VAASAN YLIOPISTO

MATTI LAAKSONEN – SEPPO PYNNÖNEN (Eds)

Contributions to Management Science, Mathematics and Modelling

Essays in Honour of Professor Ilkka Virtanen

ACTA WASAENSIA

No. 122

Mathematics 9

UNIVERSITAS WASAENSIS 2004

Ilkka Virtanen

Foreword

This volume consists of 15 contributions written by 23 authors in management science, mathematics and economics modelling dedicated to Ilkka Virtanen on the occasion of his 60th birthday. Ilkka Virtanen has had an outstanding career as a researcher, teacher, university administrator, and as an active member of the business and the wider community. This collection gives us an opportunity to express our gratitude to him.

We would like to extend our gratitude to the contributors of this volume for joining us in honouring our friend and colleague. We are particularly grateful to Seppo Hassi for his continuous help and advice during the preparation of the volume. We are grateful to John Shepherd for checking the language of this foreword and the English version of the short biography of Ilkka Virtanen. We would also like to thank Anneli Virta for her invaluable secretarial work during the process and Tarja Salo for her skilful editorial work.

Vaasa, November, 2003

Matti Laaksonen Seppo Pynnönen

Ilkka Virtanen 60 vuotta

Ilkka Virtanen syntyi Paimiossa tammikuun 4. päivänä 1944. Hän valmistui 1962 ylioppilaaksi Paimion yhteiskoulusta ja suoritti filosofian kandidaatin tutkinnon 1968 Turun yliopistossa pääaineena sovellettu matematiikka. Vuonna 1974 hän suoritti filosofian lisensiaatin tutkinnon ja väitteli filosofian tohtoriksi vuonna 1977 Turun yliopistossa sovelletussa matematiikassa.

Ilkka Virtanen toimi vuosina 1968–1978 talous- ja tilastomatematiikan assistenttina Turun kauppakorkeakoulussa. Vuosina 1978–1981 Ilkka Virtanen toimi Lappeenrannan teknillisen korkeakoulussa matematiikan lehtorina. Vuonna 1979 hänet nimitettiin Turun kauppakorkeakoulun talous- ja tilastomatematiikan dosentiksi ja vuonna 1981 Lappeenrannan teknillisen korkeakoulun talous- ja tilastomatematiikan dosentiksi. Vaasan korkeakouluun Ilkka Virtanen nimitettiin tilastotieteen apulaisprofessorin virkaan 1981 ja talousmatematiikan professorin virkaan 1983.

Ilkka Virtasen tieteellinen julkaisutoiminta käsittää yli 100 refereemenettelyn läpikäynyttä tieteellistä julkaisua. Tieteellisen uransa alkuaikoina hänen kiinnostuksensa oli luotettavuusteoriassa, josta hän laati väitöskirjansa. Sittemmin kiinnostuksen kohteena ovat olleet mm. yleiseen systeemiteoriaan, investointisuunnitteluun ja entropia-analyysiin liittyvä tutkimus. Viimeaikaisimpana kiinnostuksen kohteena Virtasella on ollut erityisesti tilinpäätösanalyysin ja osakemarkkinoiden mallintamiseen liittyvät kysymykset.

Virtasen tutkimukselle luonteenomaista on laaja yhteistyö erityisesti taloustieteiden sovellusalojen tutkijoiden kanssa. Jotta yhteistyö toimisi, on se merkinnyt matematiikkaa pääaineena opiskelleelle omaehtoista perehtymistä talouden substanssiteoriaan. Turun kauppakorkeakoulussa Virtanen paneutui professori Pentti Malaskan ohjauksessa kauppakorkeakoulun tutkimusalueisiin ja matemaattisen mallintamisen hyödyntämiseen tässä ympäristössä. Työ jatkui Ilkka Virtasen siirryttyä Lappeenrannan teknillisen korkeakouluun, jossa verrattomaksi yhteistyökumppaniksi osoittautui yrityksen taloustieteet monipuolisesti taitava Teemu Aho. Hänen kanssaan Ilkka Virtanen julkaisi lukuisia yrityksen kannattavuuteen, investointilaskelmiin ja yleisesti tilinpäätösalaan liittyviä tutkimuksia. Vaasan korkeakouluun siirryttyään Virtasen tilinpäätösalaan liittyvä tutkimus sai uusia vivahteita, kun tutkimus laajeni rahoitusalalle. Tärkeinä yhteistyökumppaneina Vaasan yliopistosta ovat olleet Reijo Ruuhela, Timo Salmi ja Paavo Yli-Olli sekä tämän ryhmän jälkipolvena syntyneet useat nuoremmat tutkijat. Vaasan yliopiston aikana kontaktipinta laajeni entisestään rahoitusmarkkinatutkimuksen myötä ulkomaisiin yhteistyökumppaneihin, joista erityisesti mainittakoon Geoffrey G. Booth ja Angappa Gunasekaran.

Kasvaneiden luottamustoimien myötä Ilkka Virtasen viimeisimmissä tutkimuksissa mukaan on tullut yliopistojen ja korkeakoulujen arviointi. Virtanen on osallistunut laajamittaisesti suomalaisten korkeakoulujen arviointiin. Lisäksi hän on ollut myös Viron yliopistojen kauppatieteellistä koulutusta arvioivan kansainvälisen asiantuntijaryhmän jäsenenä.

Ilkka Virtanen tunnetaan myös energisenä jatko-opintojen ohjaajana. Aktiivinen osallistuminen laskentatoimen ja rahoituksen jatko-opintoseminaareihin takasi monelle lisensiaattityötä ja väitöskirjaa valmistelevalle rakentavat ja täsmälliset neuvot työn edistämiseksi. Vaikka Vaasan yliopistossa ei pääaineena varsinaisesti ole matemaattisia aineita, on jatkotutkintojen suorittaminen näissä aineissa mahdollista. Virtasen ohjauksessa valmistui useita lisensiaatintöitä ja väitöskirjoja. Hänen oppilaistaan Vaasassa väitteli ensimmäisenä Jukka Perttunen, sitten Irma Luhta, Matti Laaksonen ja Olli Bräysy.

Ansiokkaan tutkijanuransa lisäksi Ilkka Virtasella on poikkeuksellisen laaja ja monipuolinen ura myös muissa korkeakoulumaailman tehtävissä. Vuosina 1984–1987 hän oli Vaasan korkeakoulun vararehtorina ja vuosina 1987–1994 Vaasan korkeakoulun/ yliopiston rehtorina. Vuosina 1983–1984 Ilkka Virtanen toimi myös Vaasan kesäyliopiston vararehtorina ja rehtorina vuosina 1985–1987. Näiden lisäksi hän on toiminut Vaasan yliopiston kaupallis-teknisen tiedekunnan dekaanina vuosina 1998–2001 ja Vaasan yliopiston teknillisen tiedekunnan (vuoden 2002 loppuun informaatioteknologian tiedekunta) dekaanina vuodesta 2002 alkaen. Näiden lisäksi Ilkka Virtanen on toiminut useaan otteeseen laitoksen johtajana. Vaasan yliopiston/korkeakoulun hallituksen jäsenenä hän oli 1982–1983, varapuheenjohtajana 1984–1987, puheenjohtajana 1987–1994 ja uudestaan jäsenenä vuodesta 1998 alkaen. Näiden lisäksi Ilkka Virtasella on ollut lukuisia toimikuntien ja työryhmien jäsenyyksiä ja puheenjohtajuuksia.

Suomen yliopistojen ja korkeakoulujen rehtorien neuvoston jäsenenä Ilkka Virtanen oli vuosina 1987–1994 ja työvaliokunnan jäsenenä 1991–1992. Hän toimii ja on toiminut myös useissa tutkimusta ja koulutusta tukevien säätiöiden hallituksissa, kuten Liikesivistysrahaston Vaasan aluetoimikunnan puheenjohtajana vuodesta 1993 alkaen, Karl Erling ja Anja Nymanin säätiön hallituksen jäsenenä vuodesta 1989 alkaen, Suomen Kulttuurirahaston Etelä-Pohjanmaan rahaston hoitokunnan ja sen työvaliokunnan jäsenenä vuodesta 2000 alkaen. Vuosina 1987–1994 hän toimi Vaasan korkeakoulu/yliopistosäätiön hallituksen jäsenenä. Vuosina 1987–1995 hän oli Vientikoulutussäätiön/Fintran (The Finnish Institute for International Trade) valtuuskunnan jäsenenä.

Ilkka Virtaselle on kertynyt myös lukuisia talouden ja yhteiskunnan luottamustehtäviä. Näistä mainittakoon: Vaasan Osuuspankin hallintoneuvoston jäsen 1990–1995 ja johtokunnan varajäsen vuodesta 1995 alkaen, Osuuskauppa KPO:n edustajiston jäsen vuodesta 1996 alkaen, Vaasan teknillisen oppilaitoksen johtokunnan jäsen 1992–1995, Vaasan hotelli- ja ravintolaoppilaitoksen johtokunnan jäsen 1993–1995 ja Kokkolan seudun osaamiskeskuksen neuvottelukunnan jäsen vuodesta 2003 alkaen.

Vaikka Ilkka Virtasen korkeakoulu- ja yhteiskuntaelämään osallistuminen on ollut näinkin laajaa, on hänellä aina riittänyt aikaa myös harrastuksille. Kulttuuri on aina ollut Ilkka Virtasta lähellä. Osallistumisajat teatteri, ooppera ja klassisen musiikin konserttiesityksiin ovat aina löytäneet paikkansa Virtasen kalenterissa, samoin kuin aktiivinen osallistuminen eri kulttuurijärjestöjen toimintaan. Kulttuuriharrastusten lisäksi vapaaehtoinen maanpuolustustyö on myös lähellä Virtasen sydäntä.

Ilkka Virtasen laajaa panosta tieteen, yliopiston, yleisesti korkeakoulun, yhteiskunnan ja talouselämän alueilla on mahdotonta sisällyttää yhden teoksen puitteisiin. Niinpä teoksessa keskitytäänkin ainoastaan Virtasen toiminnan yhteen tärkeään osa-alueeseen – tieteeseen, jossa kirjoitustensa kautta tieteenalansa merkittävät tutkijat haluavat onnitella Ilkka Virtasta hänen juhlapäivänään.

Ilkka Virtanen: 60 years

Ilkka Virtanen was born in Paimio on the 4th January, 1944. He matriculated in 1962, and was awarded his Master's degree in 1968, majoring in Applied Mathematics at the University of Turku. In 1974 he was awarded a Licentiate degree, and in 1977 defended his doctoral thesis in Applied Mathematics at the University of Turku.

Ilkka Virtanen was Assistant in the Business Mathematics and Statistics Department at the Turku School of Economics and Business Administration between 1968–1978. During 1978–1981 he was Senior Lecturer in Mathematics at Lappeenranta University of Technology. In 1979 he was named Docent of Business Mathematics and Statistics at the Turku School of Economics and Business Administration and in 1981 Docent of Business Mathematics and Statistics at the Lappeenranta University of Technology. He was appointed Associate Professor of Statistics at the University of Vaasa in 1981, and Professor of Management Sciences in 1983.

Ilkka Virtanen has produced more than 100 refereed scientific publications. At the beginning of his scientific career his interest was in reliability theory and maintenance, the area in which he defended his doctoral thesis. Later on his interest has been, among other things, in general system theory, investment planning, and entropy analysis. The most recent subjects of his interest have been in financial statement analysis and stock market applications. For instance, he has worked on classification of financial ratios, association between financial ratios and security characteristics, and equilibrium models in thin markets.

Extensive cooperation with researchers in applied economics is characteristic of Ilkka Virtanen's research. In order to make productive cooperation possible, a keen and spontaneous interest in the substance theory of economics has been necessary for a person who in fact studied mathematics as a major subject. The period spent in Professor Pentti Malaska's group at the Turku School of Economics and Business Administration was a time for Ilkka Virtanen to orientate his mathematics expertise towards research with applications in business economics. This work continued when he moved to Lappeenranta

University of Technology, where Professor Teemu Aho proved to be an excellent collaborator in the field. With him, Ilkka Virtanen published numerous papers dealing with profitability, investment calculations, and other research topics typically related to financial statement analysis. Research in the field of financial statements developed further with asset pricing applications when he moved to the University of Vaasa. Important research collaborators at the University of Vaasa have been Reijo Ruuhela, Timo Salmi, Paavo Yli-Olli, and several younger scholars nurtured by this group. At the same time, the contact enlarged further afield to foreign colleagues, of whom Geoffrey G. Booth and Angappa Gunasekaran should be mentioned as leading figures.

Along with an increasing number of positions of responsibility, research related to university evaluation has become the latest new area of Ilkka Virtanen's expertise. He has participated widely in the evaluation process of Finnish universities as well as several foreign universities. For instance, he has been a member of an international peer review team evaluating business education in Estonian universities.

Ilkka Virtanen is well-known also as an energetic supervisor of post-graduates. His active participation in the post-graduate seminars of management accounting and finance guaranteed a supply of constructive and accurate advice to the participants working on their doctoral theses. Even though at the University of Vaasa there is no major subject in mathematics and statistics at master's degree level, it is possible to take postgraduate degrees in these subjects. Under Ilkka Virtanen's supervision several Licentiate and Doctoral theses were completed. From his students Jukka Perttunen was the first to defend his dissertation in Vaasa, followed by Irma Luhta, Matti Laaksonen, and Olli Bräysy.

In addition to his excellent career as a researcher, Ilkka Virtanen has had an exceptionally versatile career with respect to other tasks in the university sector. From 1984–1987 he was Vice Rector of the University of Vaasa and Rector during 1987–1994. From 1983–1984 he was Vice Rector of the Summer University of Vaasa, and Rector from 1985–1987. In addition, he was Dean of the Faculty of Business Administration and Accounting at the University of Vaasa from 1998–2001, and has been Dean of the Faculty of Technology (the Faculty of Information Technology until the end of 2002) since 2002, and furthermore held the Department Chair for several periods in his departments. He has been a member of the Board of the University of Vaasa during 1982–1983 and since 1998, Vice-Chairman from 1984–1987, and Chairman from 1987–1994. In addition to these positions, he has been a member and chairman of numerous administrative commissions, societies, and research foundations.

Ilkka Virtanen was a member of the council of Finnish University Rectors from 1987–1994, and a member of the working committee from 1991–1992. He is a member of the administrative boards of several research foundations, holding positions such as Chair of the Vaasa Regional Committee of the Foundation of Economic Education (Liike-sivistysrahaston Vaasan aluetoimikunta) since 1993. Since 1989 he has been a member of the administrative board of the Karl Erling and Anja Nyman Foundation. From 1987–1994 he was a member of the administrative board of the Foundation of the University of Vaasa, and from 1987–1995 a member of the council of The Finnish Institute for International Trade (FINTRA).

Ilkka Virtanen also serves on several committees and administrative boards in the business sphere and in the wider community. For instance, he was a member of the administrative board of Vaasan Osuuspankki from 1990–1995, has been a member of the representatives of Osuuskauppa KPO since 1996, and a member of the administrative boards of several educational institutes in Vaasa.

In addition to his exceptionally wide contribution to university and business life and to the wider community, he has always found time for other cultural and social life as well. Cultural life has been always near to Ilkka Virtanen's heart. He is a frequent visitor to the theatre and the opera, as well as to classical music concerts and other cultural events.

Ilkka Virtanen's wide contribution to science, to the university community, to economics and to society at large is impossible to convey within the limits of only one volume. This volume focuses merely on one important aspect – science, where experts in their field want to congratulate Ilkka Virtanen on his 60th birthday through the medium of their writing.

Tabula Gratulatoria

Aaltio Iiris
Aho Teemu
Alander Jarmo Tapani
Do Thi Quynh Nhu
Elomaa Virpi
Fellman Johan
Gustafsson Christina
Hakkarainen Raija ja Erkki
Havunen Jouko
Ingström Petri
Jakobsson Matti ja Marja-Leena
Jokisalo Jorma
Jutila Matti
Jäntti Riku
Kallunki Juha-Pekka
Kantanen Helena ja Teuvo
Katajamäki Hannu
Kerttula Esa
Kinnunen Juha
Knif Johan
Kolari James
Kolehmainen Osmo
Korhonen Pekka
Koskela Merja
Kukkohovi Kari
Laakkonen Eero Veli
Laaksonen Martti
Laaksonen Pirjo
Lahtinen Aatos
Laitinen Erkki ja Teija
Lampinen Jouni
Lanne Markku
Larimo Marjatta ja Jorma
Laurén Ulla ja Christer
Lehtonen Asko
Leppiniemi Jarmo
Liljeblom Eva
Linna Matti
Liski Erkki
Lonka Heikki
Lukka Kari
Lumio Markku
Mangeloja Esa
Mannermaa Mika

Miettinen Kaisa
Mikkonen Kauko
Niemi Pentti Valtteri
Nordberg Leif
Nordblad John
Nordman Marianne
Nyblom Jukka Olavi
Nyqvist Lars
Palomäki Mauri ja Sirkku
Palonen Vuokko
Pape Bernd
Parry Christoph
Pukkila Tarmo
Puntanen Simo
Pursiheimo Ulla
Riepula Esko
Rekilä Eila
Rintanen Markku
Rosenqvist Gunnar
Routamaa Vesa
Ruotsala Pentti
Ruuhela Reijo
Salmenjoki Kimmo
Salomaa Hely Tuulikki
Sandström Jaana
Saretsalo Lauri
Skog-Södersved Mariann
Strandin Bengt
Suomela Pentti
Suutari Vesa
Swanljung Harry ja Katriina
Takala Josu
Tarvonen Sari
Tietäväinen Aimo
Tolonen Juha
Uusitupa Matti
Wanne Merja
Vartiainen Perttu
Vartiainen Pirkko
Vekara Timo
Vesalainen Jukka
Vieru Markku
Virta Anneli
Virtanen Keijo
Väänänen Keijo
Yli-Olli Paavo

ABB Oy
HM-Profiili Oy Martela
Humanistinen tiedekunta, Vaasan yliopisto
ISS Suomi Oy
Kansantaloustieteen laitos, Vaasan yliopisto
Matemaattisten tieteiden laitos, Vaasan yliopisto
Medivire työterveyspalvelut Oy
Nordea Pankki Suomi Oyj
Oy Merinova Ab
Pohjanmaan maanpuolustuskilta
Pohjanmaan Radio
Professoriliitto
Professoriliitto, Vaasan yliopiston osasto
Svenska handelshögskolan stiftelse
Teknillinen tiedekunta, Vaasan yliopisto
Tieteentekijöiden liitto
Tietotekniikan laitos, Vaasan yliopisto
Tietotekniikan osasto, Lappeenrannan teknillinen yliopisto
Tuotantotalous, Lappeenrannan teknillinen yliopisto
Vaasan kaupungin opetustoimi
Vaasan kaupunki
Vaasan seurakuntayhtymä
Vaasan Sähkö Oy
Vasa Andelsbanken
VLP Oy
Wärtsilä Finland Oy
Yleishallinto, Vaasan yliopisto

ACTA WASAENSIA

CONTENTS

ACTA WASAENSIA

Scheduling flexibility and insertion zones in routing

Olli Bräysy, Wout Dullaert, and Geert Van de Weyer

Dedicated to Ilkka Virtanen on the occasion of his 60th birthday

Abstract

Bräysy, Olli, Wout Dullaert, and Geert Van de Weyer (2004). Scheduling flexibility and insertion zones in routing. In *Contributions to Management Science, Mathematics and Modelling. Essays in Honour of Professor Ilkka Virtanen.* Acta Wasaensia No. 122, 21–36. Eds Matti Laaksonen and Seppo Pynnönen.

In this paper, scheduling flexibility and insertion zones are formally defined for less-than-truckload (LTL) and full-tuckload (FTL) routing. Scheduling flexibility refers to the flexibility that a time-constrained customer (LTL) or load (FTL) offers to a dispatcher. Insertion zones indicate the area from which a customer (LTL) or load (FTL) can be inserted into a partially finished route. For both cases, the insertion zones are proven to be elliptic.

Olli Bräysy, Department of Optimization, SINTEF Applied Mathematics, P.O. Box 124 Blindern, 0314 Oslo, Norway, e-mail olli.braysy@sintef.no.
Wout Dullaert, Institute of Transport and Maritime Management Antwerp, University of Antwerp, Keizerstraat 64, 2000 Antwerp, Belgium, e-mail wout.dullaert@ua.ac.be.
Geert Van de Weyer, Faculty of Applied Economics, University of Antwerp, Prinsstraat 13, 2000 Antwerp, Belgium, e-mail geert.vandeweyer@ua.ac.be.

Key words: Vehicle routing, time windows, insertion zones.

1. Introduction

Scheduling flexibility refers to the degree of freedom that a customer offers a dispatcher to design routes. The higher the flexibility, the more cost efficient the routes the dispatcher can design. Although some authors have reported on certain aspects of scheduling flexibility in routing before, no systematic approach has been published. In this paper, we lay the foundation for a systematic approach to scheduling flexibility.

In the case of less-than-truckload (LTL) routing, scheduling flexibility takes the form of time windows in which customers wish to be serviced. These time windows can differ as far as their moment in time and width is concerned. In the case of a full-truckload (FTL) routing problem, the dispatcher's objective is to service loads between two nodes at minimal distribution costs, instead of servicing individual customers (nodes). In this case the scheduling flexibility of a load is determined by both the time windows of the pick-up and the delivery nodes. A route no longer consists of nodes (i.e. individual unrouted customers), but of loaded route segments. A loaded route segment is defined as an arc on which a load is transported between two nodes. Unloaded arcs are used to travel between the loaded arcs and from/to the depot. The scheduling flexibility of a loaded arc is determined by the time windows of both its starting and ending node.

An insertion zone is the area in which a customer (or a load) can be inserted between two others. For LTL routing, the scheduling flexibility of each of two adjacent customers in the route determines whether a new customer can be inserted between them. If the time windows of customers i and j offer more time than needed to service i and travel directly to j, there may be time left to insert an unrouted customer u, located in an insertion zone around i and j. For the loaded arcs in FTL routing the same reasoning applies. The time available to travel from one loaded arc to another, can be used to insert a new loaded arc.

Flexibility and scheduling flexibility issues have been addressed in a number of fields such as manufacturing, computing, labor economics etc. In the literature on the Vehicle Routing Problem a number of authors have modeled situations with different levels of scheduling flexibility or they informally referred to its impact on routing costs. Therefore a brief literature review on scheduling flexibility will be presented in Section 2. In the next section, scheduling flexibility is defined for the LTL case and the insertion zones are shown to be of elliptic shape. In Section 4 the same is done for full-truckload routing with time windows. Insertion zones between loaded route segments are proven to be also elliptic. Finally, conclusions are formulated.

2. Literature review on flexibility and scheduling flexibility

Flexibility is an important issue in the contemporary, globalizing economy. Increasing competition forces companies to quickly adapt/react to changes in the environment.

Customers' (variable) demand for high-quality, differentiated products have – through concepts such as Just-In-Time and Total Quality Management – spurred academics' and practitioners' interest for *operational* or *productive flexibility*[1].

Flexible manufacturing systems are designed to combine the efficiency of large scale production with the flexibility of a job shop environment. They can produce different products, while keeping setup times short and work-in-progress inventories low. As a result, a wide range of products can be offered while keeping costs at an acceptable level. Because a higher need for flexibility imposes additional constraints on the production process to be optimized, there is a trade-off between flexibility and efficiency (i.e. production costs). Tarifa and Chiotti (1995) study this trade-off in the so-called Flexibility Problem. A bi-criterion optimization approach is used to determine the optimal size of a plant such that it satisfies all constraints for any of the parameters during the process operation.

Flexible manufacturing often requires a flexible workforce. However, not only from the employers' side there is a call for flexibility as flexibility promises improved working conditions and more varied and more interesting jobs (Dyer 1998). Daniels et al. (1999) study the operational impact of both machine flexibility and labor flexibility. In the field of computing, considerable attention is paid to ways to increase the flexibility of real-time systems (Burns and Fohler 1991; Burns et al. 2000). Also the flexible delivery of (vocational) training has been addressed in the literature (Evans and Smith 1999; Smith 2000).

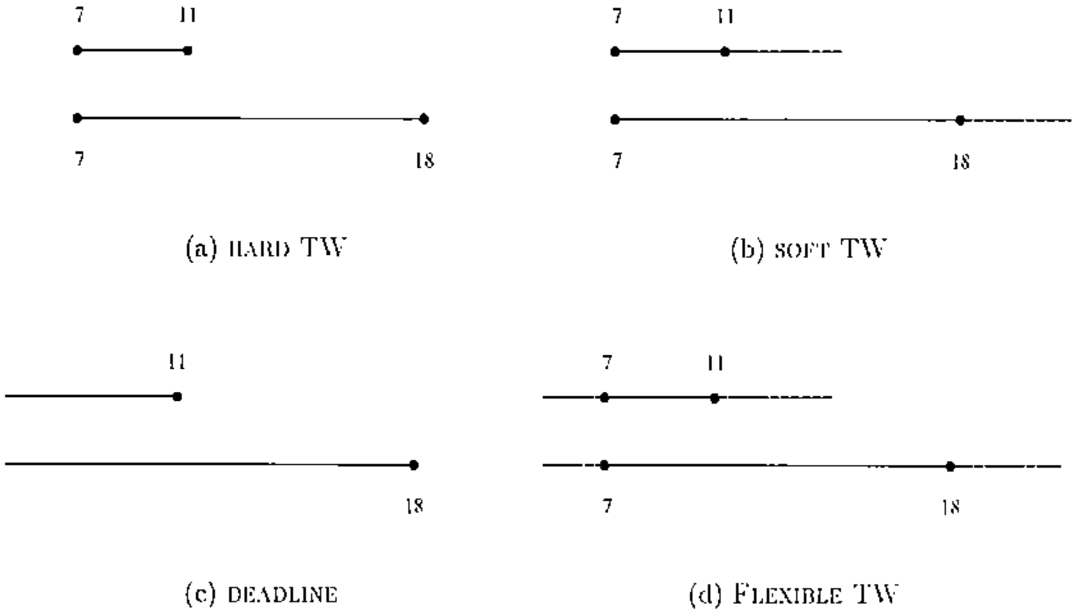
In activity scheduling models for travel demand a customer's ability (or flexibility) to revise his schedule under new circumstances and the flexibility of the schedule itself to account for new activities is addressed (Venter and Hansen 1998). While these activities can be scheduled over shorter and longer time horizons, little attention is paid to the impact of time window sizes on the flexibility of the schedule.

[1] See Martinez et al. (2000) for a study of the relation between operational flexibility, JIT and TQM. Reviews on flexibility in manufacturing can be found in Beckman (1990), Sethi and Sethi (1990), Hyun and Ahn (1992), and Upton (1995).

In transport and logistics, operational flexibility is a driving source for product innovation and cost reduction. A few examples are: trucks that can combine with different types of semi-trailers (e.g. for bulk or container transport), multi-compartment, multi-temperature semi-trailers to maximize loading flexibility for supermarkets (Clancy 2000), flexible bulk palletizing for the optimum bulk pallet for each client or job (DeFayette 1996), automated transport and inventory systems for aluminium coils (Aluminium 1994), designing space modules (Basile et al. 1998).

The impact of novel equipment or technologies on the operational flexibility of a transport firm has received more attention than the impact of customers' desiderata. Scheduling flexibility can different forms. Customers that allow their delivery to be split over two or more visits are clearly more flexible than those requesting a single visit. Customers allowing for these so-called split deliveries enable a dispatcher to achieve a higher vehicle (capacity) utilization. Customers located in the city center and/or with a narrow entrance to the company premises can only be serviced with smaller vehicles and therefore they are more difficult or more expensive to service. The most important form of scheduling flexibility is probably flexibility with respect to the moment of delivery. In the literature, little attention is paid to the cost impact of a customer's specifications on the moment on which he wants to be serviced. If customers are demanding on the moment of delivery, they offer a dispatcher little flexibility to schedule their order. As a result, rigid customers can lead to cost ineffective schedules with considerable waiting time and additional distance to be traveled. Because customers in the Vehicle Routing Problem with Time Windows (VRPTW) have to be serviced in a time window of their choice, the nature and size of the time window reflects their scheduling flexibility.

Four types of time windows have been studied – in decreasing amount of research spent on – : hard, soft, one-sided and flexible time windows. If time windows are hard, service has to start within the specified time window (see Figure 1(a)). In the soft time window case (see Figure 1(b)), a vehicle is allowed to arrive too late at a customer but a penalty is incurred (see e.g. Taillard et al. (1997)). The rationale behind soft time windows is that by allowing a few (small) time window violations, solution quality can be significantly improved. In both the hard and the soft time window case, a vehicle arriving too early has to wait until the start of the service time window. This is not the case if customers have one-sided time windows without an earliest time (Nygard et al. 1988, Thangiah et al. 1994). One-sided time windows or deadlines (see Figure 1(c)) offer more flexibility to the dispatcher in that waiting times before customers can be avoided. However, respecting the latest possible time at which service can start, remains a hard constraint. Chuin and Ming (1998) generalize soft time windows by putting a bound on the maximum waiting time and lateness. In the resulting flexible time window (see Figure 1(d)), no penalties are incurred in the original (hard) time window. Arriving too early or too late, but within the respective bounds, is penalized[2]. In this paper, we focus on scheduling flexibility for hard time windows.



(a) HARD TW

(b) SOFT TW

(c) DEADLINE

(d) FLEXIBLE TW

Figure 1. Types of time windows.

To our knowledge, Dullaert (1999), Doerner et al. (2000) and Dullaert (2001) are one of the few to draw attention to the effect of the size of time windows on routing costs. Doerner et al. (2000) notice for full-truckload routing that cost savings through larger time windows are larger if the original time windows are rather tight. They also raise the question on how much a customer should be charged depending on his time window

[2] Notice that the soft time windows in Balakrishnan (1993) can be considered as flexible time windows as they allow the start of service before the earliest deadline at the cost of a penalty. In Ibaraki et al. (2002), one or more flexible time windows can be assigned to each customer.

preferences. Independent from Doerner et al. (2000), Dullaert (1999) raise the same question and develop a framework to study the relationship between scheduling flexibility and freight rates for less-than-truckload routing (Dullaert 2001 and 2002).

3. Less-than-truckload Routing

The problem of LTL routing is extensively studied in the Vehicle Routing Problem with Time Windows literature (see e.g. Desrosiers et al. 1995). In the VRPTW capacitated vehicles, located at a depot, are required to service geographically scattered customers over a limited scheduling period (e.g. a day). Each customer i has a known demand q_i to be serviced (either for pick-up or delivery but not both) at time b_i chosen by the carrier. If time windows are hard, b_i is chosen within a time window, starting at the earliest time e_i and ending at the latest time l_i that customer i permits the start of service. In the soft time window case, a vehicle is allowed to arrive too late at a customer but a penalty is incurred. In both cases, a vehicle arriving too early at customer j, has to wait until e_j. If t_{ij} represents the direct travel time from customer i to customer j, and s_i the service time at customer i then the moment at which service begins at customer j, b_j, equals $\max\{e_j, b_i + s_i + t_{ij}\}$ and the waiting time w_j is equal to $\max\{0, e_j - (b_i + s_i + t_{ij})\}$. A time window can also be defined for the depot in order to define a `scheduling horizon' in which each route must start and end (Potvin and Rousseau 1993).

Definition 1. *LTL Scheduling Flexibility*

Given a customer i with service time s_i and a hard service time window $[e_i, l_i]$, bounded by the earliest and latest time at which service can start. The scheduling flexibility of customer i is defined as $(l_i - e_i)$.

Scheduling flexible customers enable a dispatcher to design cost-efficient routes in two ways. First, the width of their service time windows allows a dispatcher to schedule the customers efficiently in a partially finished route. Rigid customers are more difficult to schedule and often lead to schedules with considerable waiting time and distance to be traveled. Second, a flexible customer in a partially finished route facilitates inserting unrouted customers in the same route. The latter effect is demonstrated by the use of insertion distances and insertion zones.

In the VRPTW literature, a route is traditionally represented as a sequence of nodes $\{i_0, i_1, i_2, \ldots, i_m\}$ with $i_0 = i_m =$ depot. A route can also be represented as a sequence of arcs $\{(i_0,i_1),(i_1,i_2),\ldots,(i_{m-1},i_m)\}$. On each arc (i,j) the time a vehicle has to service i, s_i, and travel directly to j, is larger or equal to t_{ij}. If the customers' time windows do not allow this, the route is infeasible.

Definition 2. *LTL Insertion Distance*

Given a customer i with service time s_i and a hard service time window $[e_i, l_i]$, bounded by the earliest and latest time at which service can begin. If one unit of time equals one unit of distance[3], an insertion distance of $\{l_j - [b_i + s_i] - s_u\}$ can be defined to insert a customer u between customers i and j.

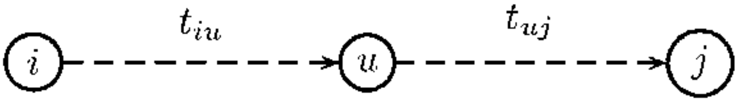


Figure 2. Inserting u between i and j.

Consider Figure 2 in which an unrouted customer u is inserted between nodes i and j. The route is feasible up to j if

$$(1) \qquad \begin{cases} b_i + s_i + t_{iu} & \leq l_u \\ b_i + s_i + t_{iu} + s_u + t_{uj} & \leq l_j \end{cases}$$

By assuming $s_i = s_u = s$ and rewriting the last inequality, the insertion distance becomes

$$(2) \qquad t_{iu} + t_{uj} \leq l_j - b_i - 2s \,.$$

[3] Without loss of generality we make this common assumption (e.g. see Solomon (1987)) to simplify the analysis.

Proposition 1. *LTL Insertion Zone*

If all customers have the same service time $s = s_i = s_u$, the insertion distance $l_j - b_i - 2s$ defines an elliptic insertion zone having customers i and j as its foci. Any unrouted customer located in the elliptic insertion zone, whose time windows are compatible with those of i and j, can be inserted between i and j if the vehicle's capacity permits and if the route remains feasible after j.

Proof 1.

Let the insertion distance $d = l_j - b_i - 2s$. The d units of time can be used to service an unrouted customer u between i and j. At each point at the boundary of the insertion area

(3) $$t_{iu} + t_{uj} = d$$

The maximum area that can be covered in d units of time is bounded by an ellipse having i and j as its foci. By the definition of the distance insertion, the route remains feasible at least up to j. Introducing a new customer between i and j can create a push forward on the begin of service of all subsequent nodes in the route. Time feasibility at the successors of j can be checked by Solomon's (1987) necessary and sufficient conditions for time window feasibility. ■

To check the time feasibility of the schedule after inserting an unrouted customer, Solomon (1987) develops necessary and sufficient conditions for time feasibility if time windows are hard.

If we denote by $b_{i_p}^{new}$ the new time at which service begins at customer i at position p after the insertion of customer u in the partially constructed route $\{i_0, i_1, \ldots, i_m\}$ and if the triangle inequality holds for both distances and travel times, then the *push forward* in the schedule at customer i_p is defined as:

(4) $$PF_{i_p} = b_{i_p}^{new} - b_{i_p} \geq 0$$

and

(5) $$PF_{i_{r+1}} = \max\{0, PF_{i_r} - w_{i_{r+1}}\}, p \le r \le m-1.$$

Solomon (1987) assumes that all vehicles leave the depot at e_0 to use the idea of the maximum possible push forward generated by inserting an unrouted customer u between two adjacent stops i_{p-1} and i_p. The necessary and sufficient conditions for time feasibility when inserting a customer u between i_{p-1} and $i_p, 1 \le p \le m$, on a partially constructed feasible route $\{i_0, i_1, \ldots, i_m\}, i_0 = i_m =$ depot, are

(6) $$b_u \le l_u \text{ and } b_{i_r} + PF_{i_r} \le l_{i_r}, p \le r \le m$$

Indeed, if $PF_{i_p} > 0$, the schedule at customer i and some of its successors, i.e. customers $i_r, p \le r \le m$ may become infeasible. These customers have to be examined one by one for time feasibility until we find a customer i_r whose waiting time and the one of its predecessors before the insertion of u, has nullified the push forward, i.e. $PF_{i_r} = 0$, or which is serviced after l_{i_r}, making the schedule infeasible.

Example 1.

Consider arc $(2,3)$ in Figure 3. The actual travel time from i to j, $t_{ij} = 7.21$. Suppose that the time available to travel between i and j, $l_j - b_i - 2s = 10$. The zone in which customers can be serviced if their time window permits is elliptic, having i and j as its foci. The general equation of an ellipse is given by

(7) $$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$$

By drawing the X-axis through the foci and by making the Y-axis perpendicular to the X-axis at the middle of the two foci, the distance between the two foci, $2c = 7.21$. The length of the major axis of the ellipse equals the insertion distance, $2a = l_j - b_i - 2s = 10$. Since

$b=\sqrt{a^2-c^2}, b=\sqrt{5^2-(3.605)^2}\approx 3.47$, all the necessary information to draw the ellipse is obtained.

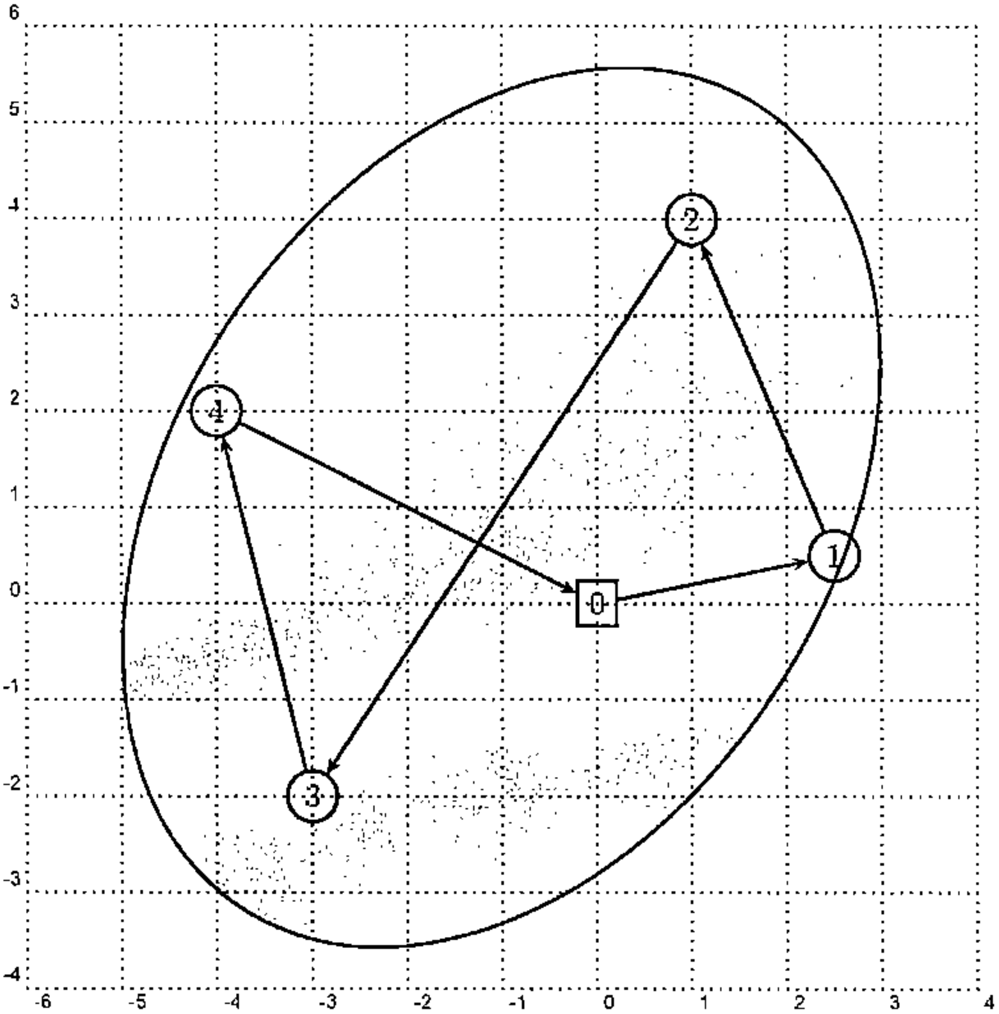


Figure 3. The insertion zone of arc $(2,3)$.

4. Full-truckload routing

In full-truckload routing, a route consists of loads (arcs) instead of individual nodes. The scheduling flexibility of an individual arc is determined by the time windows of its starting and ending node.

Definition 3. *FTL Scheduling Flexibility*

Given a loaded arc (i,j) whose nodes have a service time s and hard service time windows $[e_i,l_i]$ and $[e_j,l_j]$. The scheduling flexibility of the loaded arc (i,j) is defined as $\min\{l_i,l_j-s_i-t_{ij}\}-\max\{e_i,e_j-s_i-t_{ij}\}$.

Denote the start of service at node i as b_i. Because time windows are hard, b_i must fall within the node's service time window $[e_i,l_i]$:

(9) $e_i \leq b_i \leq l_i$

Because allowing for waiting time at node j in determining the scheduling flexibility of load (i,j) overestimates the freedom a dispatcher has in choosing b_i, and because the service time window at j is hard

(10) $e_j \leq b_i + s_i + t_{ij} \leq l_j$

Combining (9) and (10) yields

(11) $\max\{e_i,e_j-s_i-t_{ij}\} \leq b_i \leq \min\{l_i,l_j-s_i-t_{ij}\}$

and equals the scheduling flexibility of load (i,j) to

(12) $\min\{l_i,l_j-s_i-t_{ij}\}-\max\{e_i,e_j-s_i-t_{ij}\}$

Unloaded arcs provide an opportunity to insert loaded arcs. The area in which loaded arcs can be inserted depends on the length of the arc to be inserted.

Definition 4. *FTL Insertion Distance*

Given two adjacent arcs (i,j) and (p,q) in a partially finished route. If each node has a service time s, and time windows $[e_w,l_w], w=i,j,p,q$, then the insertion distance is equal to $l_p-b_j-t_{pq}-3s$.

When inserting a load (m,n) between the two adjacent arcs (i,j) and (p,q), the new route is feasible up to p if (see Figure 4).

$$(13) \qquad \begin{cases} b_j + s + t_{jm} \leq l_m \\ b_j + s + t_{jm} + s + t_{mn} \leq l_n \\ b_j + s + t_{jm} + s + t_{mn} + s + t_{np} \leq l_p \end{cases}$$

If all three inequalities are satisfied, the insertion distance can be determined by rewriting the last inequality as

$$(14) \qquad t_{jm} + t_{mn} + t_{np} \leq l_p - b_j - 3s$$

and the time feasibility of the schedule after (p,q) can be checked by applying Solomon's (1987) necessary and sufficient conditions for time feasibility on the nodes of the arcs.

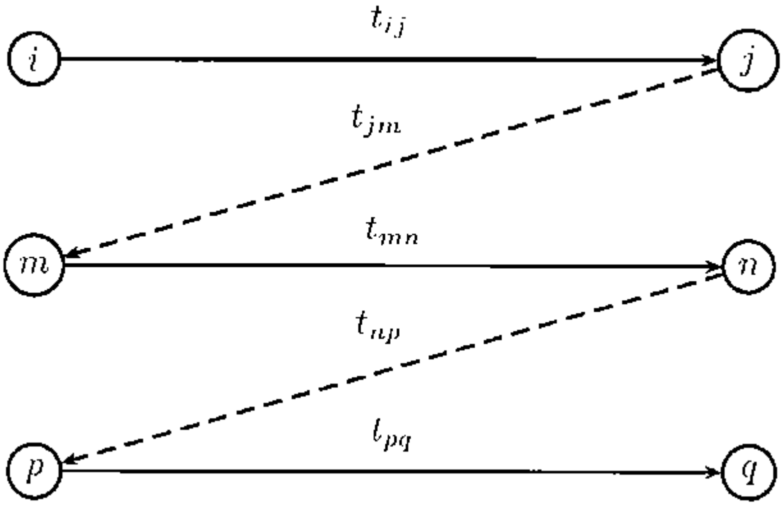


Figure 4. Insertion of (m,n) between (i,j) and (p,q).

Proposition 2. *FTL Insertion Zone*

Given two route segments (i,j) and (k,l) with each node having a service time s, and time windows $[e_w, l_w]$, $w = i,j,k,l$. The insertion distance $l_p - b_j - 3s$ defines an elliptic insertion zone, having nodes j and k as its foci. Any unrouted load which is located in the ellipse, can be inserted between j and k if the route remains feasible after j.

Proof 2.

Consider in Figure 5 a load (p,q) that falls within the insertion zone between (i,j) and (k,l). If we denote the insertion distance $l_k - b_j - 3s$ by d, then

$$(15) \qquad t_{jp} + t_{pq} + t_{qk} \leq d$$

The distance from p to k, t_{pk}, is according to the triangle inequality smaller than or equal to $t_{pq} + t_{qk}$. But then $t_{jp} + t_{pk} \leq d$ and p is contained in an ellipse with foci j and k. Along the same lines, the distance from j to q, t_{jq} is smaller than $t_{jp} + t_{pq}$ and therefore $t_{jq} + t_{qk} \leq d$. As a result also q lies within the ellipse with foci j and k. ■

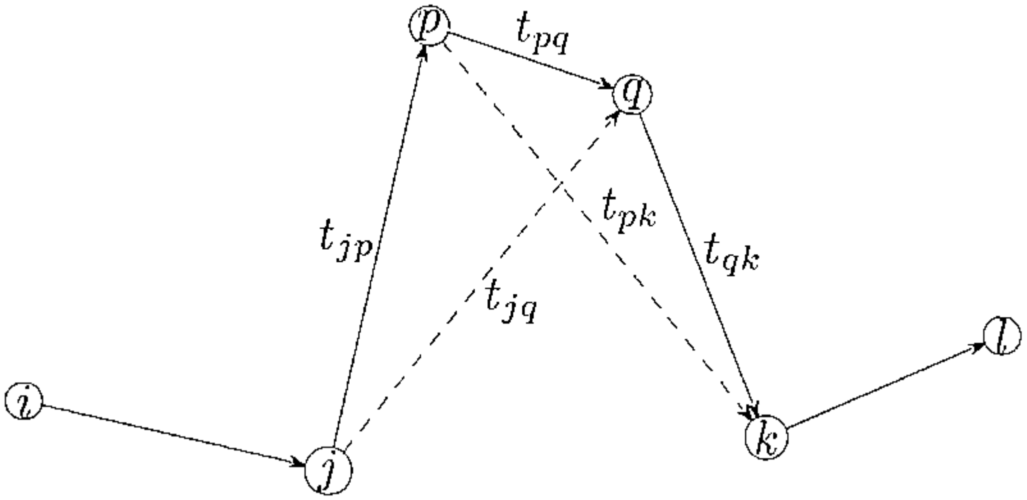


Figure 5. Inserting (p,q) between (i,j) and (k,l).

If nodes are inserted instead of loads, $t_{pq} = 0$ and only a single service time is inserted between j and k. As a result, the FTL insertion distance reduces to the LTL insertion distance, making full-truckload routing a special case of less-than-truckload routing.

5. Conclusions

Although some authors have already briefly addressed the impact of the size of time windows on solution quality, this paper contains the first formal analysis of time windows as a measure of scheduling flexibility in routing. Flexible customers are easier to schedule efficiently, but also facilitate inserting unrouted customers in a route. Insertion distances and insertion zones are used to demonstrate this effect. The insertion distance is defined as the distance that can be traveled between two nodes (two adjacent customers or the beginning or ending node of two adjacent loads). If arcs of length 0 (i.e. nodes) are inserted, the FTL insertion distance reduces to the LTL insertion distance. This makes the FTL routing problem a special case of LTL routing. The insertion distance can be used to determine the elliptic zone from which customers (or loads) can be inserted in the route. Insertion zones can be used for filtering customers eligible for insertion and can at the same time reduce the number of vertices in the partially finished route eligible for insertion.

References

Aluminium (1994). Aluminium coils flexible transportieren und lagern. *Aluminium* 70:5, 277–279.

Balakrishnan, A. (1993). Simple heuristics for the vehicle routing problem with soft time windows. *Journal of the Operational Research Society* 44:3, 279–287.

Basile, L., S. Brondolo & S. Lioy (1998). MPLM: Flexibility in pressurized cargo transportation. *Acta Astronauticai* 42:9, 565–574.

Beckman, S. (1990). Manufacturing flexibility: The next source of competitive advantage. In *Strategic Manufacturing*, 107–132. Ed. P. Moody. Irwin: Dow Jones.

Burns, A. & G. Fohler (1991). *Incorporating Flexibility into Offline Scheduling for Hard Real-time Systems*. Technical Report 3/1991, Institut für Technische Informatik, Technische Universität Wien.

Burns, A., D. Prasad, A. Bondavilli, F. Di Giandomenico, K. Ramamritham, J. Stankovic & L. Stringini (2000). The meaning and role of value in scheduling real-time systems. *Journal of Systems Architecture* 46, 305–325.

Chuin, L.H. & D.S. Ming (1998). An efficient technique for routing of vehicles reactively. In *Proceedings of the 8th ITSA Annual Meeting & Exposition*.

Clancy, S. (2000). Flexible friends and easy. *Transport Engineer* 2000 (June), 26–27.

DeFayette, J. (1996). Flexible bulk palletising – tools for creating the right package. *Handling and Packing* 73:9, 393–398.

Daniels, R.L., S.Y. Hua & S. Webster (1999). Heuristics for parallel-machine flexible resource scheduling problems with unspecified job assignment. *Computers & Operations Research* 26, 143–155.

Desrosiers, J., J. Dumas, M.M. Solomon & F. Soumis (1995). Time constrained routing and scheduling. In *Handbooks in Operations Research and Management Science 8: Network Routing*, 35–139. Eds M. Ball, T. Magnanti, C. Momna & G. Nemhauser. Amsterdam: Elsevier.

Doerner, K., M. Gronalt, R.F. Hartl & M. Reimann (2000). *Time Constrained Full Truckload Routing: Optimizing Fleet Size and Vehicle Movements*. Technical report 2000/002, Department of Production and Operations Management, Institute of Management Science, University of Vienna.

Dullaert, W. (1999). Towards a profit analysis of delivery time flexibility in road freight transport. In *Bijdragen Vervoerslogistieke Werkdagen* 1999, 271–283. Eds R. Rodenburg & A. Kruse. Delft: Connekt.

Dullaert, W. (2001). *Scheduling Flexibility and the Contribution Maximizing Vehicle Routing Problem with Time Windows*. Presented at the 9th World Conference on Transport Research, July 2001, Seoul, South Korea.

Dullaert, W. (2002). *Pricing and Scheduling Flexibility in Road Freight Transport*. Ph.D. Dissertation, University of Antwerp, Belgium.

Dyer, S. (1998). Flexibility models: A critical analysis. *Internation Journal of Manpower* 19:4, 223–233.

Evans, T. & P. Smith (1999). Flexible delivery in Australia: Origins and conceptualizations. *Federation for Information and Documentation Review* 1:2/3, 116–120.

Hyun, J.-H. & B.-H. Ahn (1992). A unifying framework for manufacturing flexibility. *Manufacturing Review* 5:4, 251–260.

Ibaraki T., S. Imahori, M. Kubo, T. Masuda, T. Uno & M. Yagiura (2002). Effective local search algorithms for routing and scheduling problems with general time window constraints. *Transportation Science*, in press.

Martínez, S.A., P.M. Pérez & U.O. Pérez (2000). Flexibilidad organizativa y relación entre JIT y calidad total. *Alta Dirreción* 36:210, 74–84.

Nygard, K.E., P. Greenberg, W.E. Bolkan & E. Swenson (1988). Generalized assignment models for the deadline vehicle routing problem. In: *Vehicle Routing: Methods and Studies*, 107–125. Eds B. Golden & A. Assad. Amsterdam: Elsevier Science Publishers.

Potvin, J.-Y. & J.-M. Rousseau (1993). A parallel routing building algorithm for the vehicle routing and scheduling problem with time windows. *European Journal of Operational Research* 66:3, 331–340.

Sethi, A. & S. Sethi (1990). Flexibility in manufacturing: A survey. *International Journal of Flexible Manufacturing Systems* 2:4, 289–328.

Smith, P.J. (2000). Flexible delivery and apprentice training: Preferences, problems and challenges. *Journal of Vocational Education and Training* 52:3, 483–503.

Solomon, M.M. (1987). Algorithms for the vehicle routing and scheduling problem with time window constraints. *Operations Research* 35:2, 254–265.

Taillard, E., P. Badeau, M. Gendreau, F. Guertin & J.-Y. Potvin (1997). A tabu search heuristic for the vehicle routing problem with soft time windows. *Transportation Science* 31:2, 170–186.

Tarifa, E. & O. Chiotti (1995). Flexibility vs. costs in multiproduct batch plant design: A calculation algorithm. *Chemical Engineering Research and Design* 73:8, 931–940.

Thangiah, S., I.H. Osman, R. Vinayagaoorthy & T. Sun (1994). Algorithms for vehicle routing problems with time deadlines. *American Journal of Mathematical and Management Science* 13:3/4, 323–355.

Upton, D.M. (1995). Flexibility as process mobility: The management of plant capabilities for quick response manufacturing. *Journal of Operations Management* 12:3, 205–225.

Venter, C. & M. Hansen (1998). Flexibility and time dependence in activity scheduling models. *Transportation Research Record* 1645, 120–126.

ACTA WASAENSIA

On the Friedrichs and the Kreĭn-von Neumann extension of nonnegative relations

Seppo Hassi

Dedicated to Ilkka Virtanen on the occasion of his 60th birthday

Abstract

Hassi, Seppo (2004). On the Friedrichs and the Kreĭn-von Neumann extension of nonnegative relations. In *Contributions to Management Science, Mathematics and Modelling. Essays in Honour of Professor Ilkka Virtanen.* Acta Wasaensia No. 122, 37–54. Eds Matti Laaksonen and Seppo Pynnönen.

Some characteristic properties for the Friedrichs and the Kreĭn-von Neumann extension of a nonnegative linear relation in a Hilbert space are established, including criteria which describe the range and the domain of these extensions as well as of their square roots. By systematically allowing linear relations in the study makes it possible to translate results from the domain to the range and vice versa by inversion of the relation. The results are applied to derive an analog for Kreĭn's uniqueness criterion concerning the nonnegative selfadjoint extensions in the case of a nonnegative linear relation.

Seppo Hassi, Department of Mathematics and Statistics, University of Vaasa, P.O. Box 700, FIN-65101 Vaasa, Finland, E-mail: sha@uwasa.fi

Keywords: Linear relation, nonnegative relation, selfadjoint extension, Friedrichs extension, Kreĭn-von Neumann extension, rigged Hilbert space.
Mathematics Subject Classification (2000): 47B25; 47B65; 47A70; 47B15.

1. Introduction

In this paper the basic notion is a nonnegative linear relation A in a Hilbert space $\mathfrak{H}$. This means that A is a linear subspace of the space $\mathfrak{H} \oplus \mathfrak{H}$ and satisfies $(f', f) \geq 0$ for all $\{f, f'\} \in A$. Equivalently, A can be interpreted as a multivalued linear mapping from $\mathfrak{H}$ into itself, whose graph belongs to $\mathfrak{H} \oplus \mathfrak{H}$. In a natural way nonnegative linear relations extend the classes of nonnegative bounded and unbounded densely or nondensely defined linear operators acting on $\mathfrak{H}$. The main interest will be in the selfadjoint extensions $\widetilde{A}$ of A inside $\mathfrak{H}$. This means that $A \subset \widetilde{A}$ in $\mathfrak{H} \oplus \mathfrak{H}$ and that $\widetilde{A} = \widetilde{A}^*$, where $\widetilde{A}^*$ stands for the adjoint linear relation of $\widetilde{A}$ in $\mathfrak{H}$.

The study of selfadjoint extensions of densely defined symmetric operators goes back to J. von Neumann, cf. Akhiezer and Glazman (1993). In the famous work of Kreĭn (1947) a complete description of all nonnegative selfadjoint extensions of a densely defined nonnegative operator A was established. In the theory originating from Kreĭn (1947) two nonnegative extensions of A play an important role, the Friedrichs extension A_F of A introduced in Friedrichs (1934) and the Kreĭn-von Neumann extension of A_K of A, cf. Kreĭn (1947), von Neumann (1929). A more recent exposition of this theory based on Kreĭn (1947), Birman (1956), and Vishik (1952) can be found in Alonso and Simon (1980). For a modern approach to the extension theory built on abstract boundary conditions the reader is referred to Derkach and Malamud (1991, 1995) and Gorbachuk and Gorbachuk (1991).

The extension theory of nonnegative operators in the case where the operator A is allowed to be nondensely defined was studied later, cf. e.g. Strauss (1970), Derkach and Malamud (1995) and the references therein. In particular, Ando and Nishio (1970) introduced the notion of positively closable operators, extending the class of densely defined nonnegative operators, and showed that such operators admit a nonnegative selfadjoint operator extension, namely the Kreĭn-von Neumann extension A_K of A. An extension of the results for the case where A is a nonnegative linear relation was obtained by Coddington and de Snoo (1978). In particular, the description of all nonnegative selfadjoint (relation) extensions analogous to Kreĭn (1947) was established in Coddington and de Snoo (1978).

In this paper a couple of results which are known about the nonnegative selfadjoint extensions of A when A is a nonnegative operator typically satisfying some additional properties, like A being densely defined or positively closable, will be generalized by removing these additional assumptions on A, and, more generally, by systematically allowing A to be a nonnegative linear relation in $\mathfrak{H}$. In particular, some characteristic properties known for the Friedrichs and the Kreĭn-von Neumann extension in the case of densely defined nonnegative and positively closable operators are established for the case where A is a nonnegative linear relation.

The paper is organized as follows. In Section 2 the basic notions on linear relations T from the Hilbert space $\mathfrak{H}$ into the Hilbert space $\mathfrak{K}$ are given and a characterization of the domain $\operatorname{dom} T^*$ of the adjoint linear relation T^* of T is established. Then simply by inverting the relation T a similar characterization for the range $\operatorname{ran} T^*$ of T^* is obtained; a result which generalizes the well-known description of $\operatorname{ran} A$ due to Shmul'yan (1967) in the case of a bounded and everywhere defined operator $T \in \mathcal{B}(\mathfrak{H})$. In the special case where T is a densely defined operator in $\mathfrak{H}$ this characterization of $\operatorname{ran} T^*$ reduces to the one given by Sebestyén (1983). In Section 3 the nonnegative selfadjoint square root of a nonnegative selfadjoint relation is shortly discussed and the main facts concerning the description of nonnegative selfadjoint extensions of a nonnegative linear relation are recalled from Kreĭn (1947); Ando and Nishio (1970), and Coddington and de Snoo (1978). Then a description of the range $\operatorname{ran} A_F^{1/2}$ of the square root of the Friedrichs extensions A_F and the domain $\operatorname{dom} A_K^{1/2}$ of the square root of the Kreĭn-von Neumann extension A_K of A is established. Moreover, some similar descriptions for $\operatorname{dom} A_F$, $\operatorname{ran} A_F$, $\operatorname{dom} A_K$, and $\operatorname{ran} A_K$ are derived. The results generalize the corresponding descriptions in Ando and Nishio (1970); Sebestyén and Stochel (1991); Prokaj and Sebestyén (1996), and Sebestyén and Sikolya (2003), to the case where A is a nonnegative linear relation in $\mathfrak{H}$. In Section 4 the Friedrichs extension A_F and the Kreĭn-von Neumann extension A_K of A are characterized among the class of all selfadjoint extensions of A in $\mathfrak{H}$. For this purpose a class of rigged Hilbert spaces associated with a selfadjoint relation H in $\mathfrak{H}$ is introduced. These characterizations together with the descriptions of $\operatorname{ran} A_F^{1/2}$ and $\operatorname{dom} A_K^{1/2}$ in Section 3 are then applied to prove an analog of Kreĭn's uniqueness criterion concerning the equality $A_F = A_K$ of the extreme extensions A_F and A_K of A again in the case of a nonnegative linear relation A in $\mathfrak{H}$.

2. Linear relations and their adjoint

2.1. Some basic facts. Let T be a linear relation (multivalued mapping) from $\mathfrak{H}$ into $\mathfrak{K}$. This means that (the graph of) T is a linear subspace of $\mathfrak{H} \oplus \mathfrak{K}$. A linear relation T is closed if it is a closed subspace of $\mathfrak{H} \oplus \mathfrak{K}$. Analogous with the case of linear operators the following basic notations will be used:

$$\operatorname{dom} T = \{\, f \in \mathfrak{H} : \{f,g\} \in T \,\}, \quad \ker T = \{\, f \in \mathfrak{H} : \{f,0\} \in T \,\},$$
$$\operatorname{ran} T = \{\, g \in \mathfrak{K} : \{f,g\} \in T \,\}, \quad \operatorname{mul} T = \{\, g \in \mathfrak{K} : \{0,g\} \in T \,\}.$$

The inverse of T is defined by $T^{-1} = \{\, \{g,f\} : \{f,g\} \in T \,\}$ and thus T^{-1} is a linear relation from $\mathfrak{K}$ into $\mathfrak{H}$. A linear relation T is (the graph of) of a single-valued linear mapping from $\mathfrak{H}$ into $\mathfrak{K}$ precisely when $\operatorname{mul} T = \{0\}$. The adjoint T^* of T is defined by

$$(1) \qquad T^* = \{ \{h,k\} : (k,f)_{\mathfrak{H}} - (h,g)_{\mathfrak{K}} = 0 \text{ for all } \{f,g\} \in T \}.$$

It is automatically a closed linear relation from $\mathfrak{K}$ into $\mathfrak{H}$ and it satisfies the usual identities familiar in the case of linear operators. For instance, $\ker T^* = (\operatorname{ran} T)^{\perp}$, $\operatorname{mul} T^* = (\operatorname{dom} T)^{\perp}$, and $(T^*)^{-1} = (T^{-1})^* =: T^{-*}$. For a closed linear relation T define $T_{\infty} = \{ \{0,g\} \in T\}$ and $T_s = T \ominus T_{\infty}$, where $\ominus$ refers to the orthogonality in $\mathfrak{H} \oplus \mathfrak{K}$. This decomposes T orthogonally in $\mathfrak{H} \oplus \mathfrak{K}$ as follows

$$(2) \qquad T = T_s \oplus T_{\infty}.$$

Here T_s is a closed linear operator with $\operatorname{dom} T_s = \operatorname{dom} T$ and $\operatorname{ran} T_s \subset (\operatorname{mul} T)^{\perp} = \overline{\operatorname{dom}}\, T^*$, while T_{∞} is a closed linear relation with $\operatorname{dom} T_{\infty} = \{0\}$ and $\operatorname{ran} T_{\infty} = \operatorname{mul} T$. If in particular $\mathfrak{H} = \mathfrak{K}$ and T is for instance a closed symmetric relation in $\mathfrak{H}$, i.e. $T \subset T^*$, then $\overline{\operatorname{dom}}\, T \subset \overline{\operatorname{dom}}\, T^* = (\operatorname{mul} T)^{\perp}$. The last inclusion guarantees that the decomposition of $T = T_s \oplus T_{\infty}$ in (2) becomes orthogonal also in $\mathfrak{H} = \overline{\operatorname{dom}}\, T^* \oplus \operatorname{mul} T$. Moreover, T_s is a closed symmetric operator in $\overline{\operatorname{dom}}\, T^*$. By definition, a linear relation T in $\mathfrak{H}$ is selfadjoint if $T = T^*$. Hence, a selfadjoint relation is always closed, linear, and symmetric.

