activities”, *Business Process Management Journal*, Vol. 11, No. 6, pp 709–723.
Hevner, A. (2007) “A three-cycle view of design science research”, *Scandinavian Journal of Information Systems*, 19(2), pp 87–92.
Keim, D.A., Mansmann, F., Schneidewind, J., Thomas, J. and Ziegler, H. (2008) “Visual Analytics: Scope and Challenges”, in: Simoff, S.J. et al (Eds.): Visual Data Mining, LNCS 4404, pp 76-90. Springer Verlag Berlin Heidelberg.
Kohonen, T. (2001) *Self-Organizing Maps,* Springer, Berlin.
Lingras, P., Hogo, M., Snorek, M. and West, C. (2005) “Temporal analysis of clusters of supermarket customers: conventional versus interval set approach”, *Information Sciences*, Vol. 172, Nos. 1–2, pp 215–240.
March, S.T. and Smith, G.F. (1995) "Design and natural science research on information technology", *Decision Support Systems*, vol. 15, no. 4, pp 251-266.
Olson, D.L. and Delen, D. (2008) *Advanced data mining techniques*, Springer Verlag.
Rajaraman, A. and Ullman, J.D. (2011) *Mining of massive datasets*, Cambridge Univ Pr.
Rygielski, C., Wang, J. and Yen, D.C. (2002) “Data mining techniques for customer relationship Management”, *Technology in Society*, Vol. 24, No. 4, pp 483–502.
Thomas, J. and Cook, K. (2006) “Visualization viewpoints”, *IEEE Computer Graphics and Applications*, Jan/Feb 2006, pp 10-13.
Vanharanta, H., Magnusson, C., Ingman, K., Holmbom, A.H., and Kantola, J. (2012) “Strategic Knowledge Services”, in Kantola, J. and Karwowski, W. (Eds.) *Knowledge Service Engineering Handbook,* CRC Press, Taylor and Francis Group.

Turku Centre for Computer Science
TUCS Dissertations

1. **Marjo Lipponen**, On Primitive Solutions of the Post Correspondence Problem
2. **Timo Käkölä**, Dual Information Systems in Hyperknowledge Organizations
3. **Ville Leppänen**, Studies on the Realization of PRAM
4. **Cunsheng Ding**, Cryptographic Counter Generators
5. **Sami Viitanen**, Some New Global Optimization Algorithms
6. **Tapio Salakoski**, Representative Classification of Protein Structures
7. **Thomas Långbacka**, An Interactive Environment Supporting the Development of Formally Correct Programs
8. **Thomas Finne**, A Decision Support System for Improving Information Security
9. **Valeria Mihalache**, Cooperation, Communication, Control. Investigations on Grammar Systems.
10. **Marina Waldén**, Formal Reasoning About Distributed Algorithms
11. **Tero Laihonen**, Estimates on the Covering Radius When the Dual Distance is Known
12. **Lucian Ilie**, Decision Problems on Orders of Words
13. **Jukkapekka Hekanaho**, An Evolutionary Approach to Concept Learning
14. **Jouni Järvinen**, Knowledge Representation and Rough Sets
15. **Tomi Pasanen**, In-Place Algorithms for Sorting Problems
16. **Mika Johnsson**, Operational and Tactical Level Optimization in Printed Circuit Board Assembly
17. **Mats Aspnäs**, Multiprocessor Architecture and Programming: The Hathi-2 System
18. **Anna Mikhajlova**, Ensuring Correctness of Object and Component Systems
19. **Vesa Torvinen**, Construction and Evaluation of the Labour Game Method
20. **Jorma Boberg**, Cluster Analysis. A Mathematical Approach with Applications to Protein Structures
21. **Leonid Mikhajlov**, Software Reuse Mechanisms and Techniques: Safety Versus Flexibility
22. **Timo Kaukoranta**, Iterative and Hierarchical Methods for Codebook Generation in Vector Quantization
23. **Gábor Magyar**, On Solution Approaches for Some Industrially Motivated Combinatorial Optimization Problems
24. **Linas Laibinis**, Mechanised Formal Reasoning About Modular Programs
25. **Shuhua Liu**, Improving Executive Support in Strategic Scanning with Software Agent Systems
26. **Jaakko Järvi**, New Techniques in Generic Programming – C++ is more Intentional than Intended
27. **Jan-Christian Lehtinen**, Reproducing Kernel Splines in the Analysis of Medical Data
28. **Martin Büchi**, Safe Language Mechanisms for Modularization and Concurrency
29. **Elena Troubitsyna**, Stepwise Development of Dependable Systems
30. **Janne Näppi**, Computer-Assisted Diagnosis of Breast Calcifications
31. **Jianming Liang**, Dynamic Chest Images Analysis
32. **Tiberiu Seceleanu**, Systematic Design of Synchronous Digital Circuits
33. **Tero Aittokallio**, Characterization and Modelling of the Cardiorespiratory System in Sleep-Disordered Breathing
34. **Ivan Porres**, Modeling and Analyzing Software Behavior in UML
35. **Mauno Rönkkö**, Stepwise Development of Hybrid Systems
36. **Jouni Smed**, Production Planning in Printed Circuit Board Assembly
37. **Vesa Halava**, The Post Correspondence Problem for Market Morphisms
38. **Ion Petre**, Commutation Problems on Sets of Words and Formal Power Series
39. **Vladimir Kvassov**, Information Technology and the Productivity of Managerial Work
40. **Frank Tétard**, Managers, Fragmentation of Working Time, and Information Systems

