BoF Online

2007 • No. 20

Kyselyaineiston tilastollisesta laadunvalvonnasta (SPC), maksutaseaineiston pilottikokeita

Heikki Hella

Tony Härkönen

Maija Salmela

Tässä julkaisussa esitetyt mielipiteet ovat kirjoittajan omia eivätkä välttämättä edusta Suomen Pankin kantaa.

SUOMEN PANKKI FINLANDS BANK

EUROJÄRJESTELMÄ
EUROSYSTEMET

Suomen Pankki

Rahoitusmarkkina- ja tilasto-osasto

31.12.2007

Sisällys

Tiivistelmä 4

Abstract 5

1 Johdanto 6

1.1 Mitä on tilastollinen laadunvalvonta? 6

1.2 Tämän pilottiraportin tavoitteet 6

2 Kyselyaineistot 8

2.1 Kyselylomake tiedonkeruutekniikkana 8

2.2 Kyselylomakkeen virheistä 8

2.3 Havaintoaineistot 9

3 Aineistojen rakenteen eksploratiiviset tarkastelut 12

3.1 Kuvailevat tilastolliset tunnusluvut 12

3.2 Graafiset esitykset 12

3.3 Autokorrelaatio(AC)- ja osittaisautokorrelaatio(PAC)-estimaatit 16

4 Kokeilussa tehtyjä tilastollisen laadunvalvonnan peruslaskelmia 18

5 Havaintojen tarkistus ja käsittely 20

5.1 Evaluoinnin kynnystarkastelut: poikittais- ja pitkittäisinfo 20

5.2 Poikkeavat havainnot 22

6 Tulokset 23

6.1 SV-kysely 23

6.2 SVA-kysely 25

7 Yhteenveto ja jatkoselvityksiä 27

7.1 Yhteenveto 27

7.2 Jatkoselvityskohteita 28

Kirjallisuutta 29

Liite 1. Taulukot 31

Liite 2. Kuviot 34

Liite 3. Laskelmissa käytetyt toimisto- ja analyysiohjelmat 37

BoF Online

Päätoimittaja

Jouko Marttila

ISSN

1796-9123 (online)

Postiosoite
PL 160
00101 HELSINKI

Käyntiosoite Snellmaninaukio
Puhelin 010 8311
Faksi (09) 174 872

Sähköposti
etunimi.sukunimi@bof.fi
www.suomenpankki.fi

Swift SPFB FI HH
Y-tunnus 0202248-1
Kotipaikka Helsinki

Kuvioluettelo

Kuvio 1. SV-aineiston muuttujien (havaintomatriisi) sarakesummat prosenttiosuuksineen suuruusjärjestyksessä 2006M12 kantatietojen perusteella 13
Kuvio 2. Box-plot -kuviot kysymyksen ulkomaiset arvopaperisaamiset muuttujista 14
Kuvio 3. Aikasarja-aineiston pylväskuvioita, joihin on väritetty eritasoisen hälytyksen saaneet havainnot (tumma harmaa: alempi taso, musta: ylempi taso). 15
Kuvio 4. Aikasarja-aineistosta differensoidut sarjat hälytysrajoinaan mediaani + standardipoikkeama (viiva) ja mediaani + 2*standardipoikkeama (katkoviiva). 16
Kuvio 5. Aikasarja-aineiston perustunnusluvut ja histogrammi erään vastaajan arvopaperisaamisten joukkolainat-muuttujalle 19
Kuvio 6. SV-aineiston (2006M12) lukujen jakautuminen vastaajakohtaisten rivisummien mukaisessa suuruusjärjestyksessä 25
Kuvio 7. Aikasarja-aineiston normaalisuutta kuvaavat qq-plot -kuviot eräiden muutossarjojen tapauksissa 34
Kuvio 8. Vastaajan 3 autokorrelaatio- ja osittaisautokorrelaatioestimaatit 35
Kuvio 9. Vastaajan 3 aikasarja sekä sen eksponentiaalisesti ja robustilla filtterillä 4253H tasoitettu sarja residuaaleineen mukaan lukien eksponentiaalisen tasoituksen ennusteita 36