2.2. A characterization of the domain and the range of the adjoint. The next result is well known for densely defined linear operators, being closely related to the definition of the adjoint operator itself. Here it is extended for linear relations T from $\mathfrak{H}$ into $\mathfrak{K}$.

Lemma 1. *Let T be a linear relation from $\mathfrak{H}$ into $\mathfrak{K}$. Then $g \in \operatorname{dom} T^*$ if and only if there exists $C_g < \infty$, such that*

$$(3) \qquad |(f',g)_{\mathfrak{K}}| \le C_g \|f\|_{\mathfrak{H}}, \quad \text{for all } \{f,f'\} \in T.$$

In this case the smallest C_g satisfying (3) is $C_g = \|g'\|_{\mathfrak{H}}$ with $\{g,g'\} \in T^*$ and $g' \in \overline{\operatorname{dom}}\, T$, i.e., $C_g = \|(T^*)_s g\|_{\mathfrak{H}}$.

Proof. First assume that $g \in \operatorname{dom} T^*$. Then $\{g,g'\} \in T^*$ for some $g' \in \mathfrak{H}$ and by the definition of the adjoint T^* in (1) one obtains for every $\{f,f'\} \in T$:

$$|(f',g)_{\mathfrak{K}}| = |(f,g')_{\mathfrak{H}}| \le \|f\|_{\mathfrak{H}} \|g'\|_{\mathfrak{H}},$$

so that one can take $C_g = \|g'\|_{\mathfrak{H}}$ in (3).

Conversely, assume that $g \in \mathfrak{K}$ satisfies the estimate (3). Define the linear relation d_g in $\mathfrak{H} \oplus \mathbb{C}$ by

$$d_g := \{\, \{f, (f', g)_{\mathfrak{K}}\} : \{f, f'\} \in T \,\}.$$

Then it follows from (3) that d_g is single-valued, since $\|f\|_{\mathfrak{H}} = 0$ implies $(f', g)_{\mathfrak{K}} = 0$. Hence, d_g is (the graph of) of a single-valued bounded linear functional defined on $\operatorname{dom} T$. Therefore, it has a continuation $\bar{d}_g$ from $\overline{\operatorname{dom} T}$ into $\mathbb{C}$ with the same norm ($\|\bar{d}_g\| \le C_g$). By the Riesz Representation Theorem there exists $g' \in \overline{\operatorname{dom} T}$ with $\|g'\|_{\mathfrak{H}} = \|\bar{d}_g\|$, such that

$$\bar{d}_g f = (f, g')_{\mathfrak{H}}, \quad \text{for all } f \in \overline{\operatorname{dom} T}.$$

Thus, $(f', g)_{\mathfrak{K}} = (f, g')_{\mathfrak{H}}$ holds for every $\{f, f'\} \in T$, and hence in view of (1) $\{g, g'\} \in T^*$. This shows that $g \in \operatorname{dom} T^*$.

The last statement is clear from the given arguments and the decomposition (2) of $T = T_s \oplus T_\infty$. □

One advantage of the characterization of $\operatorname{dom} T^*$ for linear relations is that, by inverting the relation T, one immediately obtains an analogous characterization for $\operatorname{ran} T^*$.

Corollary 2. *Let T be a linear relation from $\mathfrak{H}$ into $\mathfrak{K}$. Then $g' \in \operatorname{ran} T^*$ if and only if there exists $C_{g'} < \infty$, such that*

$$(4) \qquad |(f, g')_{\mathfrak{H}}| \le C_{g'} \|f'\|_{\mathfrak{K}}, \quad \textit{for all } \{f, f'\} \in T.$$

In this case the smallest $C_{g'}$ satisfying (4) is $C_{g'} = \|g\|_{\mathfrak{K}}$ with $\{g, g'\} \in T^*$ and $g \in \overline{\operatorname{ran} T}$, i.e., $C_{g'} = \|(T^{-*})_s g'\|_{\mathfrak{K}}$.

Proof. The result is obtained by applying Lemma 1 to the inverse linear relation T^{-1}. □

Remark 3. Corollary 2 generalizes Shmul'yan's characterization of $\operatorname{ran} T$ for bounded operators $T \in \mathcal{B}(\mathfrak{H})$, see Shmul'yan (1967), cf. also Fillmore and Williams (1971). In the special case where T is a densely defined operator in $\mathfrak{H}$ Corollary 2 gives, as a generalization of Shmul'yan's result, a characterization for $\operatorname{ran} T^*$ of a densely defined operator T in $\mathfrak{H}$, a result given in Sebestyén (1983: Theorem 1).

3. Nonnegative relations and their nonnegative selfadjoint extensions

In this section certain result in the extension theory of nonnegative operators will be generalized to the case of nonnegative relations; for the basic results in these lines see Coddington and de Snoo (1978). For this purpose the first basic notion that will be needed is the square root of a nonnegative selfadjoint relation. This will be defined here along the lines familiar from the case of (densely defined) nonnegative selfadjoint operators.

3.1. Square root of a nonnegative selfadjoint relation. Let B be a nonnegative selfadjoint relation in the Hilbert space $\mathfrak{H}$: $B = B^* \geq 0$, so that $(g, f) \geq 0$ holds for all $\{f, g\} \in B$. As in the case of nonnegative selfadjoint operators, the nonnegative selfadjoint square root C of B should satisfy the relations:

$$C^2 = CC = B, \quad C = C^* \geq 0.$$

It will be shown that these conditions determine the linear relation C uniquely. Moreover,

$$\ker C = \ker B, \quad \operatorname{mul} C = \operatorname{mul} B, \quad C_s = (B_s)^{1/2}. \tag{5}$$

Indeed, clearly $\ker C \subset \ker B$. Conversely, let $f \in \ker B$. Then $C^2 = B$ implies that

$$\{f, g\}, \{g, 0\} \in C \text{ for some } g \in \mathfrak{H}.$$

Now due to $C = C^*$ one obtains by using (1) the identities

$$0 = (g, g) - (f, 0) = (g, g).$$

This gives $g = 0$ and therefore $\{f, 0\} \in C$, which proves the reverse inclusion $\ker B \subset \ker C$. A similar argument shows that $\operatorname{mul} C = \operatorname{mul} B$. (One may also use $C^{-1}C^{-1} = B^{-1}$ and $\operatorname{mul} C = \ker C^{-1} = \ker B^{-1} = \operatorname{mul} B$.) Now the orthogonal decomposition $C = C_s \oplus C_\infty$ implies that $B = C^2 = C_s^2 \oplus C_\infty = B_s \oplus B_\infty$. In particular, the orthogonal operators parts in $\overline{\operatorname{dom}}\, C = \overline{\operatorname{dom}}\, B$ satisfy $C_s^2 = B_s$. The condition $C = C^* \geq 0$ implies that $C_s = C_s^* \geq 0$. Hence, by the result concerning nonnegative selfadjoint square root of a nonnegative selfadjoint operator one concludes that $C_s = (B_s)^{1/2}$. This proves the uniqueness of C and the identities (5). As in the operator case, the nonnegative square root of B is denoted by $B^{1/2}$:

$$B^{1/2} = B_s^{1/2} \oplus B_\infty = B_s^{1/2} \oplus (\{0\} \oplus \operatorname{mul} B). \tag{6}$$

The inverse of a nonnegative selfadjoint relation B is also a nonnegative selfadjoint linear relation. It is clear from (5) that $(B^{1/2})^{-1} = (B^{-1})^{1/2} =: B^{-1/2}$.

The formula (6) for the square root of a nonnegative selfadjoint relation gives also a possibility to establish the functional calculus for selfadjoint relations $A = A_s \oplus A_\infty$ by means of the functional calculus for the selfadjoint operator part A_s, i.e. $f(A) := f(A_s) \oplus A_\infty$ for measurable f.

3.2. Semibounded relations and forms. A (closed) linear relation A in a Hilbert space $\mathfrak{H}$ is said to be bounded from below with a lower bound $\alpha \in \mathbb{R}$ if $(f', f) \geq \alpha$ holds for all $\{f, f'\} \in A$, or equivalently, if $(A_s f, f) \geq \alpha$ for all $f \in \operatorname{dom} A$. Similarly one can define linear relations which are bounded from above. A semibounded relation is always symmetric. Let A be an operator which is semibounded from below, $A \geq \alpha I$. Then $\operatorname{dom}[A]$ stands for the closure of $\operatorname{dom} A$ with respect to the norm $\|f\|_A^2 = (1-\alpha)\|f\|^2 + (Af, f)$, $f \in \operatorname{dom} A$. The closure of the form (Af, f) with respect to the norm $\|f\|_A^2$, $f \in \operatorname{dom} A$ is denoted by $A[f, f]$. It is well known, cf. Akhiezer and Glazman (1993); Kato (1966), that $\operatorname{dom}[A] = \operatorname{dom}(A-\alpha)^{1/2}$ when A is selfadjoint. In the case of a semibounded linear relation $A \geq \alpha I$ the corresponding form and its domain are defined by $A[f, f] := A_s[f, f]$, $\operatorname{dom}[A] := \operatorname{dom}[A_s]$. A closed linear relation is nonnegative if it has a nonnegative lower bound.

3.3. The Friedrichs and the Kreĭn-von Neumann extension. Let A be a nonnegative relation in $\mathfrak{H}$. The form domain generated by $A \geq 0$ is the completion of $\operatorname{dom} A$ with respect to the inner product $(f, g) + (f', g) = (f, g) + (A_s f, g)$, where $\{f, f'\}, \{g, g'\} \in A$ and where A_s is the operator part of A. The form domain $\operatorname{dom}[A]$ can be described as follows: $f \in \operatorname{dom}[A]$ if and only if there is a sequence $(f_n) \in \mathfrak{H}$, such that

$$f_n \to f, \quad (A_s(f_n - f_m), f_n - f_m) \to 0 \quad (m, n \to \infty). \tag{7}$$

It follows from the First and the Second Representation Theorem (see e.g. Kato (1966)) when extended to nondensely defined closed forms that there is a unique nonnegative selfadjoint relation A_F in $\mathfrak{H}$ such that

$$A[f, g] = ((A_F)_s^{1/2} f, (A_F)_s^{1/2} g), \quad f, g \in \operatorname{dom}[A] = \operatorname{dom} A_F^{1/2}.$$

Clearly, A_F is a selfadjoint extension of A in $\mathfrak{H}$, the Friedrichs extension of A, which in the densely defined case goes back to Friedrichs (1934). The lower bound of A_F is equal to the lower bound of A. Moreover, A_F is the only selfadjoint extension of A whose domain is contained in $\operatorname{dom}[A]$ and the following alternative description for A_F holds:

$$A_F = \{\, \{f, f'\} \in A^* : \, f \in \operatorname{dom}[A] \,\}. \tag{8}$$

Notice that $\operatorname{mul} A_F = \operatorname{mul} A^* = \mathfrak{H} \ominus \overline{\operatorname{dom}} \, A$.

The Kreĭn-von Neumann extension A_K of A, cf. von Neumann (1929); Kreĭn (1947), can be defined via

$$(A^{-1})_F = (A_K)^{-1}, \quad (A^{-1})_K = (A_F)^{-1}. \tag{9}$$

Hence, A_K^{-1} can be constructed also by means of the representation theorems when applied to the closed form $A^{-1}[f, g]$ with $\operatorname{dom}[A^{-1}] =: \operatorname{ran}[A]$ generated by the inverse A^{-1} of A. In particular, A_K is the only selfadjoint extension of A whose range is contained in $\operatorname{ran}[A]$ and the following description holds:

$$A_K = \{\, \{f, f'\} \in A^* : \, f' \in \operatorname{ran}[A] \,\}. \tag{10}$$

Notice also that $\ker A_K = \ker A^* = \mathfrak{H} \ominus \overline{\operatorname{ran}} \, A$.

The main result concerning nonnegative selfadjoint extensions of a nonnegative relation A in $\mathfrak{H}$ can now be stated. In the densely defined case this result goes back to Kreĭn (1947), for nondensely defined A it was proved by Ando and Nishio (1970), and for nonnegative relations by Coddington and de Snoo (1978).

Theorem 4. *(Kreĭn (1947); Ando and Nishio (1970); Coddington and de Snoo (1978)) Let A be a closed nonnegative relation in $\mathfrak{H}$. Then A_F and A_K are nonnegative selfadjoint extensions of A. Moreover, $\widetilde{A}$ is a nonnegative selfadjoint extension of A in $\mathfrak{H}$ if and only if*

$$(A_F + a)^{-1} \leq (\widetilde{A} + a)^{-1} \leq (A_K + a)^{-1}, \quad a > 0. \tag{11}$$

The extremal properties (11) of the Friedrichs and the Kreĭn-von Neumann extension can be reformulated also by means of the corresponding nonnegative forms:

$$A_F[f, f] \geq \widetilde{A}[f, f] \geq A_K[f, f] \tag{12}$$

with $\operatorname{dom} A_F^{1/2} \subset \operatorname{dom} \widetilde{A}^{1/2} \subset \operatorname{dom} A_K^{1/2}$ and the corresponding operator parts are ordered by $(A_F)_s \geq (\widetilde{A})_s \geq (A_K)_s$ on their domains.

It is convenient to describe the extremal extensions A_F and A_K of A in a slightly different manner.

Lemma 5. *Let $\{f, f'\} \in A^*$ and let $\{f_A, f'_A\} \in A$. Then:*

(i) *$\{f, f'\} \in A_F$ if and only if*

$$\inf\{\, \|f - f_A\|^2 + (f' - f'_A, f - f_A) : \{f_A, f'_A\} \in A \,\} = 0; \tag{13}$$

(ii) *$\{f, f'\} \in A_K$ if and only if*

$$\inf\{\, \|f' - f'_A\|^2 + (f' - f'_A, f - f_A) : \{f_A, f'_A\} \in A \,\} = 0. \tag{14}$$

Lemma 5 is just a reformulation for the description of the form domains $\operatorname{dom}[A]$ and $\operatorname{dom}[A^{-1}] = \operatorname{ran}[A]$ used in the construction of A_F and A_K, see (7); cf. also Arlinskiĭ (1988); Arlinskiĭ, Hassi, de Snoo, and Sebestyen (2001); Hassi, Malamud, and de Snoo (2001).

The next result gives descriptions for $\operatorname{ran} A_F^{1/2}$ and $\operatorname{dom} A_K^{1/2}$ of the extreme extensions A_F and A_K in the case of a nonnegative relation A. Observe, that the descriptions of $\operatorname{dom} A_F^{1/2}$ and $\operatorname{ran} A_K^{1/2}$ are contained in their constructions.

Proposition 6. *Let A be a nonnegative linear relation in $\mathfrak{H}$ and let A_F and A_K be the Friedrichs extension and the Kreĭn-von Neumann extension of A, respectively. Then:*

(i)

$$\operatorname{ran} A_F^{1/2} = \left\{\, k \in \mathfrak{H} : |(f,k)|^2 \leq C_k(g,f) \;\; \textit{for all } \{f,g\} \in A \textit{ and some } C_k < \infty \,\right\}.$$

Moreover, for every $k \in \operatorname{ran} A_F^{1/2}$ the smallest C_k is given by $C_k = \|h\|^2$ with $\{h,k\} \in A_F^{1/2}$ and $h \in \overline{\operatorname{ran}}\, A_F$, i.e., $C_k = \|((A_F)^{-1/2})_s k\|^2$.

(ii)

$$\operatorname{dom} A_K^{1/2} = \left\{\, h \in \mathfrak{H} : |(g,h)|^2 \leq C_h(g,f) \;\; \textit{for all } \{f,g\} \in A \textit{ and some } C_h < \infty \,\right\}.$$

Moreover, for every $h \in \operatorname{dom} A_K^{1/2}$ the smallest C_h is given by $C_h = \|k\|^2$ with $\{h,k\} \in A_K^{1/2}$ and $k \in \overline{\operatorname{dom}}\, A_K$, i.e., $C_h = \|(A_K^{1/2})_s h\|^2$.

Proof. (i) Let $\{f,g\} \in A$. Then $\{f,g\} \in A_F$ and $(g,f) = \|(A_F)_s^{1/2} f\|^2$, where $(A_F)_s$ stands for the orthogonal operator part of A_F. Hence, the condition for k can be rewritten in the form

$$|(f,k)|^2 \leq C_k \|(A_F)_s^{1/2} f\|^2 \quad \text{for all } f \in \operatorname{dom} A. \tag{15}$$

In view of (13) in Lemma 5 the estimate (15) holds also for every $f \in \operatorname{dom} A_F^{1/2} = \operatorname{dom} (A_F)_s^{1/2}$:

$$|(f,k)|^2 \leq C_k \|(A_F)_s^{1/2} f\|^2 \leq C_k \|f'\|^2 \quad \text{for all } \{f, f'\} \in A_F^{1/2}. \tag{16}$$

According to Corollary 2 this means that $k \in \operatorname{ran} A_F^{1/2}$. Moreover, Corollary 2 shows that the smallest C_k is given by $C_k = \|h\|^2$ with $\{h, k\} \in A_F^{1/2}$ and $h \in \overline{\operatorname{ran}}\, A_F^{1/2} = \overline{\operatorname{ran}}\, A_F$, or equivalently, $C_k = \|((A_F)^{-1/2})_s k\|^2$.

(ii) This is obtained by applying (i) to the inverse A^{-1} of A and taking into account the relations (9). □

Corollary 7. *The following characterizations hold:*

(i) $\operatorname{dom} A_F = \{\, h \in \operatorname{dom} A_F^{1/2} : \sup \{ |(f', h)| : \{f, f'\} \in A, \|f\| \leq 1 \} < \infty \}$;

(ii) $\operatorname{ran} A_F = \{\, k \in \mathfrak{H} : \{h, k\} \in A^*$ *for some* $h \in \operatorname{dom} A_F^{1/2} \}$;

(iii) $\operatorname{ran} A_K = \{\, k \in \operatorname{ran} A_K^{1/2} : \sup \{ |(f, k)| : \{f, f'\} \in A, \|f'\| \leq 1 \} < \infty \}$;

(iv) $\operatorname{dom} A_K = \{\, h \in \mathfrak{H} : \{h, k\} \in A^*$ *for some* $k \in \operatorname{ran} A_K^{1/2} \}$.

Proof. (i) It follows from (8) that $\operatorname{dom} A_F = \operatorname{dom} [A] \cap \operatorname{dom} A^*$. Since $\operatorname{dom} [A] = \operatorname{dom} A_F^{1/2}$ the description of $\operatorname{dom} A_F$ is now obtained by applying Lemma 1.

(iii) According to (10) $\operatorname{ran} A_K = \operatorname{ran} [A] \cap \operatorname{ran} A^*$ and hence the assertion follows from Corollary 2 and the equality $\operatorname{ran} [A] = \operatorname{ran} A_K^{1/2}$.

The statement (ii) and (iv) are immediate from (8) and (10), respectively. □

Remark 8. The description of $\operatorname{ran} A_K^{1/2}$ for a nonnegative operator A which is positively closable originates from Ando and Nishio (1970); see also Sebestyén and Stochel (1991); Sebestyén and Sikolya (2003). In the case that A is densely defined the description of $\operatorname{ran} A_F^{1/2}$ can be found in Prokaj and Sebestyén (1996), see also Sebestyén and Sikolya (2003). If A is a nonnegative operator the descriptions of $\operatorname{ran} A_F^{1/2}$ and $\operatorname{dom} A_K^{1/2}$ in Proposition 6 can be rewritten by replacing g with Af and the assumption $\{f, g\} \in A$ with $f \in \operatorname{dom} A$. In this case the descriptions of $\operatorname{ran} A_F^{1/2}$ and $\operatorname{dom} A_K^{1/2}$ in Proposition 6 take the same form as the corresponding results in Ando and Nishio

(1970); Prokaj and Sebestyén (1996), without the assumptions that A is positive closable, which guarantees that A_K is an operator, and respectively that A is densely defined, which guarantees that A_F is an operator. Under the assumption that A is positively closable, the descriptions of $\operatorname{ran} A_K$ and $\operatorname{dom} A_K$ similar to those in Corollary 7 have been proved also in the recent paper of Sebestyén and Sikolya (2003: Theorem 2), by using a factorization of A_K which was established in Sebestyén and Stochel (1991).

Finally, observe that $\operatorname{ran} A_F$ and $\operatorname{dom} A_K$ admit the following decompositions:

$$\operatorname{ran} A_F = \operatorname{mul} A^* \oplus \operatorname{ran}(A_F)_s, \quad \operatorname{dom} A_K = \ker A^* \oplus \operatorname{ran} A_N^{(-1)},$$

where $(A_F)_s$ is the orthogonal operator part of A_F and $A_N^{(-1)}$ stands for the (Moore-Penrose) pseudo-inverse of the relation A_N, which is defined by

$$A_N^{(-1)} = \{\, \{g,f\} : \{f,g\} \in A_N, \quad f \perp \ker A_N \,\}.$$

4. The extreme extensions

In this section the Friedrichs extension A_F and the Kreĭn-von Neumann extension A_K of a nonnegative relation A are characterized among the class of all selfadjoint extensions of A in $\mathfrak{H}$. These characterizations together with Proposition 6 are applied to give an analog of Kreĭn's uniqueness criterion for the equality $A_F = A_K$ of the extreme extensions A_F and A_K of A in the case that A is a nonnegative linear relation.

4.1. A class of rigged Hilbert spaces. The description of the Friedrichs extension in (8) is based on the form domain $\operatorname{dom}[A]$ equipped with the inner product $(f,g) + (A_s f, g)$, $f,g \in \operatorname{dom} A$. To give some further analysis of the Friedrichs extension the notion of certain rigged Hilbert spaces, cf. Berezanski (1965), associated with a selfadjoint operator H is shortly recalled in a somewhat explicit form along the lines in Hassi and de Snoo (1998). The space $\mathfrak{H}_{+1}(H)$ is defined as $\operatorname{dom} |A|^{1/2}$ with the corresponding graph norm: $(f,g)_+ = (f,g) + (|H|^{1/2} f, |H|^{1/2} g)$, $f,g \in \operatorname{dom} |H|^{1/2}$. Here $|H| := (H^* H)^{1/2} = (H^2)^{1/2}$ is the modulus of H. The space $\mathfrak{H}_{-1}(H)$ is the completion of $\mathfrak{H}$ with respect to the inner product $(f,g)_- = ((I+|H|)^{-1} f, g)$, $f,g \in \mathfrak{H}$. This results in the chain of Hilbert spaces,

$$\mathfrak{H}_{+1}(H) \subset \mathfrak{H} \subset \mathfrak{H}_{-1}(H), \tag{17}$$

the rigged Hilbert spaces with indices $+1, -1$ associated to H. The space $\mathfrak{H}_{-1}(H)$ can be identified with the dual space of $\mathfrak{H}_{+1}(H)$. The duality between $\mathfrak{H}_{+1}(H)$ and $\mathfrak{H}_{-1}(H)$ can be expressed as

$$(18) \qquad (f,g) = (Vf,g)_- = (f,V^{-1}g)_+, \quad f \in \mathfrak{H}_{+1}(H),\, g \in \mathfrak{H}_{-1}(H).$$

Here V is the so-called Riesz operator, the unique isometric extension of $I + |A|$ from $\mathfrak{H}_{+1}(H)$ onto $\mathfrak{H}_{-1}(H)$. The bilinear form in (18) extends the inner product (f,g) in the original Hilbert space $\mathfrak{H}$, therefore the same symbol is being used. An advantage of the rigged Hilbert spaces (17) is that H as an operator from $\mathfrak{H}_{+1}(H)$ into $\mathfrak{H}_{-1}(H)$ is contractive. Moreover, for every $\lambda \in \rho(H)$ the resolvent operator $(H-\lambda)^{-1}$ is continuous. This means that after continuation H becomes a closed contractive operator from $\mathfrak{H}_{+1}(H)$ into $\mathfrak{H}_{-1}(H)$ which is defined everywhere on $\mathfrak{H}_{+1}(H)$. Similarly, after continuation the resolvent operator $(H-\lambda)^{-1}$ is closed, bounded, and boundedly invertible operator from $\mathfrak{H}_{-1}(H)$ onto $\mathfrak{H}_{+1}(H)$. These continuations satisfy the usual relations and hence they will be denoted still by the same symbols. In particular, one has

$$(19) \quad (Hf,g) = (f,Hg) = (U|H|^{1/2}f, |H|^{1/2}g) = \int_{\mathbb{R}} t\, d(E(t)f,g), \quad f,g \in \mathfrak{H}_{+1}(H),$$

where $H = U|H|$ is the polar decomposition of H (cf. Kato (1966)) and $E(t)$ stands for the spectral family of the selfadjoint operator H. Similarly, the continued resolvents satisfy for instance the identities

$$((H-\lambda)^{-1}f,g) = (f,(H-\bar{\lambda})^{-1}g), \quad f,g \in \mathfrak{H}_{-1}(H), \quad \lambda \in \rho(H),$$

$$(f,g) = ((H-\lambda)f,(H-\bar{\lambda})^{-1}g), \quad f \in \mathfrak{H}_{+1}(H), g \in \mathfrak{H}_{-1}(H), \quad \lambda \in \rho(H).$$

4.2. A characterization of the extreme extensions. Let A be a nonnegative relation in $\mathfrak{H}$. The defect subspaces of A are denoted by $\mathfrak{N}_\lambda = \ker(A^* - \lambda)$, $\lambda \in \mathbb{C}$. Moreover, $\widehat{\mathfrak{N}}_\lambda$ stands for

$$\widehat{\mathfrak{N}}_\lambda = \{\, \{f, \lambda f\} : f \in \ker(A^* - \lambda)\,\}, \quad \lambda \in \mathbb{C}.$$

The analog of von Neumann's formula for A^* in the case of linear relations reads as

$$A^* = A \dotplus \widehat{\mathfrak{N}}_\lambda \dotplus \widehat{\mathfrak{N}}_{\bar{\lambda}}, \quad \lambda \in \mathbb{C} \setminus \mathbb{R},$$

where $\dotplus$ denotes the componentwise sum in $\mathfrak{H} \oplus \mathfrak{H}$. The next result gives a characterization of the Friedrichs extension A_F of A by means of the spaces $\mathfrak{H}_{+1}(\widetilde{A})$, i.e., $\operatorname{dom}|\widetilde{A}|^{1/2}$ equipped with the graph norm of $|\widetilde{A}_s|^{1/2}$, which are associated with the selfadjoint extensions $\widetilde{A}$ of A.

Theorem 9. *Let A be a closed nonnegative relation in $\mathfrak{H}$ and let $\widetilde{A}$ be a selfadjoint extension of A. Then the following assertions are equivalent:*

(i) $\widetilde{A} = A_F$;

(ii) $\ker(A^* - \lambda) \cap \operatorname{dom} |\widetilde{A}|^{1/2} = \{0\}$ *for some (equivalently for every)* $\lambda \in \rho(\widetilde{A})$;

(iii) $\operatorname{dom} A$ *is dense in* $\mathfrak{H}_{+1}(\widetilde{A})$.

Proof. (i)$\Longrightarrow$(ii) Let $\widetilde{A} = A_F$ and assume that $f \in \ker(A^* - \lambda) \cap \operatorname{dom} |\widetilde{A}|^{1/2}$. Then (8) shows that $\{f, \lambda f\} \in \widehat{\mathfrak{N}}_\lambda \cap A_F$. This implies that $f = 0$ for every $\lambda \in \rho(\widetilde{A})$, since otherwise λ would be an eigenvalue of A_F.

(ii)$\Longrightarrow$(iii) Let (ii) be satisfied with some $\lambda \in \rho(\widetilde{A})$. Assume that there exists $g_0 \in \mathfrak{H}_{+1}(\widetilde{A})$ such that for every $f \in \operatorname{dom} A$,

(20)

$$0 = (f, g_0)_+ = (f, Vg_0) = ((\widetilde{A}_s - \lambda)f, (\widetilde{A}_s - \bar{\lambda})^{-1} Vg_0) = ((A_s - \lambda)f, (\widetilde{A}_s - \bar{\lambda})^{-1} Vg_0).$$

Equivalently,

$$(g - \lambda f, (\widetilde{A}_s - \bar{\lambda})^{-1} Vg_0) = 0, \qquad \text{for all } \{f, g\} \in A.$$

This means that $(\widetilde{A}_s - \bar{\lambda})^{-1} Vg_0 \in \ker(A^* - \bar{\lambda}) \cap \operatorname{dom} |\widetilde{A}|^{1/2}$. Hence, by the assumption $(\widetilde{A}_s - \bar{\lambda})^{-1} Vg_0 = 0$, which implies that $g_0 = 0$. Therefore, $\operatorname{dom} A$ is dense in the Hilbert space $\mathfrak{H}_{+1}(\widetilde{A})$.

(iii)$\Longrightarrow$(i) For every $\{f, g\} \in A$ one has $(\widetilde{A}_s f, f) = (g, f) \geq 0$. If $\operatorname{dom} A$ is dense in $\mathfrak{H}_{+1}(\widetilde{A})$, then also the continuation of $\widetilde{A}_s$ from $\mathfrak{H}_{+1}(\widetilde{A})$ into $\mathfrak{H}_{-1}(\widetilde{A})$ satisfies $(\widetilde{A}_s f, f) \geq 0$ for every $f \in \mathfrak{H}_{+1}(\widetilde{A})$. Thus, in view of (19) $\widetilde{A}_s \geq 0$ and now clearly the space $\mathfrak{H}_{+1}(\widetilde{A})$ coincides with the form domain $\operatorname{dom}[A]$ generated by A. Consequently, $\operatorname{dom} \widetilde{A} \subset \operatorname{dom}[A]$ and the unique selfadjoint extension with this property is $\widetilde{A} = A_F$, cf. (8). □

Remark 10. When the defect numbers of A are $(1,1)$, i.e., $\dim \ker(A^* - \lambda) = 1$, for $\lambda \in \mathbb{C}_\pm$, the equivalence of (i) and (ii) in Theorem 9 has been established in Hassi, Langer, and de Snoo (1995), even for a class of symmetric operators which is wider than the class of semibounded operators. Some equivalent results, formulated in terms of the Q-functions of A, can be found in Hassi, Langer, and de Snoo (1995); Hassi, Kaltenbäck, and de Snoo (1997); Hassi and de Snoo (1998), and Kaltenbäck

(1998). In the special case that $A \geq 0$ is densely defined this result can be derived from the results in the fundamental papers of M.G. Kreĭn, at least for the nonnegative selfadjoint extensions $\widetilde{A}$ of A, cf. Ćurgus (1989).

The above proof of Theorem 9 is based on the rigged Hilbert spaces defined in (17). The proof is explicit, namely it shows how to construct symmetric restrictions $\tilde{S} \subset \widetilde{A}$ for $\widetilde{A} \neq A_F$, so that the Q-function of the pair $\{\tilde{S}, \widetilde{A}\}$ belongs to the subclass $\mathbf{N}_1$ (see Hassi et al. (1995) for the definition) of Nevanlinna functions:

$$\tilde{S} := \{\, \{f, g\} \in \widetilde{A} : \, (f, \omega) = 0 \,\},$$

where $\omega := V g_0 \in \mathfrak{H}_{-1}(\widetilde{A})$ is defined by means of (20).
As a consequence of Theorem 9 one obtains the following characterization for the Kreĭn-von Neumann extension A_K of A.

Corollary 11. *Let A be a closed nonnegative relation in $\mathfrak{H}$ and let $\widetilde{A}$ be a selfadjoint extension of A. Then the following assertions are equivalent:*

(i) $\widetilde{A} = A_K$;

(ii) $\ker(A^* - \lambda) \cap \operatorname{ran} |\widetilde{A}|^{1/2} = \{0\}$ *for some (equivalently for every)* $\lambda \in \rho(\widetilde{A})$;

(iii) $\operatorname{ran} A$ *is dense in* $\mathfrak{H}_{+1}(\widetilde{A}^{-1})$.

Proof. In view of (9) the equality in (i) is equivalent to $\widetilde{A}^{-1} = (A_K)^{-1} = (A^{-1})_F$. The formula

$$\lambda(\widetilde{A} - \lambda)^{-1} = -I + (I - \lambda \widetilde{A}^{-1})^{-1}$$

implies that $\lambda \in \rho(\widetilde{A})$ if and only if $1/\lambda \in \rho(\widetilde{A}^{-1})$. Moreover, it is easy to check that

$$\ker(A^* - \lambda) = \ker(A^{-*} - 1/\lambda), \quad \lambda \neq 0.$$

Since $|\widetilde{A}|^{-1/2} = |\widetilde{A}^{-1}|^{1/2}$, one concludes that

$$\ker(A^* - \lambda) \cap \operatorname{ran} |\widetilde{A}|^{1/2} = \ker(A^{-*} - 1/\lambda) \cap \operatorname{dom} |\widetilde{A}^{-1}|^{1/2}, \quad \lambda \neq 0.$$

By definition the space $\mathfrak{H}_{+1}(\widetilde{A}^{-1})$ in (iii) is $\operatorname{dom} |\widetilde{A}^{-1}|^{1/2} = \operatorname{ran} |\widetilde{A}|^{1/2}$ equipped with the graph inner product of $|(\widetilde{A}^{-1})_s|^{1/2}$. Now the result is obtained by applying Theorem 9 to the inverse relation $\widetilde{A}^{-1}$. □

4.3. The equality of extreme extensions and Kreĭn's uniqueness criterion. The class of selfadjoint extensions of a nonnegative operator is always nonempty and if A itself is not selfadjoint (i.e. A has positive defect numbers (n, n), $1 \leq n \leq \infty$) there are infinitely many selfadjoint extensions of A. However, it may happen that among the selfadjoint extensions of A there is only one which is nonnegative, i.e., it is possible that the Friedrichs extensions A_F of A is the only nonnegative selfadjoint extension of A. According to Theorem 4 this situation occurs precisely when the extreme extensions A_K and A_F of A coincide. As an immediate consequence of Theorem 9 and Corollary 11 one obtains the following result for characterizing this situation.

Proposition 12. *Let A be a closed nonnegative relation in $\mathfrak{H}$ and let A_K and A_F be the Kreĭn von-Neumann and the Friedrichs extension of A, respectively. Then the following assertions are equivalent:*

(i) $A_K = A_F$;

(ii) *there is a selfadjoint extension $\widetilde{A}$ of A such that for some (equivalently for every) $\lambda \in \rho(\widetilde{A})$,*

$$\ker(A^* - \lambda) \cap \operatorname{dom} |\widetilde{A}|^{1/2} = \{0\} \text{ and } \ker(A^* - \lambda) \cap \operatorname{ran} |\widetilde{A}|^{1/2} = \{0\};$$

(iii) *there is selfadjoint extension $\widetilde{A}$ of A such that* $\operatorname{dom} A$ *is dense in* $\mathfrak{H}_{+1}(\widetilde{A})$ *and* $\operatorname{ran} A$ *is dense in* $\mathfrak{H}_{+1}(\widetilde{A}^{-1})$.

Moreover, if $\widetilde{A}$ satisfies (ii) or (iii) then $\widetilde{A} = A_K = A_F$.

By combining Proposition 6 with Theorem 9 and Corollary 11 one arrives at the following two characterizations for $A_K \neq A_F$.

Theorem 13. *Let A be a closed nonnegative relation in $\mathfrak{H}$ and let A_K and A_F be the Kreĭn von-Neumann and the Friedrichs extension of A, respectively. Then the following assertions are equivalent:*

(i) $A_K \neq A_F$;

(ii) *for some (equivalently for every) $\lambda \in \mathbb{C} \setminus [0, \infty)$ there exists $h \in \ker(A^* - \lambda)$, $h \neq 0$, such that*

$$|(g,h)|^2 \leq C_h(g,f) \text{ for all } \{f,g\} \in A \text{ and some } C_h < \infty; \tag{21}$$

(iii) *for some (equivalently for every) $\lambda \in \mathbb{C} \setminus [0, \infty)$ there exists $k \in \ker(A^* - \lambda)$, $k \neq 0$, such that*

$$|(f,k)|^2 \leq C_k(g,f) \text{ for all } \{f,g\} \in A \text{ and some } C_k < \infty. \tag{22}$$

Proof. (i)$\Longleftrightarrow$(ii) By Theorem 9 the condition $A_K \neq A_F$ is equivalent to $\ker(A^* - \lambda) \cap \operatorname{dom} A_K^{1/2} \neq \{0\}$, $\lambda \in \operatorname{Ext}[0, \infty)$. According to part (ii) of Proposition 6 this means that $h \in \ker(A^* - \lambda) \cap \operatorname{dom} A_K^{1/2}$ satisfies (21).

(i)$\Longleftrightarrow$(iii) By Corollary 11 the condition $A_F \neq A_K$ is equivalent to $\ker(A^* - \lambda) \cap \operatorname{ran} A_F^{1/2} \neq \{0\}$, $\lambda \in \operatorname{Ext}[0,\infty)$. Now, according to part (i) of Proposition 6 $k \in \ker(A^* - \lambda) \cap \operatorname{dom} A_K^{1/2}$ satisfies (22). □

Remark 14. Theorem 13, in particular part (iii), can be seen as an analog of Kreĭn's uniqueness criterion for the equality $A_K = A_F$ originally established in the famous paper of Kreĭn (1947) in the case where A is densely defined: $A_F = A_K$ if and only if

$$\sup_{f \in \operatorname{dom} A} \frac{|(f,\varphi)|^2}{(Af,f)} = \infty \quad \text{for every } \varphi \in \mathfrak{N}_{-a} \setminus \{0\} \quad (a > 0). \tag{23}$$

In this form Kreĭn's uniqueness criterion has been also extended in Hassi, Malamud, and de Snoo (2001). The proof given in Hassi et al. (2001) is via Cayley transforms. It was based on the corresponding result concerning contractive extensions of Hermitian contractions, following the original approach of M.G. Kreĭn. The direct proof given above is based on Proposition 6 and Theorem 9. Observe, that the present approach gives a simple geometric interpretation for the original criterion (23) of M.G. Kreĭn.

Since $A \geq 0$ in Proposition 13 is allowed to be a nonnegative relation in $\mathfrak{H}$, the condition (ii) can be derived also from (iii) since $A_K \neq A_F$ is equivalent to $(A^{-1})_K \neq (A^{-1})_F$.

Acknowledgements

Support from the Research Institute for Technology of the University of Vaasa is gratefully acknowledged.

References

Akhiezer, N.I. & I.M. Glazman (1993). *Theory of Linear Operators in Hilbert Space.* (Two volumes bound as one.) New York: Dover Publications Inc.

Alonso, A. & B. Simon (1980). The Birman-Kreĭn-Vishik theory of self-adjoint extensions of semibounded operators. *J. Operator Theory* 4, 251–270.

Ando, T. & K. Nishio (1970). Positive selfadjoint extensions of positive symmetric operators. *Tôhoku Math. J.* 22, 65–75.

Arlinskiĭ, Yu.M. (1988). Positive spaces of boundary values and sectorial extensions of a nonnegative symmetric operator. *Ukrainian Math. Journ.* 40 No. 1, 5–10.

Arlinskiĭ, Yu.M., S. Hassi, H.S.V. de Snoo & Z. Sebestyen (2001). On the class of extremal extensions of a nonnegative operator. *Oper. Theory Adv. Appl.* 127, 41–81.

Berezanski, Ju.M. (1965). *Expansions in Eigenfunctions of Selfadjoint Operators.* Kiev, Naukova Dymka. (Russian) [English translation: Translations of Mathematical Monographs, Volume 17, American Mathematical Society (1968)].

Birman, M.S. (1956). On the self-adjoint extensions of positive definite operators. *Mat. Sb.* 38, 431–450.

Coddington, E.A. & H.S.V. de Snoo (1978). Positive selfadjoint extensions of positive symmetric subspaces. *Math. Z.* 159, 203–214.

Ćurgus, B. (1989). Definitizable extensions of positive symmetric operators in a Kreĭn space. *Integral Equations and Operator Theory* 12, 615–631.

Derkach, V.A. & M.M. Malamud (1991). Generalized resolvents and the boundary value problems for Hermitian operators with gaps. *J. Funct. Anal.* 95, 1–95.

Derkach, V.A. & M.M. Malamud (1995). The extension theory of hermitian operators and the moment problem. *J. Math. Sciences* 73, 141–242.

Fillmore, P.A. & J.P. Williams (1971). On operator ranges. *Adv. Math.* 7, 254–281.

Friedrichs, K.O. (1934). Spektraltheorie halbbeschränkter Operatoren und Anwendung auf die Spektralzerlegung von Differentialoperatoren. *Math. Ann.* 109, 465–487.

Gorbachuk, V.I. & M.L. Gorbachuk (1991). *Boundary Value Problems for Operator Differential Equations.* Mathematics and its Applications (Soviet Series), 48. Dordrecht: Kluwer.

Hassi, S., M. Kaltenbäck & H.S.V. de Snoo (1997). Triplets of Hilbert spaces and Friedrichs extensions associated with the subclass $\mathbf{N}_1$ of Nevanlinna functions. *J. Operator Theory* 37, 155–181.

Hassi, S., H. Langer & H.S.V. de Snoo (1995). Selfadjoint extensions for a class of symmetric operators with defect numbers (1,1). *15th Oper. Theory Conf. Proc.*, 115–145.

Hassi, S., M.M. Malamud, and H.S.V. de Snoo (2001). On Kreĭn's extension theory of nonnegative operators. Preprint. Preliminary version published in: *Working Papers of the University of Vaasa, Department of Mathematics and Statistics* 1. 32 s.

Hassi, S. & H.S.V. de Snoo (1998). Nevanlinna functions, perturbation formulas and triplets of Hilbert spaces. *Math. Nachr.* 195, 115–138.

Kaltenbäck, M. (1998). Models for compositions of Q-functions. *Math. Nachr.* 195, 159–170.

Kato, T. (1966) *Perturbation Theory for Linear Operators.* Berlin: Springer Verlag.

Kreĭn, M.G. (1947). The theory of selfadjoint extensions of semibounded Hermitian operators and its applications, I. *Mat. Sb.* 20, 431–495.

von Neumann, J. (1929). Allgemeine Eigenwerttheorie Hermitescher Funktionaloperatoren. *Math. Ann.* 102, 49–131.

Prokaj, V. & Z. Sebestyén (1996). On extremal positive operator extensions. *Acta Sci. Math. (Szeged)* 62, 458-491.

Sebestyén, Z. (1983). On ranges of adjoint operators in Hilbert space. *Acta Sci. Math. (Szeged)* 46, 295–298.

Sebestyén, Z. & E. Sikolya (2003). On Kreĭn-von Neumann and Friedrichs extensions. *Acta Sci. Math. (Szeged)* 69, 323–336.

Sebestyén, Z. & J. Stochel (1991). Restrictions of positive self-adjoint operators. *Acta Sci. Math. (Szeged)* 55, 149–154.

Shmul'yan, Yu. L. (1967). Two-sided division in a ring of operators. *Math. Notes* 1, 400–403.

Strauss, A.V. (1970). Extensions and generalized resolvents of a symmetric operator which is not densely defined. *Izv. Akad. Nauk SSSR* 34 No. 1. (Russian) [English translation: Math. USSR-Izvestija, 4 No. 1 (1970), 179–208.]

Vishik, M.I. (1952). On general boundary problems for elliptic differential equations. *Trans. Moscow Math. Soc.* 1, 186–246. (Russian) [English translation: Amer. Math. Soc. Transl., 24 (1963), 107–172].

Hodrick-Prescott operator for quarterly observed economic time series

Matti Laaksonen

Dedicated to Ilkka Virtanen on the occasion of his 60th birthday

Abstract

Laaksonen, Matti (2004). Hodrick-Prescott operator for quarterly observed economic time series. In *Contributions to Management Science, Mathematics and Modelling. Essays in Honour of Professor Ilkka Virtanen*. Acta Wasaensia No. 122, 55–78. Eds Matti Laaksonen and Seppo Pynnönen.

The Hodrick-Prescott Filter is a commonly used method to extract the business cycle from an empirical economic time series. Recently the Hodrick-Prescott Filter has been criticized for spurious results. Hodrick and Presscott (1997) originally define their method as an optimization procedure. The outcome of the procedure depends on the so-called smoothing parameter. Increasing the value of the parameter makes the cycles larger and the trend less varying. A widely accepted value of the smoothing parameter is $\lambda = 1600$ for quarterly observed economic time series. In this paper we describe the HP-estimation not as a filter but as a linear operator. The eigenvalues of the operator depend on the smoothing parameter but the eigenvectors do not. Thus we get a natural basis for the generalized Fourier analysis for the selection of the optimal smoothing parameter. The value we get is $\lambda \approx 500$ and it depends on the length of the time series. The operator context makes it easy to modify the estimation process. We will introduce an HP-like band-pass operator by which we can decompose the time series into 'trend', 'business cycle' and 'seasonal + irregular' components. The joint implementation of band-pass estimation and wavelet based seasonal adjustment is also discussed.

Matti Laaksonen, Department of Mathematics and Statistics, University of Vaasa, P.O.Box 700, FI-65101, Vaasa, Finland, e-mail mla@uwasa.fi.

Key words: Time Series Decomposition, Hodrick-Prescott filter, Business cycles.

1. Introduction

The dynamical system of the national economy is normally not in the equilibrium. Although we can calculate the equilibrium state for the mathematical models of the system, real time series show more or less permanent cyclical fluctuation around the equilibrium point. Even the equilibrium point is not at rest. It moves because of growth and changes in the structure and efficiency of the production. The evolution of the equilibrium is the trend of economy.

There is no commonly accepted model of the economy, so there is also no consensus of the best decomposition of economic time series. We expect the trend to be smoother than the cycle and we expect the cycles to have zero mean. We can also decompose different time series and compare their cycles to each other and to the aggregate GDP-cycle. Some of the time series are leading and some are lagging the other. Some of the time series are pro-cyclical (positive correlation with the aggregate GDP-cycle) and some are counter-cyclical. These features of economic time series are called 'stylized facts' (Lucas 1977; Kydland and Prescott 1990; Blackburn and Ravn 1992; Englund, Persson and Svensson 1992; Fiorito and Kollintzas 1994; Christodoulakis 1995; Bjømlan 2000). The stylized facts we observe on the data depend on the detrending method we use (Bjømland 2000).

One of the mostly used methods is the so-called Hodrick-Presscot Filter. The HP-Filter has been subject to some criticism (see e.g. Canova 1994, 1998; Cogley and Nason 1995; Harvey and Jaeger 1993; King and Rebelo 1992; Schenk-Hoppe 2001). Still it is one of the most often used methods. The well-known improvement of HP-Filter is the band pass filter of Baxter and King (1995); see also Kaiser and Maravall (1999), and Harvey and Trimburg (2001). A recent promising modification is the work done by Reeves et al. (2000).

In this paper we recall the original definition of Hodrick and Prescott (1997). In Section 2 we will consider the filtering procedure in its usual form. In Section 3 the original filter is considered as a linear operator (i.e. as a matrix). The eigenvalues and the eigenvectors are determined. Eigenvalues are real and discrete. Eigenvectors do not depend on the smoothing parameter of the operator. By this generalized spectral representation we can construct new band pass operators. We demonstrate the use of these band pass operators with a real Japanese data. In Section 3.3 we consider the proper value of the smoothing parameter. Some conclusions are made in Section 4.

2. The Hodrick-Prescott Filter

Let $y=(y_1,y_2,\ldots,y_N)^{\mathrm{T}}$ be an observed empirical time series. Let $y_t=x_t+c_t$, where $\mathbf{x}=(x_1,x_2,\ldots,x_N)^{\mathrm{T}}$ is the unobserved trend component of the time series and $\mathbf{c}=(c_1,c_2,\ldots,c_N)^{\mathrm{T}}$ is the unobserved cyclical component of the time series. Hodrick and Prescott determine the trend component as a solution to the following optimization problem

$$(1)\qquad \min_{\mathbf{x}}\left(\sum_{t=1}^{N}(y_t-x_t)^2+\lambda\sum_{t=2}^{N-1}((x_{t+1}-x_t)-(x_t-x_{t-1}))^2\right),$$

where λ is the smoothing parameter. We call the problem (1) the HP optimization problem and its solution the HP estimated trend. The HP problem can be rewritten in the following equivalent form

$$(2)\qquad \min_{\mathbf{x}}\left(\sum_{t=1}^{N}c_t^{\,2}+\lambda\sum_{t=2}^{N-1}\left(\Delta^2x_t\right)^2\right),\ \lambda>0.$$

There are two parts in the objective function. The first sum in (2) is the variance of the cyclic component ($E(c_t)=0$). The second part is the sum of squared second differences of the trend measuring the curvature of the trend. So in the minimization we seek small cycles and smooth trend. The relative weight of these objectives is controlled by the smoothing parameter λ. With high value of λ we prefer smooth trend and with low value of λ the preference is put on small cycles. Hodrick and Prescott give a value of $\lambda=1600$ as a proper value of the smoothing parameter for the quarterly data of economic time series. This value is widely accepted in the literature.

The first order condition for the HP optimization problem (2) is

$$(3) \qquad \begin{cases} c_1 & = \lambda(x_1 - 2x_2 + x_3) \\ c_2 & = \lambda(-2x_1 + 5x_2 - 4x_3 + x_4) \\ c_t & = \lambda(x_{t-2} - 4x_{t-1} + 6x_t - 4x_{t+1} + x_{t+2}),\ t = 3,4,\ldots,N-2 \\ c_{N-1} & = \lambda(-2x_{N-3} + 5x_{N-2} - 4x_{N-1} + x_N) \\ c_N & = \lambda(x_{N-2} - 2x_{N-1} + x_N) \end{cases}$$

According to the first order condition there exists a symmetric $(N \times N)$-matrix $\mathbf{F}$ such that $\mathbf{c} = \lambda\mathbf{F}\mathbf{x}$, and thus because $\mathbf{y} = \mathbf{x} + \mathbf{c}$,

$$(4) \qquad \mathbf{x} = (\lambda\mathbf{F} + \mathbf{I})^{-1}\mathbf{y}$$

$$(5) \qquad \Leftrightarrow \mathbf{c} = \left(\mathbf{I} - (\lambda\mathbf{F} + \mathbf{I})^{-1}\right)\mathbf{y} \quad =_{\text{def}} \mathbf{H}\mathbf{y}.$$

The matrices $\mathbf{F}$ and $\mathbf{H} = \mathbf{I} - (\lambda\mathbf{F} + \mathbf{I})^{-1}$ will be analyzed further in Section 3, but first we consider their traditional interpretation as filters.

2.1. HP-estimation as a filter

There is a tradition to express the Hodrick-Prescott estimation process as a ‘filter’. We next shortly explain how it is done and we also indicate the difficulties hiding in this interpretation.

If we forget the first two and last two values of the cycle then (3) takes the form

$$(6) \qquad c_t = \lambda(x_{t-2} - 4x_{t-1} + 6x_t - 4x_{t+1} + x_{t+2})$$

$$(7) \qquad = \lambda(y_{t-2} - 4y_{t-1} + 6y_t - 4y_{t+1} + y_{t+2}) - \lambda(c_{t-2} - 4c_{t-1} + 6c_t - 4c_{t+1} + c_{t+2}).$$

For the Z-transforms we get

$$(8) \qquad C(z) = \lambda(z^{-2} - 4z^{-1} + 6 - 4z + z^2)Y(z) - \lambda(z^{-2} - 4z^{-1} + 6 - 4z + z^2)C(z)$$

$$(9) \qquad \Rightarrow C(z) = \frac{\lambda(1-z)^4}{z^2 + \lambda(1-z)^4}Y(z).$$

Next we approximate the Fourier gain of the cycle estimation with frequency ω as the modulus of the ratio of the Fourier transforms of $\mathbf{c}$ and $\mathbf{y}$.

$$(10) \qquad G_{\text{cycle}}(\omega) = \left|\frac{C(e^{-i\omega})}{Y(e^{-i\omega})}\right| = \frac{4\lambda(1-\cos\omega)^2}{1+4\lambda(1-\cos\omega)^2}\,.$$

In a similar way we get the Fourier gain of the trend estimation as

$$(11) \qquad G_{\text{trend}}(\omega) = \left|\frac{X(e^{-i\omega})}{Y(e^{-i\omega})}\right| = \frac{1}{1+4\lambda(1-\cos\omega)^2}\,.$$

In Figures 1(a) and 1(b) we have drawn the Fourier gains of trend and cycle estimation for HP-filter with smoothing parameter value $\lambda = 1600$. Not so orthodox but maybe more informative are the graphs of the gains as functions of the period $k = 2\pi/\omega$ drawn in Figures 1(c) and 1(d). From the Figures we see that the HP-filter is a high pass filter in cycle estimation.

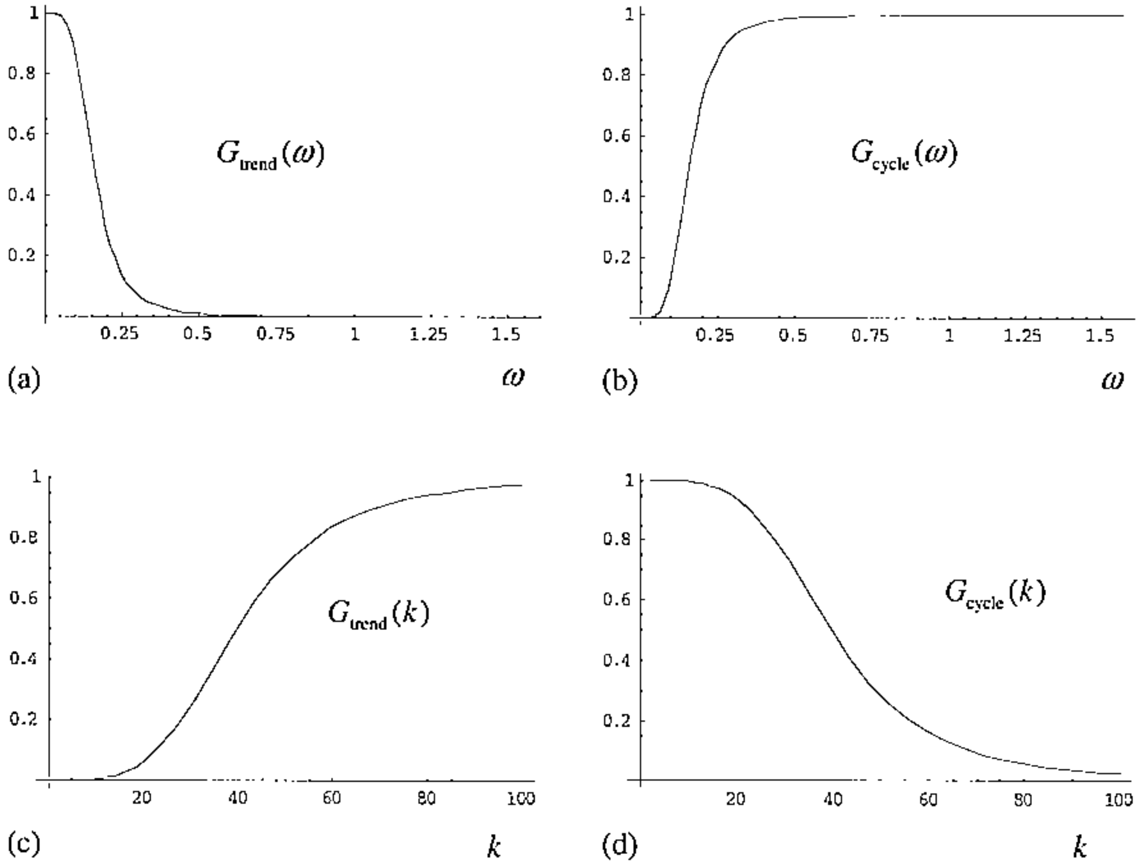


Figure 1. The Fourier gain of HP1600-filter ($\lambda = 1600$).

The main problem with the interpretation of HP as a filter comes evident when we calculate the roots of the denominator of the transfer function (10)

$$(12) \qquad z^2 + \lambda(1-z)^4 = 0$$

$$(13) \qquad \Leftrightarrow z = 1 + \frac{\eta}{2}\left(i \pm \sqrt{\frac{4i}{\eta} - 1}\right),\ \eta = \pm\lambda^{-1/2}.$$

Some of these four roots have modulus greater than 1. Thus the filter has not finite response to finite input. This is the case with any possible value of the smoothing parameter. A numerical example is drawn in Figure 2 iterating the difference equation (7) with parameter value $\lambda = 1600$ and the unit impulse input ($y_0 = 1$ and $y_t = 0$ when $t \neq 0$).

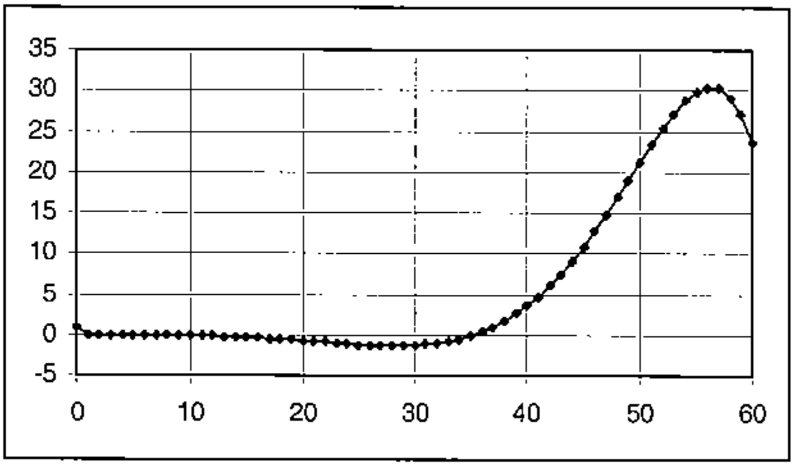


Figure 2. The unit impulse response of the filter of the equation (7).

3. HP-estimation as a linear operator

The matrix $\mathbf{H} = \mathbf{I} - (\lambda\mathbf{F} + \mathbf{I})^{-1}$ in equation (5) is symmetric but it is not of full rank. If the length of the observed time series is N then the matrix $\mathbf{F}$ is an $N \times N$ -matrix of the form

$$\mathbf{F}=\begin{pmatrix} 1 & -2 & 1 & & & & & \\ -2 & 5 & -4 & 1 & & & & \\ 1 & -4 & 6 & -4 & 1 & & & \\ & 1 & -4 & 6 & -4 & 1 & & \\ & & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & & 1 & -4 & 6 & -4 & 1 \\ & & & & 1 & -4 & 5 & -2 \\ & & & & & 1 & -2 & 1 \end{pmatrix}. \tag{14}$$

The matrix $\mathbf{F}$ is not of full rank since $(1,1,\ldots,1)^T \in \mathrm{Null}(\mathbf{F})$. In order to analyze $\mathbf{H}$ we write $\mathbf{F}$ as a product of two matrices

$$\mathbf{F}=\mathbf{L}\mathbf{L}^T, \tag{15}$$

where $\mathbf{L}$ is an $N \times (N-2)$-matrix

$$\mathbf{L}=\begin{pmatrix} 1 & & & & \\ -2 & 1 & & & \\ 1 & -2 & \ddots & & \\ & 1 & \ddots & 1 & \\ & & \ddots & -2 & 1 \\ & & & 1 & -2 \\ & & & & 1 \end{pmatrix}. \tag{16}$$

The decomposition (15) is not a Cholesky decomposition, since the matrix $\mathbf{L}$ is not a square triangular matrix. Because $\mathbf{F}=\mathbf{L}\mathbf{L}^T$ we have

$$\langle \mathbf{F}\mathbf{x},\mathbf{x}\rangle = \langle \mathbf{L}\mathbf{L}^T\mathbf{x},\mathbf{x}\rangle = \langle \mathbf{L}^T\mathbf{x},\mathbf{L}^T\mathbf{x}\rangle = \left\|\mathbf{L}^T\mathbf{x}\right\|^2 \geq 0, \tag{17}$$

$$\mathrm{Null}(\mathbf{F}) = \mathrm{Null}\left(\mathbf{L}^T\right). \tag{18}$$

So $\mathbf{F}$ is positive semidefinite (i.e. its eigenvalues are non-negative). In a similar way, we can show that $\mathbf{L}^T\mathbf{L}$ is positive definite. The smoothing parameter λ is positive and we can say a little bit more of the matrix $(\lambda\mathbf{F}+\mathbf{I})$ than of the matrix $\mathbf{F}$ alone

$$\langle (\lambda\mathbf{F}+\mathbf{I})\mathbf{x},\mathbf{x}\rangle = \lambda\left\|\mathbf{L}^T\mathbf{x}\right\|^2 + \left\|\mathbf{x}\right\|^2 > 0, \text{ if } \mathbf{x}\neq\mathbf{0}. \tag{19}$$

So the matrix $(\lambda\mathbf{F}+\mathbf{I})$ is positive definite and of full rank. By this argument $\mathbf{H}=\mathbf{I}-(\lambda\mathbf{F}+\mathbf{I})^{-1}$ is well defined for all $\lambda>0$.

3.1. Eigenvalues and eigenvectors of Hodrick-Prescott operator

Let N be the number of observation in the time series we will analyze by the Hodrick-Prescott operator **H**. Next we construct the normalized eigenvectors $\psi_k \in \mathbf{R}^N$ and eigenvalues $\theta_k \in \mathbb{C}$ of the Hodrick-Prescott operator $\mathbf{H}$. We order the eigenvalues and eigenvectors in the usual way so that

(20) $$\theta_1 \geq \theta_2 \geq \ldots$$

The motivation of this paragraph is the following. If we have N linearly independent eigenvectors of **H** then we can introduce an $N\times N$-matrix $\mathbf{\Psi}$ with k:th column vector equal to ψ_k and we can also introduce an $N\times N$ diagonal matrix $\mathbf{\Theta}$ with k:th diagonal element equal to θ_k. By these matrices we have a useful decomposition

(21) $$\mathbf{H}=\mathbf{\Psi\Theta\Psi}^{\mathrm{T}}.$$

First we show the existence of N linearly independent eigenvectors. Two vectors are of special interest. Let

(22) $$\mathbf{w}_0 \in \mathbf{R}^N,\ \ w_{0,t}=1/\sqrt{N}, \forall t=1,2,\ldots,N,$$

(23) $$\mathbf{w}_1 \in \mathbf{R}^N,\ \ w_{1,t}=\alpha(2t-N-1), \forall t=1,2,\ldots,N, \|\mathbf{w}_1\|=1.$$

Then $\mathbf{L}^T\mathbf{w}_0=\mathbf{0}$ and $\mathbf{L}^T\mathbf{w}_1=\mathbf{0}$. Now we get

Theorem 3.1. Let **H** be the $N\times N$ Hodrick-Prescott –operator and let the vectors $\mathbf{w}_0,\mathbf{w}_1 \in \mathbf{R}^N$ be as defined in (22) and (23), then

(24) $$\mathrm{Null}(\mathbf{H})=\mathrm{span}\{\mathbf{w}_0,\mathbf{w}_1\},$$

(25) $\operatorname{Rank}(\mathbf{H}) = N-2$.