41. **Jan Manuch**, Defect Theorems and Infinite Words
42. **Kalle Ranto**, Z_4-Goethals Codes, Decoding and Designs
43. **Arto Lepistö**, On Relations Between Local and Global Periodicity
44. **Mika Hirvensalo**, Studies on Boolean Functions Related to Quantum Computing
45. **Pentti Virtanen**, Measuring and Improving Component-Based Software Development
46. **Adekunle Okunoye**, Knowledge Management and Global Diversity – A Framework to Support Organisations in Developing Countries
47. **Antonina Kloptchenko**, Text Mining Based on the Prototype Matching Method
48. **Juha Kivijärvi**, Optimization Methods for Clustering
49. **Rimvydas Rukšėnas**, Formal Development of Concurrent Components
50. **Dirk Nowotka**, Periodicity and Unbordered Factors of Words
51. **Attila Gyenesei**, Discovering Frequent Fuzzy Patterns in Relations of Quantitative Attributes
52. **Petteri Kaitovaara**, Packaging of IT Services – Conceptual and Empirical Studies
53. **Petri Rosendahl**, Niho Type Cross-Correlation Functions and Related Equations
54. **Péter Majlender**, A Normative Approach to Possibility Theory and Soft Decision Support
55. **Seppo Virtanen**, A Framework for Rapid Design and Evaluation of Protocol Processors
56. **Tomas Eklund**, The Self-Organizing Map in Financial Benchmarking
57. **Mikael Collan**, Giga-Investments: Modelling the Valuation of Very Large Industrial Real Investments
58. **Dag Björklund**, A Kernel Language for Unified Code Synthesis
59. **Shengnan Han**, Understanding User Adoption of Mobile Technology: Focusing on Physicians in Finland
60. **Irina Georgescu**, Rational Choice and Revealed Preference: A Fuzzy Approach
61. **Ping Yan**, Limit Cycles for Generalized Liénard-Type and Lotka-Volterra Systems
62. **Joonas Lehtinen**, Coding of Wavelet-Transformed Images
63. **Tommi Meskanen,** On the NTRU Cryptosystem
64. **Saeed Salehi**, Varieties of Tree Languages
65. **Jukka Arvo**, Efficient Algorithms for Hardware-Accelerated Shadow Computation
66. **Mika Hirvikorpi**, On the Tactical Level Production Planning in Flexible Manufacturing Systems
67. **Adrian Costea**, Computational Intelligence Methods for Quantitative Data Mining
68. **Cristina Seceleanu**, A Methodology for Constructing Correct Reactive Systems
69. **Luigia Petre**, Modeling with Action Systems
70. **Lu Yan**, Systematic Design of Ubiquitous Systems
71. **Mehran Gomari**, On the Generalization Ability of Bayesian Neural Networks
72. **Ville Harkke**, Knowledge Freedom for Medical Professionals – An Evaluation Study of a Mobile Information System for Physicians in Finland
73. **Marius Cosmin Codrea**, Pattern Analysis of Chlorophyll Fluorescence Signals
74. **Aiying Rong**, Cogeneration Planning Under the Deregulated Power Market and Emissions Trading Scheme
75. **Chihab BenMoussa**, Supporting the Sales Force through Mobile Information and Communication Technologies: Focusing on the Pharmaceutical Sales Force
76. **Jussi Salmi**, Improving Data Analysis in Proteomics
77. **Orieta Celiku**, Mechanized Reasoning for Dually-Nondeterministic and Probabilistic Programs
78. **Kaj-Mikael Björk**, Supply Chain Efficiency with Some Forest Industry Improvements
79. **Viorel Preoteasa**, Program Variables – The Core of Mechanical Reasoning about Imperative Programs
80. **Jonne Poikonen**, Absolute Value Extraction and Order Statistic Filtering for a Mixed-Mode Array Image Processor
81. **Luka Milovanov**, Agile Software Development in an Academic Environment
82. **Francisco Augusto Alcaraz Garcia**, Real Options, Default Risk and Soft Applications
83. **Kai K. Kimppa**, Problems with the Justification of Intellectual Property Rights in Relation to Software and Other Digitally Distributable Media
84. **Dragoş Truşcan**, Model Driven Development of Programmable Architectures
85. **Eugen Czeizler**, The Inverse Neighborhood Problem and Applications of Welch Sets in Automata Theory