Taulukkoluettelo

Taulukko 1. Tarkastelussa olevat ulkomaisten saamisten ja velkojen kysymykset (rasteroituna) muuttujineen 11
Taulukko 2. Jarque-Bera -testin tulokset arvopaperisaamisten joukkolainat-muuttujalle 22
Taulukko 3. SV- ja SVA-pilottiaineistolle laskettuja tunnuslukuja 2006M12 (SV) ja 2006 (SVA) kantatietojen perusteella 23
Taulukko 4. Kolmen eritasoisen ylähälytysrajan antamien hälytysten lukumäärä joukkolainat-muuttujalle 24
Taulukko 5. Vastausten lukumääriä muuttujittain ja vastaajittain 24
Taulukko 6. Matriisitaulukko SV-poikkileikkausaineistosta (näkymä ulkomaisten arvopaperisaamisten osalta) 2006M12 kantatietojen perusteella 31
Taulukko 7. Aikasarja-aineiston autokorrelaatiot ja osittaisautokorrelaatiot joukkolainat-muuttujalle 32
Taulukko 8. Kysymyskohtaisia tunnuslukuja kyselyaineistoille, 2006M12 32
Taulukko 9. Muuttujakohtaiset vastausten lukumäärät, 2006M12 33

Tiivistelmä

Tässä pilottiraportissa tarkastelujen kohteeksi valittiin osia maksutaseen ulkomaisten saamisten ja velkojen kuukausikyselyn (SV) ja vuosikyselyn (SVA) vastausaineistoista. SV-aineisto sisälsi kuukausittaiset kyselyn vastaukset aikaväliltä 2003–2006 ja SVA-aineisto vastaavasti vuosittaiset kyselyn vastaukset vuosilta 2005 ja 2006. Tarkemmin tarkasteltiin ulkomaisten arvopaperisaamisten muuttujia: osakkeet, rahasto-osuudet, joukkolainat ja rahamarkkinapaperit. Tavoitteena oli ensinnäkin sovittaa erilaisia menetelmiä osoittamaan poikkeamia valituissa poikkileikkaus- ja aikasarja-aineistoissa. Toisena tavoitteena oli estimoida aineistoista tilastollisia tunnuslukuja, joita voitaisiin hyödyntää kynnysrajojen ja -sääntöjen muodossa aineiston tilastollisen laadunvalvonnan välineenä. Lisäksi sivutuotteena saatiin tietoa SV- ja SVA-kyselyaineistojen valittujen osien rakenteesta tarkasteluajalta.

Tarkastelut osoittivat, että ainakin valitussa aineistossa mediaania tai keskiarvoa ja standardipoikkeamaa voidaan käyttää kynnyssääntöjen luomisessa. Aineiston perusteella on suositeltavaa käyttää sekä klassisia että robusteja versioita tilastollisista menetelmistä ja myös ei-parametrisia menetelmiä tulisi käyttää, koska aineistojen tilastollisesta jakaumasta ei ole tietoa. Systemaattista otantaa voidaan hyödyntää, mikäli käytettävissä on riittävästi havaintoja. Lisäksi nähtiin, että aineiston rakenteesta saadaan nopeasti ja selkeästi tietoa graafisten esitysten, erityisesti box-plot - ja qq-plot -kuvioiden, avulla. Tarkastelut osoittivat myös, että jako suuriin ja muuhun osaan vastaajista on varsin selvä, ja että on tärkeää tarkastella sekä aikasarja- että poikkileikkausaineistoa niiden täydentäessä toisiaan.

Avainsanat: Tilastollinen laadunvalvonta, poikkeava havainto, kynnyssääntö, robustit menetelmät

Abstract

Objectives of this pilot report were explorative. Examinations were made using certain parts of the balance of payments monthly survey (SV) and annual survey (SVA) of foreign assets and liabilities. The SV data consisted of the monthly observations over the time period from 2003 to 2006 and the SVA data included the annual data of 2005 and 2006. Detailed examination focused on variables of foreign liabilities: shares, mutual fund shares, bonds and money market papers. The first objective was to fit different methods to indicate outlying observations in the selected cross-sectional and time series data. The second objective was to estimate statistics, which could be utilized as threshold values and rules in statistical process control. In addition, information about the structure of the observation data was received as an offshoot of the examinations.