Proof. Direct calculation shows that

(26) $$\mathbf{H} = \mathbf{I} - \left(\mathbf{I} + \lambda\mathbf{L}\mathbf{L}^{\mathrm{T}}\right)^{-1} = \lambda\mathbf{L}\left(\mathbf{I} + \lambda\mathbf{L}^{\mathrm{T}}\mathbf{L}\right)^{-1}\mathbf{L}^{\mathrm{T}}.$$

The matrix $\mathbf{I} + \lambda\mathbf{L}^{\mathrm{T}}\mathbf{L}$ is positive definite (cf. 19). Since matrices $\mathbf{L}$ and $\left(\mathbf{I} + \lambda\mathbf{L}^{\mathrm{T}}\mathbf{L}\right)^{-1}$ have trivial null spaces, we have $\operatorname{Null}(\mathbf{H}) = \operatorname{Null}(\mathbf{L}^{\mathrm{T}}) = \operatorname{Span}\{\mathbf{w}_0, \mathbf{w}_1\}$. Then also $\operatorname{Rank}(\mathbf{H}) = N - \dim(\operatorname{Null}(\mathbf{H})) = N-2$. □

Corollary 3.2. The $N \times N$ Hodrick-Prescott –operator **H** is nonnegative and strictly contractive, i.e. $\mathbf{0} \leq \mathbf{H} < \mathbf{I}$, and has N linearly independent eigenvectors ψ_k and corresponding eigenvalues θ_k so that

(27) $$1 > \theta_1 \geq \theta_2 \geq \ldots > \theta_{N-1} = \theta_N = 0.$$

Proof. The statement is immediate from (26). Like before it is easy to show that $\mathbf{I} + \lambda\mathbf{L}^{\mathrm{T}}\mathbf{L}$ is positive definite. Thus $\langle\mathbf{Hx}, \mathbf{x}\rangle = \lambda\left\langle\left(\mathbf{I} + \lambda\mathbf{L}^{\mathrm{T}}\mathbf{L}\right)^{-1}\mathbf{L}^{\mathrm{T}}\mathbf{x}, \mathbf{L}^{\mathrm{T}}\mathbf{x}\right\rangle \geq 0$, $\forall \mathbf{x} \in \mathbf{R}^N$. So the eigenvalues of **H** are nonnegative. Because **H** is symmetric and real, there exist an orthonormal set of eigenvectors spanning the column space. The number of these eigenvectors is $\operatorname{Rank}(\mathbf{H}) = N-2$. By augmenting this set by vectors $\mathbf{w}_0$ and $\mathbf{w}_1$ we get the final set of eigenvectors. □

The determination of the eigenvalues and the eigenvectors of **H** is not a trivial task. The empirical time series we analyze are usually so long that N is quite large and then also the number of eigenvalues is large and they are close to each other. Not all standard methods to determine the eigensystem are reliable in this case. We get a stable reliable method by using singular value decomposition (SVD) .

Let **A** be an real $m \times n$-matrix. Its singular value decomposition (SVD) is the decomposition

(28) $$\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\mathrm{T}},$$

where $\mathbf{U}$ is an real orthogonal $m \times m$-matrix, $\mathbf{V}$ is a real orthogonal $n \times n$-matrix, and $\boldsymbol{\Sigma}$ is a real diagonal $m \times n$-matrix with diagonal elements $(\sigma_1, \sigma_2, \ldots, \sigma_p)$, $p = \min(m,n)$. The column vectors $\mathbf{u}_j$ of the matrix $\mathbf{U}$ are called left singular vectors of $\mathbf{A}$ and the column vectors $\mathbf{v}_j$ of the matrix $\mathbf{V}$ are called right singular vectors of $\mathbf{A}$. The SVD decomposition always exists (Golub, van Loan 1983), and it is easy to show that the singular values are non-negative and, $\mathbf{A}\mathbf{v}_j = \sigma_j \mathbf{u}_j$ and $\mathbf{A}^{\mathrm{T}}\mathbf{u}_j = \sigma_j \mathbf{v}_j$ for all $j = 1,2,\ldots,p$. Especially this means that

(29) $$\mathbf{A}^{\mathrm{T}}\mathbf{A}\mathbf{v}_j = \sigma_j^2 \mathbf{v}_j, \quad j = 1,2,\ldots,p,$$

(30) $$\mathbf{A}\mathbf{A}^{\mathrm{T}}\mathbf{u}_j = \sigma_j^2 \mathbf{u}_j, \quad j = 1,2,\ldots,p.$$

Now using the SVD of $\mathbf{L}$ we see that the left singular vectors of $\mathbf{L}$ are eigenvectors of the matrix $\mathbf{F}$ and the corresponding eigenvalues of $\mathbf{F}$ are the squares of the singular values of $\mathbf{L}$.

Theorem 3.3. Let $\mathbf{H}$ be the $N \times N$ Hodrick-Prescott -operator with smoothing parameter λ, eigenvectors ψ_k, $k = 1,2,\ldots,N$ and eigenvalues $\theta_k, k = 1,2,\ldots,N$. Let the singular values of $\mathbf{L}$ be $\sigma_1 \geq \sigma_2 \geq \ \ldots \geq \sigma_{N-2}$ and let the left singular vectors of $\mathbf{L}$ be $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_N$. Then

(31) $$\psi_k = \mathbf{u}_k,\ k = 1,2,\ldots,N-2,$$

(32) $$\theta_k = \frac{\sigma_k^2}{\sigma_k^2 + 1/\lambda},\ k = 1,2,\ldots,N-2,$$

(33) $$\psi_{N-1} = \mathbf{w}_1,\ \theta_{N-1} = 0,$$

(34) $$\psi_N = \mathbf{w}_0,\ \theta_N = 0.$$

Proof. Let ψ be an eigenvector of $\mathbf{H}$ with eigenvalue θ. Then

$$\begin{aligned} & \mathbf{H}\psi = \theta\psi \\ (35) \qquad & \Leftrightarrow \psi - (\mathbf{I} + \lambda\mathbf{F})^{-1}\psi = \theta\psi \\ & \Leftrightarrow \mathbf{F}\psi = \frac{\theta}{\lambda(1-\theta)}\psi. \end{aligned}$$

So we see that ψ is an cigenvector of $\mathbf{F}$. If we set $\psi_k = \mathbf{u}_k$, then

$$(36) \qquad \frac{\theta_k}{\lambda(1-\theta_k)} = \sigma_k^2 \iff \theta_k = \frac{\sigma_k{}^2}{\sigma_k{}^2 + 1/\lambda}.$$

□

Some remarks: (1) **H** and **F** have the samc cigenvectors, which do not depend on the smoothing parameter. (2) Eigenvalues of **H** depend on thc smoothing parameter λ, but in all cases they arc less than 1. ($1 > \theta_1 \geq \theta_2 \geq \ldots > \theta_{N-1} = \theta_N = 0$)

In Figure 5 wc have drawn some eigenvectors of **H**. Also the corrcsponding eigenvectors of the matrix $\mathbf{G} = \mathbf{L}^T\mathbf{L}$ are exhibited. This matrix is a symmetric Toepliz matrix and its eigenvcctors are near sinusoidal functions. If one has not a suitable tool to calculate the SVD decomposition (e.g. MatLab), then the eigenvectors ψ_k^G of **G** can be determincd by the Rayleigh quotient iteration method by sinusoidal initial vector. Finally take $\psi_k = \mathbf{L}\psi_k^G$.

3.2. Generalized HP-spectrum of time series

Let $\mathbf{y} \in \mathbf{R}^N$ bc a time series. Because the eigenvectors of **H** form an orthonormal sequence the set of the eigenvectors $E = \{\psi_1, \psi_2, \ldots, \psi_N\}$ is an orthonormal basis of $\mathbf{R}^N$. Thus there exists a unique presentation of **y** in the basc E

$$(37) \qquad \mathbf{y} = \sum_{k=1}^{N} h_k \psi_k .$$

We call the sequence $(h_1, h_2, \ldots, h_N)$ as HP-coefficients of **y**. Because of Parsevall's identity

(38) $$\|y\|^2 = \sum_{k=1}^{N} h_k^{\,2},$$

we call the sequence $(h_1^{\,2}, h_2^{\,2}, \ldots, h_N^{\,2})$ as the HP energy spectrum of **y**.

In Figure 5 (in Appendix) we have plotted some eigenvectors ψ_k of **H**. It is easy to see that the function ψ_k has $N-k$ zeros. So for indices $k = 1,2,\ldots,N-2$ we can define the quasi period T_k and quasi frequency ω_k of ψ_k by

(39) $$T_k = \frac{2N}{N-1-k},\ \ \omega_k = \frac{2\pi}{T_k} = \pi\left(1-\frac{1+k}{N}\right),\ \ k = 1,2,\ \ldots,\ N\text{–}2\ .$$

For the last two indices we simply define $\omega_{N-1} = 0$, and $\omega_N = -\pi/N$. Now we get the graph of the HP energy spectrum by plotting the points $\left(\omega_k, h_k^{\,2}\right)_{k=1,2,\ldots,N}$. Often the last two coefficients are so dominant, that we ignore them from the graph of the spectrum. In this case one must remember that the constant component $h_N \psi_N$ and the constant slope component $h_{N-1}\psi_{N-1}$ are not included in the graph.

As an example we have drawn into Figures 6a and 6b two economic time series with their HP- and Fourier energy spectrum. In Figure 6a we have plots of 'Japanese metal product of building (quarterly 1993–1999)'. The data has 24 observations. In Figure 6b we have plots of 'Japanese inventories, public sectors (constant prices, quarterly 1980–2000)'. The data is made by summing the changes in inventories, there are 80 observations. In subplot (i) we will find the original time series, in the subplot (ii) we will find the generalized HP energy spectrum of the data and in the subplot (iii) we will find the classical Fourier energy spectrum of the same data. The HP energy spectrum and the Fourier energy spectrum are not identical but they are in line with each other giving similar interpretation of the frequency distribution of the data.

Let **y** be an observed time series with the representation in HP basis,

(40) $$\mathbf{y} = \sum_{k=1}^{N} u_k \psi_k .$$

Then the HP-estimate for the cyclic component is

(41) $$\mathbf{c} = \mathbf{Hy} = \sum_{k=1}^{N} \theta_k u_k \psi_k .$$

The HP energy spectrum of **y** is $\left(\omega_k, u_k^2\right)$ and the HP energy spectrum of **c** is $\left(\omega_k, \theta_k^2 u_k^2\right)$. So we can plot the HP energy gain by plotting the intensity ratios on the frequencies (i.e. plotting the points $\left(\omega_k, \theta_k^2\right)$) or plotting the intensity ratios on the periods (i.e. plotting the points $\left(T_k, \theta_k^2\right)$). In Figure 3 we have the plot of HP gain with data length $N = 80$ and the smoothing parameter with the value $\lambda = 1600$.

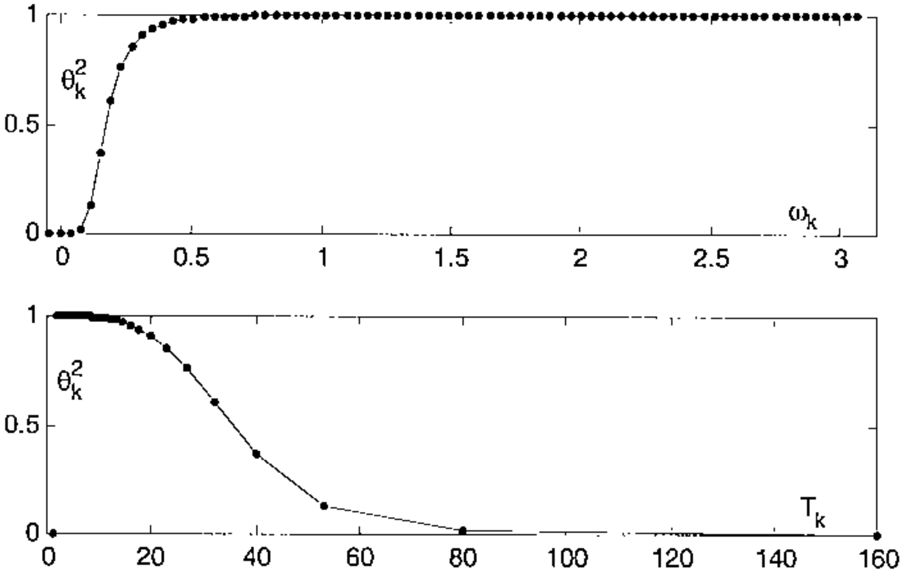


Figure 3. The HP energy gain as a function of quasi frequency (ω_k, top of the Figure) and as a function of quasi period (T_k, bottom of the Figure).

If we apply the HP operator to the data of Japanese inventories, public sectors (constant prices, quarterly 1980–2000) (see Figure 3a), we get the cyclic component drawn in the Figure 7a and the trend component drawn in Figure 7b with their HP energy spectrum and their Fourier energy spectrum.

3.3. The Classical Band Pass modification of the HP operator

The HP operator $\mathbf{H} = \mathbf{\Psi\Theta\Psi}^{\mathrm{T}}$ has the energy gain plotted in Figure 3. The gain is actually a plot of the squares of the diagonal elements of $\mathbf{\Theta}$. Thus we can easily construct classical Hodrick-Prescott type low-pass, band-pass and high-pass operators by defining

$$\text{(42)} \qquad \mathbf{H}^{\mathrm{low}} = \mathbf{\Psi\Theta}^{\mathrm{low}}\mathbf{\Psi}^{\mathrm{T}}, \begin{cases} \theta_k^{\mathrm{low}} = 1 - \theta_k, & T_k > 6, \ (\text{or } \omega_k < \pi/3) \\ \theta_k^{\mathrm{low}} = 0, & \text{otherwise} \end{cases}$$

$$\text{(43)} \qquad \mathbf{H}^{\mathrm{band}} = \mathbf{\Psi\Theta}^{\mathrm{band}}\mathbf{\Psi}^{\mathrm{T}}, \begin{cases} \theta_k^{\mathrm{band}} = \theta_k, & T_k > 6, \ (\text{or } \omega_k < \pi/3) \\ \theta_k^{\mathrm{band}} = 0, & \text{otherwise} \end{cases}$$

$$\text{(44)} \qquad \mathbf{H}^{\mathrm{high}} = \mathbf{\Psi\Theta}^{\mathrm{high}}\mathbf{\Psi}^{\mathrm{T}}, \begin{cases} \theta_k^{\mathrm{high}} = 1, & T_k \le 6, \ (\text{or } \omega_k \ge \pi/3) \\ \theta_k^{\mathrm{high}} = 0, & \text{otherwise} \end{cases}$$

The threshold period 6 comes from the definition of business cycle. The usual way is to define business cycles as variations with duration between 6 and 32 quarters. Thus the corresponding threshold frequency is $\pi/3$ and the threshold index is $k_{\mathrm{threshold}} = 2N/3-1$. Using these operators we can decompose the time series $\mathbf{y}$ to the trend component $\mathbf{t} = \mathbf{H}^{\mathrm{low}}\mathbf{y}$, to the cyclic component $\mathbf{c} = \mathbf{H}^{\mathrm{band}}\mathbf{y}$ and to the residual component $\mathbf{r} = \mathbf{H}^{\mathrm{low}}\mathbf{y}$. The gains of the operators are drawn in Figure 8. In Figure 9 we have drawn the original data of Japanese inventories, public sectors (1980–2000), its trend, cyclical and residual components by the classical Band Pass HP operator with their energy spectrums. Observe, that the residual clearly contains the seasonal component which is approximately 4-periodic.

3.4. The proper value of the smoothing parameter

By the classical definition the threshold value between the frequencies belonging to the trend and the frequencies belonging to the cycle equals to $2\pi/32 = \pi/16$. Thus we demand that

(45) $$\begin{cases}\theta^{\text{low}}(\omega)\geq \theta^{\text{band}}(\omega), & \text{if } \omega \leq \pi/16 \\ \theta^{\text{low}}(\omega)\leq \theta^{\text{band}}(\omega), & \text{if } \omega \geq \pi/16\end{cases} \Leftrightarrow \; \theta(\pi/16)=1/2.$$

The condition (45) is fulfilled if we choose a proper value for the smoothing parameter λ. This value depends on the length of the time series under analysis. The proper value of the smoothing parameter is determined numerically using cubic spline approximation of the $\theta(\omega)$ function. The graph is drawn in Figure 4.

Another way to evaluate the proper value of the smoothing parameter is a modification of the Q-measure used in Baxter and King (1995). Let us define the measure of the distance of the eigensystem of HP band pass operator and the eigensystem of the corresponding ideal operator as

(46) $$Q(\lambda)=\int_0^{\pi}\left(\theta(\omega)-\theta_{id}(\omega)\right)^2 d\omega=\int_0^{\pi/16}\theta(\omega)^2 d\omega+\int_{\pi/16}^{\pi}\left(\theta(\omega)-1\right)^2 d\omega.$$

As the continuous function $\theta(\omega)$ in the integral we use the cubic spline on the points (ω_k,θ_k), $k=1,2,\ldots,N$ (see the first picture of Figure 3). For every N we choose λ which minimizes the measure $Q(\lambda)$. In this way we get the HP operator as close to the ideal operator as possible. The results of this minimization procedure are shown in Figure 4 (dotted curve). There are two remarks we shall make from Figure 4. First the recommended value of $\lambda \approx 500$ is considerably smaller than the value $\lambda = 1600$ normally used in the literature. Secondly the proper value of λ depends on the length of the data.

This paper assumes quarterly data. Still the analysis above can be done also in the case of annual data. Then the threshold period is 8 and the threshold frequency is $\pi/4$. The minimum of the Q-measure comes out with the value of the smoothing parameter $\lambda \approx 3$. Ravn and Uhlig (2002) get by different argument an approximate value 6.5 if the quarterly value is 1600 for the smoothing parameter.

As an example we reproduced Figures 8 and 9 with smoothing parameter $\lambda = 550$. See Figures 10 and 11. The difference of these decompositions is clearly visible at the end of the cyclic component.

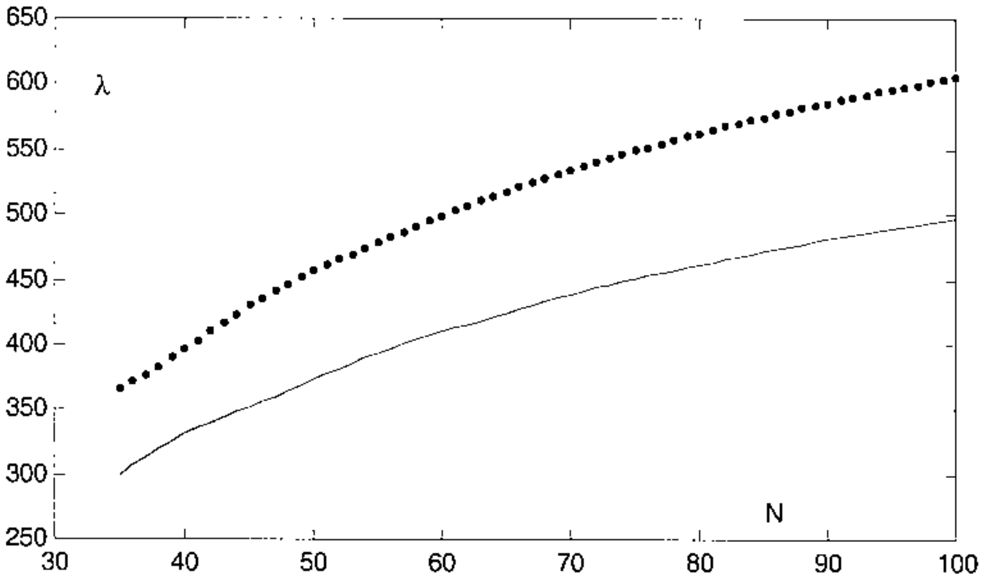


Figure 4. The proper value of the smoothing parameter λ with different values of N evaluated by the condition (45) (solid curve) and by minimizing the measure $Q(\lambda)$ in (46) (dotted curve).

3.5. The seasonal adjustment and detrending

The seasonal adjustment can be done by advanced computer packages. X-12-ARIMA and Tramo-Seats are widely used programs (see Findley et al. 1998, Gomez and Maravall 1996). The programs are complicated and advanced. At the end there is an open question: is the adjustment procedure truly linear (see e.g. Ghysels, Granger and Siklos (1996), Ghysels and Perron (1996)). Findley et al. (1998) describe clearly how the seasonal adjustment is closely combined with some trend estimation method. The estimation of the seasonal component and the estimation of the trend component are repeated in an alternating way with several iterations in the procedure.

If we are analyzing aggregate time series (e.g. in European Union) we sometimes make aggregates of seasonally adjusted series and sometimes we seasonally adjust aggregates. In these kind of cases it would be nice if the adjusting method is linear (see Astolfi et. al 2003). The HP operator is completely a linear tool. So it is ideal for the detrending method of the truly linear seasonally adjustment procedure. Also the estimation of the seasonal component can be made in a linear way by using wavelets. For wavelets see Gençay, Selçuk and Whitcher (2001) or Percival and Walden (2000), Laaksonen (2004).

4. Conclusions

We have shown that the procedure defined in Hodrick and Prescott (1998) yields a completely linear operator with real eigenvalues and eigenvectors which are independent of the value of the smoothing parameter. We introduced the Band Pass Operators by which we can decompose any quarterly data easily to the trend, to the cycle and to the irregular component. The typical seasonal component is included in the irregular component.

The value of the smoothing parameter by which the operator will be as close to the ideal operator as possible with respect to the Q-measure is estimated to be approximately 500. The exact value depends on the number of observations in the time series. For annual data the corresponding proper value of the smoothing parameter is approximately 3.

The linearity of the operators make it easy to combine these operators with wavelet based seasonal adjustment procedure. This will be one objective of some further research.

References

Astolfi, Roberto, Dominique Ladiray & Gian Luigi Mazzi (2003). *Seasonal Adjustment of European Aggregates: Direct versis Indirect Approach.* Working papers and studies, Eurostat, 2003, ISBN 92-894-5389-3.

Baxter, Marianne & Robert G. King (1995). Measuring business cycles: approximate band-pass filters for economic time series. *NBER Working Paper* no. 5022.

Bjørnland, Hilde Christiane (2000). Detrending methods and stylized facts of business cycles in Norway – an international comparision, *Empirical Economics* 25:3, 369–392.

Blackburn, Keith & Morten Ravn (1992). Business cycles in the U.K.: facts and fictions. *Economica* 59, 383–401.

Canova, Fabio (1994). Detrending and Turning-Points. *European Economic Review* 38:3/4, 614–623.

Canova, Fabio (1998). Detrending and business cycle facts. *Journal of Monetary Economics* 41, 475–512.

Christodoulakis, Nicos, Sophia P. Dimelis & Tryphon Kollintzas (1995). Comparison of business cycles in the EC: Idiosyncracies and regularities. *Economica* 62, 1–27.

Cogley, Timothy & James M. Nason (1995). Effects of the Hodrick-Prescott filter on trend and difference stationary time series: Implications for business cycle research. *Journal of Economic Dynamics and Control* 19, 253–278.

Ehlgen, Jürgen (1998). Distortionary effects of the optimal Hodrick-Prescott filter. *Economics Letters* 61, 345–349.

Englund, Peter, Torsten Persson & Lars E.O. Svensson, (1992). Swedish business cycles 1861-1988. *Journal of Monetary Economics* 30, 343–371.

Findley, David F., Brian C. Monsell, William R. Bell, Mark C. Otto & Bor-Chung Chen (1998). New capabilities and methods of the X-12-ARIMA seasonal adjustment program. *Journal of Business and Economic Statistics* 16:2, 127–157.

Fiorito, Riccardo & Tryphon Kollintzas (1994). Stylized facts of business cycles in the G7 from a real business cycles perspective. *European Economic Review* 38, 235–269.

Gençay, Ramazan, Faruk Selçuk & Brandon Whitcher (2001). *An Introduction to Wavelets and Other Filtering Methods in Finance and Econometrics*, Hartcourt Publishers Ltd.

Ghysels, Eric, Clive W.J. Granger & Pierre L. Siklos (1996). Is Seasonal Adjustment a Linear or Nonlinear Data Filtering Process? *Journal of Business and Economic Statistics* 14:3, 374–386.

Ghysels, Eric & Pierre Perron (1996). The effect of linear filters on dynamic time series with structural change. *Journal of Econometrics* 70:1, 69–98

Golub, Gene H. & Charles F. van Loan (1996). *Matrix Computations*. The Johns Hopkins University Press.

Gomez, V. & A. Maravall (1996). *Programs TRAMO and SEATS, Instructions for the User*, Bank of Spain, Working Paper n. 9628, Madrid.

Guay, Alain & Pierre St-Amant (1996). Do mechanical filters provide a good approximation of business cycles? *Technical Report* nr. 78, Bank of Canada.

Harvey, Andrew C. & A. Jaeger (1993). Detrending, stylized facts and the business cycle. *Journal of Applied Econometrics* 8, 231–247.

Harvey, Andrew C. & Thomas M. Trimbur (2001). General model-based filters for extracting cycles and trends in economic time series. *DAE Working Papers* 0113, Department of Applied Economics, University of Cambridge.

Hecq, Alain (1998). Does seasonal adjustment induce common cycles? *Economic Letters* 59, 289–297.

Hodrick, R.J. & E.C. Prescott (1997). Post-War U.S. Business Cycles: An Empirical Investigation. *Journal of Money, Credit, and Banking* 29, 1–16.

Kaiser, Regina & Augustin Maravall (1999). Estimation of the business cycle: A modified Hodrick-Prescott filter. *Spanish Economic Review* 1, 175–206.

King, Robert G. & Sergio T. Rebelo (1993). Low frequency filtering and real business cycles. *Journal of Economic Dynamics and Control* 17, 207–231.

Kydland, Finn E. & Edward C. Prescott (1990). Business cycles: Real facts and a monetary myth. *Federal Reserve Bank of Minneapolis Quarterly Review* 14, 3–18.

Laaksonen, Matti (2004). Wavelet Based Seasonal Adjustment. Working Paper (forthcoming).

Lucas, R (1977). Understanding business cycles. In: *Stabilization of the Domestic and International Economy.* Eds K. Brunner & A. Meltzer. Carnegie-Rochester Conference Series, 5, North Holland, Amsterdam.

Pedersen, T.M. (2001). The Hodrick-Prescott filter, the Slutzky effect, and the distortionary effect of filters. *Journal of Economic Dynamics and Control* 25:8, 1081–1101.

Percival, Donald B. & Andrew T. Walden (2000). *Wavelet Methods for Time Series Analysis.* Cambridge University Press.

Prescott, Edward C. (1986). Theory ahead of business cycle measurement. *Carnegie-Rochester Conference Series on Public Policy* 25, 11–66.

Ravn, Morten & Harald Uhlig (2002). On adjusting the HP-filter for the frequency of observations. *Review of Economics and Statistics* 84:2, 371–376.

Reeves, Jonathan J., Conrad A. Blyth, Christopher M. Triggs & John P. Small (2000). The Hodrick-Prescott filter, a generalization, and a new procedure for extracting an empirical cycle from a series. *Studies in Nonlinear Dynamics and Econometrics* 4:1, 1–16.

Schenk-Hoppe, Klaus Reiner (2001). Economic growth and business cycles: A critical comment on detrending time series. *Studies in Nonlinear Dynamics and Econometrics* 5:1, 75–86.

Figures 5–11

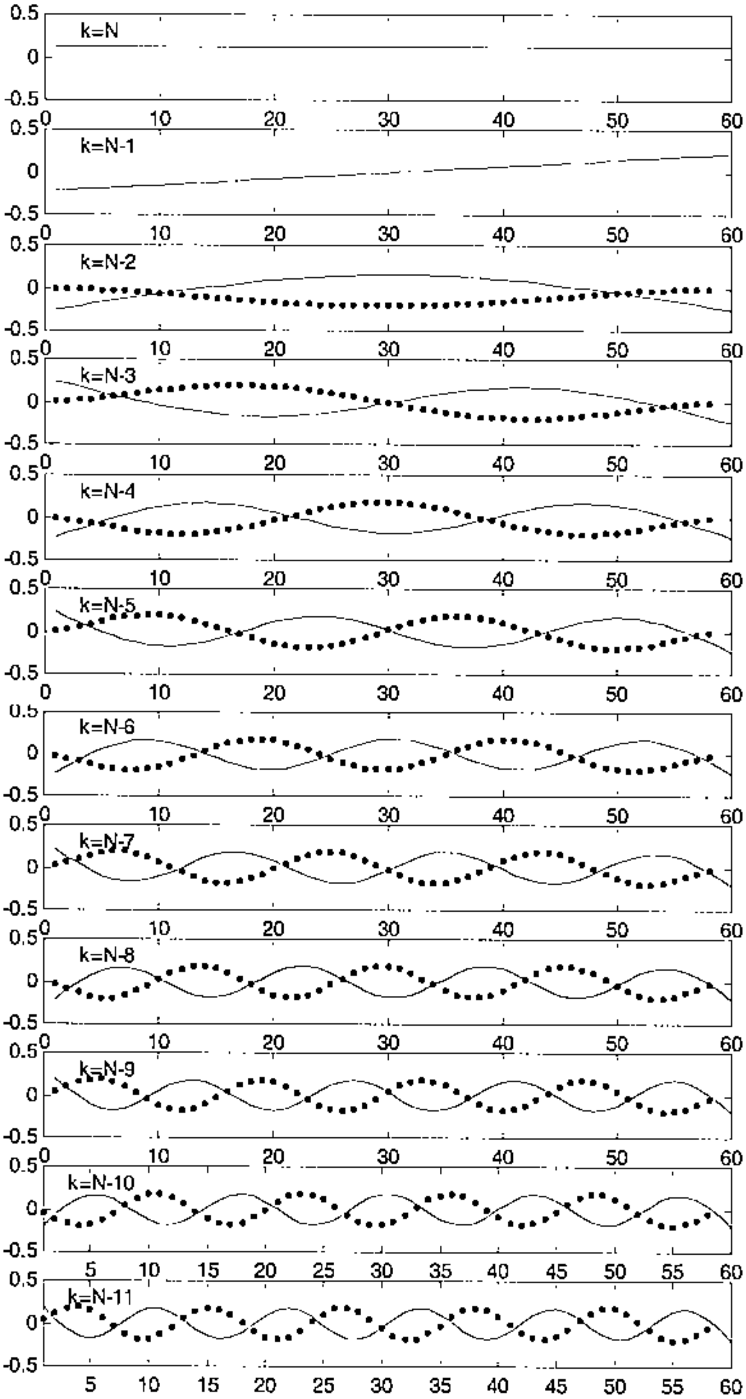


Figure 5. The eigenvectors ψ_k of $\mathbf{H}$ (solid curve) and the corresponding eigenvectors $\psi_k^G = \sigma_k^{-1}\mathbf{L}^T\psi_k$ of $\mathbf{G} = \mathbf{L}^T\mathbf{L}$ (dotted curve) in the case $N = 60$.

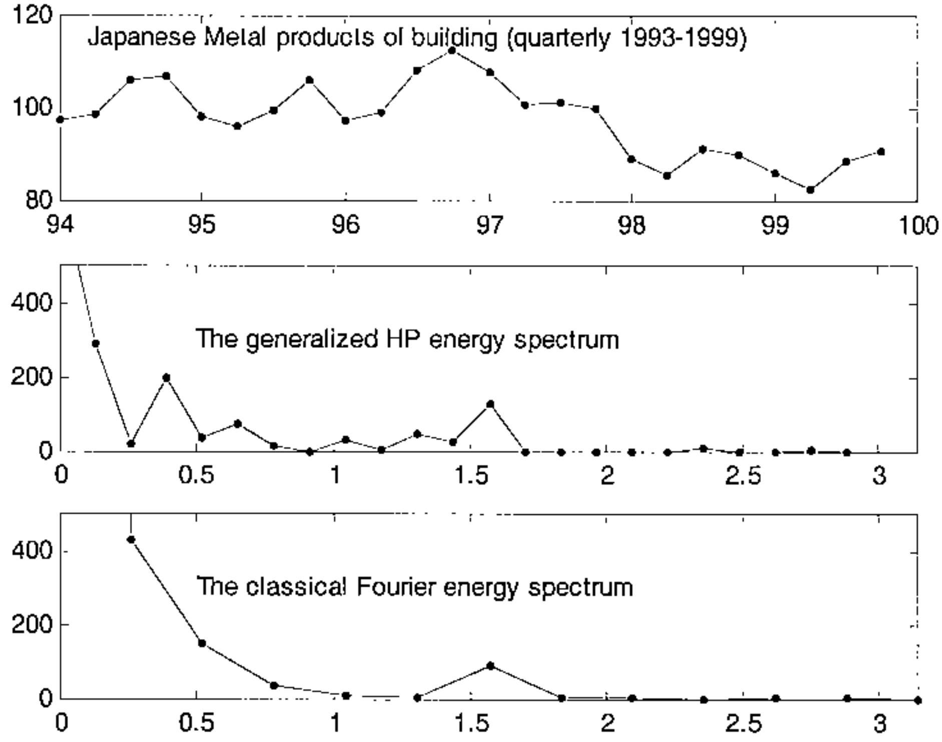


Figure 6a. The generalized HP energy spectrum and the classical Fourier energy spectrum of Japanese metal product of building (quarterly 1993–1999).

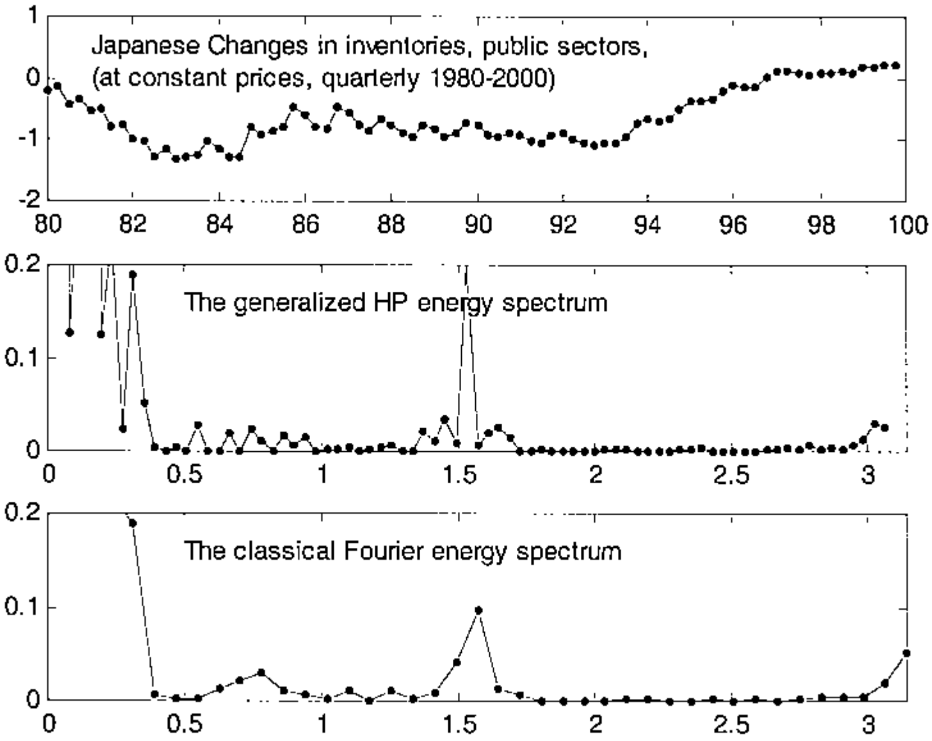


Figure 6b. The generalized HP energy spectrum and the classical Fourier energy spectrum of Japanese inventories, public sectors (constant prices, quarterly 1980–2000).

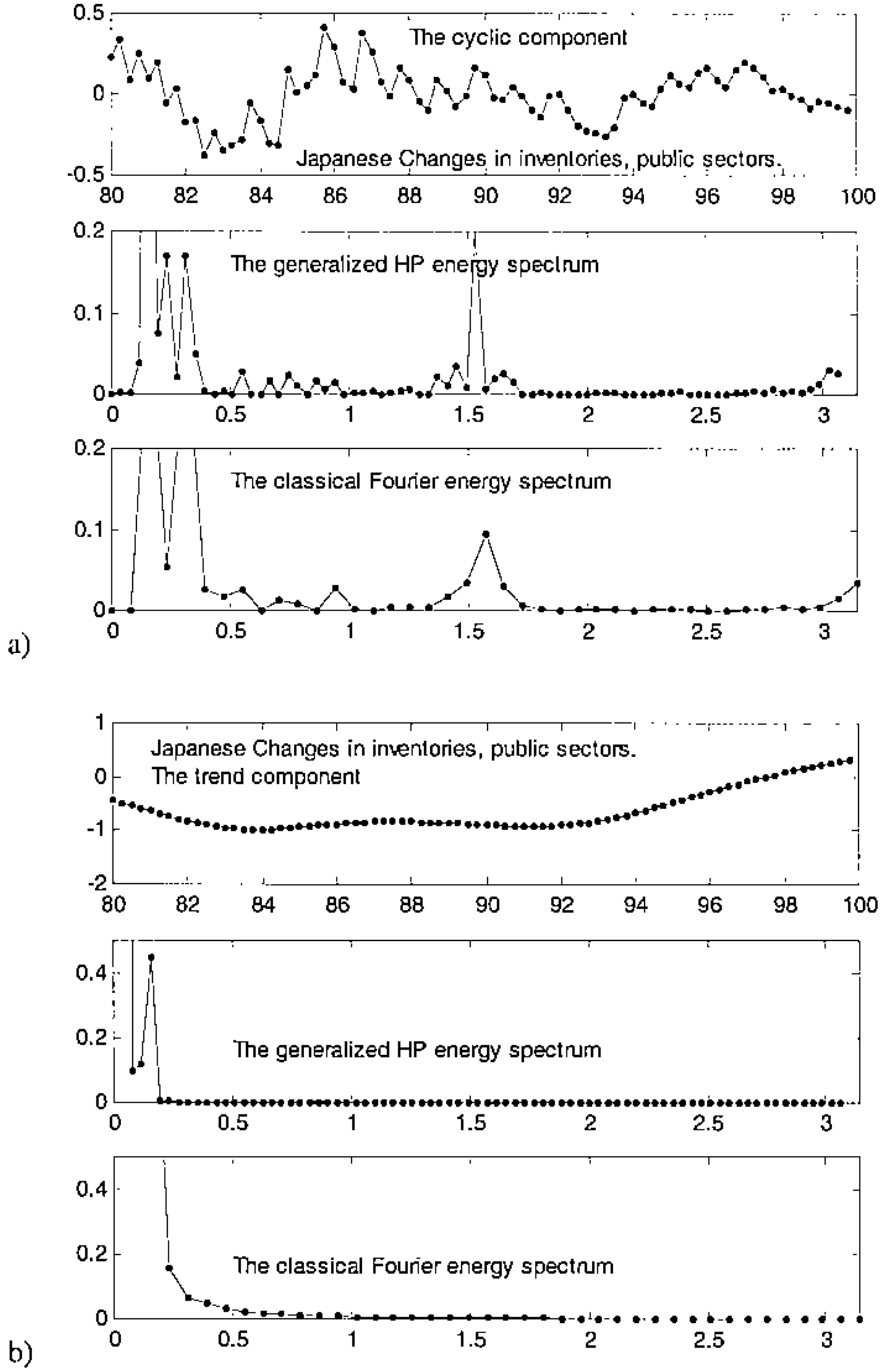


Figure 7. The generalized HP energy spectrum and the classical Fourier energy spectrum of the cyclic (a) and the trend (b) components of the Japanese inventories, public sectors (constant prices, quarterly 1980–2000).

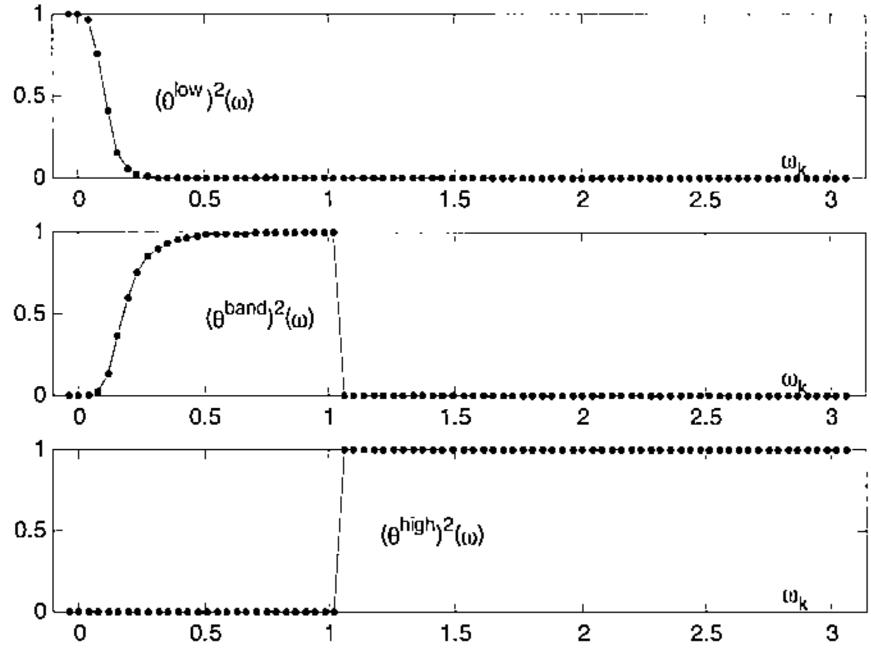


Figure 8. The gains of the low pass operator (42), the band pass operator (43) and the high pass operator (44) with the value of the smoothing parameter $\lambda = 1600$.

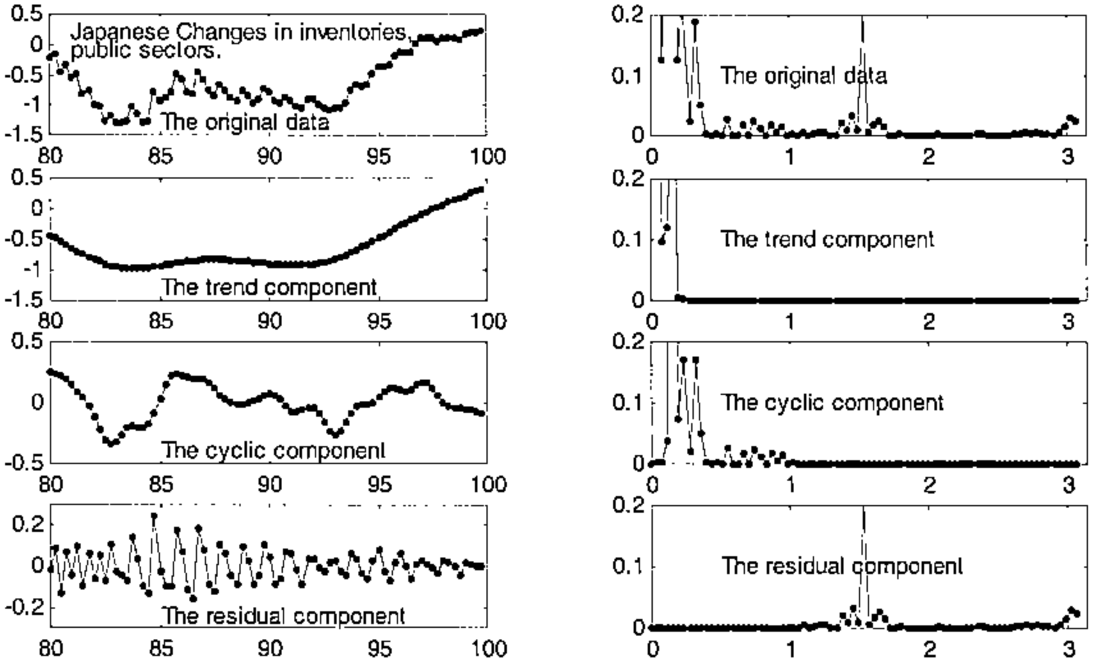


Figure 9. The decomposition of Japanese inventories, public sectors (1980–2000) by classical band pass HP operators with smoothing parameter $\lambda = 1600$. Original data (top row), the trend component (2nd row), the cyclic component (3rd row) and the residual (last row) with their HP energy spectrums (right hand side panel).

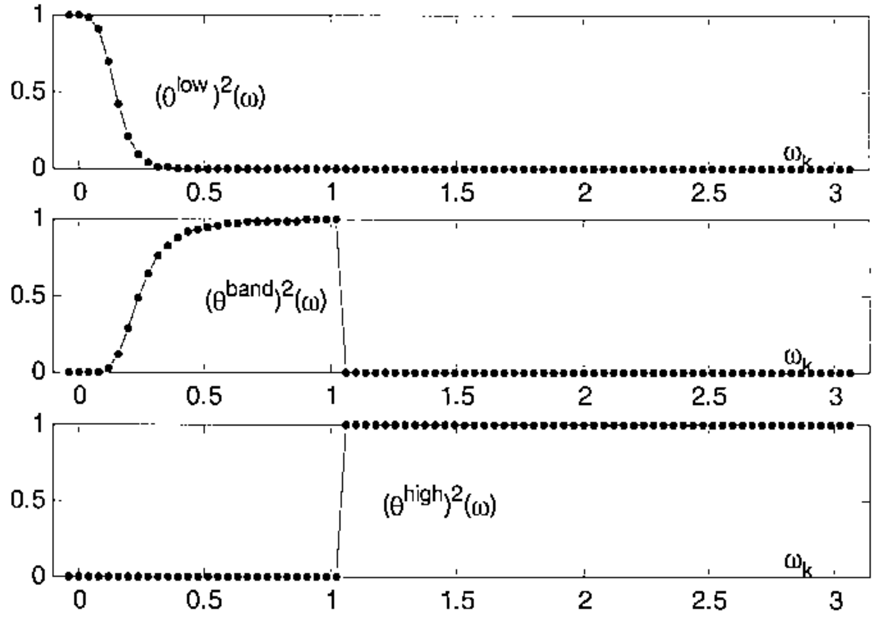


Figure 10. The gains of the low pass operator (42), the band pass operator (43) and the high pass operator (44) with the value of the smoothing parameter $\lambda = 550$.

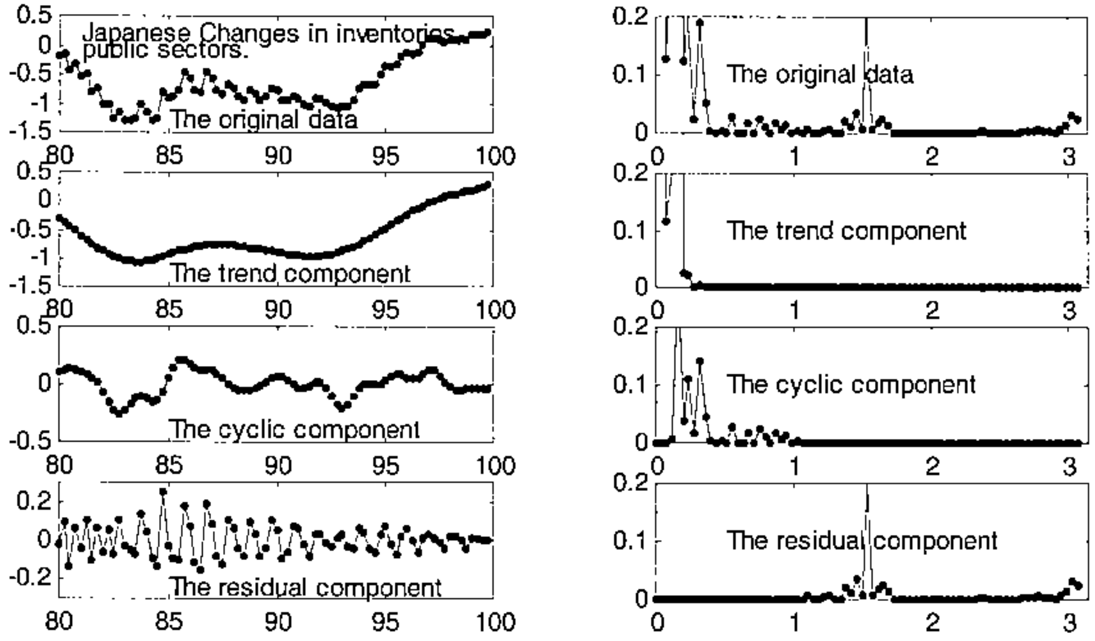


Figure 11. The decomposition of Japanese inventories, public sectors (1980–2000) by classical band pass HP operators with smoothing parameter $\lambda = 550$. Original data (top row), the trend component (2nd row), the cyclic component (3rd row) and the residual (last row) with their HP energy spectrums (right hand side panel).

Modelling nonlinear advertising policy

Irma Luhta

Dedicated to Ilkka Virtanen on the occasion of his 60th birthday

Abstract

Luhta, Irma (2004). Modelling nonlinear advertising policy. In *Contributions to Management Science, Mathematics and Modelling. Essays in Honour of Professor Ilkka Virtanen*. Acta Wasaensia No. 122, 79–94. Eds Matti Laaksonen and Seppo Pynnönen.

In this study there is investigated the effects of discrete versus continuous time modeling and time series analysis in delayed feedback dynamics. As an application we have a time delayed feedback model describing the relations between advertising and goodwill on the basis of Ostrusska-Luhta advertising capital model (Luhta 1997). The advertising function depends on the actual value and on the target value of a firm's goodwill. One version of the model is introduced and analyzed using the concepts of maximal correlation and nonparametric regression analysis.

Irma Luhta, Department of Economics, Turku School of Economics and Business Administration.

Key words: Modeling, marketing, dynamic systems, chaos theory, advertising models, maximal correlation, nonparametric data analysis.

1. Introduction

This study is based on the classical Nerlove-Arrow model. Its purpose is to study the behavior of goodwill defined in accordance with the classical model (Nerlove and Arrow 1962). The classical model and all its optimal control theory extensions are optimization models. In the state space approach of the control theory the purpose is to find the optimal control (see e.g. Virtanen 1982). Sethi (1977) made a comprehensive survey of the literature on dynamic optimal control models in advertising devoted to determining optimal advertising expenditures over time subject to some dynamics that define how advertising expenditures translate into sales and in turn, into profits for a firm or a group of firms under consideration. The survey by Sethi was organized in four model categories which are no longer sufficient. Therefore a new, broader classification was developed

(Feichtinger, Hartl and Sethi 1994). The Nerlove-Arrow model belongs in this classification to the category of capital stocks generated by advertising, price and quality.

The behavior of goodwill, without optimization, was studied by Ostrusska (1990) in an interesting way. In that model, where goodwill was chosen as a state, the development of goodwill was studied by establishing a feedback system between advertising and goodwill and introducing the time lag effect of advertising on goodwill. Luhta and Virtanen (1996) created further this advertising function with several possibilities for different advertising policies. Luhta (1997) analyzed the model when there were three different effects of advertising on the dynamics of goodwill. Another way to study the nonlinear economic relations is based on the analysis of empirical time series (see e.g. Booth, Martikainen, Sarkar, Virtanen and Yli-Olli 1994; Hibbert and Wilkinson 1994). Voss and Kurths (2003) analyzed Ostrusska–Luhta model using nonparametric regression analysis and the concept of maximal correlation.

The linear model is simple, but it doesn't take the increasing marginal cost of goodwill or a saturation level for equilibrium goodwill into consideration. These ideas are with in the nonlinear and in the bounded model. Mahajan and Muller (1986) defined five different advertising policies: blitz, pulsing, chattering, even and pulsing maintenance. The model cosidered in this paper is delayed feedback model. In the terms of Maharan and Muller it is almost "even". If the depreciation rate of goodwill is small and there is no external disturbance (competition) then the behavior of the model tends to a stable and even state.

An argument against discrete time modeling is that there does not exist any natural a priori periodization. The basic equation in a discrete dynamic model $x_{t+T} = f(x_t)$ is very special one. This means that for any value of t there exists a value T such that the knowledge of the state of the system at t is sufficient to determine the state of the system at $t + T$. Of course it is possible to introduce distributed discrete time lags instead of a single one. But by doing this the greatest advantage of the discrete time model, its simplicity, is lost.

Discrete time dynamical models lead to difference equations whereas continuous time dynamical models lead to differential equations. Discrete models yield a qualitatively different behavior than their continuous analogues. While chaotic dynamics in discrete time systems can already occur in one-dimensional systems, the equivalent phenomenon in continuous time is restricted to three and higher dimensional systems. One approach is to use the knowledge about the history of some variables and thus incorporate the delays into the models. These delay-differential equations can be looked at as infinite-dimensional systems because one requires an infinite set of independent numbers to specify initial conditions.

The paper is organized as follows. Section 2 presents the Nerlove-Arrow model and the extension of the model including the continuous advertising expenditure and time delays. In Section 3 is introduced a nonparametric approach to analyze delayed feedback dynamics as an example Ostrusska-Luhta model. The paper ends with a summary in Section 4.

2. The Nerlove-Arrow model

Advertising expenditure is in many ways similar to investments in durable plant and equipment. The latter affects the present and future net revenue of the investing firm. Advertising expenditures on their part affect the present and future demand for the product and, hence, the present and future net revenue of the firm advertising (Nerlove and Arrow 1962: 129). One way to present the temporal differences of the effects of advertising on the demand is to define a stock which is called goodwill and denoted by $g(t)$. Goodwill is supposed to summarize the effects of current and past advertising outlays on demand. If the price of a unit of goodwill is $ 1, then a dollar spent on advertising increases goodwill by an equal amount. On the other hand, a dollar spent some time ago should, according to the previous argument, contribute less. One possible way of presenting this lesser contribution is to say that goodwill depreciates. If it is further assumed that current advertising expenditure cannot be negative and that depreciation occurs at a constant proportional rate, r, we get the equation

$$(2.1) \qquad \frac{dg(t)}{dt} + r \cdot g(t) = a(t) \geq 0,$$

where $a(t)$ is current advertising outlay and $a(t)$ and $g(t)$ are continuous functions of time. Equation (2.1) states that the net investment in goodwill is the difference between the gross investment (current advertising outlay) and the depreciation of the stock of goodwill (Nerlove and Arrow 1962: 130–131).

The Nerlove-Arrow Equation (2.1) assumes that there is no time lag between advertising expenditures and increases in the stock of goodwill. As the demand for the monopolist's product is a function of the stock of goodwill, that implies that the rate of sales at time t totally adjusts itself immediately to the rate of advertising prevailing at time t (Pauwels 1977). Introducing a time lag between the rate of advertising and its effect on the rate of sales leads to a control problem in which the equation of motion is given by an integro-differential equation. There are many generalized Nerlove-Arrow models dealing with distributed time lags in this way (e.g. Bensoussan, Bultez and Naert 1973; Hartl 1984 and Mann 1975 and Pauwels 1977).

If it is assumed that the change of goodwill at time t depends on the amount of advertising in an earlier period a certain time ago, a fixed time delay τ can be incorporated into Equation (2.1) (Ostrusska 1990)

$$(2.2) \qquad \frac{dg(t)}{dt} = a(t-\tau) - r \cdot g(t).$$

In the previous models with time delays it had been assumed that advertising was independent of sales. When the advertising budget was, however, assumed to depend on sales, which, on the other hand, depended on the goodwill stock, the amount of advertising could be described as a function of goodwill.

Many different advertising strategies are possible. Ostrusska's idea was to use an advertising budget of the following type:

$$(2.3) \qquad a(g) = b \cdot g \cdot e^{-\left(\frac{g}{m}\right)^2},$$

where b and m are parameters. A target goodwill is assumed to exist and is denoted with g^*. For the target goodwill g^* it is set some given value, for example it is set to be in some relation to the competitor's goodwill. The corresponding advertising expenditure is a^*. Goodwill increases with the help of advertising and a manager of the firm increases advertising to the maximum possible amount of money $\bar{a}$. The larger the value of the parameter b (b is a scale parameter of the model) is in (2.3) the larger is an advertising budget and the value of $\bar{a}$. The maximum amount of money $\bar{a}$ to be spent on advertising is reached when the value of goodwill is estimated to be $\bar{g}$, i.e. before the target goodwill has been reached. Following model (2.3), the maximum amount of advertising expenditure $\bar{a} = b \cdot \frac{m}{\sqrt{2}} \cdot e^{-\frac{1}{2}}$ is used, when the estimated goodwill is $\bar{g} = \frac{m}{\sqrt{2}}$. It can be seen, therefore, that m is a location/shape parameter of the model.

A still easier form for interpreting the parameters is the model (Luhta 1997):

$$(2.4) \qquad a(g) = b \cdot g \cdot e^{\frac{1}{2}\left(1-\left(\frac{g}{m}\right)^2\right)},$$

where the maximum advertising expenditure $\bar{a} = b \cdot m$ is used when goodwill is m. This means that before the target goodwill g^* has been reached, i.e. when goodwill is estimated to be $g=m<g^*$, the manager invests as much as possible money in advertising.

Luhta assumed that investment in advertising will not be ceased abruptly but, on the contrary, the firm continues to advertise to get a better and better image. Then a more general exponent s can be incorporated into the function (2.4) to generate this property. In this case the advertising function becomes

$$(2.5) \qquad a(g) = b \cdot g \cdot e^{\frac{1}{s}\left(1-\left(\frac{g}{m}\right)^s\right)}.$$

With this formula (see Fig. 1) there will be possibilities for several advertising strategies and one can also look for an optimal advertising strategy by using the above function.

Functions (2.3) and (2.5) can be regarded as reaction functions for the estimated goodwill. When the parameter s is large the advertising flow $a(g)$ decreases quickly after the value m and when the parameter s is small the advertising flow decreases slowly after the value m. Setting the value for the parameter s is a long term decision in a firm, so it is important that the directors define the advertising policy of the firm in the right way when attempting to attain more permanent goodwill for the firm.

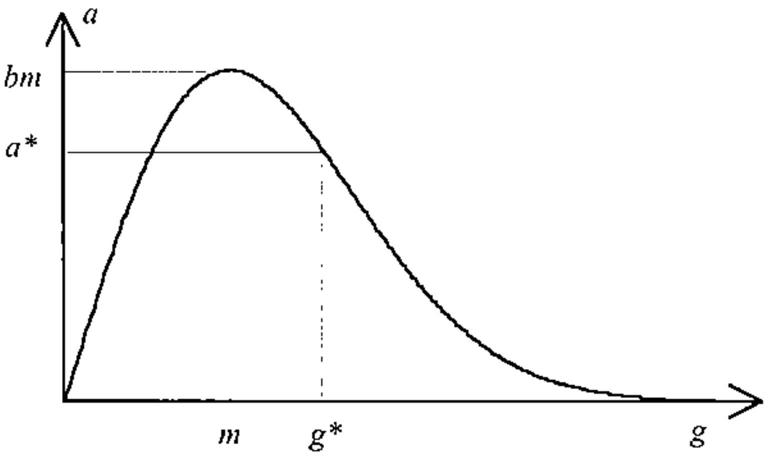


Figure 1. Advertising policy: dependence of advertising outlay on goodwill ($\overline{a} = bm$ and $\overline{g} = m$).

3. Analysis of delayed feedback dynamics

Luhta (1997) studied the classical Nerlove Arrow model e.g. modeling the system (2.2) and (2.5) by delay-difference equation

$$(3.1) \qquad \begin{cases} g_{t+1} = g_t - r \cdot g_t + a(g_{t-\tau}) \\ a(g) = b \cdot g \cdot e^{\frac{1}{s}\left(1-\left(\frac{g}{m}\right)^{s}\right)} \end{cases} .$$

The model behavior changed dramatically if a fixed time lag between change of goodwill and advertising was introduced.

Voss and Kurths (2003) used delay differential equation form of (3.1) as the basis when studying the model

(3.2) $$\dot{g}(t) = -r \cdot g(t) + b \cdot g(t-\tau) \cdot e^{-\frac{1}{2}g^2(t-\tau)}.$$

The dynamics of model (3.2) describes in an infinite-dimensional phase space (Farmer 1982), since a delay-differential equation is a functional-differential equation. This means that for fixing the state of the system at time t_0 (the state of a system determines its future evolution uniquely) it should be provided a function defined on an interval $[t_0 - \tau, t_0]$ as initial condition. Therefore, even this scalar equation can produce high-dimensional dynamics, if the time delay τ is large or the nonlinearity strong (Ikeda and Matsumoto 1987; Farmer 1982). It has even been shown that for very strong nonlinearities the solutions of delay-differential equations with periodic feedback can be described asymptotically by means of stochastic terms, thus mimicking a stochastic ordinary differential equation (Dorizzi 1987). These properties are completely lost if the delay-differential equation is numerically approximated by a simple difference scheme. In this case the attractor dimension cannot exceed the dimension of the resulting map from this kind of discretization, and the results based on such approaches (like bifurcation diagrams, fractal dimensions and statistical properties of the solutions (Dorizzi 1987)) have to be treated with greatest caution. That is the reason why Voss and Kurths (2003) used only quasi-continuous numerical solvers for differential equations in the analyses of model (3.2).

Voss and Kurths used a discrete time sample y_t of a typical solution $g(t)$of equation (3.2) with the time lag τ = 50, r = 1, and b = 6. Using the value 1 for the parameter r means that there is no reserve for goodwill. The idea of Ostrusska-Luhta model (3.1) was that the depreciation rate of goodwill r is smaller than 1. The time lag in this continuous model is really large compared to the discrete one (Luhta 1997). The sample comprised 500 data points with a time step of $\Delta t = 0.2$ (thus, t = 0.2, 0.4, . . . , 100.0) and had been obtained using a Runge-Kutta integrator of fourth order for delay-differential equations. As initial condition a sample of normalized Gaussian white noise was used and the data sample had been taken after a sufficient amount of time such that transient dynamics had been died out (Voss and Kurths 2003).