86. **Sanna Ranto**, Identifying and Locating-Dominating Codes in Binary Hamming Spaces
87. **Tuomas Hakkarainen**, On the Computation of the Class Numbers of Real Abelian Fields
88. **Elena Czeizler**, Intricacies of Word Equations
89. **Marcus Alanen**, A Metamodeling Framework for Software Engineering
90. **Filip Ginter**, Towards Information Extraction in the Biomedical Domain: Methods and Resources
91. **Jarkko Paavola**, Signature Ensembles and Receiver Structures for Oversaturated Synchronous DS-CDMA Systems
92. **Arho Virkki**, The Human Respiratory System: Modelling, Analysis and Control
93. **Olli Luoma**, Efficient Methods for Storing and Querying XML Data with Relational Databases
94. **Dubravka Ilić**, Formal Reasoning about Dependability in Model-Driven Development
95. **Kim Solin**, Abstract Algebra of Program Refinement
96. **Tomi Westerlund**, Time Aware Modelling and Analysis of Systems-on-Chip
97. **Kalle Saari**, On the Frequency and Periodicity of Infinite Words
98. **Tomi Kärki**, Similarity Relations on Words: Relational Codes and Periods
99. **Markus M. Mäkelä**, Essays on Software Product Development: A Strategic Management Viewpoint
100. **Roope Vehkalahti**, Class Field Theoretic Methods in the Design of Lattice Signal Constellations
101. **Anne-Maria Ernvall-Hytönen**, On Short Exponential Sums Involving Fourier Coefficients of Holomorphic Cusp Forms
102. **Chang Li**, Parallelism and Complexity in Gene Assembly
103. **Tapio Pahikkala**, New Kernel Functions and Learning Methods for Text and Data Mining
104. **Denis Shestakov**, Search Interfaces on the Web: Querying and Characterizing
105. **Sampo Pyysalo**, A Dependency Parsing Approach to Biomedical Text Mining
106. **Anna Sell**, Mobile Digital Calendars in Knowledge Work
107. **Dorina Marghescu**, Evaluating Multidimensional Visualization Techniques in Data Mining Tasks
108. **Tero Säntti**, A Co-Processor Approach for Efficient Java Execution in Embedded Systems
109. **Kari Salonen**, Setup Optimization in High-Mix Surface Mount PCB Assembly
110. **Pontus Boström**, Formal Design and Verification of Systems Using Domain-Specific Languages
111. **Camilla J. Hollanti**, Order-Theoretic Mehtods for Space-Time Coding: Symmetric and Asymmetric Designs
112. **Heidi Himmanen**, On Transmission System Design for Wireless Broadcasting
113. **Sébastien Lafond**, Simulation of Embedded Systems for Energy Consumption Estimation
114. **Evgeni Tsivtsivadze**, Learning Preferences with Kernel-Based Methods
115. **Petri Salmela**, On Commutation and Conjugacy of Rational Languages and the Fixed Point Method
116. **Siamak Taati**, Conservation Laws in Cellular Automata
117. **Vladimir Rogojin**, Gene Assembly in Stichotrichous Ciliates: Elementary Operations, Parallelism and Computation
118. **Alexey Dudkov**, Chip and Signature Interleaving in DS CDMA Systems
119. **Janne Savela**, Role of Selected Spectral Attributes in the Perception of Synthetic Vowels
120. **Kristian Nybom**, Low-Density Parity-Check Codes for Wireless Datacast Networks
121. **Johanna Tuominen**, Formal Power Analysis of Systems-on-Chip
122. **Teijo Lehtonen**, On Fault Tolerance Methods for Networks-on-Chip
123. **Eeva Suvitie**, On Inner Products Involving Holomorphic Cusp Forms and Maass Forms
124. **Linda Mannila**, Teaching Mathematics and Programming – New Approaches with Empirical Evaluation
125. **Hanna Suominen**, Machine Learning and Clinical Text: Supporting Health Information Flow
126. **Tuomo Saarni**, Segmental Durations of Speech
127. **Johannes Eriksson**, Tool-Supported Invariant-Based Programming