Examinations showed that, at least with the selected data, median or mean and standard deviation can be used in creating the threshold rules. The data indicated that it is recommended to use both classic and robust versions of statistical methods, and that the non-parametrical methods should be used as well when statistical distribution of the data is unknown. Systematic sampling can be used if there are enough observations. Further, it was seen that graphics, especially box-plots and qq-plots, can give fast and clear information about the structure of the data. The data showed as well that the partition made between big and other respondents is fairly clear and that both time series and cross-sectional examinations are needed because they complete each other.

Key words: Statistical Process Control, outlying observation, outlier, threshold rule, robust methods.

1 Johdanto

1.1 Mitä on tilastollinen laadunvalvonta?

Tilastollisella laadunvalvonnalla (Statistical Process Control, SPC) tarkoitetaan kerätyn kyselyaineiston tutkimista valituilla kuvailevilla tunnusluvuilla mukaan lukien graafiset tarkastelut. Päätavoite on poikkeavien havaintojen tai havaintoryhmien tunnistaminen aineistosta[1]. Prosessien tilastollinen laadunvalvonta on paljon käsitelty aihe tilastotieteessä; kirjallisuuden eräs keskeinen teos on Montgomery (2001).

Tilastollisen laadunvalvonnan tehokkuus kasvaa olennaisesti, kun sitä käytetään yhdessä sisällöllisen laadunvalvonnan kanssa. Ainoastaan substanssiasiantuntijat voivat korjata esimerkiksi luokitteluvirheet, joista ensin on saatu signaaleja puhtaasti tilastollisella analysoinnilla (esim. jotkut outlierit ja merkittävät poikkeamat %-jakaumassa). Siksi tällainen yhteistyö tuottaa tunnetusti parhaan tuloksen siis laadun. Vastaajat (respondents) luonnollisesti ovat viime kädessä vastuussa tietojen oikeellisuudesta.

1.2 Tämän pilottiraportin tavoitteet

Tämän pilottiraportin tavoitteet ovat tutkiskelevat (eksploratiiviset). Tavoitteet voidaan jakaa kahteen osaan. Ensiksikin, yhdessä substanssiasiantuntijan[2] kanssa kyselyaineistosta valittiin osa, johon sovitettiin erilaisia menetelmiä indikoimaan poikkeamia sekä poikkileikkaus- että aikasarjainformaatiossa. Menetelmien soveltuvuutta arvioitiin valitun kyselyaineiston valossa. Toiseksi, tavoitteena oli estimoida tilastollisia tunnuslukuja valitusta kyselyaineistosta, jotta estimoituja arvoja voitaisiin hyödyntää kynnysrajojen ja -sääntöjen muodossa input-aineiston mekanistisen laadunvalvonnan välineenä. Sivutuotteena nämä erilaiset estimointitulokset valottavat kyselyaineistojen (SV ja SVA) valittujen osien tämänhetkistä rakennetta vastaajittain ja muuttujittain (kyselydata, tyhjät solut ja vastausfrekvenssit havaintomatriisis-

[1] Sovellettavien tilastollisten tunnuslukujen ja estimointimenetelmien tulisi olla sekä klassisia että robusteja (ks. Maronna, Martin and Yohai, 2006). Viimeksi mainitut ovat tärkeitä siksi, että ne sallivat puutteellisuuksia aineistossa ja antavat siitä tällöin harhattomamman kuvan verrattuna ei-robusteihin kerroinestimaatteihin ja tunnuslukuihin, esimerkiksi keskiarvoon. Usein on tarkoituksenmukaista laskea estimaatit molemmilla menetelmillä ja tunnusluvuilla ja vertailla tuloksia. Esimerkiksi keskiarvon ja mediaanin estimaatissa on selvä ero, kun aineistossa on outliereita, virheellisiä ja/tai todellisia.