First Voss and Kurths applied linear techniques to analyze the simulated data, the sample autocorrelation function and the power spectral density according to Priestley (1981). Neither the sample autocorrelation function nor the power spectral density showed an unambiguous indication for a typical time scale of the time lag $\tau = 50$. The time series looked rather erratic and some remaining smoothness that could also be due to some filtering of a completely stochastic signal, it resembled just colored noise. In the following is described an analysis method to distinguish this kind of data from a random signal like colored noise (Honerkamp 1998; Priestley 1981).

3.1. A nonparametric approach to analyze delayed-feedback dynamics

Like Equation (3.2) is a special case of a simple system that exhibits complex dynamics caused by a time-delayed feedback. It belongs to the class of delay-differential equations of the form

$$(3.3) \qquad H(\dot{x}(t)) = F_0(x(t)) + \sum_{i=1}^{k} F_i(x(t-\tau_i))$$

with k different time delays. Since most economic models depend on a single delay it is considered only the single-delay model, described by the delay-differential equation

$$(3.4) \qquad H(\dot{x}(t)) = F_0(x(t)) + F_1(x(t-\tau)) \, .$$

The functions H, F_0 and F_1 are assumed to be continuous. In most models H is the identity. In contrast to systems described by ordinary differential equations, scalar delay-differential equations have a property that can easily be ex-ploited for data analysis. Although system (3.4) can produce high-dimensional chaotic dynamics. The time evolution of the triple $(\dot{x}(t), x(t), x(t-\tau))$ is always constrained to the one-dimensional invariant subspace defined by Equation (3.4). Therefore, if it is possible to estimate the time derivative $\dot{x}(t)$ from a time series y_t accurately it can be found the relationship between the values $\Delta y_t, y_t$ and $y_{t-\tau}$. Here, Δy_t denotes the estimate of the time derivative, e.g. $\Delta y_t = \frac{y_{t+\Delta t} - y_{t-\Delta t}}{2\Delta t}$. This particular estimate is of only low order in accuracy. The analysis of delay-differential systems based on these in-sights was first performed by Bünner, Meyer, Kittel and Parisi (1996, 1997). The triple $(\Delta y_t, y_t, y_{t-\tau})$ can be seen as a three-dimensional embedding vector where the first component comes from a differential embedding and the second and third components come from a delay-embedding (Packard, Crutchfield, Farmer and Shaw 1980). Therefore, for the reconstruction of a scalar delay-differential equation (supposed, the measurement function is one-to-one), such a three-dimensional mixed embedding always suffices, no matter how large the attractor dimension actually is. This is the property that will be exploited in the following (Voss and Kurths 2003).

Voss and Kurths found a relationship of the form

$$h(\Delta y_t) = f_0(y_t) + f_1(y_{t-\tau}), \tag{3.5}$$

to be used as an estimate of Equation (3.4). This equation expresses the inverse problem to equation (3.4), i.e., the problem of estimating a delay-differential equation from data. In principle, it can be used any numerical scheme for function estimation to yield estimates for h, f_0 and f_1. Voss and Kurths used a nonparametric technique that is based on multiple nonlinear regression analysis. In contrast to parametric techniques, where coefficients of a given model equation are estimated, this nonparametric approach yields function estimators through a minimization of a distance in function space rather than in coefficient space. Therefore it is quite flexible in providing model estimates also when there are few a priory assumptions about the model (Voss and Kurths 2003).

Next the multiple nonlinear regression analysis is performed using Rényi's concept of maximal correlation (Hirscfeld 1935; Gebelein 1941; Rényi 1970). This statistical quantity measures the dependence between two random variables x_0 and x_1 :

$$\Psi(x_0, x_1) = \sup_{\Phi_0^*, \Phi_1^*} \left| R(\Phi_0^*(x_0), \Phi_1^*(x_1)) \right|, \tag{3.6}$$

where R is the linear correlation coefficient

$$R(x_0, x_1) = \frac{E[x_0 x_1] - E[x_0]E[x_1]}{\sqrt{E(x_0 - E[x_0])^2 E(x_1 - E[x_1])^2}}. \tag{3.7}$$

Here $E[.]$ denotes the expectation (or mean) value. To obtain the supremum, the functions Φ_0^* and Φ_1^* are varied in the space of Borel measurable functions. It can be thought of any possible function that is not somehow pathologically defined. Also discontinuities and jumps are allowed. The only requirement or constraint onto these functions is that they have finite non-vanishing variances, i.e., $0 < \mathrm{var}\left[\Phi_j^*(x_j)\right] < \infty \ (j = 0,1)$. This ensures that trivial solutions are excluded. Finite variances can be achieved, for example, by fixing the variance of the first function to $\mathrm{var}\left[\Phi_1^*(x_1)\right] = 1$. Taking the absolute value in the definition of the maximal correlation is not necessary and has historical reasons. It reminds that $\Psi(x_0, x_1) > 0$ and that there is, therefore, nothing like anti-correlation in this generalized notion of correlation. The maximal correlation can be interpreted easily. The functions $\Phi_0(x_0)$ and $\Phi_1(x_1)$ for which the supremum is attained maximize the linear correlation between the random variables x_0 and x_1. By definition, the value of Ψ is restricted to the interval [0,1]. The maximal correlation has three important properties: (i) It vanishes if and only if the variables are independent. (ii) If there are functions Φ_0 and Φ_1 such that $\Phi_0(x_0) = \Phi_1(x_1)$, the maximal correlation attains unity. (iii) If there is a linear relationship between x_0 and x_1, it follows that $\Psi = |R|$. Thus, the maximal correlation constitutes a generalized measure of statistical dependence between two random variables (Voss and Kurths 2003).

It will also be needed a generalization for multivariate regression problems:

$$(3.8) \qquad \Psi(x_0,\ldots,x_k) = \sup_{\Phi_0^*,\ldots,\Phi^*_k} \left| R(\Phi_0^*(x_0), \sum_{i=1}^{k} \Phi_i^*(x_i)) \right|.$$

This is not the most general possibility for a definition of a generalized measure of dependence for vectorial random variables. However, the following properties are still valid: (i') If $\Psi(x_0,\ldots,x_k)$ vanishes, x_0 and x_i ($i = 1,\ldots,k$) are pairwise independent. (ii') If there are functions $\Phi_0,\ldots,\Phi_k$ such that $\Phi_0(x_0) = \sum_{i=1}^{k} \Phi_i(x_i), \Psi(x_0,\ldots,x_k)$ attains unity. (iii') If there is a multiple linear relationship between all of the x_i ($i = 0,\ldots,k$), it follows that $\Psi(x_0,\ldots,x_k) = |R'|$, where R' is the multiple correlation coefficient as defined accordingly. Voss and Kurths called $\Psi(x_0,\ldots,x_k)$ "multivariate maximal correlation".

The problem of calculating the multivariate maximal correlation (3.8) is equivalent to the estimation of optimal transformations from data. To accomplish this Voss and Kurths used the multivariate case of the alternating conditional expectation algorithm (ACE), invented by Breiman and Friedman (1985). Voss and Kurths assumed that the time lag τ is known. By solving Equation (3.8) for $x_0 = \Delta y_t, x_1 = y_t$ and $x_2 = y_{t-\tau}$, the ACE algorithm provided function estimators for h, f_0 and f_1 . In the notation there will be no difference between the estimates of a function and the function itself. The estimate for the multivariate maximal correlation became

$$(3.9) \qquad \Psi_\tau = |R(h(\Delta y_t), f_0(y_t) + f_1(y_{t-\tau})|.$$

Voss and Kurths estimated first Ψ_τ for every reasonable τ . If the maximum of Ψ_τ was close to one, its argument could be seen as the most likely time lag τ . If none of the Ψ_τ appeared to be close to one, the time series could not be the solution of a delay-differential equation of the form (3.4), or the data did not possess some minimum requirements. These are that the sampling rate should be high enough to account for proper estimates of the time derivatives, and that the noise level is not too high. Once having estimated a time lag, the model estimate was given by the function estimators h, f_0 and f_1 for the estimated time lag τ.

The validation of the resulting model could be done by introducing the estimated terms in already known models and try to improve the performance of these models, or to simply fit parameters a posteriori to the function estimators. Often already the estimated delay time is such a parameter whose value is not known exactly. Voss and Kurths (2003) applied this technique to the model (3.2).

3.2. Analysis of Ostrusska-Luhta model

Analyzing the model (3.2) Voss and Kurths (2003) first estimated the time delay τ. The maximal correlation Ψ_τ was calculated via equation (3.8) for varying values of the time delay from $\tau = 1$, $\Delta t = 0.2$ to $\tau = 480$, $\Delta t = 96$. As a result it was found $\Psi_{\tau=50} = 0.994$ as the maximum value. Thus the peak pointed exactly to the given model time delay of τ = 50.

The large values for Ψ_τ for very small τ were simply due to the fact that the very local dynamics can be described satisfactorily also by an equation without a feedback term, like equation (2.1). Voss and Kurths found that looking only at very short pieces of the solution trajectory, it is not clear whether the dynamics has been produced by an ordinary differential equation or a delay-differential equation. This is a consequence of the ill-posed nature of the inverse problem of estimating differential equations from data and not a flaw of the proposed procedure. Voss and Kurths believed that the estimates for the functions can be considered reliable, providing the data cover a large enough area of the subspace described by the delay-differential equation. This should generally be the case for long enough time series with chaotic dynamics. On the other hand, for very large values of the tested delay, Ψ_τ rose steadily. This is an effect of overfitting due to a too short data set. For $\tau = 480$, $\Delta t = 96$ there were only 19 triples $(\Delta y_t, y_t, y_{t-\tau})$ left for estimating the entire functions. It is recommended to treat results for very short data sets with care. For a similar model, Voss and Kurths (1997) have performed a study for very short data sets, but it is difficult to give rigorous rules for the minimum number of data points necessary for the analysis.

For the estimated time delay $\tau = 50$, the quantities $\Delta y_t, y_t$ and $y_{t-\tau}$ had almost entirely been transformed by the optimal transformations to a linear relationship. The remaining small scatter was caused mostly due to an inaccurate estimation of time derivatives from the data. This problem increases severely in the case of noisy data, there is no rigorous rule of how much noise may disturb the data unless the procedure fails to work (Voss and Kurths 1997; Voss, Kolodner, Abel and Kurths 1999), but it has been shown that a proper filtering of the data can yield considerable improvement (Voss 1998; Voss et al. 1999). Voss and Kurths (2003) found the optimal transformations that led to the linear relationship. The first one was to a very high accuracy simply the identity function $\Delta y_t \rightarrow \Delta y_t$, as expected from the model (3.2). The second optimal transformation gave, also to a very high accuracy, an estimate of the depreciation term $-rg(t)$. This result did not only show the correct slope of $r = 1$, but also that depreciation entered linearly into the dynamics. There are also models with a nonlinear depreciation function (Luhta 1997). This result was possible due to the nonparametric approach. For analyzing the data no assumptions have been made about the form of the depreciation function. Finally, the nonlinear advertising function $6g(t-50)e^{-\frac{1}{2}g(t-50)}$ was estimated from the data with a very good accuracy. (Voss and Kurths)

4. Summary

Starting from the classical advertising capital model, the Nerlove-Arrow model, a time delayed feedback model between advertising and goodwill was introduced. It was assumed that the effect of advertising on goodwill is linear. Then Ostrusska-Luhta version of this classical model was studied by maximal correlation and nonparametric regression analysis. The aim was to search relevant feedback times in the process underlying data. In the study was used a delay-differential equation form. It should be interesting to use the method in the future for the other versions of Ostrusska-Luhta model.

References

Bensoussan, A. & A. Bultez & P. Naert (1973). A generalization of the Nerlove-Arrow optimality condition. In *Technical Report*. European Institute of Advertising Studies in Management, Brussel.

Booth, G. & T. Martikainen, S. Sarkar, I. Virtanen & P. Yli-Olli (1994). Nonlinear dependence in Finnish stock returns. *European Journal of Operational Research*, 74.

Breiman, L. & J.H. Friedman (1985). Estimating optimal transformations for multiple regression and correlation. *J. Am. Stat. Assoc.* 80, 580–619.

Bünner, M.J., T. Meyer, A. Kittel & J. Parisi (1996). Tool to recover scalar time-delay systems from experimental time series. *Phys. Rev. E* 54, R3082–3085.

Bünner, M.J., T. Meyer, A. Kittel & J. Parisi (1997). Recovery of time-evolution equations of time-delay systems from time series *Phys. Rev. E* 56, 5083–5089.

Dorizzi, B. et al. (1987). Statistics and dimension of chaos in differential delay Systems. *Phys. Rev. A* 35, 328–339.

Farmer, J.D. (1982). Chaotic attractors of an infinite dimensional dynamical system. *Physica* D 4, 366–393.

Feichtinger, G. & Hartl R.F. & Sethi S.P. (1994). Dynamic Optimal Control Models in Advertising: Recent Developments. *Management Science* 2:2.

Gebelein, H. (1941). Das statistische Problem der Korrelation als Variations- und Eigenwertproblem und sein Zusammenhang mit der Ausgleichsrechnung, *Z. angew. Math. Mech.* 21, 364–379.

Hartl, R.F. (1984). Optimal dynamic advertising policies for hereditary processes. *Journal of Optimization Theory and Applications*, 43.

Hibbert, B. & I. Wilkinson (1994). Chaos theory and the dynamics of marketing systems. *Journal of the Academy of Marketing Science* 22:3.

Hirschfeld, H.O. (1935). A connection between correlation and contingency. *Proc. of the Camb. Phil. Soc.* 31, 520–524.

Honerkamp, J. (1998). *Statistical Physicss* Berlin: Springer.

Ikeda, K. & K. Matsumoto (1987). High-dimensional chaotic behavior in systems with time-delayed feedback. *Physica* D 29, 223–235.

Luhta, I. (1997). *Structural Changes in a Complex System. A Chaos-analytic Study of a Nonlinear Advertising Policy.* Acta Wasaensia 52. Mathematics 6. Operational Research. University of Vaasa.

Luhta, I. & I. Virtanen (1996). Non-lincar advertising capital model with time delayed feedback between advertising and stock of goodwill. *Chaos, Solitons & Fractals* 7, 2083–2104.

Mann, D.H. (1975). Optimal theoretic advertising stock models: a generalization incorporating the effects of delayed response from promotional exenditure. *Management Science*, 21.

Mahajan, V. & E. Muller (1986). Advertising pulsing policies for generating awareness for new products. *Marketing Science*, 5.

Nerlove, M. & K. Arrow (1962). Dynamical advertising policy under dynamic conditions. *Economica*, 29.

Ostrusska, D. (1990). Modelling Nonlinear Dynamical Systems in Economics. Unpublished.

Packard, N.H., J.P. Crutchfield, J.D. Farmer & R.S. Shaw (1980). Geometry from a time series. *Phys. Rev. Lett.* 45, 712–716.

Pauwels, W. (1977). Optimal dynamic advertising policies in the presence of continuously distributed time lags. *Journal of Optimization Theory and Applications*, 22.

Priestley, M.B. (1981). *Spectral Analysis and Time Series.* San Diego: Academic Press.

Rényi, A. (1970). *Probability Theory.* Budapest: Akadémiai Kiadó.

Sethi, S.P. (1977). Dynamic optimal control models in advertising: A survey. *SIAM Review* 19:4.

Virtanen, I. (1982). Optimal maintenance policy and planned sale date for a machine subject to deterioration and random failure. *European Journal of Operational Research*, 9.

Voss, H. & J. Kurths (1997). Reconstruction of nonlinear time delay models from data by the use of optimal transformations. *Phys. Lett. A* 234, 336–344.

Voss, H. (1998). *Nichtlineare statistische Methoden zur Datenanalyse.* PhD. Thesis, Universit¨at Potsdam.

Voss, H. & J. Kurths (1999). Reconstruction of nonlinear time delay models from optical data. *Chaos, Solitons & Fractals* 10, 805–809.

Voss, H., P. Kolodner, M. Abel & J. Kurths (1999). Amplitude equations from spatiotemporal binary-fluid convection data. *Phys. Rev. Lett.* 83, 3422–3425.

Voss, H. & J. Kurths (2003). *Analysis of Economic Delayed-feedback Dynamics.* Available: http://www.fdm.uni-freiburg.de/~hv/p12.pdf.

The price-volume behavior of an equity: theoretical approach

Martti Luoma, Jussi Nikkinen, and Petri Sahlström

Dedicated to Ilkka Virtanen on the occasion of his 60th birthday

Abstract

Luoma, Martti, Jussi Nikkinen, and Petri Sahlström (2004). The price-volume behavior of an equity: theoretical approach. In *Contributions to Management Science, Mathematics and Modelling. Essays in Honour of Professor Ilkka Virtanen.* Acta Wasaensia No. 122, 95–105. Eds Matti Laaksonen and Seppo Pynnönen.

In spite of the fact that the technical analysis has been used for decades to analyse stocks and other commodities, the theoretical development in the area is very limited. The purpose of this article is to present the theoretical background for the price-volume behavior of a stock which is a well-known tool in technical analysis. The theoretical background stems from the supply and demand curves of the economics literature that are used to investigate how a particular market reacts to changes in supply or demand. By applying theory of supply and demand we build a theoretical background for important rules of technical analysis. Until now the justification of the rules has been purely empirical. Moreover, the theoretical background introduced helps to understand the dynamics of the equity market where the flow of information has an essential role. Since our analysis has only a few first steps for depicting the determination of equity prices by behavior of sellers and buyers, we hope that the article will encourage work in this important research area.

Martti Luoma, Department of Mathematics and Statistics, University of Vaasa, P.O. Box 700, FIN–65101.
Jussi Nikkinen and *Petri Sahlström*, Department of Accounting and Finance, University of Vaasa, P.O.Box 700, FIN-65101.

Key words: technical analysis, supply and demand curves, volume.

1. Introduction

In spite of the fact that the technical analysis has been used for decades to analyse stocks and other commodities, the theoretical development in the area is very limited. Tools and methods in technical analysis are developed in practise and most of them are "ad hoc" in nature, lacking any theoretical justification. This is a serious drawback, since without theoretical background we cannot know, for example, how a tool used to analyse stocks works as a stock market environment changes.

From the academic point of view the lack of theoretical background has been one of the key reasons why the academic research in the area is rather limited. Moreover, the research conducted is only able to answer the question whether some tool works in a specific market environment used in empirical analysis. Consequently, the results, without any theoretical justification, are basically valueless for the users of the technical analysis since the market conditions change all the time.

The situation, however, is changing in the academic research as the so-called behavior finance literature grows. In this area, the purpose is to investigate how and why different types of market participants are doing their trades in the market-place. This is important for the users of technical analysis since the methods used are based on market information, such as trading volumes, which is caused by the traders in the market place. I.e. trading behavior is behind the information used in technical analysis.

One example of this interesting research area is a study by Gervais, Kaniel, and Mingelgrin (2001) investigating whether the trading volume of a common stock can be used to predict stock returns. Based on the *visibility hypothesis* by Miller (1977) they argue that in future the stocks with higher (lower) than normal trading volume will have better (worse) returns than other stocks. The main point in the visibility hypothesis is not the trading volume itself, but instead the *visibility* observed in the trading volume. In other words, Miller (1977) states 'In theory, high volume does not indicate that the stock will rise and merely observing heavy trading volume should not cause anyone to buy. However, if the volume does attract attention and cause more people to look at a stock, some are likely to persuade themselves that the stock should be bought.' This indicates that as the visibility increases, especially for small firms, their stock prices increase. The empirical results by Gervais, Kaniel, Mingelgrin (2001) support the visibility theory.

The purpose of this article is to present the theoretical background for the price-volume behavior of a stock. This price-volume behavior is a well-known tool in technical analysis.

The theoretical background stems from the supply and demand curves of the economics literature which are used to investigate how a particular market reacts to changes in supply or demand.

2. Empirical criteria for the price-volume behavior of a stock

A common notion in handbooks of technical analysis (e.g. Elder 1993) is that increasing trading volume strengthens trend, i.e. increases the probability that the ongoing trend will continue. Conversely, falling trading volume signals that the current trend is going to reverse, i.e. the probability that a trend reversal is going to happen increases. More precisely we analyze four empirical price-volume rules generally used in the practice (see e.g. Young 2000). The rules are as follows:

1. Increasing volume on increasing price indicates possible price increase.
2. Increasing volume on decreasing price indicates possible price decline.
3. Decreasing volume on increasing price indicates possible price reversal or sideways movement.
4. Decreasing volume on decreasing price indicates possible price reversal or sideways movement.

These rules are frequently used in technical analysis. The rules are clear and consistent but they are difficult to use in practise. Much experience is needed to use them. This is partly because volume has rather a big random component. It can be seen in any graphic presentation, for example in Figure 1. We have drawn a grid and 20 days' moving average of volume for better visualization. Healthy uplegs and healthy downlegs should have increasing respective declining volume. According to Figure 1 a volume of Tietoenator from Helsinki Stock Exchange trend has been increasing until mid-November and then it has begun to decrease. After mid-November we have to follow prices very carefully. Uplegs should have increasing or big volume and downlegs should have decreasing or small volume. Not before last downleg in January can we see unhealthy development. It follows a downward breakout with increasing volume.

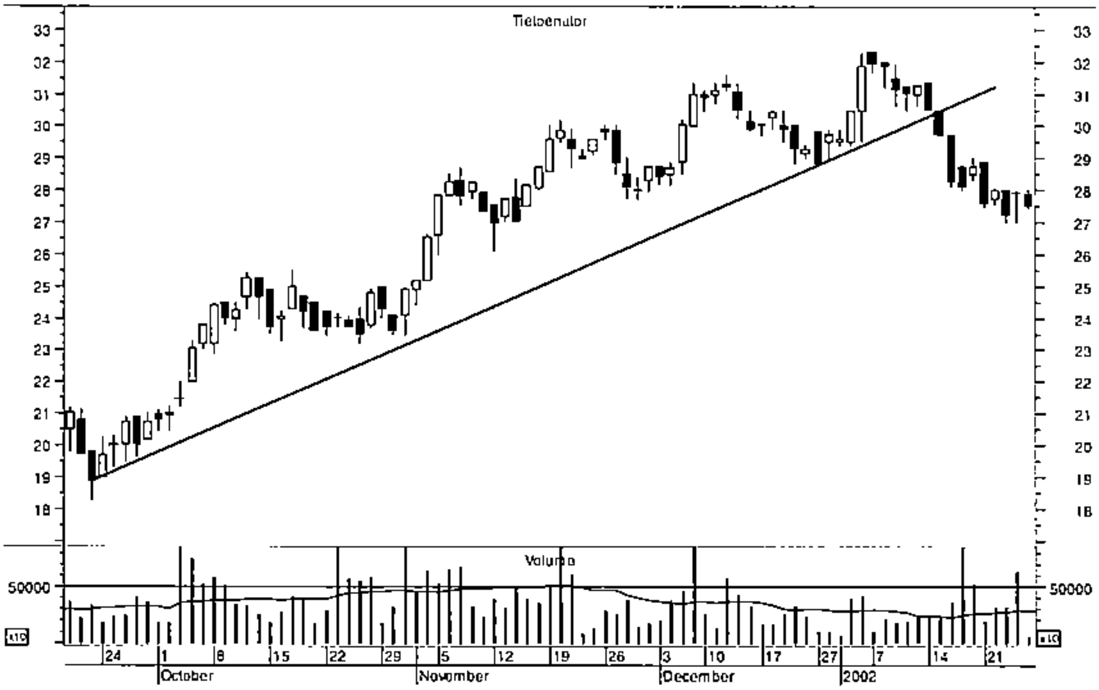


Figure 1. Tietoenator with increasing trend.

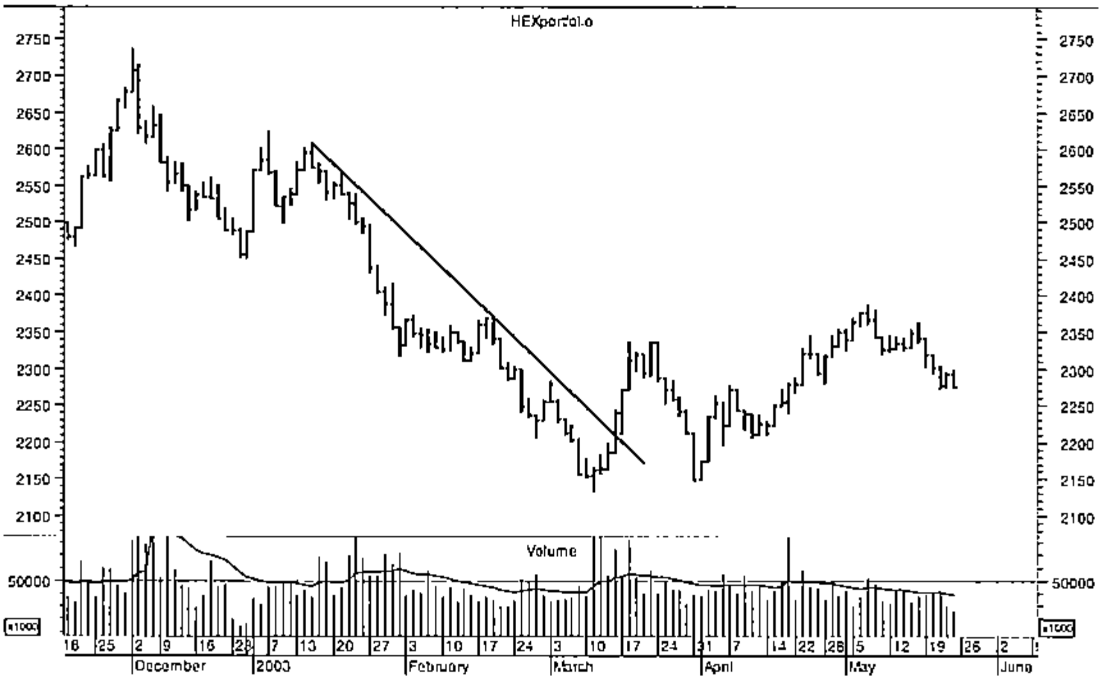


Figure 2. Hex portfolio index with a downtrend.

In Figure 2, the HEX portfolio index of Helsinki Stock Exchange has a volume trend that began to decrease from the beginning of February, signaling that the probability of a price trend reversal has grown. Trend breakout happens with increasing volume fairly soon after a volume trend has began to increase.

These two examples are rather easy to interpret using four rules presented above and other knowledge about technical analysis. Usually a graph is much more difficult to interpret. In this article, the objective is to find a theoretical background for these four rules.

3. Supply and demand curves

In economics financial assets are considered as commodities. Supply and demand of an ordinary commodity is presented in Figure 3 (see any text-book in economic analysis, for example Varian 1999). The curve "Demand" represents the dependence of the demand on the price, *ceteris paribus* and the curve "Supply" presents the dependence of the supply on the price, *ceteris paribus*. Consequently, the interpretation of supply and demand curves is based on the idea that one curve moves and the other remains same, i.e. the equilibrium point moves to the right or to the left.

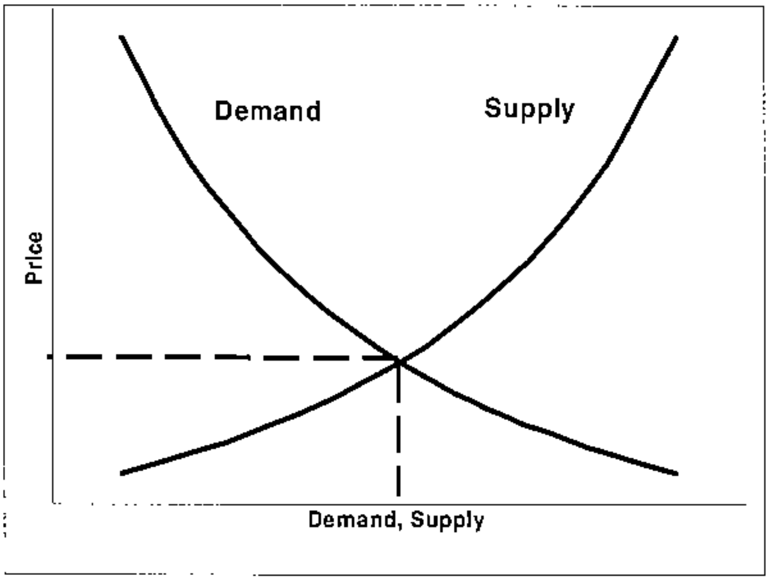


Figure 3. Typical demand-supply curves with an equilibrium point.

On asset markets, information arrives both during the trading day and overnight. As a consequence of this information arrival, the demand curve does not remain unchanged but changes. If the information is not very exceptional and the liquidity of the financial asset is high enough, the demand and the supply will be in equilibrium at the end of the trading day. Moreover, the daily changes will be approximately equal in size. If the supply is greater than the demand, the prices will fall from Pa to Pb as shown in Figure 4. If, on the other hand, the demand is greater than the supply, the prices rise from Pb to Pa. The problem is the fact that on financial markets supply and demand are not known in advance. There are always as many sales as purchases and consequently, the exchange trading volume does not help in clarifying, the supply-demand situation. A questionnaire survey could in principle be used to analyze the relation of demand and supply. In practice it is not a useful approach, however, and therefore, practical and theoretical considerations about price-volume behavior are needed.

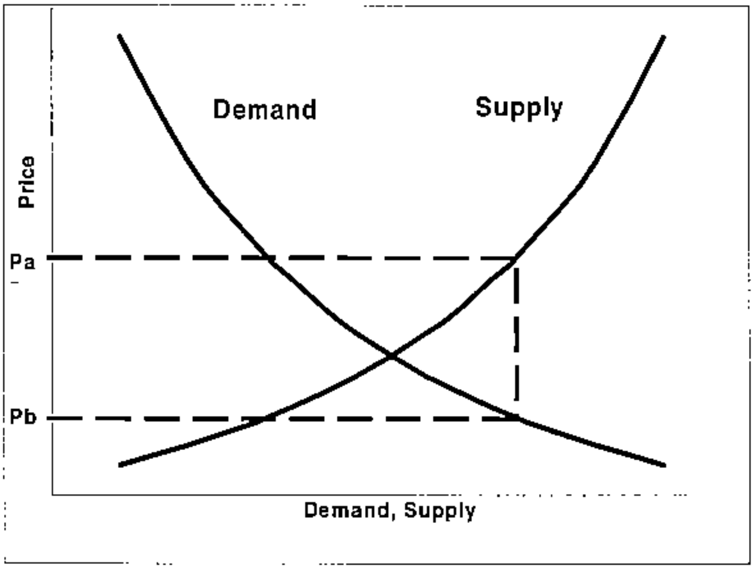


Figure 4. If supply is greater than demand the price falls (from Pa to Pb). If demand is greater than supply the price rises (from Pb to Pa).

4. Uptrend

Applying demand curves and supply curves to analyze the behavior of stock prices is problematic since along with the new information the curves may change. However, the trend represents a certain form of stability. For this reason, the trend and its evolvement can be perhaps described and can be understood by analyzing the curves. It is assumed that no unexpected news regarding either the company or the macroeconomic environment is released. The stock becomes popular and new buyers appear in a steady flow. Demand leads to further demand and the trend continues as long as new buyers appear. Figure 5 illustrates the situation of the upward trend. As can be seen in the picture, the trading volume also increases. The trend in question is often called "a healthy trend" in technical analysis because the trend will with high probability continue. This corresponds to the first empirical criterion above. It is assumed that the supply curve remains the same. This will be a reasonable assumption provided that the time span is not too long, since in the case of a rising trend the buyers' behavior determines the price level. The demand curve changes quite regularly way while the supply curve stays the same. For the sake of simplicity the random component of the price as well as the cyclical component are ignored here and also in the future examinations. Figure 1 is a representative example of an empirical uptrend. To remove noise and cyclical variation in volume figures, the 20-day moving average is presented. The figure shows that the moving average increases from the turn of September-October, i.e. nearly from the beginning of the trend, to the end of November. This part represents the healthy part of the trend and indicates that the trend will continue.

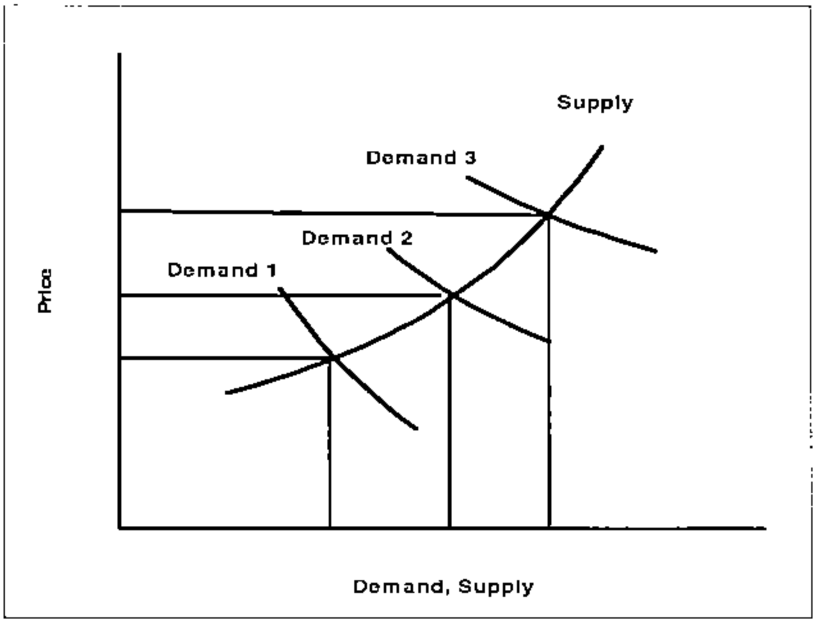


Figure 5. Healthy uptrend.

5. Downtrend

In the case of the downward trend the sellers are the ones who determine the trade price. When the market is decreasing, new sellers appear on the market in a steady flow. While nothing has essentially changed in companies, it is the attitude of the owners regarding the share which has changed. Figure 6 represents the case of a downward trend. The demand curve stays the same but the supply curve will change when the sellers' behavior determines the behavior of the financial asset prices. The figure again shows a stage of the healthy trend, the volume increases, in other words more sellers appear on the market. This corresponds to the second empirical criterion. Figure 2 shows the development of the HEX portfolio index as an example of a downward trend. It can be seen that the trend begins around the middle of January. At the same time, the trading volume increases and increases throughout January. This indicates that the trend is likely to continue. Figure 2 shows that the trend has still indeed lasted all through February.

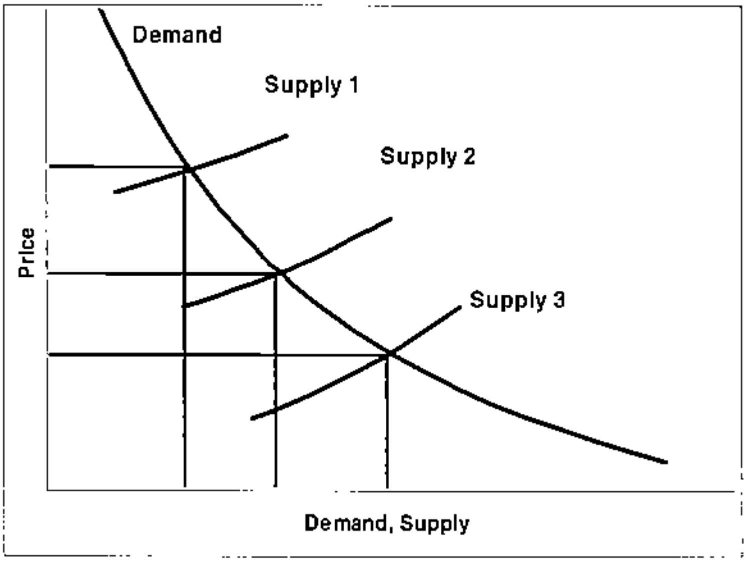


Figure 6. Healthy downtrend.

6. Ending of the uptrend

When a rising trend dominates, the new buyers flow evenly onto the market. It raises the price of a financial asset which in Figure 7 is seen as the transition of the demand curve.

At some stage of the upward trend, the share owners become aware of the rising trend of the stock. The share has become 'hot' or popular and the current shareholders become greedy and are not willing to sell at the price indicated by the current supply curve. This leads to the upward transition of the supply curve as presented in Figure 7. Soon the emergence of new buyers also will begin to decrease, which in the figure implies the smaller transition of the demand curve. From the figure it can be seen that the trading volume becomes smaller, which is in accordance with the third empirical criterion. When the investors start to extensively cash their stock positions, the stock price will begin to decrease. In our example (Figure 1) trading with Tietoenator becomes smaller from the middle of November to the middle of January, in which phase the trend line breaks.

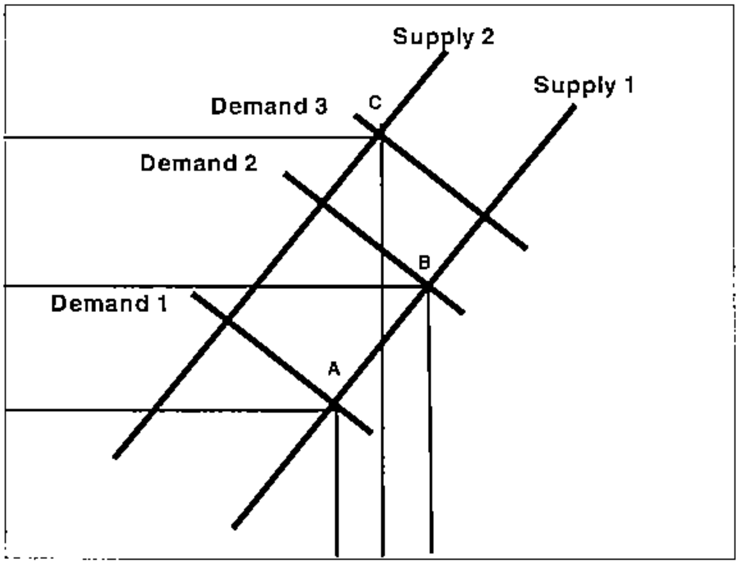


Figure 7. Ending of a uptrend.

7. Ending of the downtrend

The share will get a bad reputation after the downtrend has taken long enough and nobody will want to buy it. This leads to a decrease in the demand curve in Figure 8. Both price and volume decrease which is in accordance with the fourth empirical criterion. At some stage "wise money" considers the share underpriced and begins to buy it. At least the first stage of this phase is easily hidden by the pessimism on the market.

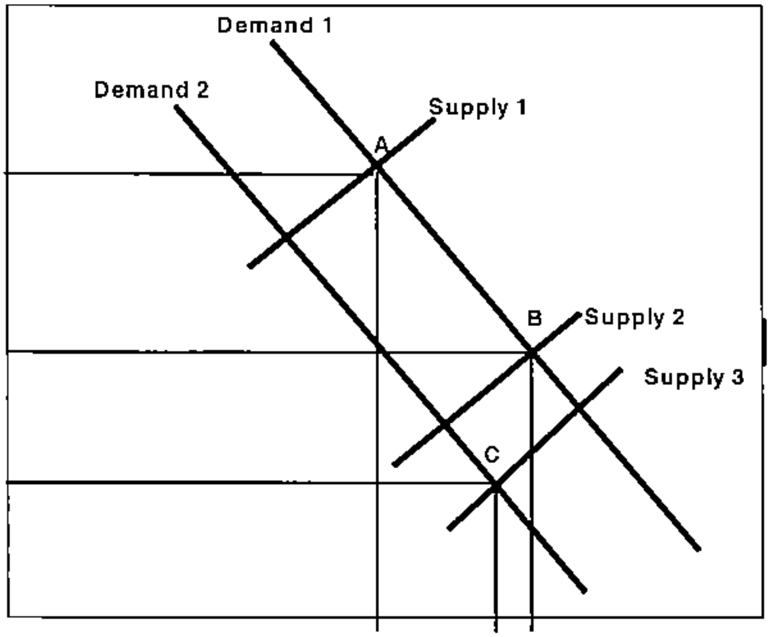


Figure 8. Ending of a downtrend.

8. Conclusions

This article has contributed the literature by presenting justification for four pricc-volume rules of technical analysis by using demand and supply curves from economics literature. In this way we have built a theoretical background for important rules of technical analysis. Until now the justification of the rules has been purely empirical. Moreover, the theoretical background presented helps to understand the dynamics of the equity market where the flow of information has an essential role. Since our analysis has only a few first steps for depicting the determination of equity priccs by bchavior of sellers and buyers, we hope that the article will encourage work in this important research area.

References

Gervais, S., R. Kaniel & D.H. Mingelgrin (2001). The high-volume return premium. *Journal of Finance* LVI, 877–919.

Elder, Alexander (1993). *Trading for living*. Kogan Page.

Miller, E. M. (1977). Risk, uncertainty, and divergence of opinion. *Journal of Finance* 32, 1151–1168.

Varian, Hal R. (1999). Intermediate microeconomics. A modern approach, fifth edition. New York. London: W.W. Norton & Company.

Young, William (2000). Parallel trading. *Active Trader* (July), 22–25.

Strategic vision:
A third level of management[1]

Pentti Malaska and Eero Kasanen

Dedicated to Ilkka Virtanen on the occasion of his 60th birthday

Abstract

Malaska, Pentti and Eero Kasanen (2004). Strategic vision: A third level of management. In *Contributions to Management Science, Mathematics and Modelling. Essays in Honour of Professor Ilkka Virtanen.* Acta Wasaensia No. 122, 107–123. Eds Matti Laaksonen and Seppo Pynnönen.

Managerial tasks are customarily divided into two: tactical and strategic. In this paper we propose a third layer of management, visionary management, which is conceptually distinct from strategy formulation and requires distinctive competencies. Visions and visionary management style are discussed in management literature but not really elaborated as a separate layer of management.

We build our case on theoretical conceptual analysis, observations on the increased complexity of the business environment, and quotes from eminent leaders. The three management layers are illustrated by two simple mathematical formalizations.

Pentti Malaska, D.Sc. (Tech.), Professor emeritus, Finland Futures Research Centre, Turku School of Economics and Business Administration.
Eero Kasanen, DBA, D.Sc. (Econ.), Professor, Rector, Helsinki School of Economics.

Introduction

Strategy and tactics are ancient concepts and topics of decision making in warfare (Sun Tzu 1963, von Clausewitz 1982); but only relatively recently, since the 1960s, have they got more seriously recognized in other business (Holstius & Malaska 2003). As a disciplined management function, separate from the day-to-day business management, strategic management was incorporated into business by *Alfred Sloan*, the president of General Motors, in the 1930s (Mintzberg et al. 1998, Whittington 1983). Sloan's policy

[1] With this joint scientific exploration we feel most happy and honored to congratulate Professor Ilkka Virtanen, our dear friend and distinguished colleague and an important contributor in national and international academic affairs, on his 60th birthday.

planning and control was exercised at the topmost level of management and it was regarded as expertise of top management. Policy decisions, i.e. strategies in current language, were aimed to constrain operational activities of the company and to give direction and coherence to allocation and combination of resources. Strategy and tactical operations became separate areas of management and – as in warfare – the tactical operations (tactics) were subordinated to strategy. But strategic competence is necessarily dependent also on the company's tactical skills and performance.

Since the 1960s strategic management has become widely implemented in practice, and an important field of management research. The number of management articles and books devoted to strategic management is overwhelming, and numerous different schools of strategic thinking exist (cf. Mintzberg op. cit., Näsi & Aunola 2001, Kaplan & Norton 2001). This study aims at contributing to conceptual clarification between the levels of tactics, strategy and visionary decisions by means of theoretical modeling.

A tactical model has a fixed strategy and variables of only tactical decisions. A strategic decision making model has both tactical and strategic decision variables in its structure. They need to be used in a hierarchical relation to each other, respecting the principle of subordination. It is also noteworthy that there is not one strategic reality of business but diverse and multiple realities, which can be modeled only by a variety of diverse logical structures.

The latest development in advanced management has shown a need yet for a third hierarchical level of decision making, i.e. the level of visions and visionary management. Broadly stated visionary management is a process which generates and renews the purpose of the company and its existence in a turbulent environment, facilitating its choice options of strategic maneuvers and directing its tactical competence. The environment is now changing in ways, which the previous approaches to strategic management cannot well cope with. Not only the company needs to change but strategic management itself needs a change when facing unfolding new challenges.

When a company's business is well established it has necessarily positioned itself strategically in the market and in relation to its investors, clients, competitors, and other stakeholders. A strategic position offers the company competitive advantages to run and vary its tactical operations. The strategic positioning is a fixed base and support for successful day-to-day transactions in order to accumulate cash flow and make profit. How well the company is able to utilize its current strategic position depends on its tactical business competence under the given circumstances. The outside business environment – markets – where the company has its position is the other most vital element of the business with its proper positioning.

Tactics, strategy and vision

Tactical competence appears within a fixed strategic positioning in the day-to-day operations and transactions and it is judged by short term value-added, cash flows, and profit making and guided by budget control. Feasible tactical operations are conditioned by established policies and the assumed strategic positioning.

Strategic competence in turn is determined by the managers' ability to foresee environmental changes and to decide on maneuvers to respond to new opportunities, and by an ability to reframe the current business and its operational tactics respectively. Strategic competence can be measured by long term growth in shareholder value, and judged by how apt managers are to transform their knowledge of the business environment or the company's capabilities into a new and better strategic position in the markets. For instance, an investment mix and diversification profile of a company is a manifestation of its strategy. Any change in these parameters demonstrates strategic maneuvers and a change of strategy.

It is widely accepted in warfare that there is a logical difference between strategy and tactics, and that tactics is subordinated to strategy and strategy conditions tactics; it is not unanimously agreed on what this means in business practice. Strategy and tactics are parts of the same totality and a demarcation between them becomes necessarily interwoven. A conceptual difference has been outlined by Holstius and Malaska (2003) as follows:

> **A strategy** is a description of allocation of available resources and positioning of the company in the competitive market, or a plan or re-allocation and new positioning as a response to anticipated changes in the business environment in order to maximally support short term tactical operations. A *strategic maneuver* is an undertaking by which the aimed change of strategic positioning or strategic renewal of the company is realized in practice. The deliberate maneuver is based on scanning the business environment and judging the company's future advantages accordingly. The new strategy materializing is deliberate and planned only partially; real strategic renewal will partially be emergent in character and an ad hoc response to unfolding challenges. The modeling approach can obviously deal only with the deliberate part of strategic decision making.
>
> **Tactics** is about conducting opportunistic day-by-day operative transactions of the company in order to accumulate cash flow and make profit. Tactics, which is a realization of the company's tactical competence, is an ad hoc interplay between competitors, clients, etc. Tactical operations are conditioned by fixed resources and policies determined by the strategic frame.

The third kind of decision making in advanced management i.e. visionary decision making refers to very long term sustainability and renewal which can be characterized as follows:

> **A vision** refers to a shared mindset held by the principal actors of the company concerning the entrepreneurial business concept and core competence areas and company's responsibilities for the long-term success. The vision is a shared success story of the company, and it conditions, directs and guides strategic choices. Visionary management generates a framework and platform for strategic maneuvers, and a vision is realized through strategic management.

The three different varieties of decision-making of the advanced management are illustrated in Fig.1. Business development is a continuous flow of opportunistic, strategic and visionary decisions and their modifications along with experience accumulated.

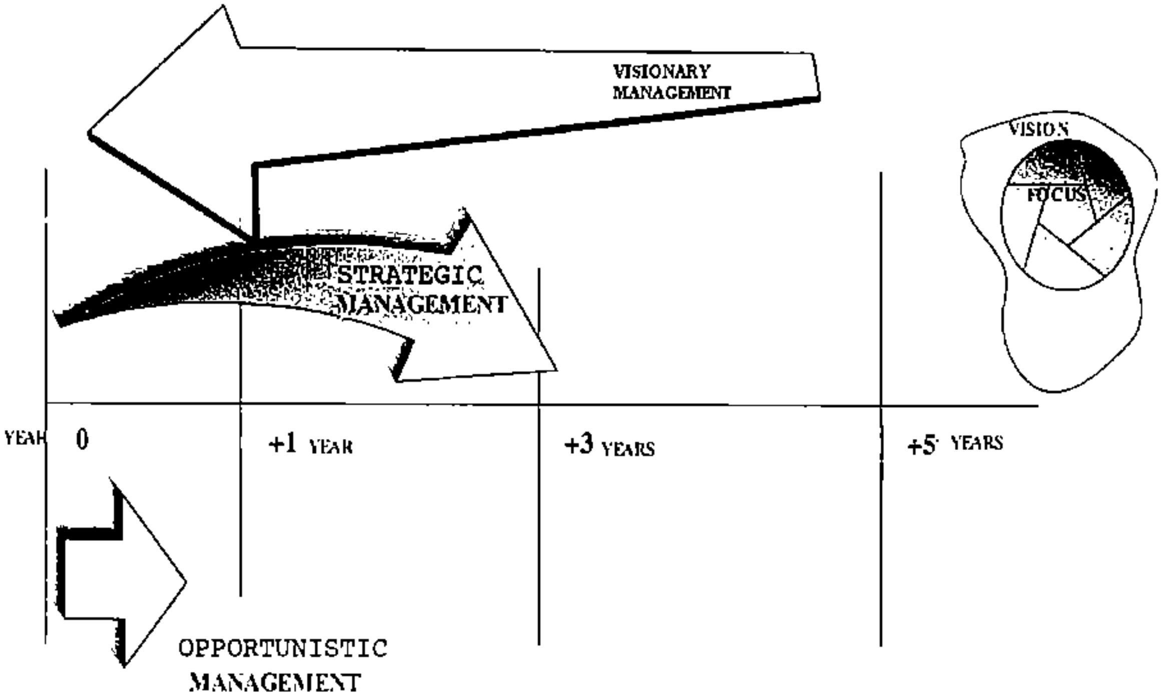


Fig. 1. Visionary, strategic and opportunistic management and decision making illustrated (Malaska & Holstius 1999: 355).

From the modeling point of view the logical separation of strategic and tactical decisions implies a model structure of two levels of decision variables within the same market frame; the hierarchical relationship between the concepts means an asymmetrical interaction between the variables in the model. The visionary model leads to a structure where the reframing of the market – market creation – provides an essential third level of decision making.

Examples of real life visionary management

We provide illustrations of the visionary, third level of management. Powerful leaders have typically stepped outside the straight forward "net present value" thinking and relied on visions to guide the organization. Strategic and tactical plans are easier to formulate in economic facts and figures whereas visions are communicated through beliefs, values, stories, slogans, and patterns. Visionary management language and competencies are distinct from those of structured strategic and tactical layers.

1) Jeroen van der Veer of Shell

Source: Annual Report and Accounts 2002. Shell 2003.

In his 2002 annual message to the company's shareholders Mr. Jeroen van der Veer, the President of Shell, writes:

"2002 was a pivotal year. We delivered robust and competitive profitability in testing conditions and made great strategic progress in pursuing our goals – making four major acquisitions and investing in organic growth. We worked hard to live up to our business principles and commitment on sustainable development. We are well placed to maintain momentum in uncertain times."

This announcement captures clearly in a nut shell the three levels of advanced strategic management studied in this paper: tactical, strategic and visionary management.

To repeat van der Veer: 'we delivered robust and competitive profitability in testing conditions' can be interpreted to refer to tactics and the company's tactical competence; '(we) made great strategic progress in pursuing our goals – making acquisitions and investing in organic growth' tells about strategy and strategic maneuvers reframing the tactical operation area. 'We worked hard to live up to our business principles and commitment on sustainable development. We are well placed to maintain momentum in uncertain times' shows visionary insight of the future chosen and valued and committed to by strategic management.

2) A change of Stalin's vision in 1942-43

Source: Antony Beavor, Stalingrad, Penguin books 1999

The vision of the war both of Hitler and Stalin since the beginning of the war was focused on Stalingrad as planned by Hitler in Barbarossa. But Stalin crucially changed his view with a new vision of Uranus during the most desperate times of bloodshed in Stalingrad.

The original idea (of the new vision) dated back to Saturday, 12 September, the day that Zhukov was summoned to Kremlin after failed attacks against Paulus's northern flank. Vasilevsky, the Chief of the General Staff, was also present in Stalin's office. Zhukov was made to explain again what went wrong.... Stalin demanded to know what was needed. "... another full-strength army, supported by a tank corps, three armoured brigades and at least 440 howitzers, all backed by an aviation army." But Zhukov and Vasilevsky agreed that another solution would have to be found. "And what does another solution mean?" asked Stalin. "Go over to General Staff and think over very carefully indeed what must be done in the Stalingrad area", he told the generals. They returned following evening to Stalin's office. "Well, what did you come up with?" he asked them. Zhukov argued:

"The city of Stalingrad should be held in a battle of attrition, with just enough troops to keep the defence alive. Then, while the Germans focused entirely on capturing the city, the Stavka (Soviet force) would secretly assemble fresh armies behind the lines for a major encirclement, using deep thrusts far behind the point of the apex." Eventually, Stalin saw the advantage of the much more ambitious operation, (i.e. the new vision). On that night of 13 September, Stalin gave this plan for deep operations his full backing. He instructed the two men to introduce a regime of the strictest secrecy. "No one, beyond the three of us, is to know about it for the time being." The offensive was to be called Operation Uranus. Zhukov was not just a good planner, he was the best implementer of plans (here a vision). (p. 220–222)

The plan (here a strategy) for Operation Uranus (the vision) was simple, yet daringly ambitious in its scope. The main assault, over a hundred miles west of Stalingrad, would be launched south-eastwards from Serafimovich bridgehead, a forty-mile-long stretch south of the Don. This point of attack was so far to the rear of the Sixth Army that German mechanized forces in and around Stalingrad would not be able to get back in time to make a difference. This would mark the encirclement of Paulus's Sixth Army and part of Holt's Fourth Panzer Army. (p. 226) Most generals back in Germany were convinced that the Soviet Union was incapable of two offensives. (p. 228)

The new vision turned out to be victorious. Germans were caught off guard and eventually Paulus's army surrendered which changed the dynamics of the war. Stalin's vision would be hard to phrase in any military equivalent of "net present value" thinking. Huge sacrifices and risks were taken because at stake was the survival of a nation.

3) Jack Welch's vision of GE

From Ian Wilson, Realizing the Power of Strategic Vision.
LRP Vol.25, 1992, 18-28

From the day that Dr. John F. Welch, Jr. took over as GE's CEO from Reginald H. Jones, he articulated, strongly, clearly and constantly, a vision for the company that stressed two elements (a visionary idea): restructured portfolio and a revitalized culture. He has consistently implemented these two elements (as strategy) in tandem, although, for the first 6 or 7 years, portfolio restructuring was the first priority. Only in the past 2 or 3 years (in 1992) has revitalizing the culture become the dominant theme.

The vision of a 'restructured portfolio' embraces a basic concept and a central image. The basic concept is that GE will be only in those businesses in which it is (or can be) number one or number two in the global market (or, in the service businesses, has a substantial position) and that are of a scale and potential appropriate to a 50 billion dollar enterprise.

The vision of a 'revitalized culture' focuses on achieving excellence and entrepreneurship, leanness and agility, and a 'boundaryless' company. It envisages paring away bureaucracy, moving faster, demanding (and making possible) the very best for everyone. The aim, as Welch is fond of saying, is to combine 'the sensitivity, the leanness, the simplicity, and the agility of a small company' with the strength, resources, and reach of a big company.

The twin elements of restructured portfolio and revitalized culture share the common thread of 'global competitiveness'. Simply, if GE is to be a world-class competitor, then competitiveness – and all that it takes – must be engrained in the corporate culture.

Welch's vision is hard to phrase in precise economic figures. His vision is more like a belief, corporate culture, world view, and a pattern. However, the vision can be rigorously argued for. Some part of the rationale for this vision reflects Welch's character and management style. But the larger part stems from changing conditions in the business environment. Of the many imperatives that these changes forced on U.S. companies, Welch has emphasized three:

With the spread and intensification of global competition, success and profitability go only to the global leaders (to Welch, the number one and two companies in the world). But not even GE can be pre-eminent in everything, so selection and focus in the business portfolio are essential.

Throughout the value chain and across virtually all companies (many of which are not, on the surface, 'high-tech'), technology is increasingly becoming the key source of competitive advantage.

More than ever before, speed and flexibility drive competition today. The only way to deal with accelerating change successfully, Welch maintains, is to change faster than the world around you. And this task, in itself, entails a significant shift in corporate thinking and organization.

4) Lou Gerstner's vision of IBM

Source: Lou Gerstner has a vision: Network-centric computing.
Business Week/Oct. 30, 1995

"One of the greatest things about this industry is that every decade or so, you get a chance to redefine the playing field. We're in that phase of redefinition now"

One of IBM's fastest-growing units is Integrated Systems Solution (of information services), which handles mega-outsourcing deals for company's biggest clients. With its order backlog worth S 30 billion, the organization is growing more than 33% a year.

"There is no question that the PC-based model is now not the future...The speed with which the Internet has emerged has caught all industries related to this technology by surprise. .. We are reconceptualizing our old businesses. We're bringing them into this new model."

Gerstner's vision opened up conceptually a huge new market for IBM to play around. The world of networking differs from hardware, software, and service industries. It calls for new types of concepts, products, and competencies.

Modeling investment decisions

Investment decisions have been a classic subject for mathematical modeling. The time value of money and the definition of proper cash flows and the discount rate have provided sufficient complexity and quantitative data to work on. Tactical and strategic issues of capital budgeting have been extensively studied. As far as we know, visionary aspects of investment decisions have not been subjected to mathematical analysis.

In its simplest form one isolated investment project is analyzed at a time. The theoretically correct criterion is to accept those projects, the net present value of which is positive. Another widely studied variation is to look at several investment projects and impose a budget constraint for the investment program. The problem can be modeled by linear programming and extended to dynamic and stochastic settings.

The above classic investment models take the strategic position and the emergence of profitable investment projects as given. Strategic issues in capital budgeting have been modeled for example by real options (Trigeorgis 1986) or by spawning models (Kasanen 1986). Using real options framework, the value of straight cash flows is augmented by the value of future modifications to the projects, for example by the strategic option to expand if things go well. The spawning framework models the interaction between strategic and tactical investment projects by looking at how current investment portfolio affects the opportunities to invest in the future. Both these strategic capital budgeting approaches rely strictly on the maximization of the net present value of the cash flows.

Modeling visionary investment decisions, if at all possible, would be a totally different challenge. Questions to be asked and formulated are: what has to be invested in (company structure, new leaders, knowledge, research, technology development, allies, markets, production, etc.) or disinvested from in order to open a new vision area and to communicate it to the possible stakeholders, what sort of a strategic frame of maneuvers a new vision would create, how to evaluate ex-ante the value space of the vision, or how to generate and formulate the practical conditions that would make the foreseen strategic maneuvers of the vision realizable. With a change of the vision the company makes itself a change agent of the market situation instead of just being a changing actor in the market.

Through these examples of investment decisions we have illustrated how the quantitative operations research models are transformed into qualitative mathematical formalisms, and how at some point these issues defy mathematical exactness. Visionary layers of management are probably best left for the most part for verbal analysis of managerial discourse. However, once a powerful new vision has emerged, the strategic and tactical investments needed to operate within it can be quantitatively analyzed.

Mathematical sketches of some issues in visionary management

Having analyzed the concept of visionary management and after some real life examples we are now in a position to use mathematical formalism for illustrative purposes. Our first model is related to three layers of investments and to the definition of the objective function. Our second model addresses the issue of learning across the three layers of management.

1. Investment layers

Let us construct a simple model of a company living in a two-period world (short term and long term). The decision variables are investments in the three layers of management realm: tactical, strategic, and visionary. Here tactical investments affect the efficiency of the operations through cost cutting, strategic investments change the competitiveness and subsequently the market share of the company, and visionary investments create new markets in the future. Mathematically we set up the following model

$$V = [p(1)-c(1)r(y_t(1))]\ \frac{m(y_s(1))}{m(y_s(1)) + M(1)}\ Q(1) - y_t(1) - y_s(1) - y_v(1)$$

$$+ f\ \{[p(2)-c(2)r(y_t(2))]\ \frac{m(y_s(2))}{m(y_s(2)) + M(2)}\ [q(y_v(1))+Q(2)] - y_t(2) - y_s(2)\}$$

where

V	=	net present value of the company over the planning horizon
$p(t)$	=	price level in period t
$c(t)$	=	cost trend in period t, if no rationalization is done
$r(x)$	=	rationalization function, $r(0)=1$, $r(x) > 0$, $r'(x) < 0$, $r''(x) > 0$
$m(x)$	=	market power function, $m(0)=0$, $m'(x) > 0$, $m''(x) < 0$
$M(t)$	=	competitors' market power in period t
$Q(t)$	=	old market size in period t
$q(x)$	=	new market creation function, $q(0)=0$, $q'(x) > 0$, $q''(x) < 0$
f	=	discount factor
$y_t(t)$	=	tactical investments in period t
$y_s(t)$	=	strategic investments in period t
$y_v(1)$	=	visionary investments in period 1.

We may elaborate the model by expanding the number of products, time periods and effects of various investment programs. A detailed model can then be used as a simulation model for investment programs. In fact, standard investment proposals often have a simplified structure written as a spreadsheet model. Adding accounting information, we may produce typical pro forma income statements and cash flow statements from the model structure. These elaborations of the model are omitted here.

Treating the model as an optimization problem, we may set up various submodels to capture the interplay between management layers. An economic modeling effort would treat the model as a global net present value maximization problem with investment levels as decision variables

$$\text{Max } V(y_o(1),y_s(1),y_v(1),y_o(2),y_s(2))$$

It is a straight forward derivation exercise to calculate the first order optimality conditions of the model. The marginal conditions for optimality will tell us that each investment category is set at such a level where the marginal cost of additional investment equals the marginal net benefit from the investment. The global approach treats all the investment layers symmetrically, and there exists only one clear global objective function across all the managerial layers.

Another approach is to look at the three management layers in a hierarchical order and define different models for different organizational contexts. The tactical layer is looking at tactical investments, given the current vision and strategy

$$\text{Max } V_o = V(y_o(1),y_o(2) \mid y_s(1),y_s(2),y_v(1)).$$

The strategic layer is trying to set the strategic investments at optimal levels, given the current vision, and assuming that the tactical layer will make optimal tactical decisions once the strategy is set up. Mathematically we can formulate this as a sequential optimization problem

$$\text{Max } V_s = V(y_s(1),y_s(2) \mid y_o(1),\ y_o(2) = \text{arg Max } V_o,\ y_v(1)).$$

The visionary layer is concentrating on the visionary investments, and assumes that strategic and tactical layers will set up their investments accordingly

$$\text{Max } V_v = V(y_v(1) \mid y_s(1),\ y_s(2) = \text{arg Max } V_s,\ y_o(1),\ y_o(2) = \text{arg Max } V_o).$$

Mathematically, global maximization of V, and sequential maximization of V_o, V_s, V_v lead here to the same solution. However, if we take the various managerial layers seriously and look at their key variables more closely we see the special character of the visionary management more clearly. Tactical and strategic investments are tied together in the same period and they both have to consider the time-value of revenues seriously. The visionary management layer is looking directly at future market creation.