128. **Tero Jokela**, Design and Analysis of Forward Error Control Coding and Signaling for Guaranteeing QoS in Wireless Broadcast Systems
129. **Ville Lukkarila**, On Undecidable Dynamical Properties of Reversible One-Dimensional Cellular Automata
130. **Qaisar Ahmad Malik**, Combining Model-Based Testing and Stepwise Formal Development
131. **Mikko-Jussi Laakso**, Promoting Programming Learning: Engagement, Automatic Assessment with Immediate Feedback in Visualizations
132. **Riikka Vuokko**, A Practice Perspective on Organizational Implementation of Information Technology
133. **Jeanette Heidenberg**, Towards Increased Productivity and Quality in Software Development Using Agile, Lean and Collaborative Approaches
134. **Yong Liu**, Solving the Puzzle of Mobile Learning Adoption
135. **Stina Ojala**, Towards an Integrative Information Society: Studies on Individuality in Speech and Sign
136. **Matteo Brunelli**, Some Advances in Mathematical Models for Preference Relations
137. **Ville Junnila**, On Identifying and Locating-Dominating Codes
138. **Andrzej Mizera**, Methods for Construction and Analysis of Computational Models in Systems Biology. Applications to the Modelling of the Heat Shock Response and the Self-Assembly of Intermediate Filaments.
139. **Csaba Ráduly-Baka**, Algorithmic Solutions for Combinatorial Problems in Resource Management of Manufacturing Environments
140. **Jari Kyngäs**, Solving Challenging Real-World Scheduling Problems
141. **Arho Suominen**, Notes on Emerging Technologies
142. **József Mezei**, A Quantitative View on Fuzzy Numbers
143. **Marta Olszewska**, On the Impact of Rigorous Approaches on the Quality of Development
144. **Antti Airola**, Kernel-Based Ranking: Methods for Learning and Performace Estimation
145. **Aleksi Saarela**, Word Equations and Related Topics: Independence, Decidability and Characterizations
146. **Lasse Bergroth**, Kahden merkkijonon pisimmän yhteisen alijonon ongelma ja sen ratkaiseminen
147. **Thomas Canhao Xu**, Hardware/Software Co-Design for Multicore Architectures
148. **Tuomas Mäkilä**, Software Development Process Modeling – Developers Perspective to Contemporary Modeling Techniques
149. **Shahrokh Nikou**, Opening the Black-Box of IT Artifacts: Looking into Mobile Service Characteristics and Individual Perception
150. **Alessandro Buoni**, Fraud Detection in the Banking Sector: A Multi-Agent Approach
151. **Mats Neovius**, Trustworthy Context Dependency in Ubiquitous Systems
152. **Fredrik Degerlund**, Scheduling of Guarded Command Based Models
153. **Amir-Mohammad Rahmani-Sane**, Exploration and Design of Power-Efficient Networked Many-Core Systems
154. **Ville Rantala**, On Dynamic Monitoring Methods for Networks-on-Chip
155. **Mikko Pelto**, On Identifying and Locating-Dominating Codes in the Infinite King Grid
156. **Anton Tarasyuk**, Formal Development and Quantitative Verification of Dependable Systems
157. **Muhammad Mohsin Saleemi**, Towards Combining Interactive Mobile TV and Smart Spaces: Architectures, Tools and Application Development
158. **Tommi J. M. Lehtinen**, Numbers and Languages
159. **Peter Sarlin**, Mapping Financial Stability
160. **Alexander Wei Yin**, On Energy Efficient Computing Platforms
161. **Mikołaj Olszewski**, Scaling Up Stepwise Feature Introduction to Construction of Large Software Systems
162. **Maryam Kamali**, Reusable Formal Architectures for Networked Systems
163. **Zhiyuan Yao**, Visual Customer Segmentation and Behavior Analysis – A SOM-Based Approach
164. **Timo Jolivet**, Combinatorics of Pisot Substitutions
165. **Rajeev Kumar Kanth**, Analysis and Life Cycle Assessment of Printed Antennas for Sustainable Wireless Systems
166. **Khalid Latif**, Design Space Exploration for MPSoC Architectures