[2] Kiitokset asiantuntevista kommenteista ja hyödyllisistä keskusteluista Maksutasetoimiston ekonomisteille Jaakko Sunille ja Pasi Ikoselle.

sa, muuttujien ja vastaajien suuruusjärjestys ja %-jakaumat). Myös outlier-ehdokkaat seuloutuvat näkyviin esimerkiksi 3-ulotteisten kuvioiden tuella.

Raportin rakenne on seuraava: kappaleissa 1-3 on tarkasteltu, mitä on tilastollinen laadunvalvonta (prosessivalvonta) sekä esitelty kyselyjen pilottiaineistot. Luvuissa 4-5 on lyhyt katsaus kuvaileviin tilastollisiin tunnuslukuihin, grafiikkaan, kynnysarvojen muodostamiseen ja poikkeaviin havaintoihin. Luvussa 6 esitellään tulokset ja luvussa 7 esitetään yhteenveto ja eräitä jatkoselvityskohteita.

2 Kyselyaineistot

Kukin kysely koostuu tavoitteidensa mukaisesti joukosta erillisiä kysymyksiä. Kysymys voi sisältää useita muuttujia, jotka voivat olla muussakin kuin rahayksikkömuodossa. Tämän pilottiraportin päätavoite on muuttuja-tason input-aineiston laadunhallintavälineistön kokeilu hälytysrajojen ja -kynnysten konstruoimiseksi. Tavoitteena on mahdollisimman yksinkertainen ja selkeä välineyhdistelmä.

2.1 Kyselylomake tiedonkeruutekniikkana

Otanta- ja kyselytutkimuksella pyritään estimoimaan perusjoukon eli populaation ominaisuuksia. Voidaan erottaa tavoite(target)perusjoukko ja kehikko(frame)perusjoukko. Suomen Pankin tilastotuotannon tiedonkeruussa käytetään kyselylomakkeita, jotka osoitetaan kyselyyn valituille populaation tilastoyksiköille (yritys, julkisyhteisö tai järjestö, joita aineiston yhteydessä kutsutaan tässä vastaajiksi). Kyselylomakkeiden vastauksista muodostetaan havaintoaineisto, joka tilastollista analyysiä varten käsitellään havaintomatriisina, missä mitattavan tilastollisen muuttujan $X_i, \{i = 1, \ldots p\}$ arvot esitetään vastaajittain $Y_j, \{j = 1, \ldots, n\}$ muodossa $\mathbf{M} = n \times p$.

2.2 Kyselylomakkeen virheistä

Sen lisäksi että kehikkoperusjoukolta kysyttävien kysymysten estimointiin liittyy tilastollisia virheitä, myös kyselytekniikkaan tiedonkeruun välineenä liittyy erityyppisiä tilastollisia virheitä. Betlehem (2005) listaa kyselytutkimukseen liittyviä tilastollisia virheitä. Kehikkoperusjoukon ja tilaston estimointiin liittyvät tilastovirheet on rajattu tarkastelun ulkopuolelle.

Vastaajan Suomen Pankille raportoiman arvon validiteetin arvioimiseksi kyselylomakkeisiin on implementoitu erilaisia tarkistuksia, jotka pyrkivät esimerkiksi eliminoimaan ristiriitaiset arvot kysymyksen muiden muuttujien suhteen. Kyselylomakkeisiin on mahdollista implementoida myös tarkistuksia, jotka pyrkivät eliminoimaan ristiriitaiset arvot eri ajankohtien suhteen sekä muuttuja- että vastaajakohtaisesti. Tilastontuotannossa havaittu empiirinen evidenssi osoittaa, että vastaajat raportoivat arvoja väärällä tarkkuudella. Mielenkiinto tässä on löytää

poikkeavat suuret ja pienet arvot. Pilottikokeilla tutkitaan aiemmin kerättyä havaintoaineistoa ja pyritään approksimoimaan parametriarvot muuttujakohtaisille tarkistuksille.