Taking one step further, if we assume that each management layer has its own expertise and separate models, and then although the goal is to maximize the net present value of the company, the cognitive tasks of the layers are different. Tactical layer looks at measurable current issues, given the strategic and visionary guidance. Strategic layer concentrates on the intertemporal interplay between strategic competitiveness and tactical efficiency, given the current overall vision of the company. Visionary layer investigates and invests in future market potential, assuming the strategic and tactical layers will act according to the vision agreed upon.

2. Growth of managerial competencies

The three layers of management call for different language, tools, competencies, and personalities. It would not be uncommon to find production engineers in the tactical layer, MBA's in the strategic layer, and entrepreneur-owners in the visionary layer. The crucial factor in the success of the company is the growth of managerial competency. The growth path crucially depends on the interaction between the managerial layers. We illustrate the situation by a simple matrix equation

$$x(t) = Ax(t-1)$$

where

$x(t) = (x_t(t), x_s(t), x_v(t))$ = tactical, strategic, and visionary competencies in period t

A = matrix of competency co-generation effects, a_{ij}

Using the matrix equation, we obtain a dynamic input-output type growth path

$$x(n) = A^n x(0)$$

which under suitable positivity and regularity conditions approaches asymptotically a steady-state growth path

$$x(n)_{asymp} = cg^n x^*,$$

where c is a constant, g is the Frobenius root of matrix A and x^* is the eigenvector.

Looking at the co-generation effects between the competencies we note that the overall steady-state growth rate and the mix of competencies in the long-run depend crucially on the shape of matrix A. Depending on the strength of strictly positive (assuming the rest to be zero) co-generation effects we will obtain quite different paths. Let us sketch some interesting cases.

> If matrix A is diagonal, then competencies develop separately without any co-generation present. In the long run, the growth rate is tied to the largest diagonal element of A and the respective competency becomes dominant and the others are left marginal.
>
> If matrix A is triangular, then competencies develop hierarchically. In the long run, the growth rate will still be determined by the largest diagonal element.
>
> If matrix A is irreducible, then competencies develop synergistically amplifying each other. In the long run, the growth rate will exceed any of the diagonal elements.

A way to interpret these stylistic growth path results is to consider coefficients a_{ij} as "learning" effects. The diagonal elements refer to learning that happens inside each managerial layer. Not surprisingly, if the layers are separated then the fastest learning layer becomes dominant. It is not hard to think of examples of companies that have been taken over by a tactical, strategic, or visionary mind set and are actually lacking other competencies. (Only the tactical mind set company can survive for a while.) Hierarchical structure leads to a more advanced mix of competencies but the growth is hampered by one-way learning. If there are learning effects (co-generation) between all managerial layers then the growth of managerial competency on all levels will be faster than any layer could achieve separately.

The matrix illustration also makes it evident how the growth and competency structure would be deficient if we constrain ourselves only to operative and strategic layers. The visionary managerial layer provides important synergistic growth potential for the company' competence.

Discussion

We have argued for the need to introduce a third conceptual layer of management to augment the established strategic and tactical layers. Our reasoning is based on conceptual analysis of the managerial tasks, literature review, and real life examples of visionary management. We have also provided stylized mathematical models to illustrate the concept of visionary management.

Visionary management uses the language and tools of stories, values, slogans, beliefs and patterns. Once a vision has been created, shared and communicated it becomes a powerful and visible tool for strategic and tactical management layers. A change of vision facilitates and necessitates a change of quantitative strategic models and tactical plans. A turbulent and complex environment makes the visionary management layer all the more crucial as the reach of structured forecasts and plans is diminishing. Qualitative mathematical models may be of some help for illustrative purposes and for clarification of new visionary concepts.

References

Beavor, Antony (1999). *Stalingrad.* Penguin Books.

Clausewitz, Carl von (1982). *On War.* Penguin Books.

Gerstner, Lou (1995). Network-centric computing. *An Interview in Business Week* (Oct.) 30, 40–49.

Holstius, Karin & Pentti Malaska (2003). *Advanced Startegic Management.* Bellagio paper (forthcoming).

Kaplan, Robert S. & David P. Norton (2001). *The Strategy Focused Organization.* Harvard Business School Press.

Kasanen, Eero (1986). *Capital Budgeting and the Control of Business Unit Growth.* (Dissertation at Harvard University Graduate School of Business Administration), Publications of the Turku School of Economics and Business Administration, Series A-4.

Malaska, Pentti & Karin Holstius (1999). Visionary Management. *Foresight* 1:4 (Aug.), 309–317.

Mintzberg, Henry, Bruce Ahlstrand & Joseph Lampel (1998). *Strategy Safari.* London: Prentice Hall.

Näsi, Juha & Manu Aunola (2002). *Strateginen johtamisen teoria ja käytäntö.* Metalliteollisuuden keskusliitto (MET), MET-julkaisuja nro 12/2001, Tampere 2002, 177.

Shell (2003). *President's Message.* Royal Dutch Petroleum Company Annual Report and Accounts 2002.

Sun Tzu (1963). *The Art of War.* Ed. Samuel Griffith. Oxford Univ. Press.

Trigeorgis, Lenos (1986). *Valuing Real Investment Opportunities: An Options Approach to Strategic Capital Budgeting.* Harvard Business School thesis.

Whittington, Richard (1993). *What is Strategy and does it matter?* London: Routledge.

Wilson, Ian (1992). Realizing the Power of Strategic Vision. *LRP* 25, 18–28.

The number of solutions of $X^p + Y^q = 1$ in certain finite fields

Marko J. Moisio

Dedicated to Ilkka Virtanen on the occasion of his 60th birthday

Abstract

Moisio, Marko J. (2004). The number of solutions of $X^p + Y^q = 1$ in certain finite fields. In *Contributions to Management Science, Mathematics and Modelling. Essays in Honour of Professor Ilkka Virtanen.* Acta Wasaensia No. 122, 125–130. Eds Matti Laaksonen and Seppo Pynnönen.

We derive a formula from which one can compute the number of solutions of $X^p + Y^q = 1$ in certain finite fields.

Marko J. Moisio, Department of Mathematics and Statistics, University of Vaasa, P.O. Box 700, FIN-65101 Vaasa, Finland.

Key words: Equations over finite fields, Jacobi sums, Gauss sums.
Mathematics Subject Classification (2000): 11T06; 11T23; 11T24; 11T71.

1. Introduction

A classical problem in number theory is the determination of the number of solutions of polynomial equations $f(X, Y) = 0$ over finite fields. It is known that this is a very difficult problem in general. In this paper we study polynomial equations $X^p + Y^q = 1$ with p, q satisfying certain conditions and obtain a formula from which we can calculate the number of the solutions of them. This problem is important also in coding theory, since it is known that one can construct good error-correcting codes based on the polynomial equations having many solutions with respect to the size of the finite field over which the polynomials are defined.

2. The formula

Let p and q be two distinct odd prime numbers and denote $N = pq$. We shall assume that the index of $< 2 >$ in $\mathbb{Z}_N^*$ $[\mathbb{Z}_N^* :< 2 >] = 2$ and $k = \phi(N)/2$. Let K be a field with 2^k elements and E_l the extension of K with 2^{kl} elements. Let $f(X,Y) = X^p + Y^q - 1 \in K[X,Y]$. Let $N(l)$ denote the number of solutions of $f(X,Y) = 0$ in E_l^2. In a forthcoming paper (Moisio and Väänänen, to appear) we proved, as a special case of Theorem 1 of that article, that

If $\mathbb{Z}_p^* =< 2 >$ and $\mathbb{Z}_q^* =< 2 >$ then

$$N(l) = 2^{kl} - \frac{(p-1)(q-1)}{2}(\alpha^l + \overline{\alpha}^l), \tag{1}$$

where

$$\alpha = 2^{h-1}(a + b\sqrt{-pq}),$$

and (a,b) is a solution of the Diophantine equation

$$a^2 + pqb^2 = 2^{k-2h+2}$$

satisfying $a \equiv 2^{k-h+1}\ (p)$, $a \equiv 1\ (2)$, and

$$h = \min\left\{S_2\left(\frac{2^k-1}{pq}\right), k - S_2\left(\frac{2^k-1}{pq}\right)\right\}$$

where $S_2(j)$ denotes the digit sum in binary expansion of $j \in \mathbb{Z}_+$. We remark that the Diophantine equation can be solved with a very fast, in fact $\mathcal{O}(k)$, algorithm introduced in Hardy, Muskat, and Williams (1990).

The aim of this article is to present a simple proof of (1) by following the classical method of A. Weil appeared in his famous paper (Weil 1949), where he expressed the number of solutions of

$$a_1x^{m_1} + \cdots + a_kx^{m_k} = 1$$

in terms of so-called Jacobi sums, and furthermore in terms of so-called Gauss sums. After that we use our knowledge of the Gauss sums appearing in the consideration of the equation $f(X,Y) = 0$. We remark that the method used in Moisio and Väänänen was totally different from the method we are using in this article.

Let χ be the generator of the group of multiplicative characters of E_l^*, i.e.

$$\begin{aligned}&\chi : E_l^* \longrightarrow \mathbb{C},\\ &\chi(ab) = \chi(a)\chi(b)\ \forall a, b \in E_l^*,\\ &\widehat{E_l^*} = <\chi> .\end{aligned}$$

We extend every character to E_l by defining $\chi^j(0) = 0$ if $0 < j < 2^{kl} - 1$ and $\chi^0(0) = 1$.

Lemma 1. *Let* $\lambda \in \widehat{E_l^*}$, $\operatorname{ord}(\lambda) = t$ *and* $a \in E_l$. *Then the number of solutions of* $X^t = a$ *in* E_l *is given by*

$$S_l(X^t = a) = \sum_{j=0}^{t-1} \lambda^j(a).$$

Proof. If $a = 0$ the equality is obvious. Let γ be a primitive element of E_l and let $a = \gamma^k$. Now $\gamma^{ti} = a$ iff $ti \equiv k\ (2^{kl} - 1)$. The congruence has exactly t solutions if $t \mid k$ and no solutions if $t \nmid k$. The claim follows now by noting that the sum is a geometric sum.

Let $\lambda, \psi \in \widehat{E_l^*}$, $\operatorname{ord}(\lambda) = p$, $\operatorname{ord}(\psi) = q$. Now, by Lemma 1, we have

$$\begin{aligned}N(l) &= \sum_{\substack{a,b \in E_l\\ a+b=1}} S_l(X^p = a) S_l(Y^q = b)\\ &= \sum_{a+b=1} \Big(\sum_{i=0}^{p-1} \lambda^i(a)\Big)\Big(\sum_{j=0}^{q-1} \psi^j(b)\Big)\\ &= \sum_{i=0}^{p-1}\sum_{j=0}^{q-1}\Big(\sum_{a+b=1} \lambda^i(a)\psi^j(b)\Big)\\ &= \sum_{i=0}^{p-1}\sum_{j=0}^{q-1} J(\lambda^i, \psi^j).\end{aligned}$$

Here $J(\lambda^i, \psi^j)$ is a Jacobi sum with the following properties (see Lidl and Niederreiter 1984 for a proof):

$$\begin{aligned}&J(\lambda^0, \psi^0) = 2^{kl},\\ &J(\lambda^0, \psi^j) = J(\lambda^i, \psi^0) = 0 \text{ if } i, j \neq 0,\\ &J(\lambda^i, \psi^j) = -1 \text{ if } \lambda^i\psi^j = \chi_0 \text{ and } i, j \neq 0,\end{aligned}$$

where χ_0 is the trivial character.

Lemma 2. *Let $0 \leq i \leq p-1$ and $0 \leq j \leq q-1$. Then $\lambda^i\psi^j = \chi_0$ if and only if $i = j = 0$.*

Proof. If $\lambda^i\psi^j = \chi_0$ then $\lambda^i \in <\psi> \cap <\lambda> =: H$. Since $\gcd(p,q) = 1$ we have $H = <\chi_0>$. If $i = j = 0$ then obviously $\lambda^i\psi^j = \chi_0$.

As a consequence,

$$N(l) = 2^{kl} + \sum_{i=1}^{p-1}\sum_{j=1}^{q-1} J(\lambda^i, \psi^j).$$

Let $G(\eta)$ denote the Gauss sum,

$$G(\eta) = \sum_{a \in E_l^*} (-1)^{Tr(a)}\eta(a),$$

where $\eta \in \widehat{E_l^*}$ and $Tr(a) = a + a^2 + a^{2^2} + \cdots + a^{2^{kl-1}}$. Gauss sums have the following properties (see Lidl and Niederreiter 1984 for a proof):

$$|G(\eta)| = 2^{kl/2} \text{ if } \eta \neq \chi_0,$$
$$J(\lambda^i, \psi^j) = \frac{G(\lambda^i)G(\psi^j)}{G(\lambda^i\psi^j)} \text{ if } \lambda^i\psi^j \neq \chi_0.$$

Now we have

$$J(\lambda^i, \psi^j) = G(\lambda^i)G(\psi^j)G(\lambda^{-i}\psi^{-j})/2^{kl}.$$

In Moisio (1998) we proved in an elementary way that

$$G(\lambda^i) = (-1)^{l-1}2^{kl/2}, \qquad G(\psi^j) = -2^{kl/2}.$$

Thus

$$J(\lambda^i, \psi^j) = (-1)^l G(\lambda^{-i}\psi^{-j}),$$

and we have

$$N(l) = 2^{kl} + (-1)^l \sum_{i=1}^{p-1}\sum_{j=1}^{q-1} G(\lambda^{-i}\psi^{-j}).$$

It is known (see Lidl and Niederreiter) that the value of $G(\eta^t)$ depends only on the 2-cyclotomic coset of t modulo order of η. It is easy to see (cf. Moisio 1998) that $\{0, \pm 1, p, q\}$ is a complete set of representatives of 2-cyclotomic cosets modulo pq.

Lemma 3.

$$G(\lambda^{-i}\psi^{-j}) = G(\chi^{\pm\frac{2^{kl}-1}{pq}})\ \forall\ 1 \leq i \leq p-1,\ 1 \leq j \leq q-1,$$

and the sign is + for exactly $(p-1)(q-1)/2$ pairs (i,j).

Proof. Let $\lambda = \chi^{\frac{2^{kl}-1}{p}}$ and $\psi = \chi^{\frac{2^{kl}-1}{q}}$. Now $\lambda^{-i}\psi^{-j} = \chi^{-\frac{2^{kl}-1}{pq}(qi+pj)}$. It is impossible to have $qi + pj \equiv p2^s\ (pq)$ for some s since otherwise $qi \equiv 0\ (p)$ or $p \mid i$. Similarly it is impossible to have $qi + pj \equiv q2^s\ (pq)$. Thus $qi + pj \equiv \pm 2^s\ (pq)$ for some s. If $qi + pj \equiv \pm 2^s\ (pq)$ then $q(p-i) + p(q-j) \equiv \mp 2^s\ (pq)$ proving the claim.

As a consequence

$$N(l) = 2^{kl} + (-1)^l \frac{(p-1)(q-1)}{2}\Big(G(\chi^{\frac{2^{kl}-1}{pq}}) + G(\chi^{-\frac{2^{kl}-1}{pq}})\Big).$$

It follows from a deep theorem of Hasse and Davenport (see Lidl and Niederreiter 1984 for a proof) that

$$G(\chi^{\pm\frac{2^{kl}-1}{pq}}) = (-1)^{l-1} G(\sigma^{\pm\frac{2^k-1}{pq}})^l,$$

where σ is a generator of $\widehat{K^*}$. Thus,

$$N(l) = 2^{kl} - \frac{(p-1)(q-1)}{2}\Big(G(\sigma^{\frac{2^k-1}{pq}})^l + G(\sigma^{-\frac{2^k-1}{pq}})^l\Big).$$

We proved in Moisio (1998) (see also van der Vlugt 1995) that

$$G(\sigma^{\frac{2^k-1}{pq}}) = \alpha$$

and so we have proved (1).

Example 1. The number of solutions of $X^5 + Y^3 = 1$ in a field with 2^{4l} elements is

$$N(l) = 2^{4l} - 4(\alpha^l + \overline{\alpha}^l),$$

where $\alpha = 1 + \sqrt{-15}$.

Example 2. The number of solutions of $X^{1061}+Y^{1019}=1$ in a field with $2^{539540l}$ elements is

$$N(l)=2^{539540l}-539540(\alpha^l+\overline{\alpha}^l),$$

where

$$\alpha=2^{269465}(a+b\sqrt{-1081159})$$

and

$a=$ 156538939224522236799623109740178972810573584249198975633413594574939868921335590498498507507,

$b=$ 608562909863905833565978580418340342441827604359378785044754804982717456405470460677418905.

References

Hardy, K., J.B. Muskat & K.S. Williams (1990). A deterministic algorithm for solving $n=fu^2+gv^2$ in coprime integers u and v. *Math. Comp.* 55, 327–343.

Lidl, R. & H. Niederreiter (1984). *Finite Fields.* Cambridge: Cambridge Univ. Press.

Moisio, M. (1998). Exponential sums, Gauss sums and cyclic codes, Dissertation. *Acta Univ. Oul. A* 306.

Moisio, M. & K. Väänänen. A comparision of the number of rational places of certain function fields to the Hasse-Weil bounds. To appear in *Applicable Algebra in Engineering, Communication and Computing.*

van der Vlugt, M. (1995). Hasse-Davenport curves, Gauss sums, and weight distributions of irreducible cyclic codes. *J. Number Theory* 55, 145–159.

Weil, A. (1949). Numbers of solutions of equations in finite fields. *Bull. of American Math. Society* 14, 497–508.

Regional growth and convergence via integration – The case of the large EU

Paavo Okko

Dedicated to Ilkka Virtanen on the occasion of his 60th birthday

Abstract

Okko, Paavo (2004). Regional growth and convergence via integration – The case of the large EU. In *Contributions to Management Science, Mathematics and Modelling. Essays in Honour of Professor Ilkka Virtanen.* Acta Wasaensia No. 122, 131–145. Eds Matti Laaksonen and Seppo Pynnönen.

The coming Eastern enlargement of the EU will be a fundamentally different step compared with the previous enlargements of the EU. It will create a new situation in which growth conditions and regional adjustment requirements of Europe are going to change, too. The theory of economic growth and regional structures has developed recently in an interesting way. Especially the endogenous growth theory and models of the new economic geography offer a relevant approach to these issues.

There is a strong tendency towards factor price equalisation and towards income convergence. But regional differences in other respects may become even deeper via this process. The enlargement of the EU is an interesting case from this point of view. There are very large income differences, which are assumed to diminish, but it seems impossible to happen without a fundamental regional restructuring.

The target of the paper is to make a survey on income convergence and regional restructuring in the case of European integration. The idea is to make an evaluation and prediction on the real convergence prospects of the large EU. Eastern enlargement is an opportunity to faster growth in Europe, but the regional specialization and restructuring is a crucial condition for materializing of this result.

Paavo Okko, Turku School of Economics and Business Administration, Department of Economics.

Key words: growth, convergence, European integration.

1. Introduction

The current enlargement of the European Union is fundamentally different compared with the five earlier steps towards the larger union. The EU is going to undertake its greatest enlargement ever. The process will result into a large and heterogeneous integration block. That is why both institutional (or constitutional) and economic issues are more demanding than in the history of the EU. The institutional questions of the EU enlargement are very important, but now we take an economic and also a regional approach to challenges and opportunities of the larger EU.

The main question is here, what kind of growth effects will emerge from integrating very different kind of economies, and what kind of regional adjustment pressures it will produce. The main group of accession countries are transition economies and their income levels are quite low. For that simple reason budgetary problems have been in the focus of the debate. That is why many citizens of the current member states have been concerned that the accession of low income/low wage countries will have adverse effects (see e.g. Boeri et al. 2002). In the long run the budget shares cannot be the crucial issue. The most important economic question now is, what is the impact of the Eastern enlargement on the long-term growth rate of the large EU and its members.

The new, so called endogenous, growth theory offers an adequate approach to these questions. It points out what is important for sustainable economic progress and it also points out, how economic growth may also benefit from the fact that the EU is coming to be a heterogeneous group of economies. There is a potential for poor countries to catch up the rich ones. A basic requirement for catching-up is to enhance the development of human capital and institutional framework. For the larger EU the hope lies in catching-up process: the average growth rate may be higher when large income differences do exist.

This paper is briefly describing the convergence debate in growth theory and applying it as a background for the analysis of EU-enlargement (see also Okko 2000 and 2003). Conclusions about the growth effect of integration are drawn from previous studies on

European integration. In particular, the role of human capital and institutions are emphasised in this process. Some conclusions concerning economic growth and regional adjustment of the larger EU are also drawn in this paper.

2. Convergence or divergence via European integration

In order to evaluate and predict effects of enlargement we should first know what is the impact of integration on growth rate and on income differences. Does integration enhance growth and what is its impact on income differences between members and regions? As an answer to this question we offer some background information from the history of European integration. There is a quite large literature on growth effects of integration and certainly a very large discussion on convergence.

Empirical work on growth effects of European integration has resulted in quite considerable positive effect for EC and EFTA members (Henrekson et al. 1997). The effect of EC or EFTA membership was around 0,6–08, percentage points in annual growth rate. The results also suggest that technology transfer was the main mechanism through which EC and EFTA membership effects on growth. Surprisingly, there were no effects of membership on investment. Even if it is not possible to draw direct conclusions from earlier experiences it seems quite obvious that new members starting from quite low relative income level compared to incumbent members will have even stronger growth impetus from their membership.

Integration means also rearrangement of institutional setup of a country. That is one important background factor for the growth (see e.g. Easterly – Levine 2003). Actually, membership of the EU is sometimes considered from a very narrow viewpoint of membership related budget expenditures and revenues. The main consequence of joining to the EU is the change of institutional environment into which a country is committing itself in the future. For accession countries even the commitment to membership negotiations has meant a strong motivation factor.

A good example about the importance of institutional factors in economic growth is offered by Finland and Estonia. At the end of 1930's those two countries were approximately at the same income level and their institutional structure was quite similar. After that they had different institutional environments for a long time. Per capita income in an accession country Estonia is now about 35 % of the level of Finland. This is also an example about the challenge of transition economies.

The regional convergence hypothesis has been in the focus of debate between the neoclassical and the endogenous growth theory approach. If the neoclassical assumption of diminishing returns holds we should have narrowing of regional income differences (convergence) via growth. Romer (1986) argued that the absence of convergence across economies throughout the world represents strong evidence against the neoclassical model and works in favour of his theory of endogenous growth. This question has been in the focus of large theoretical and empirical debate since the 1980's.

The new theory of economic growth differs from the neoclassical theory especially in respect to the endogenous treatment of technological change. According to Romer, technological change - improvement in the instructions for mixing together raw materials – lies at the heart of economic growth and that change arises in large part because of intentional actions taken by people who respond to market incentives (Romer 1990a, 72). The growth of human capital is the result of purposeful actions for increasing it, but the technological progress as an input factor is a public good. The production function may even contain increasing returns in respect to the inputs. Human capital is a special input because via it not only current production is effected but also the long-term development in technological advance is gaining more speed. Human capital has both direct and indirect or external effects (see e.g. Lucas 1988). Technological advance is an endogenous phenomenon and it may create increasing returns. Integration is adding to this a new option. It means an increase in ideas that can be used in each country in production of goods.

Romer (1986) argued that inequalities across the world show no sign of narrowing over the years. Barro (1991) produced an impressive battery of regressions showing that a negative correlation between the initial income level and the growth rate could be observed when this correlation was taken conditionally upon a set of variables (so-called conditional β -convergence), the most significant of which was the level of school enrolment.

One large empirical convergence exercise was Sala-i-Martin (1994). He produced an extensive empirical convergence analysis from the United States (48 states; 1880–1990), Canada (10 provinces; 1961–1991), Japan (47 prefectures; 1955–1990), and Europe (73 NUTS 2-regions/7 countries; mainly 1950–1990). The basic result was that there is evidence of strong forces leading to regional convergence. The estimated speeds of convergence are surprisingly similar across data sets: economies tend to converge at a speed of about 2 % per year. However, the catching up -process is quite slow: half of the original income difference is remaining after 34 years! If the speed of convergence is 3 % per year it takes 23 years to abolish a half of the original gap. The slow speed of convergence suggests that technology does not instantaneously flow across countries, and some countries are more capable to create new technologies. Anyway, integration is supposed to speed up this process.

According an unconditional beta-convergence measurement about EU14 (Luxemburg excluded) the speed of convergence has not been constant but income levels have been converting in the history of current EU members (Wagner & Hlouskova 2002: 21–22). The speed of catching up in 1961–98 was on average 2,05 % per year. It was slowest in the 1980's (0,83 %) and fastest in the 1990's (3,59 %). The result shows that even if some studies have been indicating decreasing rate of convergence in Europe, it still seems to work. All current members of the EU have not been members over the whole period of the study. It reminds that the convergence is basically a part of the growth process in general and also an aspect of the integration process because it enhances diffusion of institutions and technologies – and also growth via that way.

In addition to above-mentioned β -convergence research (conditional or unconditional) there is large empirical literature on measuring income differences by different kinds of inequality measures. So-called σ -convergence concept is based on the standard deviation of incomes. Even if poor countries are growing faster in relative terms, the income distribution may be changing adversely if original differences are large enough. Those calculations show that income inequality between the member states of the EU has been decreasing with variable speed over time (see e.g. Puga 2001 and Wagner & Hlouskova 2002).

The evolution of regional disparities within the EU seem to contain convergence among countries but not necessarily convergence among regions. There is at least some empirical evidence on that (Puga 2001 and Giannetti 2002). If international knowledge spillovers affect certain sectors only, integration and greater exchange of knowledge among countries whose regions have heterogeneous specialization spur growth and bring convergence among regions specialized in high-tech sectors, but create greater disparities within individual countries. As a result, differences in income levels among countries are decreasing, just like in the EU, because the value added of the technologically advanced regions is a rising share of GDP.

Putting it in brief, the economic integration is in favour of economic growth and growth is in favour of narrowing relative income differences among countries. But this all requires adjustment, which will change the relative position of sectors and regions within countries. That is why the conclusion is dependent on the level of regional disaggregation.

3. Integration and regional adjustment

The main hypothesis about effects of integration on regional structure has normally been concentration. The idea that larger markets mean larger concentrations has been the way of thinking. Even if the basic tendency has been working into that direction, the issue is not so simple. There are both centripetal forces and centrifugal forces functioning in integration process causing regional adjustment. The new economic geography models have offered new interpretations to these questions (see e.g. Brülhart 2001).

Economies of scale and positive external effects of concentration (agglomeration economies) are main reasons for centripetal forces. There is a home market effect meaning that the larger the home market the more attractive it is. But because of integration also peripheral areas may benefit from demand coming from foreign markets. There are immobile resources and there are transportation cost and trade barriers, too. Cost competition is willing to use also cost advantages of peripheral areas and this all creates centrifugal forces via foreign market effect.

Empirical work on the European integration (Brülhart 2001: 235–238) has resulted in some interesting results. The strongest concentration appears in traditional, low-technology industries. The technology-intensive industries are least geographically concentrated, but concentration in those industries has been increasing. Surprisingly, the scale-intensive industries are not strongly concentrated. Employment concentration has been strongest in sectors protected by high non-tariff barriers.

General conclusions drawn by Brülhart (2001: 240–241) are interesting also from point of view of accession countries and expected effects of the EU enlargement. The three main conclusions were following. First, industrial specialization has been increasing slowly but steadily. Second, the Single Market project boosted this process. Specialization accelerated after 1986 in those industries, which were strongly affected by the abolition of intra-EU non-tariff barriers. Yet, the Single Market did not affected sectoral concentration in general. Third, on the whole, specialization process reflects neither concentration in core countries nor movement towards peripheral countries; for most industries the importance of the centre-periphery dimension seems to have diminished in recent years.

This all may be interpreted that comparative advantage considerations continue to be relevant for the evolution of specialization patterns even in a relatively homogeneous area like the current EU. For the accession countries the traditional argumentation may be even more relevant. The finding that the spatial concentration of technology-intensive sectors has started to increase since the mid-1980's, however, may mean that agglomeration economics are coming to be more important in the EU.

One important new factor in future will be the impact of The Economic and Monetary Union on the income convergence. In a monetary union member states are becoming to be more like regions than nation states. Actually, it means that the principle of comparative advantage is substituted for the principle of absolute advantage and regional differences come up in the sense that member states have not any more traditional macro-policy measures to tackle their competitiveness problems. Regional adjustment requirements come up with their full power. E.g., migration flows may have bigger role in these circumstances.

There are also some doubts on the adequacy of the current regional policy instruments in the large EU. Actually, Boldrin and Canova (2003: 41) propose that the current policy should be terminated as soon as possible. They believe that labour and capital mobility are good for growth and economic convergence. Evidence from Europe and USA shows that. Labour migration is an important channel through which productive skills are acquired in advanced regions and brought into poorer regions to be applied. Regional policies should not go against this factor of convergence.

4. Income differences and growth rate

The enlargement of the EU means a change in the basic set up of growth conditions. In this sense it is surprising that the debate on the enlargement has been concentrating to large extent on the short-term budget issues. The budget of the EU is a bid over one percent of the total GDP. A small reallocation in the budget cannot be a crucial matter in a process in which the annual growth rate of the GDP may increase about to the same extent. It is not now a question of a one-shot change but a change of the growth rate. For these issues the new growth theory is capable to offer adequate insight.

The new growth theory means contributions both in the problems of economic integration and labour mobility. E.g,. these models suggest that what is important for growth is integration not into an economy with a large amount of people but rather into one with a large amount of human capital. According to Romer (1990a: 98) growth seems to be correlated with the degree of integration into worldwide markets but not closely related to population size or density. Integration means interaction of 'idea sector' and 'goods sectors'. If there is a difference in the initial endowment of countries in the level of technology the flow of goods means an extra gain in: increase in ideas that can be used in each country in production of goods. An increase in the size of the market or in the trading area in which a country operates increases the incentive for research and thereby increases the share of investment and the rate of growth of output, with no fall in the rate of return on capital (see also Romer 1990b: 366).

These models permit a distinction between a one-shot gain (a level effect) and a permanent change in the growth rate (a growth rate effect) that is important in making of estimates of the benefits of economic integration (see Rivera-Batiz and Romer 1991: 532). The results by the neoclassical model and by the new one may differ strongly. E.g., it is not obvious – like in the old theory – that a permanent increase in the investment rate could result only in a temporary change in the growth rate. Actually the opposite might be true: a temporary increase in the investment rate linked with the increase in the human capital may have a growth rate effect, at least for a considerable time. There is also criticism against the simple versions of endogenous growth models indicating that there is a possibility to raise the growth rate forever (Griffith et al. 2003).

In the case of mobility, it is very crucial whether the effects of human capital are entirely internal or whether they have external benefits that spill over from one person to another. In the latter case the wage rate of labour at any given skill level will increase with the wealth of the country in which he is employed (Lucas 1988: 40). Not at all surprising conclusion is that labour will move from poor regions to wealthy ones. But the result is interesting enough in the sense that it offers an explanation within the rigorous theory to the question why labour mobility is not equalizing wage levels. It has been a difficult question to the static neoclassical theory.

The traditional and the new growth theory give different answers also in respect of growth effect of integration. The traditional theory predicts no permanent effect of integration on the rate of growth. The new approach makes understandable the possibility of permanent change in the growth rate because of the change in dynamics of the economies. The evaluation of the creation of European single market was an interesting example about the issue. Richard Baldwin (e.g. 1989) was the first one showing medium term effects of integration in addition to static efficiency gains reported by the Cecchini Report on Single Market.

From the point view of transition economies the main message is that economic progress requires investments both into physical and human capital and that institutional framework is a crucial factor (see also Okko 2003). Institutional environment will be established via the membership in the EU. Investments into human capital need both public and private activities because market incentives are not effective in a case in which external effects are important but not compensated via markets. Transition economies have typically large investment needs. That is good for growth if investments are realised. Foreign direct investments are one way of organising that. Actually, FDIs have important role in the growth process of the accession countries integrating into the EU. Trade flows and FDIs have been in EU integration more complements than substitutes (Widgrén 2001). But until now FDIs have not had any major role in the growth of the accession countries (Boldrin – Canova 2003: 12).

Eastern Enlargement of the EU will create a union with large income differences. Countries have access to the same technology, but many of them are lagging behind. This means that the steady state income levels are near to each other but actual levels are far from each other. The crucial thing is how soon these differences will be narrowing. That will also determine the growth rate of the larger EU. There is a potential to catch-up and the R&D-input has crucial role to create an absorptive capacity of an economy (Griffith et al. 2003). The economy must be capable to create innovations but it must have also capacity to learn from others – using the main strategy to learn new things.

5. Expected income convergence in the large EU with the single currency

If these predictions hold also for the new members of the EU, it would mean that the income gap will be narrowing but it will be an issue for long time in the future. E.g., the current gap in the per capita GDP between Poland and the EU15 average is about 60%. If the convergence rate would be only 2 % per year, the difference would be still about 15 % after 70 years! It is reasonable to think that members of a single market are capable for faster convergence. Actually, the latest observations (1995–99) show that the accession countries have a higher growth rate (3,4 %) than the EU-15 (2,4 %) (see, e.g. Prime Minister's Office 2001). So it is supposed to be also in the future.

According Armstrong (1994) the convergence rate in Europe is lower if more peripheral regions are included into the analysis. This is in accordance with the original results by Romer that in the global sample including countries of very different income levels no clear over all convergence is found. The catching-up - hypothesis works only in certain circumstances. Cumulative causation may work into the both directions; there are both convergence and divergence going on. In this respect it is important to notice that according to Barro (1991: 437) those poor countries tend to catch-up the rich countries, which have high human capital per person (in relation to their level of per capita GDP), but not otherwise. The debate continues but it reveals that the accumulation of human capital has a crucial role in the process of growth and convergence.

It is possible to demonstrate how fast or slow the catching up process will be in the large EU by assuming certain growth rates and convergence speed. Calculations made by Wagner and Hlouskova (2002) are interesting, but their content depends heavily on assumptions made. In an optimistic case it will take 26 years from the ten Central and Eastern European accession countries to catch up to the level of 70 % of the EU25 (Wagner and Hlouskova 2002: 42). The variation is large: Slovenia is already about at that level and for Latvia it will take 51 years. Comparison to the current EU shows that, it will take 30 year to come to the level of 70 % of the EU15. Even if the tendency of convergence is in existence, income differences will remain forever in practice.

The prediction is that those low-income countries having the access to the same technology and investing strongly into human capital are capable to catch up high income countries. The new members of the EU are supposed to be that kind of economies. This will mean that the Eastern enlargement is going to be growth-enhancing from the point of view of the EU. The low-income entrants are often considered to be a burden to the EU budget, but the main impact is on the real side of the economy. The fact that the EU is coming to be a heterogeneous group opens up new opportunities for growth. If new members are capable to catch up, the average growth rate will increase. Transition economies entering into the EU market have urgent needs for investments and they offer new possibilities to combine new ideas and new production.

Integration is a long-term process, which tends to abolish income differences, but they will never disappear entirely. According Charles Kindleberger (1968: 194) factor price equalisation is the ultimate measure of integration, but it is like the absolute zero point in low-temperature physics: it will be never reached! Income differences are fuel of economic growth, and an integration process has not yet come to the end as long as differences still exist. This all means that regional integration – as well as global integration – is actually an ingredient of economic growth.

There are also nominal convergence criteria of the EMU process. Those requirements deal with monetary and nominal conditions on economies entering into the EMU. The real convergence considered earlier means different growth rates, and different growth rates tend to mean different inflation rates. Inflation rates in traded goods sector and nontraded goods sector will differ, too. It will create the so-called Balassa-Samuelson effect. Faster growth is bound to affect the exchange rate (see e.g. Halpern and Wyplosz 2001). There will be a tendency towards real appreciation of the currency. The Balassa-Samuelson effect is potentially problematic both in the convergence process before the third stage of the EMU and during the single currency. In the first case it is difficult to keep the required exchange rate band and in the second one inflation difference is causing adjustment problems to a sub region of a currency area. The EMU may mean problems to fast growing low-income countries showing real convergence but having problems with nominal convergence. That means a challenge to the EMU, which was established actually for a final stage of a very deep integration. Now at least some transitions economies are entering into it perhaps too early.

6. Conclusions

The coming Eastern enlargement will be a fundamentally different step in the history of the European integration. It will create a new situation in which growth conditions of Europe are going to change, too. The theory of economic growth and regional structures has developed recently in an interesting way. Especially the endogenous growth theory and models of the new economic geography offer relevant approach for interpretations.

The market-driven integration is benefiting from large income differences. There is a strong tendency towards factor price equalisation and towards income convergence. The large EU and especially new entrants are in front of a challenge. They must be capable to create an endogenous growth process by investing into physical and into human capital and maintain high growth rate even if there are strong pressures of new competition and adjustment. The endogenous growth theory points out that it requires effective transformation towards innovation-driven economy. Accession countries have also high marginal returns of physical investment. That requires capital flows within the large EU, too. It will normally mean also migration of labour. By this way the investment rate both into physical and human capital can remain high. That is the ultimate guarantee of a high growth rate. This will also contribute positively to the competitiveness of the large EU.

The speed of convergence will be quite slow even if growth rates will differ clearly. This ends up to adjustment process requiring both sectoral and regional restructuring. It is quite natural that in these circumstances regional policies of the union is required to take care of observed regional disparities. Even if the average income differences between member countries are narrowing regional inequalities will remain, but he regional policy strategy is better to be in line with long term adjustment requirements than against the basic trends.

References

Armstrong, H.W. (1994). Convergence Versus Divergence in the European Union Regional Growth Process, 1950–1990, Paper for the 34th European Congress of the Regional Science Association, August 1994.

Baldwin, R: (1989). The Growth Effect of 1992. *Economic Policy* (October).

Barro, R.J. (1991). Economic Growth in a Cross Section of Countries. *The Quarterly Journal of Economics* CVI:2 (May).

Boeri, Tito et al. (2002). Who`s afraid of the big enlargement? Centre for Economic Policy Research. *Policy Paper* 7. London.

Boldrin, Michele & Fabio Canova (2003). Regional Policies and EU Enlargement, Center for Economic Policy Research, Discussion paper series No. 3744 London, February 2003.

Brülhart, Marius (2001). Evolving geographical concentration of European manufacturing industries. *Weltwirtschaftliches Archiv* 137:2, 215–243.

Easterly, William & Ross Levine (2003). Tropics, germs, and crops: how endowments influence economic development. *Journal of Monetary Economics* 50, 3–39.

Giannetti, Mariassunta (2002) The effects on integration on regional disparities: Convergence, divergence or both? *European Economic Review* 46, 539–567.

Griffith, Rachel, Stephen Redding & John van Reenen (2003). R&D and absorptive capacity: Theory and empirical evidence. *Scandinavian Journal of Economics* 105:1, 99–118.

Halpern, Làszlò & Charles Wuplosz (2001). Economic Transformation and Real Exchange Rates in the 2000s: The Balassa-Samuelson Connection, a paper prepared for UNECE (http://heiwww.unige.ch/~wyplosz/).

Henrekson, M., J. Torstensson & R. Torstensson (1997). Growth effects of European integration. *European Economic Review* 41, 1537–1557.

Kindleberger, Charles P. (1968). *International Economics*. Fourth Edition. Homewood.

Lucas, R. E. (1988). On the mechanics of economic development. *Journal of Monetary Economics* 22, 3–42.

Okko, Paavo (2000). Growth, human capital, and agglomeration economies, management expertise for The New Millennium. Ed. Tapio Reponen. *Publications of Turku School of Economics and Business Administration, Series A*-1:2000, 227–235.

Okko, Paavo (2003). *The Large EU from the Point of View of Regional Growth and Regional Structure, The Future of Europe, Relations between the Enlarging European Union and Russia and Ukraine.* Institute for World Economics, Hungarian Academy of Sciences. Eds Gàbor Fòti & Zsuzsa Ludvig. Budapest.

Prime Minister's Office (2001). EU Enlargement and Finland. *Publications* 3, Helsinki.

Puga, Diego (2001) European regional policies in light of recent location theories. Centre for Economic Policy Research. *Discussion Paper Series* No. 2767 (April).

Rivera-Batiz, L.A. & P.M. Romer (1991). Economic integration and endogenous growth. *The Quarterly Journal of Economics* (May), 531–555.

Romer, P.M. (1986). Increasing Returns and Long Run Growth. *Journal of Political Economy* 99 (June), 500–521.

Romer, P.M. (1989). Capital accumulation in the theory of long-run growth. *Modern Business Cycle Theory,* 51–127. Cambridge Mass.

Romer, P.M. (1990a). Endogenous technological change. *Journal of Political Economy* 98:5, pt. 2, S71–S102.

Romer, P.M. (1990b). Capital, labor, and productivity. *Brookings Papers on Economic Activity,* Microeconomics 1990, 337–367.

Sala-i-Martin, X. (1994). Regional cohesion: Evidence and theories of regional growth and convergence. Centre for Economic Policy Research. *Discussion Paper Series* No. 1075, London, November 1994.

Wagner, Martin & Jaroslava Hlouskova (2002). The CEEC10's Real Convergence Prospects. Centre for Economic Policy Research. *Discussion Paper Series* No. 3318 London, April 2002.

Widgrén, M. (2001). Eastern enlargement: Trade and industrial location in Europe. *CESifo Forum* (Summer), 14–18.

The investment behavior and performance of the "men in their best age"

Jukka Perttunen

Abstract

Perttunen, Jukka (2004). The investment behavior and performance of the "men in their best age". In *Contributions to Management Science, Mathematics and Modelling. Essays in Honour of Professor Ilkka Virtanen.* Acta Wasaensia No. 122, 147–164. Eds Matti Laaksonen and Seppo Pynnönen.

Investment behavior and stock portfolio performance is evaluated in a sample of 1944-born male investors having their residence in either Southwest Finland or Southern Ostrobothnia. The empirical analysis is based on the portfolios of 657 investors who have kept a stock portfolio over the whole three-year research period on a continuous basis. The variables analyzed include the portfolio return, volatility, beta, market value, trading activity, and the number of stocks in the portfolio. The results reveal differences in investment behavior and performance within and between the two districts. Also, some of the analyses suggest positive relation between trading activity and portfolio performance. Any significant difference in the risk-adjusted investment performance could not be found between the two geographical districts.

The financial support by the OKO Bank Research Foundation and the Foundation for the Economic Education is gratefully acknowledged. We are also grateful to the Finnish Central Securities Depository for providing us with access to the data.

Jukka Perttunen, Department of Accounting and Finance, University of Oulu, P.O.Box 4600, FIN-90014 University of Oulu, Finland, E-mail: jukka.perttunen@oulu.fi.

Key words: private investor, investment behavior

1. Introduction

In recent financial literature more and more attention has been paid to the differences in the investment behavior and performance of individual investors. The traditional market efficiency research, which evaluates the ability of different investment rules to beat the market, has got on its side a new tradition of evaluating the realized performance of true investment portfolios. The availability of new large and very detailed databases has made this approach possible. From the point of view of investment research, particularly interesting have been some recent studies, which evaluate the investment behavior and performance of private investors, especially. This is due to the fact that private investors' investment decisions are largely based on their own interests and decisions only. Opposite to that, the investment portfolios of institutional investors might be distorted by the contradictory needs or preferences of their customers or owners, for instance. Also, variation in the private investors' investment behavior is remarkably strong, which enables us to evaluate the determinants that might cause differences in their behavior and performance.

There are several interesting issues that have been evaluated in recent studies concerning the private investors' investment behavior. The level of portfolio diversification has been one of the most interesting issues being evaluated. The most fundamental and robust result of the modern portfolio theory is the need of diversification, no matter what are the other preferences of an investor, besides crude risk aversion. However, most depressingly, several studies show that the private investors very seldom follow the simple rule of diversification in an acceptable way (see e.g. Blume and Friend, 1975; Kelly, 1995; Barber and Odean, 2001; Goezmann and Kumar, 2002, Polkovnichenko, 2003, and Tyynelä and Perttunen, 2003). Even the first and simplest rule of academic investment literature seems to be despised by the practitioners. The empirical results indicating how this contemptuous attitude appears in their risk and return performance have been most interesting. More research in this area is needed.

Another important issue is the effect of active trading. Traditional market efficiency literature has evaluated market timing and stock selection abilities as possible investment performance drivers. Simple trading rules like momentum or contrarian strategies have been analyzed in a wide number of empirical studies. In addition, practically oriented finance literature offers a huge number of different technical analysis tools to be applied in making investment decisions. It would be most interesting to evaluate, how the different trading strategies or rules work in private investors practical portfolio formation process. However, it is difficult, and in most cases impossible, to find out, what investment rule, if any, an individual private investor has applied in her portfolio formation. There is one indicator, however, that indirectly, with a bias, of course, measures the use of trading rules like the ones mentioned above, in general. There is large variation in the trading activity of individual investors, and one could imagine that an investor, who trades more, applies some trading rule more probably than a passive investor does. The profitability of active trading is a most interesting research issue, indeed.

Several studies have examined the impact of active trading on portfolio performance. Odean (1999) evaluates the trades of 10 000 individual investors from an U.S. brokerage house and finds that the investors trade too much. They would do better, if they traded less. Barber and Odean (2000) analyze the trading activity and performance of more than 66 000 households. The results are similar. The households seem to hurt their performace by active trading. Barber and Odean (2002) studied over 1 600 investors who switched from phone trading to online during 1990s. After switching the trading activity appeared to increase. The investors traded more speculatively, and their performance weakened. On a Finnish data, Tyynelä and Perttunen (2003) find results that are very much in line with those mentioned above.

Several studies have also evaluated the effect of different demographic factors on the investment behavior and performance of private investors. The most typical factors evaluated are age, gender, and the wealth of an investor. It has been found, that the older investors perform better than the younger ones. At least partially, this has been due to the more active trading of the younger investors, which has proved to be unprofitable compared to the passive benchmark portfolio. Also, it has been found in several studies, that women earn higher benchmark adjusted returns than men (see e.g. Barber and Odean, 2001, and Tyynelä and Perttunen, 2003). Also this result can be, at least partially, due to the trading activity differences.

The reason for the more active trading of younger and men is hypothesized to be in their relative overconfidence in their own investment ability. The theoretical models of overconfidence in investment behavior include those of Benos (1998), Odean (1998b), Cabellé and Sákovics (2003), and Gervais and Odean (2001). All these models suggest that overconfidence leads to excessive trading, and that excessive trading leads to losses.

More clearly, overconfident investors trade more actively than their rational counterparts do and this active trading leads to losses in their performance. In addition to the argument that people are overconfident in general, psychological studies also claim that there are differences in the level of overconfidence. This leads to the additional hypothesis that men and young people are more overconfident than women and older people. In the light of conventional wisdom this makes sense.

Geographical issues have also risen up in investment research. It has been found that there might be a significant home bias in the stock selection process. In other words, investors more probably invest in the stocks of the firms that have activities in their home country or district. French and Poterba (1991) find in world's five largest stock markets in 1989, that equity holdings of domestic shares are significantly greater than their relative share of world market capitalization. The same phenomenon can be found within domestic market as well. Coval and Moskowitz (1999) documented that in the United States mutual fund managers prefer to hold locally headquartered firms. Grinblatt and Keloharju (1999) discover the same phenomenon in the Finnish stock market. Investors appear to hold and trade stocks headquartered in nearby locations of their home district. An interesting and still unsolved question is, if there are geographical differences in investment behavior. For instance, do the investors from different cultures and geographical districts have the same degree of risk-aversion.

An interesting behavioral issue that has been studied in recent literature is the so-called disposition effect, which refers to investors' tendency to realize their gains at a much higher rate than their losses. Shefrin and Statman (1985) argued that because in general, a loss hurts us more than we enjoy a gain of equal size, we tend to hold on to the losing stocks (stocks that have lost value in relation to their purchase price) too long and sell the winning stocks too soon. The disposition effect is anomalous because the purchase price of a stock should not affect the selling decision; if one expects the stock will appreciate in value, one should keep it, if one expects it to decline, it should be sold. Furthermore, tax considerations should make people realize their losses rather than gains. The disposition effect has been found in both experiments and real financial markets. The pioneering field study was done by Odean (1998a), who analyzed the trading records of a large U.S.

discount brokerage house. It was founded that, overall, investors prefer selling winners to selling losers. More importantly, this kind of behavior is not justified by subsequent portfolio performance: in the following year, the winners investors chose to sell provided a higher return than the unsold losers. With a more comprehensive data set, Grinblatt and Keloharju (2001) analyze the trading behavior of all individuals and institutions in the Finnish stock market. With a variety of tests and control variables, they find that the disposition effect and tax-loss selling are the major determinants of the propensity to sell a stock one owns. In a more recent study, Using an alternative research design, Lehenkari and Perttunen (2003) also found the phenomenon among the Finnish private investors.

It is obvious that the investment behavior and performance of individual private investors is a most interesting and important issue. More research on the behavioral biases and the performance drivers of the portfolio investment process is deeply required. The present study tries to bring its modest contribution to several issues around this topic of a particular interest. We are not going to focus on all the issues discussed above, but try to shed some more light on these most important and interesting questions.

2. Research Problem

This study evaluates the differences in investment behavior and performance in a sample of Finnish private investors. In our study, we want to evaluate the investment behavior of a group of investors that is of particular interest to us. In order to control the effect of the age and the gender factors we restrict our analysis to private male investors in one specific age class. We further restrict our sample by including in our analysis investors from two limited geographical areas, only. More or less randomly we decide to focus on the investment performance of male investors born in 1944, who currently live in either of the two districts, Southwest Finland or Southern Ostrobothnia, which we call, according to the their central cities, the Turku- and the Vaasa-districts, respectively. We consider the sample selection criteria appropriate for several reasons. The age class 1944 is about to turn their 70's at the time of the preparing of this study, and thus they can be considered to be at the threshold of their best age as far as their investment analysis abilities are concerned[1]. The hasty and overconfident investment behavior, which has been found to be typical for younger investors, especially for younger men, in some earlier studies, should be left behind by this age. Therefore, our research design should control for the effect of the disabilities that are due to age or gender, and thus enable us to concentrate on the performance differences within – should we say – optimally mature private investors.

The choice of the investors from the Turku- and the Vaasa-districts is based on our deep interest in analyzing the behavior of the representatives of two Finnish tribes that have been characterized in earlier literature in a most interesting way. "Gone are the times when the Turku region was the best!" said a mighty landowner, while a sooty blacksmith disparages the Southern Ostrobothnia by asking: "What a heck can you do with those boring plains?", both according to Topelius (1875), who provides excellent and telling characterizations of the people in the different provinces in Finland.

3. Data Description

The empirical data used in this study is collected from the central register of shareholdings for Finnish stocks in the Finnish Central Securities Depository. This database contains the shareholdings of all Finnish investors on daily basis. We have selected from the database all the male investors born in 1944, and having their permanent addresses in the one-digit postal code areas of either 20000 (Turku-district) or 60000 (Vaasa-district) over the research period 1997–1999. Our research design requires us to limit our analysis to the investors who have kept a stock portfolio over the whole three-year research period on a continuous basis. This is due to the fact that for every investor, a continuous time series of portfolio returns is needed, in order to calculate the values of the risk measures. This quite strong condition leads to a sample of 657 investors, 413 of which represent the Turku-district, and the rest 244 investors the Vaasa-district. The combining of the shareholdings (or actually the changes in the shareholdings) of the investors, with the daily price and return series of the stocks listed in Helsinki Exchange, enables us to calculate the daily

[1] In the end of our research period of 1997–1999, the representatives of the age class 1944 were only 55 years old. We consider this age high enough for our purpose of controlling the possible age-disability in investment behavior.

market value weighted portfolio returns of each individual investor over the three-year research period[2]. The value weighted portfolio returns are further applied in generating a value weighted return index series for each individual investor on daily basis. These daily series have been further transfonned to the monthly returns that are used in the subsequent analyses. Also, the database allows us to calculate the number of trades of each individual investor, the market value of their portfolios, and the number of stocks in their portfolios, all on daily basis.

Our research period (1997–1999) represents exceptionally good times in the Finnish stock market and worldwide. In terms of the HEX portfolio index, the annual return over the three years period was 35,9%, on an average, and the market return was positive in ten out of the thirty-six months in that period. Regarding the generalizing of the results, the exceptional nature of the research period is not for good. The data availability, however, forces us to use this particular period in our empirical analysis.

There are six variables included in our empirical analysis. The monthly portfolio returns of each individual investor are calculated as relative differences of month-end return index values over the 36-month research period. In the subsequent analyses the *return* stands for the annualized average return over the research period. Correspondingly, the portfolio *volatility* is calculated as annualized standard deviation of monthly portfolio returns over the research period. The portfolio *beta* is estimated by the regression of the monthly portfolio excess returns on the market excess returns over the research period. The excess returns are calculated by deducting from the returns one twelfth of the one-month EURIBOR-interest rate that is applicable to the respective month. The market returns are defined on the basis of the HEX portfolio index. The average *market value* (in 1000 euros) of the portfolio of each individual investor is calculated as the three-year average of the daily portfolio market values. The *trading activity* tells how many times in a year an investor has traded on average over the research period. Finally, the *number of stocks* is the average of the number of stocks in the portfolio that is originally calculated on daily basis.

[2] It must be noticed that we are able to track the direct investments in Helsinki listed stocks only. We haven't got any information on the investments in the mutual funds or bonds, for instance.

Table 1. The distributions of the variables.

	Turku (n = 413)	Vaasa (n = 244)	All (n = 657)
Return			
5%	-0.1184	-0.0796	-0.0924
25%	0.0416	0.0565	0.0416
50%	0.1276	0.1892	0.1501
75%	0.2771	0.2771	0.2771
95%	0.6811	0.4033	0.4306
Volatility			
5%	0.1527	0.2252	0.1856
25%	0.2690	0.2647	0.2682
50%	0.3186	0.3093	0.3161
75%	0.3887	0.3422	0.3879
95%	0.6377	0.4594	0.5955
Beta			
5%	0.1653	0.4584	0.2488
25%	0.5495	0.6076	0.5610
50%	0.7881	0.8373	0.8085
75%	1.0163	1.0050	1.0150
95%	1.2043	1.1851	1.1967
Market Value			
5%	0.2974	0.1547	0.2132
25%	1.5303	0.6275	1.0861
50%	6.6255	2.3791	4.6815
75%	19.5209	11.6236	16.5721
95%	92.7259	43.9498	84.4608
Trading Activity			
5%	0.0000	0.0000	0.0000
25%	0.0000	0.0000	0.0000
50%	0.0000	0.0000	0.0000
75%	0.3333	0.3333	0.3333
95%	2.0000	2.0000	2.0000
# of Stocks			
5%	1.0000	1.0000	1.0000
25%	1.0194	1.0000	1.0000
50%	2.0000	1.8164	1.9861
75%	3.2432	2.8083	3.0972
95%	7.8178	5.7405	6.9077

The figures are 5%, 25%, 50%, 75%, and 95% quantiles of the variables.

Table 1 provides the quantiles of the variables for the whole sample, and separately for the Vaasa- and Turku-districts. As far as the portfolio returns are concerned, it can be noticed that there is strong variation in the average annualized return of individual investors' portfolios. Despite of the stock market boom in the late 90's there are investors whose stock portfolios have created losses. Furthermore, one fourth of the stock portfolios have provided returns that are somewhere around the return on a risk-free investment. There have been high returns available as well. Although the median return for all the sample investors is as moderate as 15.01 percent, quite high average annual returns have been earned by some investors. The value of the third quartile (75%) demonstrates the fact that in the sample there are investors who have kept identical portfolios over the three-year period. The quartile value of 0.2771 that is common for both districts, rises from the performance of 34 investors, who have had a position in *M-Real* stock only. Although their investments have been different in size, the percentage return is the same for all of them, of course. There are nine other cases in the sample, where there are more than three investors keeping identical single stock portfolios (in relative terms) over the whole research period. The next commonly held single stock appears to be *Outokumpu* with 32 holders, and *Lännen Tehtaat* with 16 hits. All together, there are in the sample 178 investors, who keep a totally undiversified portfolio over the whole research period.

The portfolio risk is measured using the volatility and the beta coefficient. Table 1 reveals that in terms of volatility the portfolio risk varies in quite an expected way. The median volatility appears to be a little bit over 30%. The beta coefficients, on the other hand, indicate that three fourths of the sample investors keep quite conservative portfolios that have the beta less or equal to one. In both volatility and beta, the Turku-district investors seem to have more observations on the higher risk tail than their Vaasa-district counterparts.

The average portfolio value varies a lot between the investors as well as between the two districts. While the median portfolio value of all the sample investors is 4.7 thousand euros, the third quartile is three and a half times larger than that, but the 95%-quantile already about 18 times higher, i.e. 84.5 thousand euros. The wealth is not equally distributed over the investors. Regarding the difference between the districts, as a rule of thumb, one can say that the investors in the Turku-district are more than twice as rich as those in Vaasa-district.

In addition to the risk and size variables above, there are two variables available to describe the investment behavior of the investors. The first of them, the trading activity, tells, how many times an individual investor has traded a stock over the research period. It can be noticed that the investors included in the sample have been very passive. The third quartile value of 0.3333 annual trades corresponds to one trade in the whole three-year period, and the 95%-quantile of 2.0000 six trades in the whole period. There is no difference between the districts with this respect. The average number of stocks in the portfolio once again shows the fact that private investors very often hold undiversified portfolios. The median number of stocks in the portfolio is around two, and the third quartile of about three still indicates quite a high level of unsystematic risk in the portfolio. After that the degree of diversification seems to rise to "academically acceptable" levels, however.

Table 2 provides the correlation matrix of the variables. The portfolio return appears to be correlated with all the other variables except the portfolio value. The correlation is highest with the beta, and positive with the trading activity and the number of stocks, as well. The negative correlation between the return and the volatility is a most interesting result, having in mind the high positive correlation between the return and the beta. The volatility, itself, appears to be positively correlated with the beta, and negatively with the number of stocks in the portfolio. This is, of course, a very expected result, indicating high unsystematic risk in undiversified portfolios. The market value and the trading activity do not seem to have any connection with the volatility. In the case of the beta, the trading activity appears to be significantly correlated with risk, but the market value still remains insignificant. This variable, however, correlates positively with the trading activity and the number of stocks in the portfolio. The larger the amount of wealth in the portfolio, the more actively it is taken care of, and the more carefully it is diversified. This can be seen in the positive correlation between the trading activity and the number of stocks in the portfolio, as well.

Table 2. The correlation matrix.

	Return	Volatility	Beta	Market Value	Trading Activity	# of Stocks
Return	1.000	-0.130	0.392	0.051	0.173	0.182
	(0.000)	(0.001)	(0.000)	(0.191)	(0.000)	(0.000)
Volatility	-0.130	1.000	0.424	-0.001	-0.005	-0.126
	(0.001)	(0.000)	(0.000)	(0.987)	(0.895)	(0.001)
Beta	0.392	0.424	1.000	0.014	0.087	0.034
	(0.000)	(0.000)	(0.000)	(0.721)	(0.026)	(0.384)
Market Value	0.051	-0.001	0.014	1.000	0.176	0.666
	(0.191)	(0.987)	(0.721)	(0.000)	(0.000)	(0.000)
Trading Activity	0.173	-0.005	0.087	0.176	1.000	0.206
	(0.000)	(0.895)	(0.026)	(0.000)	(0.000)	(0.000)
# of Stocks	0.182	-0.126	0.034	0.666	0.206	1.000
	(0.000)	(0.001)	(0.384)	(0.000)	(0.000)	(0.000)

The plain figures are Pearson correlation coefficients and the figures in the parentheses the *p*-values of the test of the significance of the correlation coefficients.

Besides the original six variables, we analyze the portfolio performance using two traditional risk-adjusted measures of performance. The first of them, the *Sharpe's Measure,* relates the realized excess return on a portfolio to its volatility. Thus, the portfolio return is being related to the total risk of the portfolio. It must be noticed that this measure directly corresponds to the slope coefficient of the capital allocation line between the risk-free security and the risky portfolio in the modern portfolio theory. The second measure is the *Treynor's Measure*, which relates the portfolio excess return to its beta. Here only the systematic, i.e. the beta-related risk, is seen as an acceptable benchmark in evaluating the portfolio performance. The measure doesn't pay any attention to the diversifiable risk, and assumes that the portfolio is well diversified, and that the return performance can be rated in the light of beta, alone. In the empirical analysis the performance of the investors is evaluated by applying these two measures, and the key statistics concerning the variables are reported there.

Finally, in some models we include in our analysis a third risk variable, i.e. the *unsystematic risk*. It is measured in terms of the annualized residual standard deviation of the regression model above, where the excess portfolio return is regressed on the excess market return. This measure covers the part of the volatility of the portfolio that is not related to the market risk, i.e. the beta-risk of the portfolio.

4. Empirical Results

The big question ahead of us is if there are systematic differences in the investors' performance. Is there difference between the two districts with respect the risk or return, or trading behavior? What drives the possible performance differences? Table 3 starts our analysis. The table provides the means and the standard deviations of the six original variables as well as the respective differences in the mean between the two districts. Also the probability values of the *t*-test of the equality of the means are reported. It can be noticed that there is a difference, indeed, between the mean returns of the Turku- and the Vaasa-districts. The negative and statistically significant difference of nearly 3 percentage units indicates the on an average better investment performance of the Vaasa-district investors. Furthermore, when it comes to the total risk of the portfolio, i.e. the volatility, their portfolios appear to be less risky! However, in terms of systematic risk, i.e. beta, the result turns more understandable, the Vaasa-district investors' portfolios being of higher risk. As noticed already before, there is a huge difference in the average market value of the portfolios between the two districts. Actually, the trading activity appears to be the only variable where there is no difference between the districts. Finally, the average number of stocks in the portfolio seems to be higher for the Turku-district investors, which is a bit surprising, their average volatility being higher, as well.

There seem to be differences between the investors' performance and behavior between the two districts included in our analysis. From our point of view it is important to find out, if the difference in the portfolio performance is due to the risk differences or if the Vaasa-district investors have superior abilities to build their portfolios. For this purpose the performance is next evaluated using the Sharpe's and the Treynor's Measures. The results reported in Table 4 show that after the risk adjustments the difference between the two districts turns insignificant. In terms of the Sharpe's Measure, the observed difference is still negative but clearly insignificant. In the case of the Treynor's Measure, the observed difference is positive, but also here the *t*-test fails to find any statistically significant difference between the districts. The Turku-district people can sigh with relief!

Table 3. The comparison of the districts.