167. **Bo Yang**, Towards Optimal Application Mapping for Energy-Efficient Many-Core Platforms
168. **Ali Hanzala Khan**, Consistency of UML Based Designs Using Ontology Reasoners
169. **Sonja Leskinen**, m-Equine: IS Support for the Horse Industry
170. **Fareed Ahmed Jokhio**, Video Transcoding in a Distributed Cloud Computing Environment
171. **Moazzam Fareed Niazi**, A Model-Based Development and Verification Framework for Distributed System-on-Chip Architecture
172. **Mari Huova**, Combinatorics on Words: New Aspects on Avoidability, Defect Effect, Equations and Palindromes
173. **Ville Timonen**, Scalable Algorithms for Height Field Illumination
174. **Henri Korvela**, Virtual Communities – A Virtual Treasure Trove for End-User Developers
175. **Kameswar Rao Vaddina**, Thermal-Aware Networked Many-Core Systems
176. **Janne Lahtiranta**, New and Emerging Challenges of the ICT-Mediated Health and Well-Being Services
177. **Irum Rauf**, Design and Validation of Stateful Composite RESTful Web Services
178. **Jari Björne**, Biomedical Event Extraction with Machine Learning
179. **Katri Haverinen**, Natural Language Processing Resources for Finnish: Corpus Development in the General and Clinical Domains
180. **Ville Salo**, Subshifts with Simple Cellular Automata
181. **Johan Ersfolk**, Scheduling Dynamic Dataflow Graphs
182. **Hongyan Liu**, On Advancing Business Intelligence in the Electricity Retail Market
183. **Adnan Ashraf**, Cost-Efficient Virtual Machine Management: Provisioning, Admission Control, and Consolidation
184. **Muhammad Nazrul Islam**, Design and Evaluation of Web Interface Signs to Improve Web Usability: A Semiotic Framework
185. **Johannes Tuikkala**, Algorithmic Techniques in Gene Expression Processing: From Imputation to Visualization
186. **Natalia Díaz Rodríguez**, Semantic and Fuzzy Modelling for Human Behaviour Recognition in Smart Spaces. A Case Study on Ambient Assisted Living
187. **Mikko Pänkäälä**, Potential and Challenges of Analog Reconfigurable Computation in Modern and Future CMOS
188. **Sami Hyrynsalmi**, Letters from the War of Ecosystems – An Analysis of Independent Software Vendors in Mobile Application Marketplaces
189. **Seppo Pulkkinen**, Efficient Optimization Algorithms for Nonlinear Data Analysis
190. **Sami Pyöttiälä**, Optimization and Measuring Techniques for Collect-and-Place Machines in Printed Circuit Board Industry
191. **Syed Mohammad Asad Hassan Jafri**, Virtual Runtime Application Partitions for Resource Management in Massively Parallel Architectures
192. **Toni Ernvall**, On Distributed Storage Codes
193. **Yuliya Prokhorova**, Rigorous Development of Safety-Critical Systems
194. **Olli Lahdenoja**, Local Binary Patterns in Focal-Plane Processing – Analysis and Applications
195. **Annika H. Holmbom**, Visual Analytics for Behavioral and Niche Market Segmentation

Turku Centre *for* Computer Science

Joukahaisenkatu 3-5 B, 20520 Turku, Finland | www. tucs.fi

University of Turku
Faculty of Mathematics and Natural Sciences
- Department of Information Technology
- Department of Mathematics and Statistics

Turku School of Economics
- Institute of Information Systems Science

Åbo Akademi University
Faculty of Science and Engineering
- Computer Engineering
- Computer Science

Faculty of Social Sciences, Business and Economics
- Information Systems

ISBN 978-952-12-3204-6
ISSN 1239-1883

Annika H. Holmbom

Visual Analytics for Behavioral and Niche Market Segmentation