Olkoon $x_{i,j}, \{i = 1,\ldots,p\}, \{j = 1,\ldots,n\}$ vastaajan *j* raportoima arvo kyselyn muuttujaan *i*, ja

$$E\left[x_{i,j}(t) \mid \Omega(t)\right] \qquad (1)$$

odotusarvo arvolle $x_{i,j}(t)$, missä $\Omega(t)$ on hetken *t* informaatiojoukko, joka on substanssiasiantuntijan käytettävissä. Informaatiota on esimerkiksi vastaajan kokoluokka tasearvon suhteen ja muiden vastaajien vastausten jakauma muuttujakohtaisesti. Raportoidun arvon validointitarve syntyy esimerkiksi tilanteissa

$$x_{i,j}(t) > E\left[x_{i,j}(t) \mid \Omega(t)\right], \qquad (2)$$

$$x_{i,j}(t) < E\left[x_{i,j}(t) \mid \Omega(t)\right], \qquad (3)$$

missä raportoitu arvo $x_{i,j}(t)$ poikkeaa odotusarvosta (1).

2.3 Havaintoaineistot

Havaintoaineistona käytettiin kyselyn "Selvitys ulkomaisista saamisista ja veloista maksutasetilastointia varten" tuloksia, jotka on kerätty kahdelta erityyppiseltä kehikkoperusjoukolta. SV-kehikkoperusjoukon vastaajat raportoivat Suomen Pankille kyselylomakkeella kuukausittain ja SVA-kehikkoperusjoukon vastaajat raportoivat vuosittain. SV-kehikkoperusjoukon vastaajat ovat yleisesti ottaen tasearvoiltaan suurempia kuin SVA-kehikkoperusjoukon vastaajat, joten karkeasti voidaan nähdä kahtiajako suuriin ja pieniin vastaajiin. Vastaajat koostuvat erilaisista yhteisöistä, joilla on ulkomaisia saamisia tai velkoja.

Tarkasteluja on tehty SV- ja SVA-aineistojen ulkomaisten saamisten ja ulkomaisten velkojen kysymyksille arvopaperisaamiset, muut saamiset, arvopaperivelat ja muut velat. Nämä kysymykset valittiin yhdessä substanssiasiantuntijan kanssa. Perusteena valinnalle oli muuttujakohtaisten arvojen absoluuttinen suuruus ja toisaalta raportoitujen arvojen vastausintensiteetti. Analyysi toteutettiin laskemalla havaintomatriisin **M** rivi- ja sarakesummat ja arvioimalla edellä mainittujen kriteerien osuutta muuttujakohtaisesti. Liitteen 1 taulukossa 6 on esitetty osa SV-kyselyn poikkileikkausaineiston mukaisesta havaintomatriisista.

Arvopaperisaamiset on sarakesummaltaan suurin valituista kysymyksistä ja tämän vuoksi osa tarkasteluista on tehty vain sen osalta. Tarkempi tarkastelu rajoittuu siten ulkomaisten arvopaperisaamisten muuttujiin: osakkeet, rahasto-osuudet, joukkolainat ja rahamarkkinapaperit. Taulukko 1 havainnollistaa valittuja kysymyksiä ja muuttujia. Kuviossa 1 on lisäksi esi-

tetty tarkastelussa mukana olleet muuttujat sarakesummien mukaisessa suuruusjärjestyksessä.

Sekä SV- että SVA-poikkileikkausaineistoissa ovat mukana vuoden 2006 kyselyyn vastanneet. SV- kyselyn osalta huomioon on otettu ne 92 vastaajaa, joilla ainakin yhden muuttujan arvo poikkeaa nollasta. Vastaavasti SVA-kyselyssä on huomioitu kaikki 302 vastaajaa, joilla on vähintään yksi nollasta poikkeava vastaus. Tämä ei kuitenkaan tarkoita välttämättä sitä, ettei muilla vastaajilla olisi lainkaan lukuja, ne vain saattavat olla niin pieniä, että näkyvät havaintomatriisissa nollana, kun luvut kirjataan miljoonina euroina. Tarkemmin sanottuna pienin kirjautuva arvo on 0,01 miljoonaa euroa eli 10 000 euroa pienemmät summat näkyvät nollina. Tästä syystä muuttujien minimiarvoksikin tulee nolla (0,00 milj. euroa).