	Turku	Vaasa	Difference
Return	0.1516	0.1814	-0.0298
	[0.1749]	[0.1575]	(0.0249)
Volatility	0.3409	0.3232	0.0177
	[0.1245]	[0.0876]	(0.0334)
Beta	0.7609	0.8272	-0.0663
	[0.3131]	[0.2476]	(0.0028)
Market Value	24.1900	10.4036	13.7864
	[71.1761]	[22.0900]	(0.0003)
Trading Activity	0.5174	0.4877	0.0297
	[2.3079]	[2.6869]	(0.8857)
# of Stocks	2.7561	2.2947	0.4614
	[2.6709]	[1.7300]	(0.0075)

The plain figures in the first and the second column are the means of the variables in the two districts, and in the third column the differences in the mean between the two districts. The figures in the brackets are standard deviations in the first two columns, the figures in the parentheses *p*-values of the *t*-test (2-tailed) of the equality of the means in the third column.

Table 4. The performance differences between the districts.

	Turku	Vaasa	Difference
Sharpe's Measure	0.4225	0.4692	-0.0467
			(0.2432)
Treynor's Measure	0.1725	0.1608	0.0117
			(0.6802)

The plain figures in the first and the second columns are the means of the variables in the two districts, and in the third column the differences in the mean between the two districts. The figures in the parentheses are *p*-values of the *t*-test (2-tailed) of the equality of the means.

The most interesting issue to be evaluated is the possible behavioral factors behind the return performance. This is analyzed by estimating the regression of the portfolio return, the Sharpe's Measure, and the Treynor's Measure, against the beta, the unsystematic risk, the *log* of the market value of the portfolio, the trading activity of the investor, the number of stocks in the portfolio, and a Vaasa-dummy, indicating if the investor comes from the Vaasa-district. The three models corresponding to the three dependent variables, respectively, are estimated separately for two different samples. In the first phase, the full sample of all the 657 investors is used. Next, the estimation is carried out in a restricted sub-sample, where the investors who hold a single stock over the whole research period, or don't trade at all, are dropped out of the sample. This restricted sample of 197 investors is used in order to find out the robustness of the results with respect to possible outlier observations.

Table 5. The regression results.

	Full Sample (n = 657)			Restricted Sample (n = 197)		
	(1)	(2)	(3)	(4)	(5)	(6)
Dependent Variable	**Return**	**Sharpe's Measure**	**Treynor's Measure**	**Return**	**Sharpe's Measure**	**Treynor's Measure**
Intercept	0.076	0.497	0.447	0.061	0.498	0.298
	(0.001)	(0.000)	(0.000)	(0.146)	(0.000)	(0.015)
Beta	0.237	0.406		0.304	0.547	
	(0.000)	(0.000)		(0.000)	(0.000)	
Unsystematic Risk	-0.126	-0.487	-0.312	-0.138	-0.532	-0.283
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)
Market Value	-0.008	-0.014	0.012	0.007	0.030	0.069
	(0.035)	(0.216)	(0.268)	(0.388)	(0.220)	(0.026)
Trading Activity	0.008	0.024	0.006	0.006	0.019	0.003
	(0.001)	(0.001)	(0.403)	(0.024)	(0.012)	(0.706)
# of Stocks	0.009	0.033	-0.001	-0.003	-0.003	-0.015
	(0.001)	(0.000)	(0.944)	(0.463)	(0.813)	(0.301)
Vaasa Dummy	0.000	-0.024	-0.034	0.009	0.015	0.046
	(0.984)	(0.510)	(0.315)	(0.710)	(0.836)	(0.593)
R-square	0.298	0.267	0.091	0.386	0.363	0.082

The plain figures are OLS-regression coefficients, and the figures in the parentheses the corresponding *p*-values. The last row reports the explanatory power.

The Model (1) in Table 5 the portfolio return is regressed on the variables listed above. The intercept term suggests a 7.6% base return, which is a bit high to represent the average risk-free return. We consider the level acceptable, however, and are not afraid of having left out of the model any variables of crucial importance. The first explanatory variable, the beta, appears to be very significant, and has a coefficient that represents a very acceptable price of market risk. The unsystematic risk, i.e. the residual risk over the beta risk appears to be significant and negative in sign, suggesting that the extra risk only hurts one's performance. The market value gets a marginally significant negative coefficient in this model. However, this result is not very robust, as we will see later. The trading activity also gets a significant coefficient, indicating better performance of more active traders. One additional trade per year appears to provide a 0.8 percentage units additional return. This result contradicts the results from some earlier studies, where the performance effect of extra trading is evaluated in more heterogeneous samples. Could it be so, that extra trading is for good for more experienced investors like our 1944-born men here in this study? Maybe, but some of our models below indicate that the extra trading premium doesn't exist after the beta-adjustment, anymore. The next variable, the number of stocks in the portfolio also gets a significant positive coefficient. An extra stock in the portfolio provides a 0.9 percentage units return premium. This result is not very robust, either, as we will see later. Finally, our last variable, the insignificant Vaasa-dummy confirms the releasing result of non-existing Southern Ostrobothniatic return premium.

In the Model (2), the variable to be explained is changed to be the volatility-adjusted return, i.e. the Sharpe's Measure. The estimation results are very much in line with those of the previous model. Also here, the beta appears to the most significant factor behind the return. Again, the unsystematic return only hurts the performance. The trading activity and the number of stocks in the portfolio have their positive premiums also in this model. The market value, however, is not significant anymore. All in all, the volatility adjustment does not seem to change the results too much from those obtained for the pure returns.

The Treynor's Measure is the variable to be explained in the Model (3). In this model, the beta is dropped out of the explanatory variables. This is because the dependent variable has the beta in its denominator. It can be noticed that a significant and negative effect of the carried unsystematic risk can be found also in this case. The other explanatory variables, however, lose their significance. The market value was insignificant already in the previous model, but the number of stocks in the portfolio and the trading activity now behave in a totally different way. A deeper analysis of the result suggests that the betas of totally undiversified portfolios are systematically and statistically significantly higher than those of the diversified ones. At the same time, the returns of the undiversified portfolios are lower! Such being the case, the number of stocks in the portfolio apparently corrects the effect of the low beta premium to be higher for the diversified portfolios in the first two models. In the third model the non-linear beta effect doesn't exist any more, and the variable loses it significance.

The second variable, the trading activity, appears to be positively correlated with both the portfolio return and beta. The active traders build portfolios of higher risk, but are at the same time able to earn higher returns. According to the results of the first two models, the return they earn is even higher than could be expected on the basis of the risk in their portfolios. The third model, however, doesn't show any significant extra trading premium. This demonstrates in quite an interesting way the different nature of the alternative performance measures.

The next three models replicate the analysis in the restricted sample. Leaving out from the sample the undiversified and passive investors changes the results in some respects. In the Model (4), where the portfolio return is explained, the beta remains very significant with a reasonable positive coefficient. The unsystematic risk also behaves in a very similar way as in the original sample. The market value, however, isn't significant anymore. The trading activity has a positive and statistically significant coefficient, which indicates better performance of more active traders also in this sample, where the totally passive traders are dropped out. In this sample, the unsystematic risk variable correctly captures the risk effect of undiversified portfolios, and the number of stocks isn't significant anymore. Also, the Vaasa-dummy remains most insignificant.

The Model (5) mirrors the results from the original sample in a very systematic way. The number of stocks in the portfolio is the only variable, which changes its behavior in the model. This is due to the phenomenon discussed above. Finally, the Model (6) completes our analysis. Also here the results are very similar to those from the original sample. The unsystematic risk appears to have a negative effect on the portfolio performance, the other variables being insignificant. Only the market value variable seems to be marginally significant with a positive return premium. As a conclusion, it can be said that the results are quite robust with respect to the sample selection criteria.

5. Conclusions

In the present study we evaluate the portfolio performance of a sample of elderly male investors from two geographical districts in Finland, the Southwest Finland and Southern Ostrobothnia. It is found that there are differences in the investment behavior of the sample investors within and between the two districts. The investors typically hold undiversified portfolios and are passive traders. A certain part of the investors, however, hold acceptably diversified portfolios and trade more actively. At the same time, there is quite large a variation in the return performance and both the systematic and the total risk of the stock portfolios. The analysis of returns, the volatility adjusted returns, i.e. the Sharpe's Measures, and the beta adjusted returns, i.e. the Treynor's Measures, suggests that the carried systematic risk is the most important return driver. Furthermore, the carried unsystematic risk clearly hurts the performance of the investors. The results regarding the effect of trading activity are mixed. The analysis of pure returns and the volatility adjusted returns suggests that activity in trading improves the performance. The beta adjusted returns, however, do not confirm the results, anymore. Most importantly, after the controlling of the risk, there seems to be no difference in the average performance of the investors from the Southwest Finland and the Southern Ostrobothnia.

References

Barber, M. & T. Odean (2000). Trading is hazardous to your wealth: the common stock investment performance of individual investors. *Journal of Finance* 55, 773–806.

Barber, M & T. Odean (2001). Boys will be boys: gender, overconfidence, and common stock investment. *Quarterly Journal of Economics* 116:1, 261–292.

Barber, M. & T. Odean (2002). Online investors: do the slow die first? *Review of Financial Studies* 15:2, 455–487.

Benos, A.V. (1998). Aggressiveness and survival of overconfident traders. *Journal of Financial Markets* 1:3-4, 353–383.

Blume, M.E. & I. Friend (1975). The asset structure of individual portfolios and some implications for utility functions. *Journal of Finance* 30, 585–603.

Cabellé, J. & J. Sákovics (2003). Speculating against an overconfident market. *Journal of Financial Markets* 6:2, 199–225.

Coval, J.D. & T.J. Moskowitz (1999). Home bias at home: local equity preference in domestic portfolios. *Journal of Finance* 54, 2045–2073.

French, K.R. & J.M: Poterba (1991). Investor diversification and international equity markets. *American Economic Review* 81, 222–226.

Gervais, S & T. Odean (2001). Learning to be overconfident. *Review of Financial Studies* 14:1, 1–27.

Goetzmann, W.N. & A. Kumar (2002). Equity Portfolio Diversification. NBER Working Papers 8686.

Grinblatt, M. & M. Keloharju (1999). Distance, language, and culture bias: the role of investor sophistication. Working Paper. Yale International Center for Finance.

Grinblatt, M. & M. Keloharju (2001). What makes investors trade? *Journal of Finance* 56, 589–616.

Kelly, M. (1995). All their eggs in one basket: portfolio diversification of US households. *Journal of Economic Behavior and Organization* 27, 87–96.

Lehenkari, M. & J. Perttunen (2003). Holding on to the losers: Finnish evidence. To appear in Journal of Behavioral Finance.

Odean, T. (1998a). Are investors reluctant to realize their losses? *Journal of Finance* 53, 1775–1798.

Odean, T. (1998b). Volume, volatility, price and profit when all traders are above average. *Journal of Finance* 53, 1887–1934.

Odean, T. (1999). Do investors trade too much? *American Economic Review* 89, 1279–1298.

Polkovnichenko, V. (2003). Household portfolio diversification, Working Paper, University of Minnesota.

Shefrin, H. & M. Statman (1985). The disposition to sell winners too early and ride losers too long: Theory and evidence. *Journal of Finance* 40, 777–790.

Topelius, S. (1875). Maamme kirja. Frenckell.

Tyynelä, M. & J. Perttunen (2003). Trading behaviour of Finnish households: activity, performance and overconfidence. *Finnish Journal of Business Economics* 52:2, 157–178.

Different personality – Different business controller. The enneagram as a tool for analysis

Pekka Pihlanto

Dedicated to Ilkka Virtanen on the occasion of his 60th birthday

Abstract

Pihlanto, Pekka (2004). Different personality – Different business controller. The enneagram as a tool for analysis. In *Contributions to Management Science, Mathematics and Modelling. Essays in Honour of Professor Ilkka Virtanen*. Acta Wasaensia No. 122, 165–185. Eds Matti Laaksonen and Seppo Pynnönen.

The Enneagram is a practical approach, which classifies people into nine basic personality types. The objective of this article is to evaluate the compatibility of the selected Enneagram types with the business controller's tasks, and describe what special contents and features these types tend to imprint on their work and accounting information they prepare. We also describe the tendencies of the types to change their behavior within various situations, and how these changes affect the possibilities of a person to act in the role of controller. In this manner, we come to understand the many possible ways that persons may deal with the controller's role. This examination may also increase awareness with respect to the relative nature of accounting information.

Pekka Pihlanto, Professor Emeritus, Turku School of Economics and Business Administration, Rehtorinpellonkatu 3, FIN–20500 Turku, Finland, e-mail pekka-pihlanto@tukkk.fi.

1. Introduction: The Enneagram and a New Business Controller's Role

The Enneagram is a practical theory or approach, which classifies people into nine basic personality types (in Greek "ennea" means nine and "gram" figure or letter). This approach has been used for centuries, and during the latest twenty years or so, it has become popular in the U.S. in the business world (Palmer 1995; Valtonen & Valtonen 1996; Baron & Wagele 1996; Wagner 1996; Voutilainen 1997; Lindeman et al. 1997; Daniels & Price 1998; Lindeman et al. 1998; Pihlanto 1998, 1999; Hogue 2003). In the U.S., a special journal called the *Enneagram Monthly* is published on the Enneagram field.

Recent management accounting studies have reported a change in the tasks of accounting people towards the new *business controller* (Coad 1996; Hrisak 1997; Friedman & Lyne 1997; Granlund 1997; Granlund & Lukka 1998, 1998 a; Partanen 2001; Järvenpää 2002). In this article, the Enneagram is applied to analyzing the various possible emphases in the work roles of this business controller.

The former accounting person dominating in a firm is often called "bean counter," which describes the mechanical nature of this role very well. The work of this accounting person consists typically of book-keeping and other history-recording types of tasks, and the activities are usually implemented only in the accounting department.

In contrast to this, the activities of business controller are directed outside the accounting department, to business units of the firm and even customers. In this role, which even exceeds the functional borders of the organization, the controller offers the economic viewpoint – e.g., cost consciousness and general efficiency thinking – for the decision-making of the business managers and customers. In many cases, the controller acts as an agent for change as well as an active member in management groups inside the firm.

All these outward-directed features of the role mean that the controller is required to possess certain skills, new in the accounting function, such as communication, education and people skills. It should be mentioned that recent management literature stresses, in a similar way, the relevance of these skills in managerial work in general. All this implies that, to some degree, different personality straits than before are now becoming relevant within business organizations.

The relevance of individuals and personality have been examined to some degree in accounting literature (e.g., Wikman 1993; Granlund 1998; Birkin et al. 1999). In particular, behavioral accounting studies have dealt with an individual, but usually from a very mechanical point of view, treating personality as a variable (see Pihlanto 2002 and literature referred to). In this article, the individual is approached in terms of the notions of *person* and *personality*, and moreover, this is executed in a non-mechanical way, i.e., considering an individual controller as a "genuine" human being.

2. The Notions of a Person and Personality

As background framework, the so-called *Holistic individual image* is applied in this article for defining the nature of human actor, herein the controller. The Holistic individual image has been constructed by philosopher and psychologist Lauri Rauhala (Rauhala 1972, 1973, 1986, 1995; for applications of this framework see Carr & Pihlanto 1998; Vanharanta, Pihlanto & Chua 1997; Pihlanto 2000, 2003).

According to this philosophically based theory, every individual is realized in three modes of existence: *consciousness* (or existence as a psychic-mental phenomenon, as experiencing), *situationality* (or existence in relation to reality, i.e., the world) and *corporeality* (or existence as an organism with organic processes).

In the processes of the consciousness, an individual is forming *meanings* about objects located in his or her situation, and on the basis of these meanings, understands the objects to be something. The totality of the understanding accumulated as meaning structures during one's lifetime is called the *world-view* of an individual. In the process of understanding, all three modes of existence are involved.

Nevertheless, all the modes with problem types peculiar to each of them can also be analyzed separately. For instance, it is possible and often practical as well to analyze meanings in the consciousness without necessarily referring to the brain, a part of the corporeality (which is, of course involved all the time). This is because a person is usually aware of the meanings (thoughts, so to speak) which he or she is forming and can also influence them, but is not aware of what happens at the same moment in the brain, i.e., in the corporeality dimension.

The Holistic individual image defines an individual as a free-willed actor who is able to make choices, and is therefore responsible for his or her actions. This idea – which differs dramatically from the overly used notion of person as stimulus-response automation – is quite relevant in a business firm context in particular.

Every personality typology differentiates people according to their way of concentrating attention on the objects in their situation, i.e., forming meanings about different objects. In this article, the Enneagram is used as a main framework and it defines the *personality*, whereas the Holistic individual image as a background framework determines the notion of *person*. Every individual is assumed to have the features described by the Holistic image. The Enneagram, then, allows for certain personal varieties in describing the activities of individual actors. Thus the psychologically oriented Enneagram and the philosophically based Holistic individual image complete each other and form a totality: the former describes various alternative personalities for a holistic person, and all this is applied to the case of the business controller.

3. The Basic Features of the Enneagram

The basic idea of the Enneagram is the notion that whatever object a person concentrates his or her attention on, it is that very object upon which that person also concentrates his or her resources. This again defines the way of thinking, feeling and acting of the person, i.e., *personality*.

The nine basic personality types of the Enneagram are (Valtonen & Valtonen 1996; Palmer 1995; Wagner 1996; Levine 1999; see also Pihlanto 1998, 1999, 2000 a):

1) the Perfectionist (the "Good Person"),
2) the Giver (the "Loving Person"),
3) the Performer (the "Effective Person"),
4) the Romantic (the "Original Person"),
5) the Observer (the "Wise Person"),
6) the Trooper (the "Loyal Person"),
7) the Epicure (the "Joyful Person"),
8) the Boss (the "Powerful Person"), and
9) the Mediator (the "Peaceful Person").

Because these titles respective to each type are more or less arbitrary, many Enneagram specialists use the number of each type instead of a title. The Enneagram divides the nine types into three subcategories according to their basic orientation in their action: the *instinctive* types (types 8, 9 and 1), the *feeling* types (2, 3 and 4) and the *reason* types (5, 6 and 7).

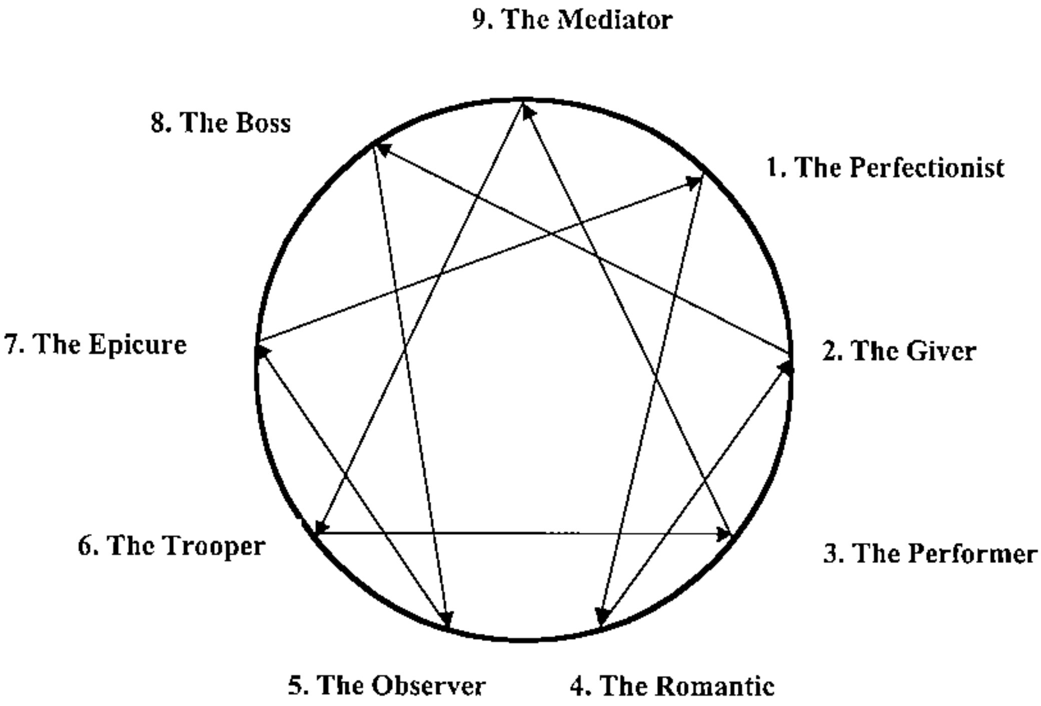


Figure 1. The Enneagram types.

In the Enneagram, every personality type also has – in addition to certain typical basic characteristics – a tendency to adopt both in a stressful situation and under relaxed conditions, certain features (weaknesses and strengths) from the two other specified types. This is an example of the influence that situationality of an individual has on the way of forming meanings, or personality.

In a *stressful situation*, every Enneagram type has the tendency to adopt features from the type to which the arrow goes from the first-mentioned type in Figure 1 (e.g., Type 1 tends to adopt under stress, from Type 4). Analogously, in a *relaxed situation*, the adoption tendency is described by the arrow, which *derives from* the type, which is the "source of adoption" over to the adopting type (e.g., Type 1 tends to adopt, in a relaxed situation, from Type 7).

In addition to these changes of a basic type, each type has "*wings*". This means that every type has the possibility to adopt features from either of the types located right beside the type in the Enneagram circle in Figure 1 (the wings of Type 1 are 9 and 2).

Because of these dynamics, the Enneagram offers potentiality for the self-development of an individual. A person who knows his or her type, the wings and the tendencies to change in stressful and relaxed situations may try to strengthen the positive and avoid negative tendencies. This means that an individual may try to change the meanings emerging in his or her consciousness into a direction of positive qualities, which in turn would cause positive effects emerging in the corporeality. In addition, the perceived situationality would also change correspondingly.

4. The Enneagram and the Role of Controller

4.1. Different Personality Means Different Controller

A general motivation for using the Enneagram in a management and accounting context is based on the above self-developing possibilities offered by it. When a person in an organization is aware of the characteristics of his or her own type and its strengths and weaknesses as well as its characteristic path to development, he or she is able to act more consciously and effectively and fulfil his or her potentiality.

Further, by being aware of the Enneagram type of people in and around the organization, a controller is better equipped by being in communication with these. The controller is now able to realize and take into consideration the fact that his or her own way of approaching problems and issues is not necessarily the only correct one, but there are other ones equally well-motivated. The controller is now able to present the message in a way, which the receiver has the best possibilities to understand. In the same way, a user of accounting information can adjust his or her own behavior according to the personality of the controller he or she is dealing with.

How, then, the personality becomes visible in the role of controller, and how it reflects on an individual's possibilities to act as a controller? The *objective* of this article is, in particular, to evaluate the compatibility of the selected Enneagram types with the controller's tasks, and describe what special contents and features these types tend to imprint on their tasks. With this objective in mind, we also describe the tendencies of the types to change within various situations, and how these changes affect the possibilities of a person to act in the role of controller. In this manner, we come to understand the many possible ways that persons may act in the controller's role.

Thus far, the role of a controller has usually been presented by generally describing the relevant *tasks*. All the persons concerned are assumed to adapt to the task in a uniform fashion, and consequently, the personality of the controller is not considered relevant at all, and therefore ignored.

In contrast, this article suggests that the role of a controller is, in practice, shaped by the individual's way of forming meanings in the current situation, i.e., by his or her Enneagram type. As mentioned above, the way a person shapes his or her actions is assumed to be dynamic. Therefore, our thesis is that a different personality means a different controller.

The *method* of the article is conceptual (see Pihlanto 1994). The result, therefore, is a kind of conceptual system describing the Enneagram-based personality variations as these affect the controller's role. On the basis of this conceptual understanding, it should be possible to approach the controller's role in further studies on an empirical basis.

In this article, we deal with a selection of types, trying to contribute the idea of utilizing the Enneagram in an accounting context. The Enneagram types of example range from the Boss (8), the Performer (3) and the Mediator (9) to the Romantic (4). In addition, all the

Enneagram types are briefly examined by speculating on the possible impact of the personality on the content of calculations.

To begin with, the Boss and the Performer seem typical in business organizations, at least in considering the view of a manager as presented in both management literature and in everyday discussions.

4.2 The Boss is a Different Controller from the Performer

4.2.1 The Boss or the Powerful Person is Willing to Influence

The Boss or the Powerful person (8) is concerned about justice and the fair use of power, and tends to avoid weakness. This instinctive type is characterized in Enneagram literature as forceful, strong-willed and autonomous, as well as eager to influence and take the lead as well as responsibility. The Boss also works hard and tends to be fair and fearless in relationships to other people who are relevant to his or her situationality. It could be questioned as to whether this kind of strong and assertive person is the best possible choice for the communicative, people-centered and educative task of the controller.

Still worse – as in the case of every type – the Boss also has negative characteristics which, when realized, further weaken his or her possibilities to deal with the controller's work. At worst, the Boss can be possessive, insensitive, non-listening and dominating, and may also become very confrontational. This kind of excessively "power-hungry" or even dictatorial kind of person would appear not to be a good choice for any position in the modern organization.

However, according to the Enneagram, it is possible that features adopted from the "wings" – here the *Epicure* (7) or the *Mediator* (9) – may soften the attitude of the Boss in relation to other people. The Epicure or the Joyful person is optimistic, friendly, creative, visionary and extroverted. These features would make the Boss people-oriented and, as such, more successful as a controller. This however, assumes that the negative sides of the

Epicure – superficiality, irresponsibility or unreliability – do not surface. The Mediator or the Peaceful person wing may then produce patience, diplomacy, objectivity and harmony in the work of the Boss, which would make the controller successful in regard to people skills. Again, there is the danger of the other side of the coin being realized: the Boss may adopt the Mediator's negative features and become indecisive, uncommitted, neglectful, overly adaptable and detached.

In a stressful situation, the Boss (as with every type) has a tendency to first drive in the direction of the weaknesses of his or her own type. In this case, the use of power takes on greater severity, and the behavior of the Boss may be insensitive, suspicious and overly aggressive. It is self-evident that this weakens the possibilities of acting successfully as a controller.

If this stress continues, the Boss has a tendency to adopt from the *Observer* or the Wise person (5): in the beginning, the weak points of this type. In this case, the controller may operate alone, avoid commitment, postpone action and become uncommunicative, even critical – clearly not welcome trends. But in the event that the positive characteristics of the Observer are realized, the Boss will change to become observant, reflective and logical as well as non-intrusive and even philosophical. These gifts from the Observer – a 'reason' type – provide the instinctive Boss with the possibility of balanced application of reason, in both power-oriented and other kinds of behavior.

Under relaxed situation, however, the Boss has a tendency to adopt features from the feeling type – the *Giver* or Loving person (2). The Giver's favorable features make the Boss helpful, other-centered, supporting, sensitive and sympathetic. In this event, the controller accepts more willingly the suggestions of others. This relationship-oriented attitude makes the Boss better equipped in dealing with people, which is crucial to the role of controller.

If, however, the negative features of the Giver win, the Boss may become manipulative, possessive and complaining. The Giver may be also overprotective and may adopt the role of a martyr. This means problems for a controller.

In the end, as a result of a process of growth, the Boss may find the positive features of his own type and give the proper attention to other people. Exaggerated self-confidence and aggression diminishes, and the lust for power is reconciled. The Boss is now fair, and takes other people's wishes in consideration. This mature person is both a better human being and a more successful controller.

4.2.2 The Performer or the Effective Person Attempts to Be the Best One

The Performer or the Effective person (3) is a highly performance-oriented type who wants to be valued on the basis of his or her performance. The Performer tries to be the best one, and honestly believes that people appreciate top-performers. This type tends to be efficient and get things done. The Performer is also an enthusiastic, pragmatic and goal-oriented manager type who considers impressions and images as very important. At best, the Performer is a popular, dynamic and self-assured team builder.

All these are clearly favorable features for a business controller. According to Enneagram literature, the Performer is often a successful business manager or sport-team leader. However, this type is not necessarily an ideal controller, because a strong performance-orientation may act as a disturbance with respect to cooperation and negotiation with other people. Moreover, images are not enough in controller's work; rather, real substance is needed.

Negative features naturally diminish the possibilities of the Performer in the tasks of controller. If these become realized, the controller is impatient – even 'workaholic' in character – and excessively image-conscious. Such a person becomes success-driven and a political overachiever who ignores feelings. Perhaps surprisingly, the Performer is a feeling type, but usually keeps his or her feelings hidden.

The Performer's wings are the Giver (2) and the Romantic (4). As mentioned above, at best the Giver tends to be helpful, sensitive and sympathetic, which strengthens people skills and makes the type better equipped as a controller. The Giver's negative features may make the Performer manipulative and complaining, which mean that giving and helping are no longer genuine.

The Romantic wing, then, gives the Performer sensitivity, creativity and intuition, as well as a certain nostalgic and aesthetic flavor. The Romantic may not be the best possible source of features for a controller, even if creativity and intuition might be required in change and development processes in a firm. Again, negative features may be realized, such as overly dramatic, exaggerating, eccentric and overly sensitive attitudes.

In a stressful situation the Performer also initially uses the weaknesses particular to this type and works increasingly in the manner of a workaholic, trying to give a still more efficient impression than before. The Performer may start new projects and new contacts in a compulsive fashion, etc. This kind of controller may at first seem efficient, but the lack of authenticity is detected sooner or later. In the second stage of the stress process, the Performer tends to shift towards the negative features of the Mediator (9). This makes the person indecisive and uncommitted or overly adaptable. Also, the energy level may sink. These are not the characteristics of a communicator or agent of change, i.e., a controller.

In case the Performer experiences stress as a possibility, he or she adopts the positive strains of the Mediator. This means an inwards-bound orientation and peaceful reflection on things in terms of new meanings. The Performer is now patient and diplomatic, more objective and permissive. The person lives in harmony and is calm and easygoing. This kind of controller is mature and down-to-earth, and not the 'compulsive performance-automaton' he or she was previously. Even if the controller does not give a very energetic and dynamic impression, the results of the work are probably now better – and the stressful situation may be effectively resolved.

Under peaceful situation the Performer may be the recipient of positive features from the Trooper or Loyal person (6). These tend to make the Performer cautious and reliable in addition to being loyal. Furthermore, conscientiousness, charm and tenaciousness may be the gifts of this thinking type. The Performer may begin to suspect the overwhelming relevance of images and performance, and begins to understand his or her own feelings better than before. The controller is now a loyal partner in cooperation, and respects other people's views.

However, the negative features of the Trooper may weaken the chances for the successful work of a controller. In this case, the Performer tends to become suspicious, indecisive and conservative as well as overly security-conscious. This does not relate very well to the attitude the role of an agent of change would require.

In the end, when the Performer finds the strong sides of his or her own type, he or she shall also find the energy necessary for developing the mental dimension – i.e., consciousness. Resultantly, this type finds and expresses real feelings instead of representing the feelings he or she believes the role requires.

The Performer is now capable of accepting failure, and does not endeavor to build a 'side scene' of wrong impression, which attempts to belittle the failure and transfer the guilt to other parties. The Performer is no longer forced to act mechanically and in the manner of one driven by success. The point of view of society and other people become relevant. This maturity gives the controller a long-term competitive advantage, which clearly exceeds the earlier results-oriented image strategy.

The above descriptions, which are based on Enneagram literature, show the different possibilities of development the Boss and the Performer possess in their work as a controller. In this way, they can change their consciousness and the contents of their world-view in the process of adjusting to changes in their situationality. Neither of them is predestined to the limits of their basic types, but can develop as human beings and controllers on the basis of alternatives offered by their wing types as well as types activated under conditions of a stressful or peaceful situationality. Because the human actor is a totality, the possible positive effects realized in the consciousness and situationality also reflect in corporeality. Concretely, this may result in better health on the part of the person concerned.

In the following, we take two additional types – the Mediator and the Romantic – under short analysis. Both of them were briefly described above – the Mediator both as wing of the Boss and stress type of the Performer, and the Romantic as a wing of the Performer.

These may not be the types one expects to meet in a controller's role or in management positions in general. It is, however, possible that these personality types are also acting as a controller in practice, and therefore it may be interesting to analyze their possibilities in this role.

4.3 The Mediator is at Home in the Role of Business Controller – But How About the Romantic?

4.3.1 The Mediator – One of the Best Choices for a Controller

Even if the Mediator or Peaceful person (9) may not be management-oriented and strive for positions in business, it can be actually assumed that this type would be one of the best possible choices for the position of controller. This is because the Mediator typically resolves and even avoids conflicts and is therefore skilful in personal relationships, which are crucial in this role.

The chances of this type to act successfully as a controller may still increase if the deficient efficiency, optimism and goal-orientation required are adopted from the Performer or the Effective person (3). This tendency is realized – according to the Enneagram – in a peaceful situation.

On the other hand, the Mediator can acquire strength and fearlessness from the Boss or the Powerful person wing (8), or precision and ethics from the other wing, the Perfectionist or the Good person (1). Furthermore, from the stress-type, the Trooper or the Loyal person (6), the Mediator may adapt reliability and responsibility, which as mentioned are useful features in the work of controller – and in crisis or stressful situations in particular.

Finally, when growing towards the positive features of one's own type, the Mediator acts in a peaceful, patient and diplomatic way. He or she is now modest and objective, as well as in harmony with the own preferences and feelings. This kind of person is reassuring and capable of taking the views of other people into consideration, which is important in the controller's role.

4.3.2 The Romantic – Too Sensitive for a Role of a Controller?

As to the Romantic or the Original person (4), it could be argued that this type would not be the best possible person in the role of the controller, because of this type's tendency to be sensitive as well as individualistic and to long for something unattainable. In particular, the negative features of this type would be harmful in the role of controller, such as dramatic and eccentric behavior.

Again, there is the possibility for positive change: this occurs for instance, by adopting a certain precision, realism and desire for high standards from the Perfectionist or the Good person (1), which may be realized in a peaceful situation. Moreover, the Romantic may seek welcomed performance orientation and effectiveness from the Performer (3) wing, or a thinking type's reflectivity and logic from the Observer (5) wing. These tendencies may complete the sensitive and caring characteristics of the Romantic to more appropriately fit the controller's accounting-oriented and, as such, calculating role.

If the Romantic obtains positive features from the Giver (2) in a stressful situation and becomes helping, supporting and other-centered, he or she seems rather satisfactory in the controller's role. Of course, the negative features of the Giver – such as manipulative, complaining and overprotecting behavior – lead to an opposite conclusion.

Finally, in case the Romantic succeeds in developing the positive features of his or her own type, he or she overcomes the emotional problems. "Here and now" kind of realism overdrives the overly romantic attitude to work and other people. This kind of person is not a bad alternative in a controller's role at all. It is, however, rather improbable that the

Romantic chooses a controller's position, but as referred to above, it is not totally excluded.

4.4. Impact of the Personality on the Content of Calculations

The suitability of various personality types for the work of controller was described above. In particular, the analysis showed how the structure of a person's consciousness – world-view included – or personality may affect the way a person behaves in a controller's role, and therefore, how well the type concerned is adapted for this role. In addition, it was also discussed how the dynamics inherent in human character may change the consciousness of a person who appears at first sight to poorly fit the controller's task and situation, in such a manner that the chance to succeed appears.

However, does the same apply in the *results* of the work of a controller or accountant, i.e., is it possible that various personality types produce, at least to some degree, differently accentuated calculations? I would answer affirmatively. The various Enneagram types show a concentration of attention on different things and therefore form various kinds of meanings in a given situation. This is why it is rather evident that in cases where there are options in formulation, with respect to the details of the calculation – as in the planning context in particular, but to some degree also in ex post calculations – different choices are made by different personality types. On the basis of the characteristics of the various types, it is possible to speculate on the probable tendencies the types imprint in the accounting information they prepare.

The *Perfectionist* (1) is willing to do everything in a correct way. Therefore, this type deals with the details of the calculation in order to produce a reliable and "correct" result, i.e., the best possible or a "perfect" result. The standards applied are typically of high quality and every detail of the calculation and its background assumptions are scrutinized.

The *Giver* (2) is willing to help people and may therefore reflect on how the calculation would serve and support the decision-maker in the best possible way. Because the Giver wants to help people, the calculation tends to be just such as the Giver assumes the user of it would desire.

The performance-oriented *Performer* (3) would typically tend to create an optimistic and goal-oriented calculation. In addition, this type may consider efficiency in respect to the contents of the calculation and in its formal side as well – at least providing an *impression* of efficiency. This type may use the calculation as a means for demonstrating that he or she is efficient ("the best one"), and in this way to promote his or her career.

The sensitive and original *Romantic* (4) may offer creative and intuitive solutions. The form and content of the calculation may be aesthetic but nevertheless expressive as well. However, the calculation may also be exaggerated, and the Romantic may try to produce a dramatic effect by it.

Being observant and reflective by nature, the *Observer* (5) tends to construct thoughtfully analytical and objective solutions – somebody might call them "witty and pithy". The solutions made in the process of calculation may also be cautious and conservative, and the calculation work tends to take a lot of time.

The *Trooper* (6) may criticize the assumptions of the calculation, but for that very reason, the resulting calculations are highly reliable and also cautious. The solutions offered may tend to prove certain loyalty in relation to the user(s) of the calculation.

The *Epicure* (7) – also called the *Optimist* – may release his or her full imagination and may arrive spontaneously at an optimistically formulated and perhaps creative solution. Completing the task may be a highly singular business, which may be visible in the results as well.

The *Boss* (8) typically acts in a self-confident way, and may keep an eye on possibilities in order to exert an influence by means of the calculation. Quite appropriately, the power-oriented use of an accounting system as an "ammunition machine" has been reported in the literature (Hopwood 1974). However, this is without reference to the personality of the accountant in question. This kind of action means that a calculation is not a pure "fact" or the best possible estimate, but fabricated in order to influence other people in a way that is desirable for the accountant.

The *Mediator* (9) may formulate the calculation at ease and in a diplomatic way, which does not harm anybody concerned. Therefore, this type's calculations would not include any seeds of conflict, but provide an impression of steady development with respect to the business concerned.

Of course, these brief characterizations should not be taken too seriously, because every individual – even those of the same Enneagram type – is unique as to consciousness, situationality and corporeality. In addition, every type may vary in behavior according to the situation, in the above manner.

These descriptions, however, suggest the plausible assumptions about the way in which each type may imprint his or her characteristics on the results and process of the accounting task. The knowledge of this possibility might help people in an accounting context to act in the appropriate manner, e.g., to try to eliminate the possible personality-driven biases included in accounting information – by themselves or by other people.

5. Conclusion: Everybody Can Change

The person is defined above in terms of the holistic individual image, which means that every individual person is realized in three modes of existence, consciousness, situationality and corporeality. In this general realm, individual personalities were analyzed in terms of Enneagram types and their variations. This provides the possibility for a more detailed and rich description of an individual than the holistic concept as such. The role of the latter is to provide a general framework for describing the ways of people's actions.

In the above manner, the pros and cons of the features of every Enneagram type can be analyzed in light of their suitability to a controller's role. On the basis of this, some understanding is gained about the spectrum of possible variations and dynamics in these roles as possible in practice. It should be realized, however, that human actors are much more many-sided than the above analysis may reveal. In any event, the knowledge of these tendencies – both in regard to one's own Enneagram type and the other relevant people at the work place – may prove extremely useful if properly utilized. This applies both to controllers and to those dealing with them.

On a general level, this knowledge about different Enneagram types and their possible effects on the work of the controller may also increase awareness with respect to the relative nature of accounting information. All too often, non-accounting personnel in particular consider mere quantitative and numerical information as objective and truthful. In reality, however, this information is also more or less influenced by the subjectivity of individuals – both producers and users of accounting information – who behave according to their personality straits, their unique world-views and situationalities when formulating and "reading" accounting information.

One of the general positive messages of this article is that every individual, irrespective of their Enneagram type, has the possibility to change in order to more appropriately fit the work role at hand. This change happens by means of the active role of the person in question in terms of the attempt to realize the natural tendencies for positive development of his or her own type. This result emphasizes those individual – and as such behavioral – aspects in the role of business controller, which have been mostly ignored in the current accounting literature in spite of their crucial relevance in practice.

References

Baron, R. & E. Wagele (1996). *Yhdeksän hyvää tyyppiä. Enneagrammi itsetuntemuksen ja kanssakäymisen oppaana* (The original title: The Enneagram Made Easy: Discover the Nine Types of People). Jyväskylä: Arena.

Birkin, F., P. Edwards & D. Woodward (1999). The Accounting Craftsperson: A Response to Contemporary Developments. A paper presented at the 22nd Annual Congress of the European Accounting Association, Bordeaux, France 5–7 May 1999.

Carr, A. & P. Pihlanto (1998). From homo mechanicus to the holistic individual: A new phoenix for the field of organisation behaviour? In *Current Topics in Management*, Vol. 3, 69–91. Eds M. Afzalur Rahim, Robert T. Golembiewski & Craig C. Lundberg. Stamford, Conneticut, USA: JAI Press.

Coad, A. (1996). Smart work and hard work: Explicating a learning orientation in strategic management accounting. *Management Accounting Research* 7:4, 387–408.

Daniels, D.N. & V.A. Price (1998). *Stanford Enneagram Discovery Inventory and Guide.* Redwood City Ca. USA: Mind Garden Inc.

Friedman, A.L. & S.R. Lyne (1997). Activity-Based Techniques and the Death of the Beancounter. *The European Accounting Review* 6:1, 19-44.

Granlund, M. (1997). From Bean-Counters to Change Agents: The Finnish Management Accounting Culture in Transition. A paper presented at a work-shop in Vesterås May 28, 1997.

Granlund, M. (1998). *The Challenge of Management Accounting Change. A Case Study of the Interplay between Management Accounting, Change and Stability.* Publications of the Turku School of Economics and Business Administration, Series A-7:1998.

Granlund, M. & K. Lukka (1998). It's a Small World of Management Accounting Practices. *Journal of Management Accounting Research* 10:1, 153-179.

Granlund, M. & K. Lukka (1998 a). Towards Increasing Business Orientation: Finnish Management Accountants in a Changing Cultural Context. *Management Accounting Research* 9:2, 185-211.

Hogue, H. (2003). The Enneagram. http://www.prosperity.com/enneagram/

Hopwood, A.G. (1974). *Accounting and Human Behaviour.* London: Haymarket Publishing.

Hrisak, D.M. (1997). Controllers viewpoint: Ethics training in cyberspace? *Corporate Controller* 10:2, 42-44.

Järvenpää, M. (2002). *Johdon laskentatoimen liiketoimintaan suuntautuminen laskentakulttuurisena muutoksena – vertaileva case-tutkimus.* Summary: Business Orientation of Management Accounting as a Cultural Change – A Comparative Case Study. Publications of the Turku School of Economics and Business Administration, Series A-5:2002.

Levine, J. (1999). *The Enneagram Intelligences. Understanding Personality for Effective Teaching and Learning.* Westport, Conneticut: Bergin & Garvey.

Lindeman, A. & K. Valto & E. Voutilainen (1997). *Yhdeksänsärmäinen työyhteisö. Enneagrammi erilaisuuden karttana* ("The Nine-Angled Working Community. The Enneagram as a Map of Personal Differences"). Espoo: EV-kehitysyhtiöt.

Lindeman, A. & L. Valtonen & E. Voutilainen (1998). *Enneagrammiopas* ("The Enneagram Guide"). Espoo: EV-kehitysyhtiöt.

Palmer, H. (1995). *The Enneagram in Love & War. Understanding Your Intimate & Business Relationships.* San Fransisco, New York: Harper.

Partanen, V. (2001). *Muuttuva johdon laskentatoimi ja organisatorinen oppiminen: Field-tutkimus laskentahenkilöstön roolin muutoksen ja uusien laskentainnovaatioiden käyttöönoton seurauksista.* Summary: The Changing Management Accounting and Organisational Learning. Publications of the Turku School of Economics and Business Administration, Series A-6:2001.

Pihlanto, P. (1994). The action-oriented approach and case study method in management studies. *Scandinavian Journal of Management* 10:4, 369–382.

Pihlanto, P. (1998). *Yritysjohtajan ja controllerin toiminta enneagrammiteorian valossa.* Summary: Roles of a Manager and Business Controller in the Light of the Enneagram Theory. Publications of the Turku School of Economics and Business Administration, Series A-8:1998.

Pihlanto, P. (1999). *Perfektionisti, seikkailija ja romantikko. Jazzin uudistaja John Coltrane enneagrammin valossa* ("The Perfectionist, Epicure and Romantic. John Coltrane, An Innovator of Jazz, in the Light of the Enneagram"). Publications of the Turku School of Economics and Business Administration, Series Discussion and Working Papers 6:1999.

Pihlanto, P. (2000). *An Actor in an Individual Situation: The Holistic Individual Image and Perspectives on Accounting Research.* Publications of the Turku School of Economics and Business Administration, Series Discussion and Working Papers 4: 2000

Pihlanto, P. (2000 a). *Nine Types of Controller. The Role of Business Controller in the Light of the Enneagram Theory.* Publications of the Turku School of Economics and Business Administration, Series Discussion and Working Papers 10: 2000.

Pihlanto, P. (2002). *Understanding Behaviour of the Decision-Maker in an Accounting Contex. The Theater Metaphor for Conscious Experience and the Holistic Individual Image.* Publications of the Turku School of Economics and Business Administration, Series A-1: 2002.

Pihlanto, P. (2003). The role of the individual actor in different accounting research perspectives. The holistic individual image as a tool for analysis. *Scandinavian Journal of Management* 19:2, 153–172.

Rauhala, L. (1972). The hermeneutic metascience of psychoanalysis. *Man and World* 5, 273–297.

Rauhala, L. (1973). *The Regulative Situational Circuit in Psychic Disturbance and Psychotherapy*. Studia Philosophica in Honorem Sven Krohn. Publications of Turku University, Humaniora, Turku, Serie B.

Rauhala, L. (1986). *Ihmiskäsitys ihmistyössä* ("The Conception of Human Being in Helping People"). Helsinki: Gaudeamus.

Rauhala, L. (1995). *Tajunnan itsepuolustus* ("Self-Defense of the Consciousness"). Helsinki: Yliopistopaino.

Wagner, J.P. (1996). *The Enneagram Spectrum of Personality Styles. An Introductory Guide*. Portland, OR: Metamorphous Press.

Wikman, O. (1993). *Yrityksen investointiprosessi ja siihen vaikuttavia tekijöitä. Toiminta-analyyttinen tutkimus*. Summary: The Company Investment Process and Factors that Influence it. Publications of the Turku School of Economics and Business Administration, Series A-7:1993.

Valtonen, L. & O. Valtonen (1996). *Yhdeksänkulmainen peili. Paranna itsetuntemustasi enneagrammin avulla* ("The Nine-Angled Mirror. Increase Your Self-Knowledge With the Help of the Enneagram"). Helsinki: Kirjapaja.

Vanharanta, H., P. Pihlanto & A-M. Chang (1997). Decision Support for Strategic Management in a Hyperknowledge Environment and the Holistic Concept of Man. In *Proceedings of the Thirtieth Hawaii International Conference on System Sciences, Volume V, Advanced Technology Track*, 307–316. Ed. Ralph H. Sprague Jr. Los Alamitos, California: IEEE Computer Society Press.

Voutilainen, E. (1997). *Persoonallista kasvua. Enneagrammi elämänvalintojen karttana* ("Personal Growth. The Enneagram as a Map of the Choices in Life"). Espoo: Enneagrammi Suomi Oy/EV-kehitysyhtiöt.

Dynamic equilibrium correction modelling of credit spreads. The case of yen Eurobonds

Seppo Pynnönen[1], Warren Hogan, and Jonathan Batten

Dedicated to Ilkka Virtanen on the occasion of his 60th birthday

Abstract

Pynnönen, Seppo, Warren Hogan, and Jonathan Batten (2004). Dynamic equilibrium correction modelling of credit spreads. The case of yen Eurobonds. In *Contributions to Management Science, Mathematics and Modelling. Essays in Honour of Professor Ilkka Virtanen.* Acta Wasaensia No. 122, 187–204. Eds Matti Laaksonen and Seppo Pynnönen.

This study specifies an equilibrium correction model of the credit spreads on quality Japanese yen Eurobonds. In an important paper by Longstaff and Schwartz (1995) the authors derive a closed form solution of the arbitrage-free value on risky debt in continuous time. However, in discrete time real world data series it is common that many non-stationary economic and financial time series are cointegrated. Nevertheless, until now there is no theory for continuous time cointegration. In addition, the existence of cointegration in prices leads to incomplete markets so that the arbitrage-free valuation should not apply. Instead one must rely on equilibrium pricing, where the markets clear in the equilibrium via a potentially complicated adjusting process. In this paper the important factors driving the credit spreads are introduced into the equilibrium relation and the adjusting process are investigated. The results indicate that the corporate bond yields are cointegrated with the otherwise equivalent Japanese Government Bond (JGB) yields, with the spread defining the cointegration relation. Furthermore, the results indicate that the equilibrium correction term is highly statistically significant in modelling spread changes. The other important factor is the risk-free interest rate with the negative sign as predicted by the Longstaff and Schwartz (1995). On the other hand there is little evidence of the contribution of the asset return to the spread behaviour. The estimated coefficient of the equilibrium correction term indicates that the adjustment process is fairly slow, which indicates that the clearing process in the markets takes time.

Seppo Pynnönen, Department of Mathematics and Statistics, University of Vaasa,
P.O.Box 700, FIN-65101, Vaasa, Finland, e-mail sjp@uwasa.fi.
Jonathan Batten, College of Business Administration, Seoul National University, San 56-1, Sillim-Dong, Kwanak-gu, Seoul, Korea, e-mail: jabatten@snu.ac.kr.
Warren Hogan, School of Finance and Economics, University of Technology, Sydney
NSW 2007 Australia, e-mail warren.hogan@uts.edu.au.

Key words: Credit Spreads; Eurobonds; Japan; Equilibrium Correction

[1] *Corresponding author.*

1. Introduction

While the Japanese Government generally does not issue securities in Eurobond, or foreign bond markets, Japanese Yen denominated Eurobond issues by the Japanese corporate sector are now the second largest in terms of new issues and outstandings after issues in U.S. dollars, and comprise nearly USS508 billion worth of outstanding bonds. From the viewpoint of corporate issuers in yen and international portfolio managers holding yen securities, the growth in the yen bond market provides important opportunities for diversification of funding and risk. However, in spite of the importance of the markets, little is known of the behaviour of yen denominated securities generally, and there are few empirical studies that investigate the relation between risky and riskless yen bonds, or specifically, the credit spreads which represent the difference in yield between these two risk classes of security. Batten, Hogan, and Pynnonen (2003a) investigate the time series relationships of Yen Eurobond spread changes to the most important factors predicted by Longstaff and Schwartz (1995) model and some additional factors found important mainly on U.S. markets. Batten, Hogan, and Pynnonen (2003b) further investigates the credit spread behaviour, and find that the equilibrium correction, time varying volatility and correlation factors are potentially important factors affecting the spread behaviour.

Recent theoretical developments on the valuation of risky debt proposed by Longstaff & Schwartz (1995), Das and Tufano (1996) and Duffie and Singleton (1999), predict a negative correlation between changes in default-free interest rates, the return on risky assets and changes in credit spreads. The empirical evidence in support of this relation is mixed. Originally, in U.S. bond markets Longstaff and Schwartz (1995) found evidence of a negative relation for both interest rate and asset changes. A weak but significant negative relation between changes in credit spreads and interest rates was also found by Duffee (1998) and Collin-Dufresne, Goldstein, and Martin (2001), while Neal, Morris, and Rolph (2000) identified a negative short-term relationship with credit spreads that reversed to positive in the long-run. For non-U.S. markets, Kamin and von Kleist (1999) find little evidence of a short-term relation between industrial country interest rates and emerging market bond spreads.

While the above papers focus directly on the modelling under the restriction of static equilibrium relation, the potential cross dynamics between the series demands a far more versatile approach. The strategy in this paper is to start off with an unrestricted setup using vector autoregression (VAR) of the important factors found in the earlier papers, particularly in Batten, Hogan, and Pynnonen (2003a,b), and proceed to identify the cross-dynamics between the series. The obvious advantage of this approach is that we can identify the important feedback relations between the credit spread and related main factors. This allows us to determine the time lags in the adjustment process to the equilibrium once shocks have driven the yields out of the equilibrium.

There is surprisingly much evidence that the credit spreads, measured in terms of yield differences over the corresponding government bonds, per se are non-stationary (for example Pedrosa and Roll 1998, Mansi & Maxwell 2000). In spite of these, it is fairly implausible that the yields could wander without bounds apart from each other in the long run. This was also confirmed in Batten et al. (2003b), where strong evidence was found for stationarity of the AA and AAA rated Yen Eurobonds.

Batten et al. (2003b) find that the most important factor driving the credit spread changes, as predicted by Longstaff and Schwartz (1995), is the changes in the risk-free rate. On the other hand the second important factor, firm asset return, implied by the Longstaff-Schwartz model did not show as being statistically significant. The reason for this may be that the stock market general index returns, that are usually utilized as proxies for the firm asset returns, do not properly reflect the true asset position, particularly when restricted to AA and AAA rated bonds. The other important factor found in Batten et al. (2003b) was the change in the slope of the term structure of Japanese government bonds proxied by the change in the yield spread of 20 year and 2 year bonds. Two other factors were the cointegration relation (spread), and conditional volatility of the spread changes. The importance of the conditional volatility as an explanatory variable in the mean equation of credit spread changes is again a factor that is not predicted by the theoretical model by Longstaff and Schwartz (1995). However, Engle, Lillien, and Robins (1987) find that the excess yield of the long bond depends on the conditional variance rather than being a constant. The empirical results of Batten et al. (2003b) indicate also that the conditional volatility is potentially an important factor in determining the credit spreads. The Longstaff and Schwartz (1995) model predicts that asset return volatility should be an important determinant of the credit spread. In the case of the corporate bonds it may well be that the conditional volatility of the credit spread replaces the more traditional stock return volatility as a surrogate for the firm asset return volatility.

Utilizing these findings we focus on dynamic modelling of the equilibrium relation of credit spreads on the yen Eurobonds. The paper contributes to the existing literature by specifying the dynamic structure of the equilibrium correction and identifying the important feedback links between the factors related to the credit spreads.

2. Methodology

Let $\mathbf{y}_t = (y_{1,t},\ldots,y_{p,t})'$ denote a column vector of p time series, where the prime denotes transposition. Denote the general unstructured vector auto regression VAR model as

$$(1) \qquad \mathbf{y}_t = \mathbf{\Phi}_1\mathbf{y}_{t-1} + \cdots + \mathbf{\Phi}_k\mathbf{y}_{t-k} + \mathbf{B}\mathbf{x}_t + \boldsymbol{\varepsilon}_t = \mathbf{\Phi}(L)\mathbf{y}_t + \mathbf{B}\mathbf{x}_t + \boldsymbol{\varepsilon}_t,$$

where $\mathbf{\Phi}_j$, $j = 1,\ldots,k$ are $p \times p$ autoregressive coefficient matrices, $\mathbf{\Phi}(L) = \mathbf{\Phi}_1 L + \mathbf{\Phi}_2 L^2 + \cdots + \mathbf{\Phi}_k L^k$ is the matrix lag-polynomial, $\mathbf{x}_t$ is a q-vector of predetermined variables including the intercept term unit vector, possible trends, seasonal terms etc., $\mathbf{B}$ is the $p \times q$ coefficient matrix of the predetermined variables, and $\boldsymbol{\varepsilon}_t$ is independent normally distributed error term with contemporaneous correlation matrix $\mathbf{\Sigma}$. Partition $\mathbf{y}_t = (\mathbf{y}'_{1,t}, \mathbf{y}'_{2,t})'$ such that $\mathbf{y}_{1,t}$ is a p_1-vector $(1 \le p_1 \le p)$ where the series are $I(1)$, and $\mathbf{y}_{2,t}$ is a $p - p_1$ vector of stationary variables. Using the partition, write (1) as

$$(2) \qquad \begin{pmatrix} \mathbf{y}_{1,t} \\ \mathbf{y}_{2,t} \end{pmatrix} = \begin{pmatrix} \mathbf{\Phi}_{11,1} & \mathbf{\Phi}_{12,1} \\ \mathbf{\Phi}_{12,1} & \mathbf{\Phi}_{22,1} \end{pmatrix} \begin{pmatrix} \mathbf{y}_{1,t-1} \\ \mathbf{y}_{2,t-1} \end{pmatrix} + \cdots + \begin{pmatrix} \mathbf{\Phi}_{11,k} & \mathbf{\Phi}_{12,k} \\ \mathbf{\Phi}_{12,k} & \mathbf{\Phi}_{22,k} \end{pmatrix} \begin{pmatrix} \mathbf{y}_{1,t-k} \\ \mathbf{y}_{2,t-k} \end{pmatrix} + \begin{pmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \end{pmatrix} \mathbf{x}_t + \begin{pmatrix} \boldsymbol{\varepsilon}_{1,t} \\ \boldsymbol{\varepsilon}_{2,t} \end{pmatrix}.$$

Because all the I(1) variables are in $\mathbf{y}_1$, the equilibrium correction (ECM) representation can be written as

(3a)

$$\Delta\mathbf{y}_{1,t} = \mathbf{\Pi}_1\mathbf{y}_{1,t-1} + \mathbf{\Gamma}_{1,1}\Delta\mathbf{y}_{1,t-1} + \cdots + \mathbf{\Gamma}_{1,k-1}\Delta\mathbf{y}_{1,t-k+1} + \mathbf{\Phi}_{12,1}\mathbf{y}_{2,t-1} + \cdots + \mathbf{\Phi}_{12,k}\mathbf{y}_{2,t-k} + \mathbf{B}_1\mathbf{x}_t + \boldsymbol{\varepsilon}_{1,t}$$

and

(3b)

$$\mathbf{y}_{2,t} = \mathbf{\Pi}_2\mathbf{y}_{1,t-1} + \mathbf{\Gamma}_{2,1}\Delta\mathbf{y}_{1,t-1} + \cdots + \mathbf{\Gamma}_{2,k-1}\Delta\mathbf{y}_{1,t-k+1} + \mathbf{\Phi}_{21,1}\mathbf{y}_{2,t-1} + \cdots + \mathbf{\Phi}_{21,k}\mathbf{y}_{2,t-k} + \mathbf{B}_2\mathbf{x}_t + \boldsymbol{\varepsilon}_{1,t},$$

where $\mathbf{\Pi}_1 = \mathbf{\Phi}_{11,1} + \cdots + \mathbf{\Phi}_{11,k} - \mathbf{I}$, $\mathbf{\Gamma}_{1,j} = -\sum_{l=j}^{k}\mathbf{\Phi}_{11,l}$, $\mathbf{\Pi}_2 = \mathbf{\Phi}_{21,1} + \cdots + \mathbf{\Phi}_{21,k}$, and $\mathbf{\Gamma}_{2,j} = -\sum_{l-j}^{k}\mathbf{\Phi}_{21,l}$, $j = 2,\ldots,k$.

Now if the I(1) series are cointegrated, the rank, r, of $\mathbf{\Pi}_1$ is less than p_1, and there exists a decomposition $\mathbf{\Pi}_1 = \boldsymbol{\alpha}_1\boldsymbol{\beta}_1'$, where $\boldsymbol{\alpha}_1$ and $\boldsymbol{\beta}_1$ are $p_1 \times r$ full rank matrices. The matrix $\boldsymbol{\beta}_2$ contains the stationary cointegration relationships of the I(1) variables, and $\boldsymbol{\alpha}_1$ contains the short term adjustment coefficients of discrepancies from the long term equilibrium determined by the cointegration vectors. In order for (3b) to be balanced in the sense that both sides of the equation are stationary there must exist a $(p - p_1) \times k$ matrix $\boldsymbol{\alpha}_2$ (that can be a zero matrix) such that $\mathbf{\Pi}_2 = \boldsymbol{\alpha}_2\boldsymbol{\beta}_1'$. Thus $\boldsymbol{\alpha}_2$ contains the response coefficients of the stationary variables to the discrepancies of the I(1) variables from the long term equilibrium. In most cases this, however, may not be plausible.

In any case, the important point here is that the rank of $\mathbf{\Pi}_1$ determines the dimension of the cointegration space. As discussed above, for identification of the equilibrium relationships, it may, however, be important to include besides the predetermined terms given in $\mathbf{x}_t$, also the stationary variables of the system in to the cointegration relation. Johansen (1995: 80–84) is a useful reference for discussion concerning the deterministic terms in the cointegration relations. Here, we focus purely on the inclusion of the endogenous stationary series. This approach is particularly useful when the cointegration vectors can be fixed on the basis of prior knowledge, like in our case of credit spreads. For the purpose write the whole system (2) in the ECM form

$$\Delta \mathbf{y}_t = \boldsymbol{\alpha}\boldsymbol{\beta}'\mathbf{y}_{t-1} + \boldsymbol{\Gamma}_1 \Delta \mathbf{y}_{t-1} + \cdots + \boldsymbol{\Gamma}_{k-1}\Delta \mathbf{y}_{t-k+1} + \mathbf{B}\mathbf{x}_t + \boldsymbol{\varepsilon}_t, \tag{4}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are $p \times r$ matrices with $\boldsymbol{\beta}' = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')$ of rank r, the cointegration rank of the system. The matrix $\boldsymbol{\beta}_1$ determines the cointegration relations of the I(1) variables in the system and $\boldsymbol{\beta}_2$ includes the coefficients of the stationary variables in the equilibrium relations. In principle the contribution of the stationary variables into the equilibrium relation can be estimated at the same time as the I(1) variables. Nevertheless, it is more efficient to solve the problem separately first for the integrated variables and then map the cointegration VAR into I(0) space, where one can purely deal with stationary variables and utilize the advanced estimation techniques developed therein.

Let $\mathbf{ci}_t = \boldsymbol{\beta}'\mathbf{y}_t = \boldsymbol{\beta}_1'\mathbf{y}_{1,t} + \boldsymbol{\beta}_2'\mathbf{y}_{2,t}$, then denoting $\mathbf{s}_t = \boldsymbol{\beta}_1'\mathbf{y}_{1,t}$, multiplying (4) from the left by $\boldsymbol{\beta}'$, and collecting terms, we get

$$\Delta \mathbf{s}_t = \tilde{\boldsymbol{\alpha}}\mathbf{s}_{t-1} - \boldsymbol{\beta}_2'\Delta \mathbf{y}_{2,t} + \boldsymbol{\eta}\mathbf{y}_{2,t-1} + \tilde{\Gamma}_1 \Delta \mathbf{y}_{t-1} + \cdots + \tilde{\Gamma}_{k-1}\Delta \mathbf{y}_{t-k+1} + \tilde{\mathbf{B}}\mathbf{x}_t + \boldsymbol{\upsilon}_t, \tag{5}$$

where $\tilde{\boldsymbol{\alpha}} = \boldsymbol{\beta}'\boldsymbol{\alpha}$, $\tilde{\Gamma}_i = \boldsymbol{\beta}'\Gamma_i$, $\tilde{\mathbf{B}} = \boldsymbol{\beta}'\mathbf{B}$, and $\boldsymbol{\upsilon}_t = \boldsymbol{\beta}'\boldsymbol{\varepsilon}_t$. Regarding parameter $\boldsymbol{\eta}$, it should be noted that, because $\mathbf{y}_{2,t}$ is stationary, it is possible that in the same manner as the exogenous variables, $\mathbf{x}_t$, the stationary variables may play in two roles; in the one hand inside and in the other hand outside the equilibrium correction relation. What we have in formula (5) is the aggregate contribution. If $\mathbf{y}_{2,t}$ solely contributes to the equilibrium correction term, then coefficient vector $\boldsymbol{\eta}$ obeys the restriction $\boldsymbol{\eta} = \tilde{\boldsymbol{\alpha}}\boldsymbol{\beta}_2'$. Consequently the portion $\tilde{\boldsymbol{\eta}} = \boldsymbol{\eta} - \tilde{\boldsymbol{\alpha}}\boldsymbol{\beta}_2'$ reflects the contribution of $\mathbf{y}_2$ to the spread, outside of the equilibrium correction relation.

In order to further interpret the parameters, let us assume that the system is in the equilibrium, which is achieved when the variables are in their means. That is, when $E(\mathbf{s}_t)=\boldsymbol{\mu}_s$, $E(\mathbf{y}_t)=\boldsymbol{\mu}_y$, $E(\mathbf{x}_t)=\boldsymbol{\mu}_x$, and $E(\boldsymbol{\upsilon}_t)=0$, in which case $E(\Delta\mathbf{s}_{1,t})=0$, $E(\Delta\mathbf{y}_{2,t})=0$, and $E(\Delta\mathbf{y}_{t-j})=0$ for all $j=1,\ldots,k-1$, and we get $0=\tilde{\boldsymbol{\alpha}}\boldsymbol{\mu}_s+\boldsymbol{\eta}\boldsymbol{\mu}_{y_2}+\tilde{\mathbf{B}}\boldsymbol{\mu}_x$. Now $\tilde{\boldsymbol{\alpha}}$ is a $p_1\times p_1$ full rank matrix so that the inverse exists, an because in our case $\mathbf{s}$ represents the spreads, we find that the average spread is determined as

$$\boldsymbol{\mu}_s=-\tilde{\boldsymbol{\alpha}}^{-1}(\boldsymbol{\eta}\boldsymbol{\mu}_{y_2}+\tilde{\mathbf{B}}\boldsymbol{\mu}_x). \tag{6}$$

Thus, particularly in our case, where $\mathbf{s}_t$ is the credit spread, we find that parameter $\boldsymbol{\eta}$ relates the long run average spread to the long run averages of the stationary variables.

3. Empirical results

The data consists of AA2, AA5, AA10, AAA2, AAA5, and AAA10 Japanese corporate bonds and corresponding maturity Japanese Government bonds (JGBs). The sample period is from January 2, 1995 to October 21, 1998. Following Longstaff and Schwartz (1995), we adopt the long yield of the 20-year JGB to represent the risk-free rate. The asset factor is measured in terms of the Nikkei return. It is common in the empirical finance literature to work with continuously compounded variables, which leads to log-returns. Particularly, because we have daily observations, we follow this practice and define the yields as

$$y_t=100\times\log(1+Y_t), \tag{7}$$

where Y_t is the yield to maturity on a bond. The credit spreads of maturity n are then simply defined as

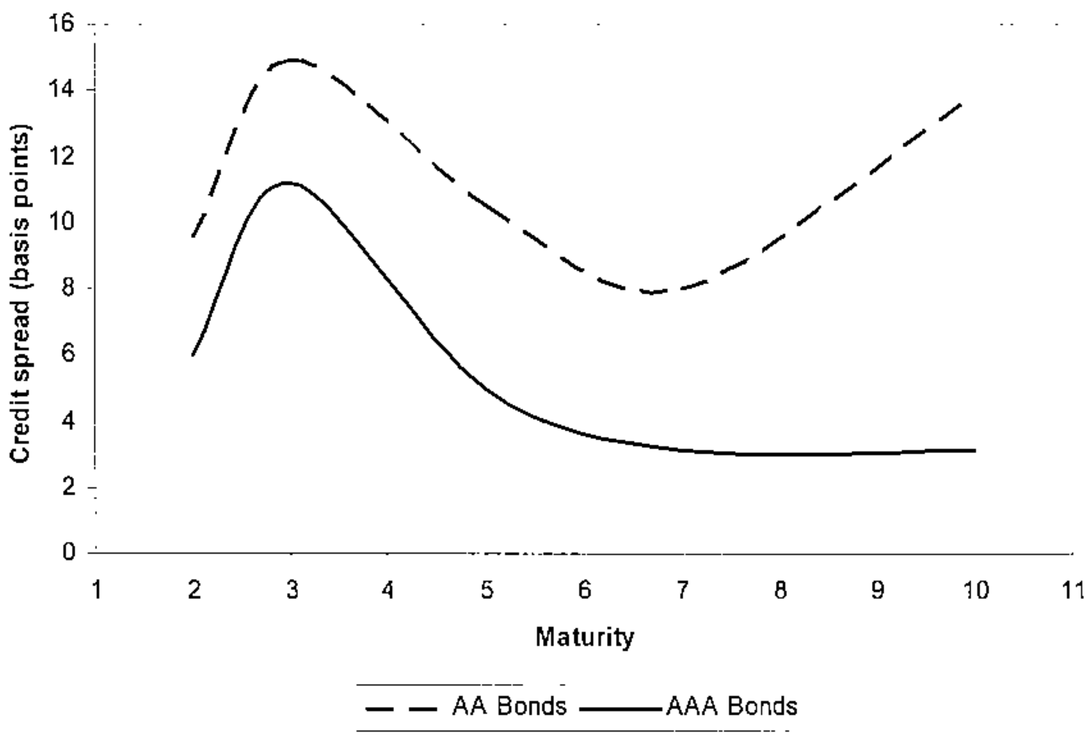


Figure 1. Log credit spread term structure of Japanese AA and AAA rated Eurobonds with maturities 2, 3, 5, 7 and 10 years, estimated from daily observations with sample period January 2, 1995 to October 21, 1998.

(8) $$s_{n,t} = y_{CBn,t} - y_{JGn,t} ,$$

where $y_{CBn,t}$ is the yield of a corporate bond with maturity n, and $y_{JGn,t}$ is the corresponding maturity Japan Government Bond yield.

Figure 1 depicts the term structure of the (log) credit spreads of the AA and AAA rated Japanese Eurobonds calculated as averages of the daily observations over the sample period for maturities 2, 3, 5, 7, 10 and 20 years.

The term structure is humped shaped with maximum spread occurring for the 3-year corporate bonds in both credit classes. While the spread is decreasing for the AAA rated longer maturity bonds the spread again increases for the AA rated 10-year corporate bonds.

Table 1. Summary statistics for log credit spreads of Japanese AA and AAA rated Eurobonds.

	AA2	*AA5*	*AA10*	*AAA2*	*AAA5*	*AAA10*	*ΔJGB20*	*Nikkei ret*
	(bp)	*(bp)*	*(bp)*	*(bp)*	*(bp)*	*(bp)*	*(%)*	*(%)*
Mean	9.50	10.50	13.82	6.03	4.98	3.15	-0.003	-0.033
Std	7.78	6.85	7.95	7.25	7.43	7.85	0.045	1.450
Kurtosis	0.82	1.33	5.01	0.05	1.39	6.82	5.750	2.654
Skewness	0.29	-0.31	0.59	0.05	-0.46	-0.15	-0.356	0.083
Minimum	-20.93	-18.61	-19.36	-21.93	-28.37	-32.99	-0.317	-5.957
Maximum	42.19	40.10	77.15	30.43	36.24	68.50	0.230	7.660
N	993	993	993	993	993	993	993	993

The sample period is daily observations from January 2, 1995 to October 21, 1998. The yield spreads are computed as $sp_t = y_{c,t} - y_{g,t}$, where y_c is the corporate daily yield and y_g is the corresponding government bond yield as defined in formula (7). The spreads are measured in basis points (1 bp = 0.01%).

Summary statistics for log credit spreads of the yen denominated Eurobonds, Nikkei Daily returns and JGB 20-year log yield changes are presented in Table 1. The mean credit spreads indicate numerically the information presented in Figure 1 for maturities investigated in this paper. The standard deviations are between 6.85 and 7.85 basis points, so within the range of one basis point. This indicates that the volatility is about the same over the maturities. The negative minima of the spreads indicate that there are periods where the yields of the corporate bonds are less than those of the otherwise equivalent government bonds.

From an econometric modelling point of view, adoption of the 20-year JGB yield as representative of the risk-free rate could be worked out through including the yield (which is an I(1) series) into the cointegration relation. Theoretically, this should be justified, because the expectation hypothesis or the liquidity premium hypothesis predicts that the term spread should be stationary. That is, short and long term JGB yields should be cointegrated (see, e.g., Hall, Anderson, and Granger 1992). The second column of Table 2 reports the Augmented Dickey-Fuller (ADF, Dickey and Fuller 1978, 1981) unit root tests for the yields. The null hypothesis of unit root is accepted for all yield series, but strongly rejected for the first differences of the series (fourth and fifth columns). Thus the empirical results support that the series are I(1), i.e., integrated of order on.

Table 2. I(1) tests for the log yields and spreads, and cointegration tests for the terms spreads of various maturity Japanese government bonds (JGBs) with the for the 20-year JGB.

Series	ADF[1)] (level)	p-val	ADF (1. diff)	p-val	Spread	ADF	p-val
AA2	-3.16	0.093	-9.97	0.000	AA2-JGB2	-6.45	0.000
AA5	-3.13	0.100	-8.46	0.000	AA5-JGB5	-6.11	0.000
AA10	-2.10	0.546	-12.65	0.000	AA10-JGB10	-8.89	0.000
AAA2	-3.13	0.100	-12.18	0.000	AAA2-JGB2	-7.66	0.000
AAA5	-3.29	0.068	-8.26	0.000	AAA5-JGB5	-5.68	0.000
AAA10	-2.32	0.421	-35.48	0.000	AAA10-JGB10	-7.80	0.000
JGB2	-2.89	0.165	-14.18	0.000	JGB2-JGB20	-2.02	0.590
JGB5	-2.73	0.224	-34.61	0.000	JGB5-JGB20	-1.77	0.721
JGB10	-2.33	0.418	-15.55	0.000	JGB10-JGB20	-4.04	0.008
JGB20	-2.55	0.304	-6.36	0.000	-	-	-

Johansen (1991, 1995) Trace and Max Eigenvalue test for cointegration of JGB yields

	Trace statistic	Max eigenvalue statistic
JGB2-JGB20	16.43	11.97
JGB5-JGB20	12.52	9.37
JGB10-JGB20	25.76*	22.30*
5% critical value	25.32	18.96
1% critical value	30.45	23.65

* = significant at the 5% level

[1)]The augmented Dickey & Fuller (1979, 1981) (ADF) test is based on the regression $\Delta y_t = \mu + \gamma y_{t-1} + \delta t + \phi_1 \Delta y_{t-1} + \cdots + \phi_m \Delta y_{t-m} + \varepsilon_t$ with null hypothesis that the series are I(1), which implies the to testing that $\gamma = 0$. The lag-length m of the differences is determined by Akaike's (1978) information criterion. The trend and intercept terms are allowed to eliminate their possible effect from the series.

Columns seven and eight report the unit root test results for the credit spreads of the corporate bonds and term spreads of the government bonds against the long maturity (20 years) bond. In all cases of the term spreads, the null hypothesis of a unit root is strongly rejected. These results, along with the above unit root findings, suggest that the individual series are best modelled as I(1), and so it is possible to infer that the corporate bonds and otherwise equivalent government bonds are cointegrated with the spread defining the cointegration relation. On the other hand in the case of term spreads, the unit root tests, reported in three last lines in columns seven and eight, indicate that JGB2 and JGB5 spreads with respect to JGB20 are non-stationary. The JGB10 is an exception for which the test results support stationarity. Even if we allow a more general cointegration relation than the term spread, the hypothesis of cointegration is rejected for the 2 and 5-year bonds as can be seen from the lower panel of Table 2, where the Johansen (1995, 1998) cointegration test results are reported. Again with the 10-year bond the null hypothesis of cointegration is accepted, supporting the above unit root test result of the spread.

Cointegration of the 10-year JGB with the 20-year JGB with the spread as the cointegration relation has an implication for the modelling of the 10-year corporate bonds. For example, because the spreads define the cointegration relation for both the AAA and JGB 10-year bonds and JGB 20 and 10-year bonds, the dimension of the cointegration space of the yields of AAA10, JGB10 and JGB20 bonds has the rank equal to two, where the spreads identify the cointegration vectors. The implication for modelling of the credit spread then is that the term spread of JGB 20-year and 10-year may have an effect on the dynamics of the credit spreads of AA and AAA 10-year bonds. However, this finding was not confirmed in the subsequent regressions and so it is ignored in the final model of the 10-year spreads.

Using the Longstaff and Schwartz (1995) and the empirical results found in Batten et al. (2003b), discussed in Section 1, we use in regression (5) besides the lagged spread, the JGB 20-year bond yield change, Nikkei return, r_t, the term structure slope change measured in terms of the JGB20 and JGB2 spread change, $\Delta(y_{JG20} - y_{JG2})_t$, where y_{JG20}, and y_{JG2} are the log-yields of the 20-year and 2-year JGBs. In addition to account for autocorrelation in the residuals we allow for ARMA structure in the residuals, and, furthermore, to model the possible conditional heteroscedasticity, we use GARCH specification, and allow its possible effect on the mean equation as well. These are the major advantages to estimate the possible contribution of the stationary series to the equilibrium relation of the yields, and hence to the spreads. Thus augmenting (5) with these factors the estimated models for the spread changes are of the form

$$\begin{aligned}\Delta s_t &= \delta_0 + \delta_1\sqrt{h_t} + \alpha\, s_{t-1} + \beta_1 \Delta r_t + \beta_2 \Delta^2 y_{JG20,t} + \beta_2 \Delta^2 (y_{JG20} - y_{JG2})_t \\ &\quad + \eta_1 r_{t-1} + \eta_2 \Delta y_{JG20,t-1} + \eta_3 \Delta(y_{JG20} - y_{JG2})_{t-1} + u_t \\ u_t &= \phi u_{t-1} + \theta\varepsilon_{t-1} + \varepsilon_t \\ h_t &= \omega + \gamma_1 \varepsilon_{t-1}^2 + \gamma_2 h_{t-1}\end{aligned} \tag{9}$$

where, we have dropped for the sake of simplicity the maturity index from the subscript, and ε_t is the residual term with a GARCH(1,1) variance process. There are of course other parameterisations of (9), but we find the above choice the most convenient, where we directly identify the equilibrium correction coefficients of the stationary series, and then find the long-run mean spread using formula (6).

The Longstaff and Schwartz (1995) model predicts that the signs of β_1 and β_2 should be negative, which imply the increased asset return should decrease the spread and an increase in risk-free return should also decrease the credit spread. The former is intuitive since improved asset return drifts the company asset value up and improves the company's solvency. The negative impact of the risk-free return on the spread is explained in Longstaff and Schwartz (1995) in terms of the risk-neutral pricing process implied by the no-arbitrage pricing mechanism of the bonds and consequently yields. Because the risk-free return serves as the drift in the risk neutral asset value process, the increase of the risk-free return drives the risk neutral asset process away from the default boundary, and hence decreases the (risk-neutral) probability of default. Consequently the spread over the risk-free rate should decrease. The slope of the yield curve is related to state of the economy. Collin-Dufrense, Goldstein, and Martin (2001) argue that as the economy moves into recession, the steepness of the yield curve declines. In such an economic phase the asset returns are expected to decrease and hence the firm values decline closer to the default boundary, and therefore increase the default risk. Consequently, the credit spread can be expected to increase, which should shows up in (9) with a negative β_3.

Estimation results and related diagnostic statistics for the selected maturity credit spreads are reported in Appendix 1 and 2. From the tables we find that the cointegration term, the spread s_{t-1}, is highly statistically significant in all cases with a negative coefficient. In the AA case the coefficient ranges from -0.14 to -0.095, which indicates a fairly slow adjustment in the spread towards the equilibrium. In terms of these estimates the model predicts that a shock of 100 basis points deviation from the equilibrium results to a correction of 9.5 to 14 basis points the next day in the spread. In the AAA case the coefficient estimates are bit lower, ranging from -0.111 to -0.048.

The other important findings from the regression results are that the contribution of the risk-free interest rate to the equilibrium can be obviously inferred to be negative (estimate of coefficient β_2), exactly as predicted by the Longstaff and Schwartz (1995) model. On the other hand the contribution of the asset factor to the equilibrium does not show up in the estimation results. Its regression coefficient (β_1) estimates close to borderline statistically significant at the 10 percent level for AAA5 and AAA10 spreads, and even in these cases the signs are opposite to that predicted by the Longstaff and Schwartz (1995) model. However, the significance of the estimate of η_1, which is the parameter measuring the asset factor's mean contribution to the average spread is highly statistically significant. Again the sign is opposite to what could be expected from the Longstaff and Schwartz (1995) model. A partial explanation to this may be that the sample period included a rather exceptional time episode in Japanese economy; the negative average daily stock return of -0.033% (about -8.3% p.a.). This implies that the ultimate contribution of the asset return to the average spread becomes negative as predicted by the Longstaff and Schwartz (1995) theory. More generally, we can write equation (6) as

$$(10) \qquad \mu_s = \mu_{s,c} + \mu_{s,a} + \mu_{s,r} + \mu_{s,y} = -(\mu_c + \eta_1\mu_a + \eta_2\mu_r + \eta_3\mu_y)/\alpha \,,$$

where $\mu_{s,c} = -\mu_c/\alpha$, $\mu_{s,a} = -\eta_1\mu_a/\alpha$, $\mu_{s,r} = -\eta_2\mu_r/\alpha$, and $\mu_{s,y} = -\eta_3\mu_y/\alpha$ are the asset, risk-free rate, and yield curve contributions to the mean spread with μ_c the constant term in the regression and μ_a, μ_r, and μ_y asset return, risk-free interest rate and yield curve change long run averages, respectively. Thus, for example, in the case of the AAA rated 10-year bond, we get an estimate for the asset factor equal to $\hat{\mu}_{s,a} = -0.0020 \times (-0.033)/(-0.111) = -0.0006$, or -0.06 basis points. This is obviously small and economically non-insignificant when compared to the total average spread of 4.10 basis points. Thus, in summary, the asset factor's impact cannot be identified clearly in the Japanese markets as a determinant of the credit spread.

The estimate of the asset factor showed up as being weaker than the interest rate factor in Longstaff and Schwartz (1995) and Collin-Dufresne et al. (2001), but was clearly statistically significant and negative. In both of these studies monthly data were used. It may well be that the daily data used in our study is too noisy to measure accurately the asset factor. In any case the role of this factor as a proxy for the economic state in the model remains unclear.

The third important factor is the change in the slope of the yield curve. The estimates are highly significant, but again of opposite sign to what is predicted by Collin-Dufresne et al. (2001). Again this may be due the sample period covering an exceptional period in Japanese economy. The average yield curve change, as measured in terms of the spread change of 20-year and 2-year Japan Government bonds, has been slightly negative.

In the GARCH volatility process all the estimates, except the constant term, are highly significant. Thus there is obvious conditional heteroscedasticity in the credit spread. The sum of the GARCH parameter estimates of γ_1 and γ_2 is in most cases close to one, indicating that the volatility process is close to being integrated with a weight 0.05 for the latest shock and 0.95 for the persistence. With these weights the volatility process is in any case fairly smooth. The changing volatility, however, does not show up in the mean equation.

4. Summary and Conclusions

In an important paper Longstaff and Schwartz (1995) derive a closed form solution for the price of risky bond under the arbitrage-free assumption. Their model predicts that the yield spreads should be a negative function of both the firm asset return and the risk-free interest rates. The model is an equilibrium solution, where price adjustments are assumed to take place immediately. Nevertheless, in real discrete time trading there are delays and frictions that constitute feedbacks between integrated price series. This implies that a the non-stationary series may become cointegrated. That is, a linear combination of the series is stationary. This study shows that there is strong empirical evidence that the Japanese Yen Eurobond yields are cointegrated with the equivalent maturity Japanese Government Bonds (JGBs) with the spread defining the cointegration relation.

Because cointegration leads to incomplete markets, we must give up arbitrage-free pricing and rely on equilibrium pricing, where the markets clear via a potentially complicated adjustement process. Taking the spread as the core of the equilibrium cointegration relation we derive a model, where the contribution of stationary series, like the asset return and the change in the risk-free rates, to the equilibrium relation can be estimated and tested. The equilibrium correction term of the cointegration relation is an important determinant in the adjustment process of the spread to the equilibrium in each of the investigated series. Furthermore, the results suggest that the adjustment process is fairly slow. Of the stationary series the most important factor with the predicted sign by Longstaff and Schwartz (1995) is the risk-free interest rate. The asset return, on the other hand, does not show up as a significant factor in the equilibrium. The slope of the yield curve of Government bonds, which is intended to reflect the phase of the economy, turned out to also be statistically significant, but with the opposite sign to what would be expected. Thus the real role of this variable remains unclear.

References

Akaike, Hirotugu (1973). Information theory and an extension of the maximum likelihood principle. In *Proceedings of the 2nd International Symposium on Information Theory*, 267–281. Eds B.N. Petrov & F. Czáki, Budapest, Akadémia Kiado.

Batten, Jonathan, Warren Hogan & Seppo Pynnönen (2003a). The time-varying behaviour of credit spreads on Yen Eurobonds. In *International Financial Review*, Vol. 4, *Japanese Finance: Corporate Finance and Capital Markets in Changing Japan*, 383–408. Eds Jay Choi & Takato Hirako. Elsevier Ltd.

Batten, Jonathan, Warren Hogan & Seppo Pynnönen (2003b). Modelling Credit Spreads on Yen Eurobonds. Submitted for publication.

Bollerslev, Tim & Jeffrey M. Wooldridge (1992). Quasi-maximum likelihood estimation and inference in dynamic models with time varying covariances. *Econometric Reviews* 11, 143–172.

Collin-Dufresne, Pierre, Robert S. Goldstein & J. Spencer Martin (2001). The determinates of credit spread changes. *Journal of Finance* 56, 2177–2207.

Das, S. & P. Tufano (1996). Pricing credit sensitive debt when interest rates, credit ratings and credit spreads are stochastic. *Journal of Financial Engineering* 5, 161–198.

Dickey David A. & Wayne A. Fuller (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association* 74, 427–431.

Dickey David A. & Wayne A. Fuller (1981). Likelihood ratio statistics for autoregressive time series with unit root. *Econometrica* 49, 1057–1072.

Duffee, Gregory, R. (1998). The relation between Treasury yields and corporate bond yield spreads. *Journal of Finance* 53, 2225–2241.

Duffie, Darrell & Kenneth Singleton (1999). Modelling term structures of defaultable bonds. *Review of Financial Studies* 12:4, 687–720.

Engle, Robert, F., David M. Lilien & Russell P. Robins (1987). Estimating time varying risk premia in the term structure: The ARCH-M model. *Econometrica* 55, 391–407.

Hall, Anthony, D., Heather M. Anderson & Clive W.J. Granger (1992). A cointegration analysis of Treasury bill yields. *Review of Economics and Statistics* 74, 116–126.

Johansen, Søren (1991). Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models. *Econometrica* 52, 1551–1580.

Johansen, Søren (1995). *Likelihood-based Inference in Cointegrated Vector Autoregressive Models*. Oxford University Press.

Kamin, S. & K. von Kleist (1999). The evolution and determinates of emerging market credit spreads in the 1990s. Board of Governors of the Federal Reserve System, *International Finance Discussion Paper* 653 (November), 1–45.

Longstaff, Francis A. & Eduardo S. Schwartz. (1995). A simple approach to valuing risky fixed and floating rate debt. *Journal of Finance* 50, 789–819.

Mansi, Sattar, A. & William F. Maxwell (2000). The stochastic nature and factors affecting credit spreads. Working Paper, Texas Tech University, College of Business Administration, Lubbock Texas.

Neal, R., D. Rolph & C. Morris (2000). Interest rates and credit spread dynamics, Working Paper IUPUI.

Pederosa, M. & R. Roll (1998). Systematic risk in corporate bond credit spreads. *Journal of Fixed Income* 8, 7–26.

Appendix 1. Estimates of the parameters of model (9) for AA rated Japanese Eurobond corporate bond spreads.

		$\Delta s2$			$\Delta s5$			$\Delta s10$	
	Coeff	std err	p-val	Coeff	std err	p-val	Coeff	std err	p-val
Constant [δ_0]	0.0064	0.0017	0.000	0.0110	0.0028	0.000	0.0163	0.0038	0.000
Sqrt(h_t) [δ_1]	-	-	-	-	-	-	-	-	-
s_{t-1} [α]	-0.095	0.017	0.000	-0.111	0.024	0.000	-0.140	0.030	0.000
Δ^2log(Nikkei)$_t$ [β_1]	-	-	-	-	-	-	-	-	-
$\Delta^2 y_{JG20,t}$ [β_2]	-0.723	0.053	0.000	-0.665	0.050	0.000	-0.684	0.047	0.000
$\Delta^2(y_{JG20} - y_{JG2})_t$ [β_3]	0.800	0.045	0.000	0.411	0.052	0.000	0.078	0.036	0.031
Δlog(Nikkei)$_{t-1}$ [η_1]	0.0021	0.0007	0.002	-	-	-	0.0015	0.0008	0.072
$\Delta y_{JG20,t-1}$ [η_2]	-0.665	0.062	0.000	-0.559	0.054	0.000	-0.599	0.058	0.000
$\Delta(y_{JG20} - y_{JG2})_{t-1}$ [η_3]	0.738	0.057	0.000	0.358	0.056	0.000	-	-	-
Residual equation									
ar(1) [ϕ]	-	-	-	-		-	-		-
ma(1) [θ]	-	-	-	-0.217	0.045	0.000	-0.180	0.065	0.006
Variance Equation									
Constant (x 1 000) [ω]	0.008	0.007	0.218	0.013	0.010	0.174	0.067	0.033	0.042
ε^2_{t-1} [γ_1]	0.055	0.017	0.002	0.060	0.015	0.000	0.139	0.039	0.000
log(h_{t-1}) [γ_2]	0.943	0.017	0.000	0.935	0.016	0.000	0.845	0.038	0.000
Diagnostics									
N of outliers removed	1	na	na	0	na	na	1	na	na
Observations	991	na	na	991	na	na	990	na	na
Adj. R^2	0.500	na	na	0.439	na	na	0.480	na	na
s(e)	0.039	na	na	0.046	na	na	0.049	na	na
Skew z	-0.03	p-val	0.737	-0.09	p-val	0.224	-0.09	p-val	0.259
Kurt z	6.70	p-val	0.000	5.22	p-val	0.000	6.88	p-val	0.000
Jarque-Bera	566.0	p-val	0.000	205.1	p-val	0.000	623.1	p-val	0.000
Q(2) z	3.52	p-val	0.172	-0.02	p-val	0.350	0.17	p-val	0.684
Q(5) z	4.16	p-val	0.527	-0.01	p-val	0.912	3.10	p-val	0.541
Q(10) z	9.16	p-val	0.517	0.02	p-val	0.133	16.35	p-val	0.060
Q(2) z^2	1.60	p-val	0.449	-0.04	p-val	0.125	0.47	p-val	0.492
Q(5) z^2	2.47	p-val	0.780	0.01	p-val	0.607	3.20	p-val	0.525
Q(10) z^2	6.16	p-val	0.801	0.02	p-val	0.813	5.52	p-val	0.787
Estimates of the long run mean spreads [see formulas (6) and (10)]									
μ_s (Basis points)	8.42			11.34			13.01		

Dashes indicate variables whose p-values were more than 0.15, and were removed from the final model. Outliers with large residuals were removed. Because the sample size is large the outliers do not materially change the regression results, but they may potentially disturb the ARMA-GARCH residual structure estimation. In the GARCH variance process we have retained the constant term even tough the p-value is larger than 0.15. The standard errors are Bollerslev and Wooldridge (1992) heteroscedasticity corrected.

Appendix 2. Estimates of the parameters of model (9) for AAA rated Japanese Eurobond corporate bond spreads.

	$\Delta s2$			$\Delta s5$			$\Delta s10$		
	Coeff	std err	p-val	Coeff	std err	p-val	Coeff	std err	p-val
Constant [δ_0]	-	-	-	0.0034	0.0013	0.011	0.0029	0.0011	0.009
Sqrt(h_t) [δ_1]	-	-	-	-	-	-	-	-	-
s_{t-1} [α]	-0.048	0.011	0.000	-0.075	0.019	0.000	-0.111	0.024	0.000
$\Delta^2\log(\text{Nikkei})_t$ [β_1]	-	-	-	0.001	0.001	0.092	0.001	0.001	0.122
$\Delta^2 y_{JG20,t}$ [β_2]	-0.732	0.041	0.000	-0.640	0.049	0.000	-0.637	0.050	0.000
$\Delta^2(y_{JG20} - y_{JG2})_t$ [β_3]	0.768	0.041	0.000	0.387	0.050	0.000	0.095	0.036	0.009
$\Delta\log(\text{Nikkei})_{t-1}$ [η_1]	0.0023	0.0007	0.001	0.002	0.001	0.007	0.0020	0.0011	0.064
$\Delta y_{JG20,t-1}$ [η_2]	-0.666	0.050	0.000	-0.519	0.052	0.000	-0.489	0.055	0.000
$\Delta(y_{JG20} - y_{JG2})_{t-1}$ [η_3]	0.688	0.057	0.000	0.303	0.054	0.000	-	-	-
Residual equation									
ar(1) [ϕ]	-	-	-	-		-	-		-
ma(1) [θ]	-0.105	0.042	0.013	-0.270	0.043	0.000	-0.278	0.044	0.000
Variance Equation									
Constant (x 1 000) [ω]	0.010	0.008	0.221	0.010	0.009	0.250	0.013	0.010	0.213
ε^2_{t-1} [γ_1]	0.050	0.019	0.007	0.057	0.014	0.000	0.046	0.013	0.001
$\log(h_{t-1})$ [γ_2]	0.945	0.022	0.000	0.939	0.016	0.000	0.948	0.016	0.000
Diagnostics									
N of outliers removed	0	na	na	0	na	na	4	na	na
Observations	991	na	na	991	na	na	897	na	na
Adj. R^2	0.495	na	na	0.425	na	na	0.571	na	na
s(e)	0.039	na	na	0.047	na	na	0.045	na	na
Skew z	-0.19	p-val	0.015	0.00	p-val	0.958	-0.42	p-val	0.000
Kurt z	7.71	p-val	0.000	4.65	p-val	0.000	5.65	p-val	0.000
Jarque-Bera	921.8	p-val	0.000	112.6	p-val	0.000	318.2	p-val	0.000
Q(2) z	0.21	p-val	0.649	0.78	p-val	0.377	0.80	p-val	0.371
Q(5) z	6.61	p-val	0.158	2.21	p-val	0.697	2.47	p-val	0.651
Q(10) z	11.82	p-val	0.224	6.08	p-val	0.732	10.40	p-val	0.319
Q(2) z^2	1.12	p-val	0.289	1.55	p-val	0.212	0.96	p-val	0.327
Q(5) z^2	1.48	p-val	0.829	2.67	p-val	0.615	6.50	p-val	0.164
Q(10) z^2	5.50	p-val	0.789	8.93	p-val	0.443	9.85	p-val	0.362
Estimates of the long run mean spreads [see formula (6) and (10)]									
μ_s (Basis points)	3.31			6.41			4.10		

Dashes indicate variables whose p-values were more than 0.15, and were removed from the final model. Outliers with large residuals were removed. Because the sample size is large the outliers do not materially change the regression results, but they may potentially disturb the ARMA-GARCH residual structure estimation. In the GARCH variance process we have retained the constant term even tough the p-value is larger than 0.15.

How confidential should strategic information be?

Tapio Reponen

Dedicated to Ilkka Virtanen on the occasion of his 60th birthday

Abstract

Reponen, Tapio (2004). How confidential should strategic information be? In *Contributions to Management Science, Mathematics and Modelling. Essays in Honour of Professor Ilkka Virtanen.* Acta Wasaensia No. 122, 205–224. Eds Matti Laaksonen and Seppo Pynnönen.

The study considers the nature of information and knowledge in strategic decision-making. Increasingly strategic information will be processed with information technology. Security, confidentiality, and reliability requirements are normally very high in this context. The research objective is to study the requirements for confidentiality in modern network environments. The concrete research problem was to study factors influencing a choice between an open and closed network environment in processing strategic information. With an action research project in the public sector, the study indicates the relevant factors in this decision-making. The main conclusion is that what was earlier strategic and confidential, is now openly informed to stakeholders. Instead tactical and operative actions are increasingly competitive and therefore subject to strict security requirements.

Tapio Reponen, Turku School of Economics and Business Administration, Rehtorinpellonkatu 3, FIN–20500, Turku, Finland, e-mail tapio.reponen@tukkk.fi.

1. Introduction: The nature of strategic decision-making

Strategic thinking has changed over time towards emphasizing the elements of expert and knowledge organizations. Consequently a key challenge is to create shared values among the experts (Senge 1994). This is a very demanding task since specialists and experts have different motivation factors than earlier "industrial workers" (Sveiby 1990). Human-centered approaches are therefore needed in designing and implementing strategies.

The concept of strategy has been defined in literature in many different ways, but it is essential in strategic thinking to find ways of acting differently to other organizations. In a competitive environment, acting better, faster and intuitively may result in success. Therefore, a key objective in strategic thinking is to create synergy between people to unite all their human potential to differentiate one organization's behavioral models from another's.

This requires that all the members of the organization should know its strategic objectives and actions. Thus the objectives of "the business strategy" should be openly presented within the organization. Everybody should know the joint goals and objectives in order to be able to behave accordingly. The requirements of confidentiality are, therefore, changing from earlier thinking, where business strategy was highly secret to most stakeholders.

Over the last decade there has been an obvious trend of the globalization of leading organizations, as they have moved from multinational to global operations (Reponen 2003). Recently we have also seen an increasing demand for combining locally customized services with the economies of the scale of worldwide operations. One of the management challenges is to cope with this new customer need.

In this environment, competitiveness calls for integrating the potential of information technology with well functioning global logistics. Leading companies have already created integrated information and communication technology (ICT) architectures to master their internal information flows, but there are still problems linked to customer service. Networking, joint ventures and mergers bring their new challenges into homogenizing the ICT environment.

ICT will have an increasing and essential strategic role in the present society of networks (Applegate et al. 2000). This role consists of coordinating activities and keeping interorganizational groups together, thus demanding the level of service and quality of ICT to be very high.

Data, information and knowledge are processed in both human and electronic ways. Exact data easily gives an impression of reliability and correctness. For instance, numbers in financial analysis are considered facts; but recent cases of manipulating profitability indicators in global companies have resulted in the collapse of this confidence. We have clearly seen that in addition to the numbers, you should know the rules and ways in which these numbers have been produced.

From this perspective it is important to consider different aspects of security, reliability and confidence in strategic decision-making. It is possible to build relatively high technical security into data processing, to ensure that the numbers as such are correct. The interpretation of the data into meaningful information is, however, very sensitive to human inference. In that area the possibility of intended or non-intended misunderstandings is highly evident.

The reliability of information systems is thus based on a combination of human and technical data processing. Information flows have, therefore, several risky points in the confidentiality of data and information.

2. The research problem and the research methodology

The research problem of this study is to consider different factors that have an influence on the choice of network environment in processing strategic information. Security, confidentiality, and reliability requirements are normally very high in this context. The risk of information leaks cannot be avoided totally, and decisions have to be made on reaching an acceptable level of confidentiality.

This study concentrates on public sector administration, where the role of information is increasing in a similar way as in business organizations. Effectiveness and efficiency are emphasized in decision-making and actions are taken to strengthen them. One special feature of ICT deployment in the public sector is considered here, namely the confidentiality requirements for a networked environment.

The empirical research project was to study information management in the defence sector in Finland. The research methodology was action research. Action science offers one possible starting point for conceptualizing the interactive learning process in strategy generation. Argyris et al. (1985) state that:

"Action scientists engage with participants in a collaborative process of critical inquiry into problems of social practice in a learning context. The core feature of this context is that it is expressly designed to foster learning about one's practice and about alternative ways of constructing it."

Action research is client-centered and contextual. The researcher who wishes to investigate an organization consults with its members. Research goals are negotiated between the client group and researcher. The research involves a learning cycle that is a continuous process resulting in outcomes acceptable to both client group and researcher (Reponen, Wood-Harper, von Hellens 1992). Action research, therefore, aims pragmatically to produce practical outcomes for immediate use.

Its assumptions are that each social context is unique rather than an instance of a general case. It does not produce law-like generalizations about organizations in the same way as positivism (Checkland 1991). But in the practical work of planning strategies, action science-like procedures are exactly what is needed.

Gummesson (1991: 103–106) states eight criteria, which an action research project should meet to be a scientific process. These requirements are:

(1) Action science always involves two goals: solve a problem for a client and contribute to science.
(2) During an action science project, those involved should learn from each other and develop their competence.
(3) The understanding developed during an action science project is holistic.
(4) Action science requires cooperation between the researcher/consultant and the client personnel, feedback to the parties involved, and continuous adjustment to new information and new events.

(5) Action science is primarily applicable to the understanding and planning of change in social systems.

(6) There must be a mutually acceptable ethical framework within which action science is used.

(7) Pre-understanding of the corporate environment and of the conditions of business is essential when action science is applied to management.

(8) Action science should be governed by the hermeneutic paradigm although elements from the positivistic paradigm may be included.

According to Riordan (1995), action research seems to offer new possibilities by incorporating a form of practice and research, which is aimed at understanding meaning, while at the same time retaining enough of the characteristic of the ideal scientific reliability.

The data elements in action science are action and talk. There are different methods of collecting this data (Argyris et al. 1985):

- observations accompanied by audio taping,
- interviews,
- action experiments,
- participant written cases.

The researchers from Turku School of Economics and Business Administration have been facilitators in designing the information management strategy for the Finnish defence sector. The researchers have been actors in the planning process and have influenced decisions made in the object organization. Using all the above-mentioned data collection methods, the researches have, together with experts from the Ministry and defence forces, made a proposal for a strategy.

The nature of the research is such that the empirical case study has been used as a source of information in generating a preliminary framework of the decision situation. The results of the study are propositions of factors influencing the decision-making.

3. The increasing role of information systems in decision-making

Information is one of the key resources in future organization structures. People are working in an interactive way in teams and groups performing tasks that are designated to them. These networks and groups may be both intraorganizational and interorganizational. Customer service requires good coordination and integration of an organization's information flows.

In order to fulfill the coordination function, the quality of information systems should be very high. In practice there are, however, numerous examples of gaps existing between management, user and supplier views of information systems and their quality (e.g. Auer, 1995; Davis, 1989). Thus the increasing interdependence relies on common understanding of quality and service issues.

At the corporate level, one of the main problems management has to face is the division of scarce resources to its units, either different profit centers or nodes of a networked organization. Each of these units has its own competitive strategies and is in different competitive situations. Corporate management has to evaluate these strategies and coordinate them into a synergic strategy (Collins and Porras 1996, Senge 1994).

The competitive situations of profit centers depend on many environmental and internal factors, which have simultaneous effects on the performance of these strategic units (Porter, 1998). There are factors that are controllable to the management of the profit center while the others are not. The corporate management should, however, be able to separate the results of the operation of the profit center from its environmental success, for which purpose it needs as much support as possible.

The nature of strategic thinking has changed over the decades. The planning of a corporate strategy may, however, still be based on the concept of a "strategic portfolio" whose elements are strategic business units. The dimensions of this portfolio may be defined in many different ways, but in each case the main objective is to describe the competitive position of each strategic business unit in its market. One of the objectives of strategic planning is to achieve a balanced portfolio of strategic units, so that there are enough cash generating units to finance the growth of units in areas of high profit potential. The position of the strategic business unit in this portfolio will thus also influence its measures of profitability and financial position.

In strategic decision-making you should also be aware of short-term changes in competition, the strategic moves of competitors and the success of your own operations. For these purposes you need measures that react immediately to changes in competitive situations. The Balanced Scorecard approach has been developed to catch the different quantitative and qualitative dimensions of business units.

Increasing competition will stabilize profits towards the theoretical situation of perfect competition. In the forces driving industry there are many factors, which will together determine the expected profitability of the firm. Above average profitability requires a special position in the industry. The old concepts of cost-leadership, differentiated products, or customer focus are still valid to describe the generic alternatives.

The factors influencing competition may be divided into environmental and internal. The environmental factors may be divided further into those which are or are not controllable by the management of the SBU or that are difficult to control. Among these environmental factors are such as:

Non-controllable

- Industry evolution, the stage in the industry lifecycle
- Number of competitors
- Number of entries or exits in the industry
- The nature of the industry (consumer/durable/services/etc.)

- Growth rate of the industry and market
- Inflation

Controllable

- Industry concentration ratio
- Capital intensity in the industry
- Customers
- Suppliers
- The number of mergers and acquisitions
- The type of entry (pioneer-follower-late entrant)
- Market shares
- Growth rate of the firm etc.

The division of factors into controllable and non-controllable is difficult because each firm in the market exerts some kind of influence on the competition. The amount of influence depends on the characteristics of the firm in question. Information is needed from all of these factors to make balanced decisions.

The main strategic decisions the managers may make are

(1) Entry into a new business
(2) Major capacity expansion within the present business
(3) Vertical integration
(4) Harvesting present business

These decisions may be supported by different tactical moves, such as price competition, advertising, customer services, warranties, buyer selection and new product introductions.

The public sector is increasingly operating under different market mechanisms, and the analysis of competitive forces is important. Strategic thinking requires information collection and the processing of information with certain security requirements.

For planning and deciding strategic and tactical moves, organizations need information from many different sources. This information has certain requirements to meet the needs of the decision-makers. The following offers a short categorization of different forms of information.

4. The concepts of data, information, and knowledge

The roles of data, information, and knowledge are strengthening in organizational activities. Creating plans, making decisions and taking actions rely heavily on the information available to the actors. The success of any organization depends on the knowledge and competence it possesses. Knowledge and information are essential in both leadership and management, and in keeping up with "state of the art" development.

Lester Thurow (1996) has stated, "Generally speaking a larger share economy shifts away from activities that are routine, repetitive and based on someone else's instructions to production that is invention, customization and personalization".

We are moving towards a working environment where human actors and information systems form a strictly coupled system of mutual interdependence. Information and knowledge is stored both in peoples' minds and in computers. Both are needed, which generates the problem of maintaining confidentiality in both areas.

The concepts of information processing may be defined as follows:

- **Data** is representation of a fact, number, word, picture or sound
- **Information** is data that is meaningful or useful to someone
- **Knowledge** is information transformed into capabilities for effective action
- **Human intelligence** is knowledge linked to a human value structure, by understanding the meaning of "right" and "wrong".

Polanyi (1966) and later by Nonaka and Takeuchi (1995) have further divided knowledge into "explicit" and "tacit". According to their definition, "explicit" knowledge is transmittable in formal, systematic language. "Tacit" knowledge is personal, context-specific, and therefore hard to communicate. Human knowledge is created and expanded through social interaction between tacit and explicit knowledge.

The concepts of "tacit" and "explicit" knowledge have been criticized recently; nevertheless, they draw a good distinction between computerized and non-computerized knowledge. This distinction is significant in considering the security requirements of a whole information system, including its human and electronic parts.

The requirements for the quality of the information are (Berry, Parasuraman, and Zeithaml, 1985):

- **Reliability**, which means consistency, dependability and lack of sloppiness
- **Responsiveness,** which is willingness, readiness and timeliness
- **Competence,** which is possession of the required skills and knowledge
- **Access,** which is approachability, ease of contact
- **Courtesy,** which is politeness, respect, consideration, friendliness
- **Communication,** keeping the customer informed
- **Credibility,** which is trustworthiness, believability, honesty
- **Security,** freedom from danger, risk, doubt
- **Understanding the customer,** making an effort to understand
- **Tangible physical evidence of the service,** like documents.

The list above has been derived from customer and user perspectives. The list may be categorized according to the dimensions human and automatic data processing. There are technical requirements such as reliability and security, user requirements like responsiveness, access, courtesy and communication, and human requirements such as competence, credibility and understanding the customer and service view. High quality information includes all these aspects; confidentiality requires both technical and human measures of information protection.

5. Case study: Information systems of the defence sector in Finland

5.1. Description of the research project and problem formulation for the case study

The results of this paper are based on an action research project with the Finnish Ministry of Defence. Three researchers from Turku School of Economics and Business Administration (Tapio Reponen, Hannu Salmela and Johanna Holm) were facilitators in a process of designing an information management strategy for the defence sector in Finland.

The empirical objective of the project was defined as follows:
The objective of the study is to design an information management strategy, that

- Offers guidelines for the deployment of ICT in the defence sector
- Supports the overall strategy of the defence sector
- Describes the general development trends of the sector and the requirements it presents to the ICT deployment
- Describes the present status of information management and sets objectives for its development
- Introduces key development projects
- Makes a proposal for organizing information management
- Presents an overall cost/benefit analysis of the investments
- Considers the international dimension of defence cooperation and its requirements for compatibility.

The research was carried out in the year 2002, starting in February and ending in November. The research data was collected from different printed sources, such as publications, reports, statistics and the minutes of meetings; and by interviewing decision-makers at different levels of defence administration.

A project group was nominated to realize the project. It comprised one expert from the Ministry, three from the headquarters of the Finnish Defence Forces, and three facilitators.

The strategy was generated in close cooperation between these internal specialists and outside facilitators.

The objective was to bring together empirical and theoretical knowledge. The methods used consisted of a combination of interviews, meetings, producing reports and reading printed documents. This approach had two objectives, namely to collect the internal knowledge from experts, and inform them of the new possibilities ICT offers.

Based on numerous empirical information management projects, researchers in the Institute of Information Systems Science at Turku School of Economics and Business Administration have developed a special framework for strategy generation, called the Evolutionary Model for Information Systems Strategies (EMIS, Reponen 1994, 1998; Reponen and Auer 1997). That framework was employed in this process to give structure to the project.

A strategy proposal was made within the schedule decided at the beginning of the project. The proposal contained recommendations for all the sub-areas mentioned above. This paper deals with one special feature of the problem area: Can commercial networks be used in transmitting strategic information?

5.2. Different categories of ICT systems in the defence sector

The basic components of Finland's security policy are as follows (Ministry of Defence, 2001):

- Maintenance and development of a credible defence capability
- Remaining militarily non-allied under the prevailing conditions
- Participation in international cooperation to strengthen security and stability.

The goal of Finland's defence is to guarantee the country's independence, secure the livelihood of its citizens, prevent Finnish territory from being seized and secure the functioning of the state leadership.

Finland's defence system is based on broad cooperation between various authorities and the private sector. It consists of both military and civilian crises management. One of the challenges is to build synergy between organizations handling these quite different situations.

A modern defence system is heavily dependent on information systems, which can be divided into the following categories:

- Decision support systems
- Operative systems
- Administrative systems.

These system areas are different in nature and they have their own security and reliability requirements. Decision support systems offer information for planning defence organizations and for making decisions in different crises situations. Operative systems are built for using the civilian and military infrastructure of crises management. Administrative systems are support systems for materials handling, financial management, personnel administration, office automation, electronic customer service and so on.

Strategic information is mainly processed in decision support and operative systems. Handling this information calls for careful consideration of the security and reliability levels needed. The research objective of this study is to find argumentation for and against using open commercial networks to transmit strategic data.

5.3. Empirical argumentation concerning open network architectures

In the context of developing information management strategy for the defence sector, seventeen high level decision-makers were interviewed. The interviewees represented the following organizations: Ministry of Defence (6), Ministry of Foreign Affairs (1), units of the Finnish Army (9), and an IT security company (1). The interviews were semi-structured with the following questions linked to network security:

> "Effective networking and international cooperation with multiple partners requires the ability to design, decide on and implement ICT platforms in a controlled manner within the defence sector.
>
> - Whose responsibility is it to decide on cooperation with outside shareholders?
> - Who designs the ICT guidelines needed for this cooperation?
> - How can we secure sufficient decision power within the collaboration network?"
>
> "Which ICT systems in the defence sector should be compatible for exchanging information between partners?"
>
> "What kinds of requirement for collaboration do public authorities have during peace and conflict situations?"
>
> "Are electronic links feasible solutions in transactions with different suppliers?"
>
> "Is the outsourcing of ICT services possible within the defence sector? How should long-term reliability be secured in this case? How does foreign ownership influence outsourcing decisions?"
>
> "What are the strengths and weaknesses of having your own closed ICT network? Is it realistic to maintain in the long run?"
>
> "Do security requirements differ in different categories of information systems?"
>
> "What is the sufficient level of security? How much should be invested in maintaining a high level of security?"

The interviewees think that integration is increasing and deepening in Europe. Finland will have closer and more active links with foreign partners than ever before. In order to create a reliable defence policy Finland must, however, have a sufficiently strong defence capability. Threat scenarios have also changed and extended recently, moving emphasis from traditional war operations towards a holistic security policy.

Information wars and threats to information society are increasing. This requires a policy where traditional military forces are only one part. That increases demands for high security in all information systems, but especially in strategic applications.

The following offers some direct excerpts from the interviews to offer an authentic picture of the opinions presented:

"If somebody wants to hurt us, it is a waste of resources to send military troops here; there are much more cost-efficient alternatives."

"We should open our systems to achieve compatibility with other systems in society."

"The defence sector is already a significant buyer in peacetime. We have to ensure that there is a balance between operationality and security. We should not build our own exotic systems."

"Strategic partnership is a fact, which should be accepted. We should find those areas, where we have to operate cost-efficiently, and build interorganizational systems there."

"In principle we could lease all of our ICT platform and software. In these partnerships costs will first increase, but in the longer run the solutions are feasible."

"The required level of security can only be achieved with your own closed networks. It may be an expensive solution and difficult to maintain, but it is the only solution."

"We should aim at the highest possible security level and then use commercial networks."

"Your own closed network creates a false sense of security. Even when we have a closed architecture there are people involved and printed reports moving around. Investing the huge amounts of money now paid to operators to increase security, would give us a more operational environment. Our secrets are no greater than those of global companies."

"We are in any case heavily dependant on commercial civilian operators. Most of the lines have been leased from them. We should concentrate on building this environment to be as secure as possible."

"Information systems strategy should be very straightforward and totally reject self-made systems. It is enough to have a good knowledge of buying services."

"According to present knowledge every firewall is breakable."

"In strategic decision-making and operations it is necessary to keep the key processes in your own hands. This is the only way to guarantee defence capability in all circumstances. Recent examples of attacks on commercial networks show their vulnerability."

"Logical and physical security are different concepts. The main concern is to keep the network logically secure."

> "Human involvement in different parts of the information chain is the greatest risk to security."
>
> "Commercial services do not cover all the geographical areas. It is necessary to have your own networks in certain parts of the country."
>
> "By withdrawing from network administration, it would be possible to concentrate resources to key military applications."
>
> "The army is interested in its security levels, not in data transmission. It is impossible to move completely away from having your own networks, but with a partnership agreement it might be possible to reach a sufficient security level."
>
> "The decision-makers have difficulty in deciding the right level of security as nobody is able to offer zero risk."
>
> "In data transmission there are still challenges to meet. Our information travels through some of the existing lines anyway, and we should aim at the highest possible security there."
>
> "Breaking into information systems is more of a human than a technical risk. The situation with hackers shows, however, that we cannot use the present standards of the civilian world."

Based on these interviews, on discussions in the project group and on written materials, the following results have been derived.

5.4. Factors influencing decisions on the security levels of strategic information

An information society based on electronic data processing is rapidly developing and increasingly commercial networks are available. This development adds pressure also to public organizations like the defence sector to accept new models in their operations.

A defence force has its own closed network to maintain high security in information processing. Logistics and materials handling are however changing so that electronic documents become almost inevitable. Consequently the need to link defence information systems to supplier's systems is increasing. There is also a need to implement citizens' services through networks. These trends are changing the way networks will be designed and operated in the future.

The first finding in this study is the necessity to define strictly which part of the information processing activity really is competitive and strategic and should, therefore, be highly secured. The temptation for overprotection is very high and therefore this analysis is one of the key success factors. If the confidential data is defined in a relevant way, its processing may still be done in a closed network, which is anyway the most secure technical platform.

The second finding is that in a modern networked environment it is too expensive to design and implement all information systems within your own organization. Outsourcing and collaboration in some areas is evidently needed. As described in the interviews, strategic partnership is one possible solution. This partnership means very close cooperation with partners and calls for undeniable trust between them.

Globalization has an influence on strategic alliances in the sense that ownerships of companies may change through mergers. Thus the decision-making of partners may shift to other countries or geographical areas. This creates some risks in confidentiality, but internationalization is a natural development. It is therefore important to be able to create international alliances in order to maintain a modern technical level in information systems. Security aspects need strict agreements, but also trust between actors.

International co-operation is increasing also in military operations. This gives rise to the problem that the information systems should be sufficiently compatible to make joint decision-making possible. One of the challenges is to maintain national security in this global environment. This calls for a systems architecture that allows both for co-operation and closed processing.

Referring to the earlier categorization of the quality of information, the following observations may be made. Technical measures to meet reliability and security needs are continuously improving, but will never be perfect. Much attention should be directed into developing technical security, but severe risks remain. Technical security is in conflict with user requirements such as responsiveness, access, courtesy and communication. The higher the protection measures, the more difficult the systems are to use. Human requirements like competence, credibility, and understanding the customer and service view are continuously increasing. The problem is to find a proper balance between all these requirements.

To summarize, we have seen that the question of securing strategic and competitive data and information is a very complicated task. The situation is problematic because even the experts have different views of what can and cannot be done. There is conflict as at the same time security requirements are high and the development towards open network society is rapid. To simplify the case, the answer to the research question is that open commercial networks are the only relevant solution in the future, also in public administration. In this environment, the confidential data should be very carefully analyzed and you should concentrate on protecting this carefully selected key strategic data.

6. Conclusion

This study deals with information processing in the public sector, but it has some implications also for other organizations. The main conclusion from this exercise is that the nature of strategic thinking is changing. Planning and decision-making has traditionally been divided into three different levels: strategic, tactical and operational. Earlier the strategic objectives and actions were highly secret and secured. Nowadays strategies, market share objectives, expected new products and even some location plans are widely known. Analysts speculate with the success factors of all significant companies and management speaks openly of the general goals of its organization.

Instead of what was earlier called strategy and strategic, competitive actions and moves are nowadays confidential. Organizations are able to gain competitive advantage from the way they are realizing their strategies, and no longer from the strategies themselves. The role of product and service development, location decisions, mergers and other similar aspects is strengthening. The processing of this information should be strictly confidential, which adds pressure to designing reliable information systems.

The final conclusion is that competitive information should be secured in a strict way. Both human and electronic processing should be considered in this context. Technical security can be high, but never perfect. Therefore the role of human information processing is extremely important. Experience and human knowledge is difficult to imitate, and much emphasis should be placed on developing sustainable organizational culture. This culture should be so strong that even losing some of the key persons does not destroy the competitiveness of the organization. This objective may be met by sharing responsibility widely to maintain continuity.

References

Applegate, Lynda M., Robert D. Austin & Warren M. McFarlan (2001). *Creating Business Advantage in the Information Age.* New York, NY: McGraw-Hill Irwin.

Argyris, C., R. Putnam, D. McLain Smith (1987). *Action Science.* Jossey-Bass Publishers.

Auer, Timo (1995). *Information Systems Related Organizational Maturity: A Conceptual Framework and an Assessment Method.* Publications of Turku School of Economics and Business Administration, A7-1995.

Berry, Parasuraman and Zeithaml (1985). Quality Counts in Services, Too. *Business Horizons* (May–June), 44–52.

Checkland, P. (1991). From framework through experience to learning: The essential nature action research. In: *Information Systems Research.* Eds Nissen, Klein, Hirschheim. North-Holland.

Collins, James C. & Jerry L. Porras (1996). Building your companies vision. *Harvard Business Review* (September–October).

Davis, F.D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly* (Sept.), 319–339.

Gummesson, Evert (1991). *Qualitative Methods in Management Research.* Newbury Park, CA: Sage Publications.

Liautaud, Bernard (2000). *e-Business Intelligence, Turning Information into Knowledge into Profit.* McGraw-Hill.

Ministry of Defence of Finland (2001). Finnish Security and Defence Policy 2001, *Report by the Government to Parliament* on 13 June 2001.

Nohria, N. & S. Ghosal (1997). *The Differentiated Network, Organizing Multinational Corporations for Value Creation.* San Francisco, USA: Jossey-Bass Inc.

Nonaka, I. & H. Takeuchi (1995) *The Knowledge Creating Company: How Japanese Companies Create the Dynamics of Innovation.* Oxford: Oxford University Press.

Polanyi, M. (1966). *The Tacit Dimension.* London: Routledge & Paul Kegan.

Prahalad, C.K. & Y.L. Doz (1987). *The Multinational Mission: Balancing Local Demands and Global Vision.* New York: Free Press.

Quinn, J.B. & P.C. Paquette (1990). Technology in services: Creating organizational revolutions. *Sloan Management Review* 33:2, 67–78.

Porter, M. E. (1998). *On Competition.* Boston: Harvard Business School Press.

Reponen, Tapio (1994). Organizational information management strategies. *Information Systems Journal* 4.

Reponen, Tapio (1998). The role of learning in information systems planning and implementation. *Information Technology and Organizational Transformation,* 134–149. Eds R.D. Galliers & W.R.J. Baets. John Wiley & Sons Ltd.

Reponen, Tapio (ed.) (2003). *Information Technology – Enabled Global Customer Service.* IDEA Group Publishing

Reponen, Tapio & Timo Auer (1997). Information systems strategy formation embedded into a continuous organizational learning process. *Information Resources Management Journal* 10:2, 32–43.

Reponen, T., T. Wood-Harper & L. von Hellens (1992). *Action Research as a Bridge between Academic World and Business Life.* Information Systems Science, The Turku School of Economics.

Riordan, Patrick (1995). The philosophy of action science. *Journal of Management Psychology* 10:6.

Senge, P. (1994). *The Fifth Discipline: The Art & Practice of the Learning Organization.* Doubleday & Company.

Sergeant, A. & S. Frenkel (2000). When do customer contact employees satisfy customers? *Journal of Service Research* 3:1, 18–34.

Sveiby, K.E. (1990). *Valta ja johtaminen asiantuntijaorganisaatioissa.* Ekonomia, Weilin & Göös.

Thurow, Lester (1996). *The Future of Capitalism: How Today's Economic Forces Shape Tomorrow's World.* William Morrow and Company, Inc.

You've got email ... again!
Protecting one's emailbox from spam with automatic filtering

Timo Salmi

Dedicated to Ilkka Virtanen on the occasion of his 60th birthday

Abstract

Salmi, Timo (2004). You've got email ... again! – Protecting one's emailbox from spam with automatic filtering. In *Contributions to Management Science, Mathematics and Modelling. Essays in Honour of Professor Ilkka Virtanen.* Acta Wasaensia No. 122, 225–244. Eds Matti Laaksonen and Seppo Pynnönen.

This paper outlines and exemplifies a spam (unsolicited commercial and other undesirable email) foiling system developed by the author. This automatic email filtering system is based on the concepts of whitelisting and blacklisting as adopted by the author in 1994. The paper suggests the idea of email password requirement returned automatically to the sender before accepting email as adopted by the author in 1997 for his own email protection. The paper also contributes the author's own method for testing the filtering recipes for procmail autonomous mail processor, and in an appendix a WWW-page based variant of the main password method.

"Looks like Timo had the right idea years ago!" (Berkes 2003). "For me, the day I read Professor Timo Salmi's webpage was a good day." (Clifford 2003).

Timo Salmi, Professor, Department of Accounting and Finance, University of Vaasa, P.O. Box 700, FIN–65101 Vaasa, Finland, e-mail ts@uwasa.fi, http://www.uwasa.fi/~ts/.

Acknowledgments

My thanks are due to Prof. Seppo Pynnönen for his comments in writing this paper. I also thank System Manager Hannu Hirvonen for many interesting computer related problem solving discussions over the years.

1. Introduction

It is well agreed that the main direction of the impact of computers on the then future society would have been very difficult to predict with any accuracy in the computers' early years in the 1950's. Originally regarded as mainly fast scientific calculators and commercial database processors, while also this aspect still is there, the computers' major domain at the beginning of the new millennium lies elsewhere. Their most profound impact has been in personal communication and services, and the consequent impact on society. This fact is based on two, originally unforeseen developments. The first was the personal computer (IBM 1981), considered just a curiosity at the outset. The second was the emergence of the Internet. Together these two have brought about a totally new era of global communication with all its benefits, but also its unfortunate, persistent abuses.

The paper at hand focuses on the most serious abuse and threat that has challenged the new flow of global communication almost since its outset (BBC News 2003, Templeton 2003 b). The phenomenon is Unsolicited Commercial Email, familiarly known as spam (Templeton 2003 a). With (according to some estimates) up to a half of the world-wide Internet bandwidth abused this way, spam poses a serious choking threat to the very texture of global communication. (Lemos 2002, Roberts 2002, Sounders 2002, Krim 2003, Paukku 2003, Siponen & Alatalo 2003). Even at best, spam and scams, such as the reinvented chain letter pyramid schemes (dmoz: Chain Letters, Osgoodby 2002, Watrous 2003) and the Nigerian Advance Fee Fraud (Arolainen 2003, US Secret Service), are huge nuisances for anyone with an email address on the Internet. Anyone with some experience on the Internet will have received such junk email to one's mailbox. A sobering indication of the extent of spam is a semi-random sample at the system-level at the University of Vaasa, Finland. On Monday, the 18th of August, 2003 the university's total email traffic amounted to 351713 messages. The collective filter at the university level intercepted 272656 messages (77.5% of the total email traffic) from 6564 different IP addresses as spam! And this figure still leaves unaccounted the spam that got through to the individual users. Such high figures are explained by the fact that the spammers systematically collect and generate target email lists. Quoting e.g. from the Federal Trade Commission (2002) spammers typically use computer programs that search public areas on the Internet to harvest lists of email addresses extensively from web pages, newsgroups and chat rooms. (An even more comprehensive list of sources is listed e.g. by Raz.)

Another, often also unsolicited email related, but a much more sporadic modern threat is the spreading of the computer viruses/worms. This related, considerable threat would merit a consideration of its own. It is limited outside the scope of the current paper. However, the methods described for filtering one's email are directly applicable to the consequences of the email worms. (Such as e.g. Sobig.F which caused an unprecedented amount of bogus email traffic starting the 19th of August, 2003. See Appendix 3.)