Aikasarjatarkasteluita on tehty vain SV-aineistolle. Kohteina ovat olleet kantatietojen 2006M12 eli poikkileikkausaineiston perusteella valitut 11 vastaajaa. Nämä kaikki ovat 14 suurimman vastaajan joukossa, kun poikkileikkausmatriisista on laskettu kaikkien vastaajien rivisummat. Lähtökohtaisesti ideana oli valikoida mukaan kaksikymmentä kohdetta siten, että mukaan otetaan 14 suurinta, jotka muodostivat rivisummien perusteella 75 prosenttia vastausten kokonaissummasta ja lisäksi kuusi vastaajaa systemaattisella satunnaisotannalla. Kuitenkin vain tarkastelussa mukana olleilla 11 vastaajalla oli lukuja edellä mainituissa ulkomaisten arvopaperisaamisten muuttujissa. Lisäksi tarkastelujen rajoittuessa joukkolainat-muuttujaan jäi tarkasteltavia vastaajia jäljelle vain kymmenen. Aikasarjat kattavat aikavälin tammikuusta 2003 joulukuuhun 2006.

Koska SVA-vuosikyselystä oli tietoja vain vuosilta 2005 ja 2006, aineistolle tehtiin yleisempiä, lähinnä vuoden 2006 poikkileikkausaineistoon perustuvia tarkasteluja. Näissä tarkasteluissa olivat mukana kaikki noin 300 vastaajaa, jotka muodostivat liitteen 1 (taulukko 6) poikkileikkausaineistoa vastaavan havaintomatriisin.

Taulukko 1. Tarkastelussa olevat ulkomaisten saamisten ja velkojen kysymykset (rasteroitu-na) muuttujineen

Ulkomaiset saamiset		Ulkomaiset velat	
Arvopaperi-saamiset	***Muut saamiset***	***Arvopaperivelat***	***Muut velat***
Osakkeet	Lainat	Osakkeet	Lainat
Rahasto-osuudet	Vuokrausluotot	Joukkolainat	Vuokrausluotot
Joukkolainat	Kauppaluotot	Rahamarkkina-paperit	Kauppaluotot
Rahamarkkina-paperit	Talletukset		Muut velat
	Muut saamiset		

3 Aineistojen rakenteen eksploratiiviset tarkastelut

3.1 Kuvailevat tilastolliset tunnusluvut

Kyselyaineiston tutkiminen alkaa perustunnuslukujen laskemisella. Näitä ovat keskiarvo, mediaani, standardipoikkeama, huipukkuus, vinous, minimi, maksimi, vaihteluväli (range), kvartiilit, IQR (kvartiiliväli, Q3-Q1) ja variaatiokerroin. Näistä eräille on kehitetty robusti versio, kuten keskiarvolle ja standardipoikkeamalle. Aikasarjatarkastelun tapauksessa lasketaan valitun aikasarjan havaintojen perusteella kynnysarvo, esimerkiksi kynnysarvo = 2 * sarjan std. poikkeama. Kynnysarvo voisi myös olla 3*std. poikkeama. Laadunvalvonnan perusmuoto: keskiarvo ja sen ympärillä ylä- ja alahälytysraja, voidaan perustaa joko klassisiin tai robusteihin keski- ja hajontatunnuslukuihin. Käytettävissä on myös puhtaasti robusteja tunnuslukuja (esim. trimmattu keskiarvo ja hajonnan estimaattori, MAD = median of absolute median deviations).