This paper presents the ideas of an automatic spam filtering system, and exemplifies the methods on a UNIX-based email system. The concepts (in particular: whitelisting, blacklisting, and the email password requirement autoresponder) to be discussed can't definitely be pinpointed to any single source of origin or a point of time. The present author does not claim proprietorship or strict originality on any of these concepts. Nevertheless, the methods, as presented in this paper, have been developed independently without known precedents by the current author. (Salmi 1997, 1999).

2. The outline of the automatic filtering design

When spam and other junk email pile in a person's mailbox at rates such as over 30 messages a day even at a collectively protected location it is advisable to consider resorting to automatic presorting of the incoming messages also at the individual user-level. Today many ISPs (Internet Service Providers) offer collective email protection already at the system level. The big problem in the collective method is finding the right balance between what is to be stopped as spam and what is to be let through. In the end it is the actual user who knows best. This fact speaks for the need of also individualized solutions.

Very generally speaking the idea is to direct the incoming email to different folders depending on where it comes from and what it contains. Such a system can grow very complicated, and therefore a basic framework for thinking is needed. The method proposed here is made up of the following three layers (and then a considerable number of practical refinements).

1. Whitelisting: All email that appears on the database of whitelisted sources is let through. The whitelisted email can typically be further presorted to different folders (including one's main mailbox) according to e.g.
   - the sender,
   - the recipients (e.g. is it a mailing list),
   - the subject matter.
2. Blacklisting: All email that appears on the database of blacklisted sources (or subjects) is stopped. The blacklisted email can typically be further processed to
   - discard it outright ("/dev/null'ing"),
   - return it to the sender only, or
   - return it also to the sender's postmaster or abuse service, whilst
     - keep yourself a copy in a specified folder, or
     - keep yourself no copy.

3. Passwording: All the rest of the email is returned to sender with a password to be used in any further email, whilst
   - no copy is made, or
   - a copy made to a dedicated folder, while
   - additional, selective actions can be taken.

A note on terminology: On the Usenet news new words and meanings emerge with the new developments. (As a matter of fact "spam" is a good example of such a neologism.) In general terms methods where the sender is asked to take some action before getting through have become to be called C-R (challenge-response) systems or colloquially prove-you-love-me texts. Another interesting concept and method greylisting (Harris 2003) came to the fore in June 2003 in a thread in the news:comp.mail.misc Usenet newsgroup. It signifies measures where the delivery of a new, unidentified message is "falsely" bounced to require the common, automatic delayed resending. This may well be a good idea, but compared to the password requirement the logic is a bit harder to follow if one is not well familiar with the details of the protocols which email programs use in handling bounces.

3. Getting started with the filtering

Given the general outline of the filtering system, to apply the delineated logic, the different cases have to be programmed as explicit rules. Some email programs, such as Microsoft Outlook Express (Microsoft Corporation) or Netscape Messenger (Netscape Marketing 2000) include the option for entering filtering rules. While useful, these options are somewhat limited in scope and in the range of actions that can be taken. The other possibility is to have a dedicated, separate filter program in front of the actual email program to do the filtering. Furthermore, complicated actions, once the incoming email is identified by the filter, can be carried out by various system tools.

In this paper the route of a separate filter program is taken. The platform for the examples to be presented will be a UNIX operating system. The email filtering program to be used in the examples is procmail ("autonomous mail processor"). Most of the actions will be programmed within procmail (procmail.org) as procmail recipes, but in more complicated cases the activities are piped to specific scripts using Bourne Shell scripting (Eriksson). Of course, in a way these are arbitrary choices, but they give a high flexibility of email filtering and they also are convenient for presenting the programming tasks involved. The downside of this flexibility will be that using such a system will require a fair amount of general understanding programming and a familiarity with the specific "power" tools chosen. However, they are much used by advanced users on the net community and they make a good demonstration of how to apply the outlined filtering design.

3.1. Setting up procmail

Provided that (as is usual) the email host system has the procmail program, the first step in installing the spam filtering system in a UNIX system is to redirect all incoming email to thc procmail program. The minute details and requirements of such setting up are better given in procmail manuals and e.g. advice WWW pages such as Salmi (1999) and its references. In a UNIX system email can be forwarded by creating a ~/.forward file with the following kind of contents

```
"|IFS=' ' && exec /usr/local/bin/procmail || exit 75 #myid"
```

The principle of piping all the incoming to thc procmail is essential here, not so much the actual details of the arrangements and syntax, even if in actual practice they naturally must be entered correctly. In other words, this paper is not a program manual. The main purpose is to present the principles of spam foiling. Therefore, many minute details necessary for actually using such a procmail email filtering system, but not essential for the delineating the broad principles, are skipped and not commented hcre.

The crucial file in the procmail email filtering system is the ${HOME}/.procmailrc "procmail rcfile" configuration file which contains the "recipes" to process the incoming email. The recipes are made up of "rules". A simple case of .procmailrc is the following set, which blankly discards all "make money fast" email and puts all other incoming email into the user's mailbox.

```
:0                              # A new recipe starts with ":"
* ^Subject:.*Make money fast
/dev/null                       # This email is discarded

:0:                             # The second recipe starts
${DEFAULT}                      # The rest go to the user's mailbox
```

The rules of selection start with the "*" token. They use approximately the same syntax as UNIX egrep ("Search a file for a pattern using full regular expressions", see e.g. Stearns 1995) without an upper/lower case sensitivity. For example, were the email subject actually "how to make money fast **now**", the first recipe's rule would find a (regular cxpression) match and thus the first recipe's action would be taken.

3.2. Creating a procmail test instrument

As explained, the user enters the desired procmail recipes into his/her ${HOME}/.procmailrc (that is ~/.procmailrc) file to take different actions on the incoming email. For testing purposes this is not a good solution. The procmail recipes can grow fairly complicated, and furthermore email does not necessarily come in suitably at will. Therefore, a detached test method is needed to develop and evaluate individual procmail filtering recipes without disturbing one's regular email handling in the process. Full originality can hardly be claimed in creating such a testing system. Nevertheless, the one to be presented below has been independently put forward by the present author in Salmi (1999).

The testing system can be set up as follows. Create the following "proctest" file, preferably at the path. Make it executable using "chmod u+x proctest". Thus a new command "proctest" will always be available. The recipes to be tested are entered into the proctest.rc file, and the corresponding email to be tested is copied to the mail.msg file (e.g. if one uses UNIX elm "interactive mail system" or mutt "The Mutt Mail User Agent" the copying of an existing message can be done with the "C" copy command from within elm or mutt). Only one email message should be subjected to testing at a time to avoid confusing results.

```
#!/bin/sh
# The executable UNIX script file named "proctest".
# Bourne shell (the original UNIX shell /bin/sh) syntax is used.
#
# A test directory is needed.
# The exact location depends on the user's environment.
# The directory must be created if it does not already exist.
TESTDIR=/home/myid/test
#
# Feed an email message to procmail. Apply proctest.rc recipes file.
# First prepare a mail.msg email file to use for the testing.
procmail ${TESTDIR}/proctest.rc < ${TESTDIR}/mail.msg
#
# Display the outcome.
ls -lF ${TESTDIR}/Proctest.*
less ${TESTDIR}/Proctest.log
#
# Clean up.
rm -i ${TESTDIR}/Proctest.*
```

The procmail configuration file to be tested

```
# The proctest.rc test configuration file
#
# Setting some environment variables used by procmail
VERBOSE=yes
```

```
LOGABSTRACT=all
MAILDIR=${HOME}/test
LOGFILE=${HOME}/test/Proctest.log
#
# The recipes
:0:                         # The first recipe starts
* ^From:.*itv@.*uwasa\.fi
Proctest.itv                # A folder for email from Ilkka Virtanen
#
:0:                         # The second recipe starts
Proctest.rest               # All other email to this folder
```

An example email message copied to the mail.msg file

```
From itv@uwasa.fi Fri Jun 13 08:39:06 2003
Date: Fri, 13 Jun 2003 08:39:05 +0300
From: Ilkka Virtanen <itv@UWasa.Fi>
To: Timo Salmi <ts@uwasa.fi>
Subject: A new link

(The body of the message)

--
Ilkka Virtanen.
Professor of Operations Research and Management Science
Dean of the Faculty of Technology
University of Vaasa, POB 700, FIN-65101 Vaasa, Finland
Tel. 358-6-3248256, 358-50-5377909, Fax 358-6-3248557
E-mail: itv@uwasa.fi          http://www.uwasa.fi/~itv/
```

Captured output from the simple example test

```
poiju> proctest
procmail: [13340] Sat Jun 14 10:12:28 2003
procmail: Assigning "LOGABSTRACT=all"
procmail: Assigning "MAILDIR=/home/myid/test"
procmail: Assigning "LOGFILE=/home/myid/test/Proctest.log"
procmail: Opening "/home/myid/test/Proctest.log"
-rw-------   1 ts   ktt   958 Jun 14 10:12 /home/myid/test/Proctest.itv
-rw-------   1 ts   ktt   335 Jun 14 10:12 /home/myid/test/Proctest.log
procmail: Match on "^From:.*itv@.*uwasa\.fi"
procmail: Locking "Proctest.itv.lock"
procmail: Assigning "LASTFOLDER=Proctest.itv"
procmail: Opening "Proctest.itv"
procmail: Acquiring kernel-lock
procmail: Unlocking "Proctest.itv.lock"
From itv@uwasa.fi Fri Jun 13 08:39:06 2003
 Subject: A new link
  Folder: Proctest.itv                                                  479
rm: remove /home/myid/test/Proctest.itv (yes/no)? n
rm: remove /home/myid/test/Proctest.log (yes/no)? y
```

The email from Ilkka Virtanen used for the presented testing now resides in the Proctest.itv file and can be processed with any email program (elm is used in these tests). The essential tool to develop and test various situations at will is now readily available.

4. Whitelisting and blacklisting

4.1. A whitelist example

The first step in installing the spamfoiling system under observation is setting up a whitelist and a blacklist. Setting them up differs so little from each other that for the general idea it is sufficient to present setting up a whitelist and to observe that in blacklisting the email just would be discarded (to /dev/null) instead of directing it to a folder (Proctest.white) or (as would be more usual) directly to the user's email box (${DEFAULT}). The option of returning blacklisted email is taken up in the next section.

```
# The proctest.rc test configuration file
#
# Environment variables for procmail
VERBOSE=yes
LOGABSTRACT=all
MAILDIR=${HOME}/test
LOGFILE=${HOME}/test/Proctest.log
#
# Auxiliary definitions
# Get the sender's bare email address using formail mail reformatter program
# Ignore the Reply-To header-field:
FROMADDR=`formail -c -I"Reply-To:" -rt -xTo: \
          | expand | sed -e 's/^[ ]*//g' -e 's/[ ]*$//g'`
#
# Extract the sender's name from the first From: field
FROMNAME=`sed -e '/^$/ q' \
          | expand | egrep '^From: ' | head -1 \
          | sed -e 's/<//g' -e 's/>//g' \
          | sed -e 's/From: //' \
          | sed -e 's/^[ ]*//g' -e 's/[ ]*$//g'`
#
# The recipes
# Accept based on the potential appearance on either of the two whitelists
:0
* 1^0 $ ? echo "${FROMADDR}" | egrep -is -f /home/myid/test/whiteaddr.lst
* 1^0 $ ? echo "\"${FROMNAME}\"" | egrep -is -f /home/myid/test/whitename.lst
{
  :0
  { RULE="Accepted based on the generic whitelists" }
  :0:
  Proctest.white
}
#
:0:                             # The second recipe starts
Proctest.rest                   # All other email to this folder
```

Example contents of whiteaddr.lst

```
itv@.*uwasa\.fi
ts@.*uwasa\.fi
```

Example contents of whitename.lst

```
Ilkka.Virtanen
Timo.Salmi
```

It is easy to read in the example above the fact that setting up such a system is prohibitively complicated for an non-specialized user. The help of computer support personnel or some sort of preprocessed packages are needed for a more widespread application of these ideas. This probably goes to explain why the kind of filtering described in this paper seems to be fairly rare despite the huge problem posed by the unrelenting spam-situation on the Internet. Fortunately, once the system is in place, it is reasonably easy for any user to write out the address lists such as whiteaddr.lst and whitename.lst.

Although the separate lists are convenient, the whitelisted and blacklisted names/ addresses could, of course, as well be inserted into the ~/.procmailrc configuration file. In fact, when filtering by the subject (see the "Make money fast" example in Section 3.1) all the parts would usually be contained within the configuration file, only.

4.2. Returning blacklisted email

As pointed out in Chapter 2, one of the options with blacklisted email is not just to discard it, but also to return it to the sender, or even make a copy to the sender's postmaster. Returning a message is of interest here because it demonstrates UNIX scripting (called by the "| piping") in conjunction with procmail. It is true that in actual practice much, if not most, of junk email on Internet comes from forged email addresses. Nevertheless, the example at hand shows the realization of the basic principle of returning a rejected message. An appropriate example recipe is given below:

```
# The recipes
# Safeguard against the possibility of email loops
:0:
* ^X-Loop:.*myid@myhost\.mydom
XLoop.mail

# Return blacklisted email
:0
* ^From:.*(\
BUSINESS.*TRAVEL.*LTD|\
MRS MURIYN JONAS SANIMBI|\
Vile.*Spammer)
{
  # First make a temporary file of the message to be returned
  :0c:formail.lock
  # Discard whitespaces from the said incoming email, insert a leading blank
  | expand | sed -e 's/[ ]*$//g' | sed -e 's/^/ /' > return.tmp
  #
  # Use formail -r mail reformatter program to resolve a suitable return address
  # Add a return subject and a loop safeguard to the outgoing email header
  # Construct and then send with sendmail program the rejection prepared
  #
  :0:formail.lock
  | (formail -r -I"Subject: Rejected mail: Recipient refusal" \
    -A"X-Loop: myid@myhost.mydom" ; \
    echo "--- begin rejected mail ---" ; \
    cat return.tmp ; \
    echo "--- end rejected mail ---" ; \
    rm -f return.tmp) \
    | /usr/lib/sendmail
}
```

In the above neither a separate blacklist file nor a separate script file is used. Instead all the rules and the actions are embedded into the recipe file for the current demonstration. In actual practice there are pros and cons to such a choice. All the handling stays in one file, which makes for a more concentrated documentation. On the other hand this way a real-life procmail configuration file tends to grow quite large and possibly quite convoluted.

Below is the current test rejection response that goes out. In actual practice the sender's address often would be masked and replaced by a non-returnable address to further guard against unwanted undeliverable email announcements and as a further safeguard against the potential email loops. Furthermore, an automatic copy could be sent to the blacklisted user's postmaster. However, those are practical details (sometimes a bit complicated), which do not add anything crucial to the ideas under observation here, and thus need not be demonstrated in the actual text. (Formulating the postmaster address is given in the Appendix 2. For more information on those aspects, see the procmail links in the references section of this paper, and the further links within those references.)

```
From ts@mail.uwasa.fi  Thu Jun 19 13:59:08 2003
Date: Thu, 19 Jun 2003 13:59:08 +0300 (EET DST)
From: <ts@mail.uwasa.fi>
To: ts@uwasa.fi
X-Loop: myid@myhost.mydom
Subject: Rejected mail: Recipient refusal

--- begin rejected mail ---
 From ts@uwasa.fi Thu Jun 19 12:59:14 2003
 Date: Fri, 13 Jun 2003 08:39:05 +0300
 From: Vile E. Spammer <forged@nowhere.com>
 To: ListOfSpamTargets <hidden@nowhere.com>
 Subject: Spam test
 Status: RO

 (The body of the test "spam" message)

    All the best, Timo;
    (posing as "Vile E. Spammer" for testing purposes)

 --
 Prof. Timo Salmi ftp & http://garbo.uwasa.fi/ archives 193.166.120.5
 Department of Accounting and Business Finance  ; University of Vaasa
 mailto:ts@uwasa.fi <http://www.uwasa.fi/~ts/>  ; FIN-65101,  Finland
 Timo's  FAQ  materials  at   http://www.uwasa.fi/~ts/http/tsfaq.html

--- end rejected mail ---
```

5. Requiring and identifying a password

The example in the previous section on returning blacklisted email to a great extent also gives the technical framework for sending out the password requirement. Only a few adjustments are needed. Thus part of the documentation comments can be omitted.

```
# Set your public email password
EMAILPASSW="abcd"

# Safeguard against the possibility of email loops
:0:
* ^X-Loop:.*myid@myhost\.mydom
XLoop.mail

# The whitelist and blacklist recipes here
#     (Those recipes)

# Accept to your mailbox all email having the password on the subject line
:0:
* $ ^Subject:.*${EMAILPASSW}
${DEFAULT}

# For the rest of the incoming email, return the password requirement
:0
{
  SUBJ=`formail -xSubject: \
       | expand | sed -e 's/^[ ]*//g' -e 's/[ ]*$//g'`
  FROM=`formail -rt -xTo: \
       | expand | sed -e 's/^[ ]*//g' -e 's/[ ]*$//g'`
  :0:formail.lock
  | (formail -r -I"Subject: Returned email: Password required" \
    -A"X-Loop: myid@myhost.mydom" ; \
    echo "*************************************************" ; \
    echo "* This is a computer-generated response message *" ; \
    echo "*************************************************" ; \
    echo "" ; \
    echo "Dear ${FROM}" ; \
    echo "" ; \
    echo "Thank you for your email to me about" ; \
    echo "${SUBJ}" ; \
    echo "" ; \
    echo "To reach me, please include my public email password" ; \
    echo "${EMAILPASSW} anywhere on your subject line.") \
    | /usr/lib/sendmail
}
```

Using exactly the same test message as in the previous section the simplified requirement returned to the sender of the incoming email would look something like this:

```
From ts@mail.uwasa.fi  Thu Jun 19 23:31:40 2003
Date: Thu, 19 Jun 2003 23:31:40 +0300 (EET DST)
From: <ts@mail.uwasa.fi>
To: ts@uwasa.fi
X-Loop: myid@myhost.mydom
Subject: Returned email: Password required

*************************************************
* This is a computer-generated response message *
*************************************************

Dear forged@nowhere.com

Thank you for your email to me about
Spam test

To reach me, please include my public email password
abcd anywhere on your subject line.
```

In principle, that's all there is to the general outline of the password requirement method combined with whitelisting and blacklisting.

6. Conclusion

About five years of the author's practical experience with effectively the described password requirement method in place has turned out to be practically foolproof against unsolicited commercial email, i.e. the spam. (By a very rough approximation an order of 0.01% of the incoming spam has made it to the author's mailbox over the said period.) If one considers the layers of the logic outlined in Chapter 2, this is not at all surprising. If the sender has a forged address, as is common in spam, s/he'll never see the public password and thus will not get past the password requirement. In fact, s/he'll be unaware of its very existence. On the other hand, if s/he does get the password requirement response, it is extremely rare, even if not impossible, that the spammer would actually use the required public password. Why? As stated in Salmi (1997) "By its very nature spamming is a huge mass activity. There is no way the spammers have the resources to customize in order to bypass a single individual's public password protection. Furthermore, contrary to a getting a new, spam-free user id, an email password is easy to change anytime."

The principles described in this paper to solve the prohibitive junk email problem are straightforward and simple. Unfortunately, programming and setting them up is obviously far too complicated for an lay-person PC user to do unaided. "Easy to describe, complicated to install." Therefore, practical methods to set up the ideas presented are needed. The incentives to do so fall into the realm of commercial programming. In fact, at the time of writing this, there are some movements in evidence on the Internet towards this direction. One of such examples is Spam Arrest (2003) which is based on directing the email sender to an identification requirement acted out on a dedicated web page. That idea has much in common with the variant presented in Appendix 1. Another similar, recent example is Spam Catcher (2003). Another consequence of the complicated nature of filtering is that quality ISPs increasingly will have to offer improved, collective spam protection already at the system level.

There also is another dilemma involved with the password requirement that has to be observed. The downside of such a junk email prevention method is that since it is very effective it also will curb some legitimate email contacts. This is not a prohibitive problem, when whitelisting is a practical proposition, but for example commercial firms with huge potential customer bases just can't afford losing the potential contacts because of a password requirement that some users will find excessive or even offensive. The system is much better fitted for a private user or a user where the contacts typically remain e.g. within, say, a university environment circle.

A third obvious dilemma definitely worth further consideration is the logic of initiating a contact between two users of the email password system when the parties are previously unknown to each other (see however, Appendix 1). Because the password system still is rather rare this problem has not yet risen more to the fore. Some common protocol is probably in order just like in the exchange of public keys in PGP exchanging encrypted messages (see e.g. Singh 1999). The options include publicizing one's public email password on one's web page, in one's email signature, or accepting email sent through one's webpage (since then special arrangements are easy to set up). In fact, the author utilizes such measures.

Be the problems with the presented method as may, it is unavoidable that one way or the other the junk email problem will eventually have to be solved before it manages to choke the usefulness of the global Internet. It only is to be hoped that the ideas presented in this paper and on the net by the present author might make up one modest step towards a resolution.

References

Arolainen, Teuvo (2003). Nigerialaiskirjeiden tulva jatkuu. *Helsingin Sanomat* 5.5.2003, A7.

Berkes, Jem (2003). *Subject: Re: anti spam* [online] [cited 28-May-2003]. A Usenet news posting in the newsgroup news:comp.mail.misc. Date: Tue, 27 May 2003 15:59:05 GMT. Message-ID: Xns93886FCFED449jbuserspc9org@205.200.16.73

BBC News [online] (2003). Spam celebrates silver jubilee. 4-May-2003 [cited 6-May-2003]. Available from Internet: http://news.bbc.co.uk/2/hi/technology/2996319.stm

Clifford, Alan (2003). *Subject: Re: The greylisting idea, effective?* [online] [cited 25-June-2003]. A Usenet news posting in the newsgroup news:comp.mail.misc. Date: Wed Jun 25 01:38:17 EET DST 2003. Message-ID: Pine.LNX.4.53.0306242249540.19119@ mundungus.clifford.ac

dmoz [The Open Directory Project] [online]. Logical search path: Top: Society: Issues: Fraud: Internet: Make Money Fast: Chain Letters [cited 14-May-2003]. Available from Internet: http://dmoz.org/Society/Issues/Fraud/Internet/Make_Money_Fast/Chain_Letters/

Eriksson, Era. *An Introduction to the Unix Shell; An HTMLized version of Steve Bourne's original shell tutorial.* [online] [cited 17-May-2003]. Available from Internet: http://rhols66.adsl.netsonic.fi/era/unix/shell.html

Federal Trade Commission (2002). Consumer Alert. *Email Address Harvesting: How Spammers Reap What You Sow.* [online] [cited 3-September-2003]. Available from Internet: http://www.ftc.gov/bcp/conline/pubs/alerts/spamalrt.pdf

Harris, Evan (2003). *The Next Step in the Spam Control War: Greylisting* [cited 24-June-2003]. Available from Internet: http://projects.puremagic.com/greylisting/

IBM (1981). Personal computer announced by IBM. *IBM (Information Systems Division, Entry Systems Business) Press Release* August 12, 1981 [online] [cited 23-June-2003]. Available from Internet: http://www-1.ibm.com/ibm/history/documents/pdf/pcpress.pdf

Krim, Jonathan (2003). Spam's Cost To Business Escalates: Bulk E-Mail Threatens Communication Arteries. *Washington Post* Thursday, March 13, 2003; Page A01 [online] [cited 5-May-2003]. Available from Internet: http://www.washingtonpost.com/wp-dyn/ articles/A17754-2003Mar12.html

Lemos, Robert (2002). Spam hits 36 percent of e-mail traffic. *CNET News.com* August 29, 2002, 4:48 AM PT [online] [cited 5-May-2003]. Available from Internet: http://zdnet.com.com/2100-1106-955842.html

Microsoft Corporation. *Home Page for Microsoft Outlook*. [online] [cited 17-May-2003]. Available from Internet: http://www.microsoft.com/office/outlook/

Netscape Marketing (2000). *Netscape Messenger*. [online] [cited 17-May-2003]. Available from Internet: http://wp.netscape.com/communicator/messenger/v4.0/index.html

Osgoodby, Bob (2002). What Is A Ponzie? *INSIDER REPORT* Monday, October 07, 2002 [online] [cited 14-May-2003]. Available from Internet: http://www.insiderreports.com/ storypage.asp_Q_ ChanID_E_MR_A_StoryID_E_20001095

Paukku, Timo (2003). Roska@joukko tukkii sähköpostit. *Helsingin Sanomat* 5.7.2003, C15.

Raz, Uri. How do spammers harvest email addresses? [online] [cited 4-September-2003]. Available from http://www.private.org.il/harvest.html

procmail.org *The home page of the procmail mail processing and SmartList mailing list suites*. [online] [cited 17-May-2003]. Available from Internet: http://www.procmail.org/

Roberts, Paul (2002). Holidays Bring a Whole Lot of Spam. *PCWorld.com* Monday, December 23, 2002 [online] [cited 5-May-2003]. Available from Internet: http://www.pcworld.com/news/article/0,aid,108174,00.asp

Salmi, Timo (1997). *Foiling Spam with an Email Password System*. [online] [cited 5-May-2003]. Available from Internet: http://www.uwasa.fi/~ts/info/spamfoil.html

Salmi, Timo (1999). *Timo's procmail tips and recipes*. [online] [cited 5-May-2003]. Available from Internet: http://www.uwasa.fi/~ts/info/proctips.html

Saunders, Christopher (2002) Study: E-mail to Double by 2006. *AtNewYork.com* [online] [cited 5-May-2003]. Available from Internet: http://www.atnewyork.com/news/article.php/1471801

Singh, Simon (1999). *Code Book. The Secret History of Codes & Code-breaking*. Fourth Estate, London.

Siponen, Mikko T. & Toni Alatalo (2003). Roskapostilta voi suojautua. Vieraskynä, *Helsingin Sanomat* 20.6.2003, A5.

Spam Arrest (2003). [online] [cited 24-June-2003]. Available from Internet: http://spamarrest.com/

Spam Catcher (2003). [online] [cited 16-August-2003]. Available from Internet: http://www.softouch.on.ca/spatcher/

Stearns, Bob (1995). UNIX regular expressions. *The UCNS Computer Review* Spring Quarter [online] [cited 23-June-2003]. Available from Internet: http://www.uga.edu/ ~ucns/tti/Computer_Review/Spring95/Regular_expressions.html

Templeton, Brad (2003 a). Reflections on the 25th Anniversary of Spam. [online] [cited 5-May-2003]. Available from Internet: http://www.templetons.com/brad/spam/spam25.html

Templeton, Brad (2003 b). Origin of the term "spam" to mean net abuse. [online] [cited 5-May-2003}. Available from Internet: http://www.templetons.com/brad/spamterm.html

United States Secret Service. *[Nigerian] Advance Fee Fraud Advisory*. [online] [cited 5-May-2003]. Available from Internet: http://www.secretservice.gov/alert419.shtml

Watrous, Donald (2003). *Chain letters* [cited 14-May-2003]. Available from Internet: http://www.cs.rutgers.edu/~watrous/chain-letters.html

Appendix 1: The WWW-page variant of spam avoidance

There is a fairly simple variation to complement (rather than replace) the public email password system. One could call it the WWW-page variant or even the Guestbook rendition. It goes as follows: in building a WWW page the HTML code for sending email via the WWW page is

```
<A HREF="mailto:myid@myhost.mydom">Click to send me email</A>
```

If one changes it to

```
<A HREF="mailto:myid@myhost.mydom(FirstName LastName Passwd)">Click to send me email</A>
```

Then the procmail recipe for accepting such email simply is e.g.

```
:0:
* ^To:.*FirstName.*LastName.*Passwd
WWW.mail
```

Naturally, this method is not a guarantee against spam. But in the author's experience it can be sufficiently effective in actual practice. The author has had this complementary variation on the side of the main email password spam foiling method on several web pages since 1998. Almost no additional spam has resulted via this route. A big advantage of this method is that it at least partly avoids the first contact conundrum if both the parties are using a challenge-response email password system.

The same contribution situation goes of the WWW-page variant as for the principal password-based challenge-response spam foiling method presented in this paper. The present author does not claim an actual originality of the idea. Nevertheless, the method presented in this appendix has been developed independently without known precedents by the current author.

Appendix 2: An advanced example recipe for detecting Korean email

Much spam seems to originate from Korea. Detecting if the incoming email is in Korcan (or Chinese or Cyrillic) is given below as one example of a complicated task and the corresponding heuristic recipe made possible in procmail filtering. The original idea owes to a WWW page (no longer available at the original address) by Walter Dnes. The demonstration below also includes getting the sender's postmaster's address. The example is an cxtract from Salmi (1999).

```
# Get the sender's address, ignore Reply-To:
FROM_=`formail -c -I"Reply-To:" -rt -xTo: \
  | expand | sed -e 's/^[ ]*//g' -e 's/[ ]*$//g'`

# Get the sender's host
FHOST_=`echo "${FROM_}" | awk -F@ '{ print $2 }'`

# Your path to sendmail
SENDMAIL="/usr/lib/sendmail"

# Reject probable Korean, Chinese or Cyrillic email using character scoring
:0
* ! ^X-Loop:.*myid@myhost\.mydom
* ! $ ? echo ${FHOST_} | fgrep -is 'myhost.mydom'
* $ ? echo ${FHOST_} | fgrep -is '.'
{
  :0BD
  *  -1^1 .
  *   2^1 =[0-9A-F][0-9A-F]
  *  20^1 [¡¢£¥_§¨©ª«¬_®¯°±__´µ¶·._º»___¿]
  *  20^1 [ÀÁÂÃÄÅÆÇÈÉÊËÌÍÎÏ_ÑÒÓÔÕÖ_ØÙÚÛÜ__ß]
  *  20^1 [àáâãäåæçèéêëìíîï_ñòóôõö÷øùúûü__ÿ]
  *  20^1 =[89A-F][0-9A-F]
  * -20^1 [åÅäÄöÖàáâçèéêë]
  * -20^1 =(E5|C5|E4|C4|F6|D6|E0|E1|E2|E7|E8|E9|EA|EB)
  {
    :0
    { RULE="Probable Korean email" }
    #
    :0c:${HOME}/procmail.lock
    | expand | sed -e 's/[ ]*$//g' \
      | sed -e 's/^/ /' > ${HOME}/procmail.reject.korean
    #
    :0:${HOME}/procmail.lock
    | (formail -r -I"Subject: Autorejected email" \
      -I"To: ${FROM_}" \
      -I"Cc: postmaster@${FHOST_}" \
      -A"X-Loop: myid@myhost.mydom" ; \
      echo "--- begin rejected probable Korean email -- " ; \
      echo "" ; \
      cat ${HOME}/procmail.reject.korean ; \
      echo "--- end of rejected probable Korean email ---" ; \
      rm -f ${HOME}/procmail.reject.korean) \
        | ${SENDMAIL}
  }
}
```

Appendix 3: Identifying a virus/worm in the email

Consider as an example avoiding the Sobig.F virus/worm which is spread by email. In terms of the techniques presented in the current paper, Sobig.F email is blacklisted. This virus and its consequences caused a then unprecedented load on the Internet in the last third of August 2003. In a way such email is easier to detect and avoid than the ordinary spam, since there is a high regularity to the pattern. In Sobig.F email the subject always is one of a limited selection, there always is an unvarying "X-MailScanner:" header, and the body of the message always contains text from just two alternatives. These facts enable a simple, efficient detection recipe with a reasonably low probability of false positives. The example is from the author's own procmail configuration file. The "RULE" variable in the recipe is for helping to separately record which of the many recipes in the configuration file has been enacted.

```
# Discard Sobig.F
# Look at the headers
:0H
* ^Subject.*(\
Re: Thank you\!|\
Thank you\!|\
Your details|\
Re: Details|\
Re: Re: My details|\
Re: Approved|\
Re: Your application|\
Re: Wicked screensaver|\
Re: That movie)
* ^X-MailScanner: Found to be clean
{
  # Look at the message body
  :0B
  * (Please)? see the attached file for details
  {
    :0
    { RULE="Sobig.F" }
    :0
    /dev/null
  }
}
```

Maximum likelihood vs. method of moment when estimating the non-integer degrees of freedom of student t-distribution

Juuso Töyli and Antti Kanto

Dedicated to Ilkka Virtanen on the occasion of his 60th birthday

Abstract

Töyli, Juuso and Antti Kanto (2004). Maximum likelihood vs. method of moment when estimating the non-integer degrees of freedom of student t-distribution. In *Contributions to Management Science, Mathematics and Modelling. Essays in Honour of Professor Ilkka Virtanen*. Acta Wasaensia No. 122, 245–251. Eds Matti Laaksonen and Seppo Pynnönen.

Here we will show that the method of moment estimation will produce positively biased estimates for the degrees of freedom parameter of Student t-distribution. With the help of a simulation study, we illustrate that maximum likelihood method produces non-biased estimates when sample size is large and with smaller sample sizes the bias is substantially smaller than in case of the method of moments. The bias of the method of moments also decreases when the distribution approaches the normal distribution.

Juuso Töyli, Senior Researcher, Laboratory of Computational Engineering, Helsinki University of Technology, P.O. Box 9400, FIN-02015 HUT, Finland and Researcher International Business, Turku School of Economics and Business Administration, Rehtorinpellonkatu 3, FIN-20500 Turku, Finland, e-mail juuso.toyli@hut.fi.
Antti Kanto, Professor, Department of Economics and Management Science, Helsinki School of Economics and Business Administration, P.O. Box 1210, FIN-00101 Helsinki, Finland, e-mail antti.kanto@hkkk.fi.

1. Introduction

The Student t-distribution is well-known in statistics and it has numerous applications. It is especially useful model in financial data that shows fat tails compared to normal distribution (see Blattberg and Gonedes 1974) and in particular for the applications of Value-at-Risk with high probability levels (Heikkinen and Kanto 2002; see also Pant and Chang 2001; McNeil 1999; McNeil and Frey 2000). Originally, it was Blattberg and Gonedes

(1974) who introduced the Student t-distribution as a model for asset returns. Since then, this model has become a standard benchmark for new and more complex models (see, e.g., Kon 1984). The Student t-distribution seems to be rather attractive model since it is able to capture the excess kurtosis observed, it is easy and fast to estimate, and its mathematical properties are well known.

The Student t-distribution can be obtained from a normal distribution whose variance is random variable. In contrast to the symmetric stable distribution, the variance is now drawn from inverted gamma distribution (Hsieh 1991; Blattberg and Gonedes 1974). This specification is in line with the observation that the variances of financial time series might be non-stationary. Traditionally, textbooks deal with a situation, where the degrees of freedom parameter v is a positive integer although there is no mathematical reason to assume that. Thus, when comparing different models, it has been common practise to maximum likelihood fit Student t-distribution and allow v to have non-integer values (see, e.g., Töyli et al. 2002; Blattberg and Gonedes 1974; Tucker 1992). Due to Heikkinen and Kanto (2002) tables for Student t-distribution's percentiles with non-integer degrees of freedom are also available.

In Value-at-Risk analysis, computational issues are often important because of huge amount of data to be analysed. It is also know that the variance of the asset return distribution is finite but time-dependent in a complex non-linear manner (Perry 1983). If Student t-distribution's parameters are estimated with the help of method of moments, the computation is fast and different kinds of exponential weighting schemes for variance and kurtosis are easily applicable. Maximum likelihood method in turn is computationally more demanding although the fitting is very fast. When the model gets more complicated, the likelihood function becomes more difficult to handle and the fittings take considerably longer. The purpose of this paper is to compare maximum likelihood method and method moments when estimating the parameters of the Student t-distribution.

Our results indicate that the method of moment estimation produces positively biased estimates for the degrees of freedom parameter of Student t-distribution. The bias of the method of moments decreases when the distribution approaches the normal distribution. Since the bias seems to behave rather systematically, it might well be possible to define a correction multiplier. In the next section we will review the Student t-distribution and its estimation in detail. We also review the simulation algorithm we used. After that, we discuss the results and then shortly summarise our main findings.

2. Methods

The density of non-central Student t-distribution equals

$$f(x)=\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu\beta}}\left(1+\frac{(x-\mu)^2}{\beta\nu}\right)^{-(1+\nu)/2}$$

(Abramowitz and Stegun 1991) and the distribution function

$$F(x)=\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu\beta}}\int_{-\infty}^{x}\left(1+\frac{(u-\mu)^2}{\beta\nu}\right)^{-(1+\nu)/2}du$$

where μ is location parameter, β is dispersion parameter and ν is the shape parameter. Now assuming that $\nu > 4$, the distribution has a finite absolute moment of order $k > 4$ and it can be shown (Heikkinen and Kanto 2002) that:

$$\hat{\nu}=4+\frac{6}{kur}$$

where *kur* stands for excess kurtosis calculated from the sample.

In order to compare the maximum likelihood method and method of moments, we constructed a simulation study, where zero mean unit variance Student t-distributed random number were generated with different values for the degrees of freedom parameter (ν). We chose ν = (4.5, 5, 6, 8, 10 , 15). In the maximum likelihood estimation, we used the same numerical algorithm as earlier (see Töyli et al. 2002). It needs to be pointed out that this algorithm was originally designed and used with financial data where the fat tailed distributions are of interest. Thus, with large values of degrees of freedom parameter (say $\nu > 15$) - i.e., the shape of the Student t-distribution is almost indistinguishable from normal distribution - the numerical precision of the algorithm is not very good. We run the simulation with three different sample sizes n = (1 000, 10 000, 100 000) and each simulation included 10 000 repetitions.

The algorithm was as follow (repeat 10 000 times for each ν):

1. Draw a sample of n random numbers form Student t-distribution with degrees of freedom parameter ν
2. Calculate the method of moment estimate for ν
3. Calculate maximum likelihood estimate for ν

In the next section we will discuss the results of these simulations.

3. Results

Figure 1 shows histograms for the degrees of freedom parameters extracted from the simulations (true value ν = 4.5). The left column reports the maximum likelihood estimates and the right column the method of moment estimates. The upper row show a simulation where the sample size was set to n = 10 000 and the lower row n = 100 000. The data indicates that the maximum likelihood estimates have smaller variance and they seem to be rather unbiased. In contrast, method of moments results in larger variance and positively biased estimates.

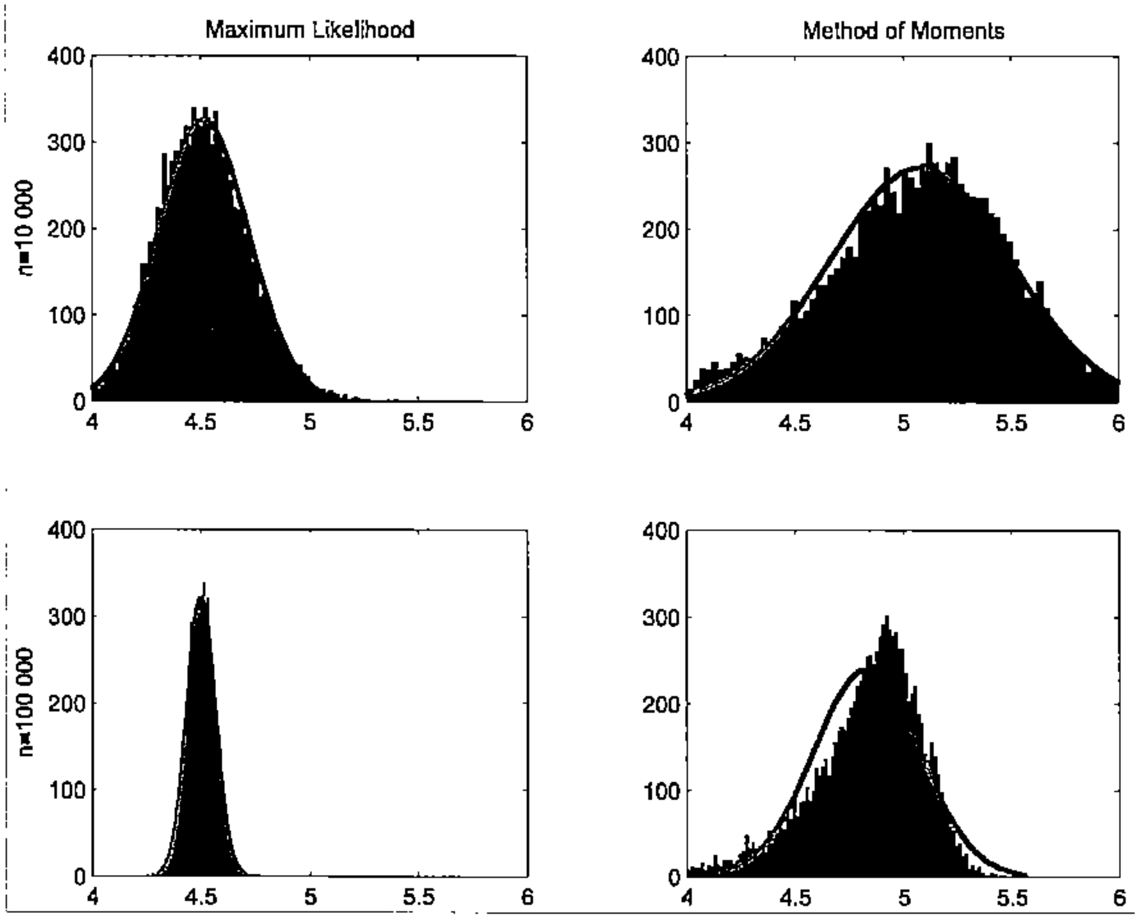


Figure 1. The degrees of freedom parameter ν extracted from the simulations ($\nu = 4.5$).

The medians of parameter ν extracted from each simulation are given in Table 1. We report the medians here because the means are not very informative because of large standard errors. It may happen that every now then the drawing distribution is close to normal distribution (i.e. $\nu \to \infty$) and these few very large values increase the variance. The data in Table 1 clearly implies that the maximum likelihood method results in virtually unbiased estimates even with small (n = 1 000) sample sizes. The method of moment estimation produces positively biased estimates for the degrees of freedom parameter of Student t-distribution when v is small. The sample size has also more visible effect on the bias than in case of the maximum likelihood method. The bias of the method of moments decreases when the distribution approaches the normal distribution and when $v > 6$, and when the sample size is large the estimates are almost unbiased. Nevertheless, the bias seems to behave in a systematic manner such that it might well be possible to define a correction multiplier.

Table 1. Medians.

v	Maximum Likelihood Method			Method of Moments		
4,5	4,53	4,51	4,50	5,60	5,10	4,87
5	5,03	5,01	5,00	6,03	5,49	5,27
6	6,08	6,00	6,00	7,00	6,37	6,16
8	8,09	8,02	8,00	9,05	8,29	8,08
10	10,23	10,03	10,00	11,11	10,26	10,05
15	15,62	15,08	15,01	16,39	15,30	15,06
n=	**1 000**	**10 000**	**100 000**	**1 000**	**10 000**	**100 000**

4. Discussion

Here we have considered the maximum likelihood vs. the method of moments estimation of Student t-distribution's degrees of freedom parameter. We allowed the parameter to receive also non-integer values and our special interest was on fat tailed distributions (i.e., small values of degrees of freedom parameter). The Student t-distribution is widely used in statistics and statistical applications, particularly in finance. Especially useful it is in value-at-risk analysis.

With the help of a simulation study, we illustrated that the method of moments estimation will produce positively biased estimates for the degrees of freedom parameter of Student t-distribution when small values are of interest. The bias decreases when sample size grows as well as when the distribution approaches the normal distribution. Nevertheless, we concluded that the bias seems to behave such systematically that it might well be possible to define a correction multiplier. The results also imply that, even when using biased estimates, the Student t-distribution is likely to outperform the traditional model of normal distribution in value-at-risk analysis.

References

Abramowitz, M. & I.A. Stegun (1971). *Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables*. New York: Dover Publications, Inc.

Blattberg, R. & N. Gonedes (1974). A comparison of stable and student distributions as statistical models for stock prices. *Journal of Business* 47, 244–280.

Heikkinen, V-P. & A. Kanto (2002). Value-at-risk estimation using non-integer degrees of freedom of Student's distribution. *The Journal of Risk* 4:4, 77–84.

Hsieh, D. A. (1991). Chaos and nonlinear dynamics: Application to financial markets. *The Journal of Finance* 46:5, 1839–1878.

Kon, S. J. (1984). Models of stock returns – a comparison. *The Journal of Finance* 39:1, 147–165.

McNeil, A. & R. Frey (2000). Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach. *Journal of Empirical Finance* 7, 271–300.

McNeil, A. (1999). Extreme value theory for risk managers. In *Internal Modeling and CADII*, 93–113. London: Risk Books.

Pant, V. & W. Chang (2001). An empirical comparison of methods for incorporating fat tails into value-at-risk models. *Journal of Risk* 3:3, 99–119.

Perry, P.R. (1983). More evidence on the nature of the Distribution of Security Returns. *Journal of Financial and Quantitative Analysis* 18:2, 211–221.

Töyli, J., K. Kaski & A. Kanto (2002). On the shape of asset return distribution. *Communications in Statistics–Simulation and Computation* 31:4, 489–521.

Tucker, A. L. (1992). A reexamination of finite- and infinite-variance distributions of daily stock returns. *Journal of Business and Economic Statistics* 10:1, 73–81.

ACTA WASAENSIA

Time inversion result for self-similar diffusions on finite number of rays

Juha Vuolle-Apiala

Dedicated to Ilkka Virtanen on the occasion of his 60th birthday

Abstract

Vuolle-Apiala, Juha (2004). Time inversion result for self-similar diffusions on finite number of rays. In *Contributions to Management Science, Mathematics and Modelling. Essays in Honour of Professor Ilkka Virtanen.* Acta Wasaensia No. 122, 253–260. Eds Matti Laaksonen and Seppo Pynnönen.

The result of T. Shiga and S. Watanabe, that if (R_t,Q^0) is a Bessel diffusion on $[0, \infty)$, starting at 0, then (R_t) and $(tR_{1/t})$ are equivalent diffusions under Q^0 is shown to be valid if the state space, instead of $[0, \infty)$, consists of n rays which meet at 0. The generalization for α-self-similar diffusions, $\alpha>0$, is also given.

Juha Vuolle-Apiala, Department of Mathematics and Statistics, University of Vaasa, P.O. Box 700, FIN–65101 Vaasa, Finland, e-mail jmva@uwasa.fi.

Key words: time inversion, self-similar, Bessel diffusion, AMS Mathematics Subject Classification (2000): 60 G 18, 60 J 60, 60 J 25.

0. Introduction

In Shiga and Watanabe (1973) the authors proved that if (R_t, Q^0) is a Bessel diffusion (including the reflecting Brownian motion) on $[0, \infty)$ starting at 0 then

(1) (R_t) and $(tR_{1/t})$ are equivalent diffusions under Q^0.

The respective result for the standard Brownian motion on R^d, $d \geq 1$, is well-known. (1) was in Graversen and Vuolle-Apiala (2000) showed to be valid for symmetrized Bessel processes on the whole real line R. In Vuolle-Apiala (2002) the corresponding problem in the case of rotation invariant α-self-similar diffusions on R^d having 0 as a polar set was studied.

In this note we will generalize the result of Graversen and Vuolle-Apiala (2000) to the case where the state space, instead of the real axis, is $F \cup \{0\}$, F consisting of n rays l_1, l_2, ... , l_n on $R^2\backslash\{0\}$, meeting at 0. Strong Markov processes on $F \cup \{0\}$ were studied in Vuolle-Apiala (2001); see also Barlow, Pitman and Yor (1989) and Walsh (1978).

Suppose (Y_t, P^0) is a diffusion process on $F \cup \{0\}$, starting at 0, such that (Y_t) on $l_k \cup \{0\}$, k=1,2, .. , n, behaves like a given Bessel diffusion (R_t) on $[0, \infty)$. Then (Y_t, P_0) is a $\frac{1}{2}$- self-similar diffusion on $F \cup \{0\}$, that is (see Lamperti 1972),

(2) (Y_t, P^0) and $(a^{-1/2}Y_{at}, P^0)$ are equivalent diffusions for all a>0.

Remark 0: (2) is, of course, also valid for (R_t, Q^0).

Remark 1: A well-known example is the Walsh's Brownian motion, first introduced by J. B. Walsh (1978). Let (Y_t, P^0) be a diffusion process on $F \cup \{0\}$, starting at 0, F consisting of n rays l_1, ... , l_n on $R^2\backslash\{0\}$ which meet at 0. (Y_t) starts at 0 and right away it chooses one of the rays such that the probability for the ray l_k to be chosen is equal to $\alpha_k \geq 0$, $\sum_{k=1}^{n} \alpha_k = 1$. Then it moves like the standard Brownian motion on the chosen ray until it again hits 0. Thereafter it immediately reflects from 0 to one of the rays l_k, with respective probabilities α_k, k=1,2, ... , n. Again it behaves like the standard Brownian motion on the chosen new line until it again hits 0 and reflects immediately from 0 to one of the n rays, randomly. This is repeated forever. If n=2 we have a special case called Skew Brownian motion (see also Barlow, Pitman and Yor 1989 and Vuolle-Apiala 2001).

Remark 2: In general, we have for some non-negative numbers $\alpha_1, \alpha_2, \ldots, \alpha_n$,

$\sum_{k=1}^{n}\alpha_k = 1$, $P^0\{Y_k \in l_k\} = \alpha_k$, k=1,2, ... ,n, for some t>0 and thus, because of (2), for all t>0. (Y_t) is symmetric iff $\alpha_k = \frac{1}{n}$ for all k=1,2, ... ,n.

Conversely, starting from a given Bessel diffusion (R_t, Q^0) and any set of non-negative numbers $\alpha_1, \alpha_2, \ldots, \alpha_n$ for which $\sum_{k=1}^{n}\alpha_k = 1$, one can construct a $\frac{1}{2}$-self-similar diffusion (Y_t, P^0) having $F \cup \{0\}$ as a state space such that $P^0\{Y_t \in l_k\} = \alpha_k$, k=1,2, ... ,n, for all t>0 (see more about the behaviour at 0 in Vuolle-Apiala 2001) and (Y_t) on $l_k \cup \{0\}$ behaves like (R_t) on $[0, \infty)$.

Remark 3: Obviously, the processes described in Remark 2 form precisely the class of $\frac{1}{2}$-self-similar diffusions on $F \cup \{0\}$ which have the given Bessel diffusion as the radial part (see Revuz and Yor 1991, Ex. (2.16), p. 449).

We will in this note show that if (Y_t, P^0) is a $\frac{1}{2}$- self-similar diffusion on $F \cup \{0\}$ having a given Bessel diffusion (R_t, Q^0) on $[0, \infty)$ as the radial part, then (1) is valid for (Y_t), that is, (Y_t) and $(tY_{1/t})$ are equivalent diffusions under P^0. The proof is analogeous to that in Graversen and Vuolle-Apiala (2000).

An obvious generalization to α-self-similar diffusions on $F \cup \{0\}$ is the following:

(3) If (Y_t, P^0) is is an α-self-similar diffusion on $F \cup \{0\}$, starting at 0, then (Y_t) and $(t^{2\alpha} Y_{1/t})$ are equivalent diffusions under P^0.

By an α-self-similar diffusion we mean (see Lamperti 1972) that

(4) (Y_t, P^0) and $(a^{-\alpha}Y_{at}, P^0)$ are equivalent diffusions for all a>0.

1. The main result

Theorem 1: Let (Y_t, P^0) be an α-self-similar diffusion starting at 0 with an infinite lifetime such that the state space is $F \cup \{0\}$, where F consists of n rays $l_1, l_2, \ldots, l_n$ on $R^2\backslash\{0\}$ meeting at 0, the corresponding angular coordinate on each ray l_k is $\theta_k \in [0, 2\pi)$, k=1, 2, ... , n. Assume further that (Y_t, P^0) has (R_t, Q^0) as the radial part, where (R_t, Q^0) is a given α-self-similar diffusion on $[0, \infty)$. Then (Y_t) and $(t^{2\alpha}Y_{1/t})$ are equivalent diffusions under P^0.

Remark 4: Processes described in Th.1 are often called Walsh's Brownian motion-type of diffusions (see Vuolle-Apiala 2001). They form precisely the class of all diffusions on $F \cup \{0\}$ starting at 0 and having (R_t) as the radial part (see Revuz and Yor 1991, Ex. (2.16), p. 449).

The proof is analogeous to the proof in Graversen and Vuolle-Apiala (2000) for symmetric α-self-similar diffusions on the whole real line R:

Proof: If T_0 is the first hitting time to 0 for (Y_t, P^0) then, according to Lamperti (1972), either $T_0<\infty$ a.s. or $T_0=\infty$ a.s., with respect to P^0. If $T_0=\infty$ a.s. then Y_t moves all the time on one of the rays l_k, k=1,2, ... ,n and thus the situation is the same as in Shiga and Watanabe (1973) and Watanabe (1975). So we only need to consider the case $T_0<\infty$ a.s..

We will first show that the d-dimensional distributions of (Y_t) and $(t^{2\alpha}Y_{1/t})$ are equal for all d=1,2, ... Assume for simplicity that d=2, $\alpha = \frac{1}{2}$, the general case is analogeous. Denote $Y_t = [R_t, \phi_t]$, where R_t and ϕ_t are the radial and the angular parts of Y_t, respectively. Similarly as in Graversen and Vuolle-Apiala (2000) we have for s<t, for Borel subsets of $[0, \infty)$ I_1, I_2 and for $\theta_i, \theta_j \in \{\theta_1, \theta_2, \ldots, \theta_n\}$

$P^0\{R_s \in I_1, R_t \in I_2, \phi_s=\theta_i, \phi_t=\theta_j\}$ =
$P^0\{R_s \in I_1, R_t \in I_2, \phi_s=\theta_i, \phi_t=\theta_j;$ R_s and R_t belong to the same excursion of $(R_u)\}$+
$P^0\{R_s \in I_1, R_t \in I_2, \phi_s=\theta_i, \phi_t=\theta_j;$ R_s and R_t belong to different excursions of $(R_u)\}$ =

$$P^0\{R_s \in I_1, R_t \in I_2, \phi_s=\theta_i, \phi_t=\theta_j; R_u \neq 0 \text{ for all } u \in (s,t)\} +$$
$$P^0\{R_s \in I_1, R_t \in I_2, \phi_s=\theta_i, \phi_t=\theta_j; R_u = 0 \text{ for some } u \in (s,t)\} =$$

Denote $\alpha_i=P^0\{Y_t \in I_i\}$ for some, and thus because of self-similarity, for all t>0, i=1,2, ...,n. Then $\alpha_1+\alpha_2+...+\alpha_n=1$.

Denote further $m_{ij}=P^0\{\phi_s=\theta_i, \phi_t=\theta_j|R_s \in I_1, R_t \in I_2, R_u \neq 0 \text{ for all } u \in (s,t)\}$. Obviously, m_{ij} does not depend on s or t. Then

$$m_{ij}=\alpha_i \text{ if } i=j \text{ and } =0 \text{ if } i \neq j.$$

Because the excursions from 0 are independent, we also have

$$P^0\{\phi_s=\theta_i, \phi_t=\theta_j|R_s \in I_1, R_t \in I_2, R_u = 0 \text{ for some } u \in (s,t)\} = \alpha_i\alpha_j$$
for all $i,j = 1,2, ... ,n$, $\forall s,t>0$.

Thus

$$P^0\{R_s \in I_1, R_t \in I_2, \phi_s=\theta_i, \phi_t=\theta_j\} =$$
$$m_{ij}Q^0\{R_s \in I_1, R_t \in I_2, R_u \neq 0 \text{ for all } u \in (s,t)\} + \alpha_i\alpha_jQ^0\{R_s \in I_1, R_t \in I_2, R_u = 0 \text{ for some } u \in (s,t)\}.$$

Because the time-inversion property is valid for the radial process (R_t, Q^0) (see Shiga and Watanabe 1973 and Watanabe 1975), we can similarly as in Graversen and Vuolle-Apiala (2000) conclude that this is equal to

$$m_{ij}Q^0\{sR_{1/s} \in I_1, tR_{1/t} \in I_2, uR_{1/u} \neq 0 \text{ for all } u \in (s,t)\} +$$
$$\alpha_i\alpha_jQ^0\{sR_{1/s} \in I_1, tR_{1/t} \in I_2, uR_{1/u} = 0 \text{ for some } u \in (s,t)\} =$$
$$m_{ij}Q^0\{R_{1/s} \in \frac{1}{s}I_1, R_{1/t} \in \frac{1}{t}I_2, R_u \neq 0 \text{ for all } u \in (\frac{1}{t},\frac{1}{s})\} +$$
$$\alpha_i\alpha_jQ^0\{ R_{1/s} \in \frac{1}{s}I_1, R_{1/t} \in \frac{1}{t}I_2, R_u = 0 \text{ for some } u \in (\frac{1}{t},\frac{1}{s})\}.$$

Reversing now the above procedure (see Graversen and Vuolle-Apiala 2000: 71) we get this equal to

$$P^0\{\phi_{1/s}=\theta_i,\phi_{1/t}=\theta_j|R_{1/s}\in \frac{1}{s}I_1,R_{1/t}\in \frac{1}{t}I_2,R_u\neq 0 \text{ for all } u\in (\frac{1}{t},\frac{1}{s})\}Q^0\{R_{1/s}\in$$

$$\frac{1}{s}I_1,R_{1/t}\in \frac{1}{t}I_2,R_u\neq 0 \text{ for all } u\in (\frac{1}{t},\frac{1}{s})\}+$$

$$P^0\{\phi_{1/s}=\theta_i,\phi_{1/t}=\theta_j|R_{1/s}\in \frac{1}{s}I_1,R_{1/t}\in \frac{1}{t}I_2,R_u=0 \text{ for some } u\in (\frac{1}{t},\frac{1}{s})\}Q^0\{R_{1/s}\in$$

$$\frac{1}{s}I_1,R_{1/t}\in \frac{1}{t}I_2,R_u=0 \text{ for some } u\in (\frac{1}{t},\frac{1}{s})\} =$$

$$P^0\{R_{1/s}\in \frac{1}{s}I_1,R_{1/t}\in \frac{1}{t}I_2,\ \phi_{1/s}=\theta_i,\phi_{1/t}=\theta_j;\ R_u\neq 0 \text{ for all } u\in (\frac{1}{t},\frac{1}{s})\} +$$

$$P^0\{R_{1/s}\in \frac{1}{s}I_1,R_{1/t}\in \frac{1}{t}I_2,\ \phi_{1/s}=\theta_i,\phi_{1/t}=\theta_j;\ R_u=0 \text{ for some } u\in (\frac{1}{t},\frac{1}{s})\} =$$

$$P^0\{R_{1/s}\in \frac{1}{s}I_1,R_{1/t}\in \frac{1}{t}I_2,\ \phi_{1/s}=\theta_i,\phi_{1/t}=\theta_j;\ R_{1/s} \text{ and } R_{1/t} \text{ belong to the same excursion}$$

$$\text{of } (R_u)\}+P^0\{R_{1/s}\in \frac{1}{s}I_1,R_{1/t}\in \frac{1}{t}I_2,\ \phi_{1/s}=\theta_i,\phi_{1/t}=\theta_j;\ R_{1/s} \text{ and } R_{1/t} \text{ belong to different}$$

$$\text{excursions of } (R_u)\} = P^0\{sR_{1/s}\in I_1,\ tR_{1/t}\in I_2,\ \phi_{1/s}=\theta_i,\phi_{1/t}=\theta_j\}.$$

Thus the d-dimensional distributions of (Y_t) and $(t^{2\alpha}Y_{1/t})$ under P^0 are equal for d=1,2, ...

It remains to show that $(t^{2\alpha}Y_{1/t})$ is a diffusion under P^0, that is, strong Markov with continuous paths. We need the following Lemma:

Lemma: (Y_t, P^0) is a Feller process.

Proof of the Lemma: Follows from the continuity of the paths $t \to Y_t(\omega)$, the self-similarity property and the strong Markov property (see also Lamperti 1972). □

Our arguments are similar as in Graversen and Vuolle-Apiala (2000). More precisely (assume again for simlicity that $\alpha=\frac{1}{2}$):

The continuity of the paths: It only suffices to show that

$$P^0\{\lim_{t\to 0} tY_{1/t} = 0\} = 1.$$

But this is a consequence of the corresponding property of the radial process (R_t)

The strong Markov property: Define $Z_t = tY_{1/t}$ if $t>0$ and $Z_0=0$. Let $(P_t(\ ,\ ))_{t\geq 0}$ be the transition function corresponding to $(Y_t)_{t\geq 0}$. According to Lemma, $(P_t(\ ,\ ))$ is a Feller transition function and (Y_t) is a Feller process. Now for a bounded measurable function f: $F \cup \{0\} \to R$ we have

$$E^0\{f(Z_{t+s})|\sigma(Z_s,s \leq t)\} = E^0\{f(Y_{t+s})|\sigma(Y_s,s \leq t)\}$$

because (Z_u) and (Y_u), as showed above, have the same finite dimensional distributions under P^0. Using the (ordinary) Markov property of (Y_u) we get the right-hand side equal to $P_sf(Y_t)$ which is further equal to $P_sf(Z_t)$ a.s. P^0.

Thus we can conclude

$$E^0\{f(Z_{t+s})|\sigma(Z_s,s \leq t)\} = P_sf(Z_t) \text{ a.s. } P^0,$$

that is, (Z_t) is a Markov process with (P_t) as its transition function. Because (P_t) is a Feller transition function, (Z_t) is a Feller process and thus strong Markov. □

References

Barlow, M.T., J.W. Pitman & M. Yor (1989). On Walsh's Brownian motions. *Sém. Prob. XXIII, Lecture Notes in Mathematics*, Vol. 1372, 275–293. Berlin, Heidelberg, New York: Springer.

Graversen, S.E. & J. Vuolle-Apiala (2000). On Paul Lévy's arc sine law and Shiga-Watanabe's time inversion result. *Probability and Mathematical Statistics* 20:1, 63–73.

Lamperti, J.W. (1972). Semi-stable Markov processes I. *Z. Wahrschein. verw. Gebiete* 22, 205–225.

Revuz, D. & M. Yor (1991). *Continuous Martingales and Brownian Motion*. Springer.

Shiga, T. & S. Watanabe (1973). Bessel diffusions as one-parameter family of diffusion processes. *Z. Wahrschein. verw. Gebiete* 27, 37–46.

Vuolle-Apiala, J. (2001). Walsh's Brownian motion-type of extensions. *Journal of Theoretical Probability* 14:1, 115–124.

Vuolle-Apiala, J. (2002). Shiga-Watanabe's time inversion property for self-similar diffusion processes. Working Papers of the University of Vaasa, Department of Mathematics and Statistics, December 2002.

Walsh, J.B. (1978). A diffusion with a discontinuous local time. *Temps Loaux, Asterisque* 52 and 53, 37–45.

Watanabe, S. (1975). On time inversion of one-dimensional diffusion processes. *Wahrschein. verw. Gebiete* 31, 115–124.