3.2 Graafiset esitykset

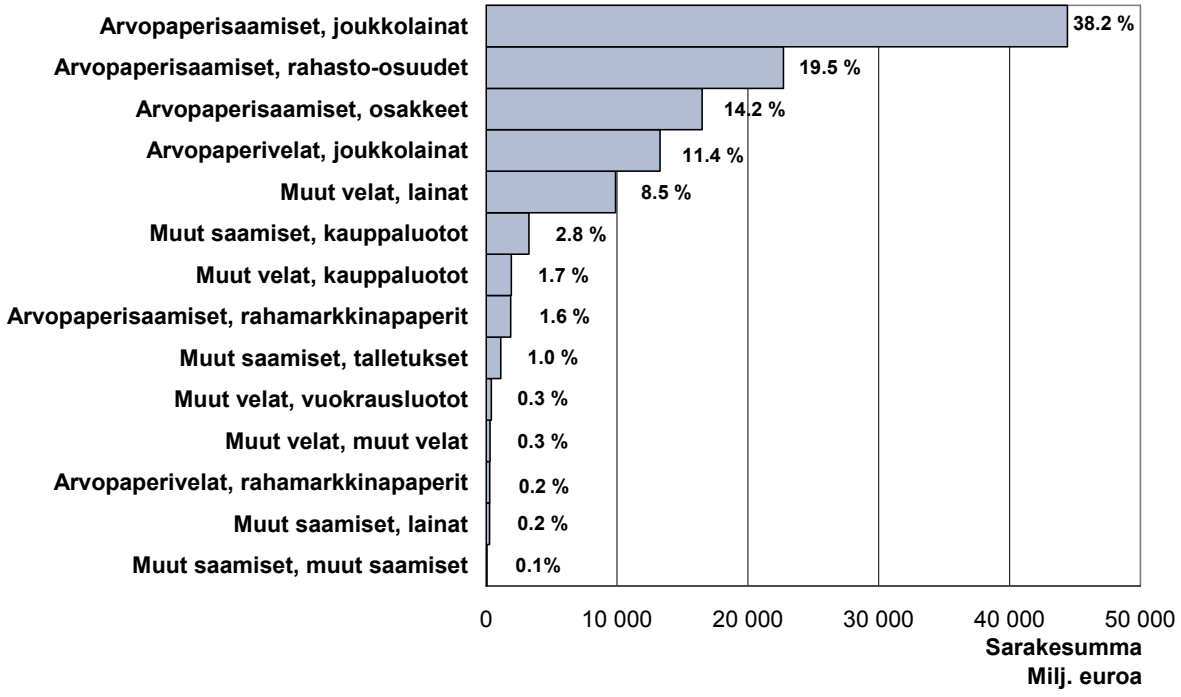
Kuvioesitys on usein nopein tapa saada yleiskuva aineiston rakenteesta. Tämä koskee sekä aikasarja- että poikkileikkausaineistoja. Riippuu aineistosta minkälaisia muunnoksia siitä esitetään, taloudellisissa aineistoissa käytetään usein logaritmista muunnosta (vain positiiviset arvot) tai itseisarvomuunnosta. Yksittäisten 2D- ja 3D-kuvioiden ohella "tehoa" graafiseen esittämiseen saadaan monikuvioesityksestä; esimerkiksi paljon käytetty box-plot - (= viiksikuvio/laatikkokuvio) monikuvioesitys on usein hyvin sopiva tulosten esittämiseen. Nykyiset tilastoanalyysipaketit tarjoavat tutkijoille hyvät mahdollisuudet monipuolisen graafisen analyysin yhdistelmiin. Aineiston normaalijakautuneisuudesta saadaan nopeasti kuva tunnetun qq-plot -kuvion avulla (ks. harjoitus liitteessä 2).

Molemmille SV- ja SVA-poikkileikkausaineistoille on paitsi laskettu erilaisia tunnuslukuja myös piirretty tunnuslukuja kuvaavia graafisia esityksiä. Kuviossa 1 on SV-poikkileikkausaineistosta laskettu muuttujien suurusjärjestyksen esittävä pylväskuvio, jota vastaavat kuviot on piirretty myös kaikkien vastausten mediaanille, keskiarvolle ja maksimille molempien ky-

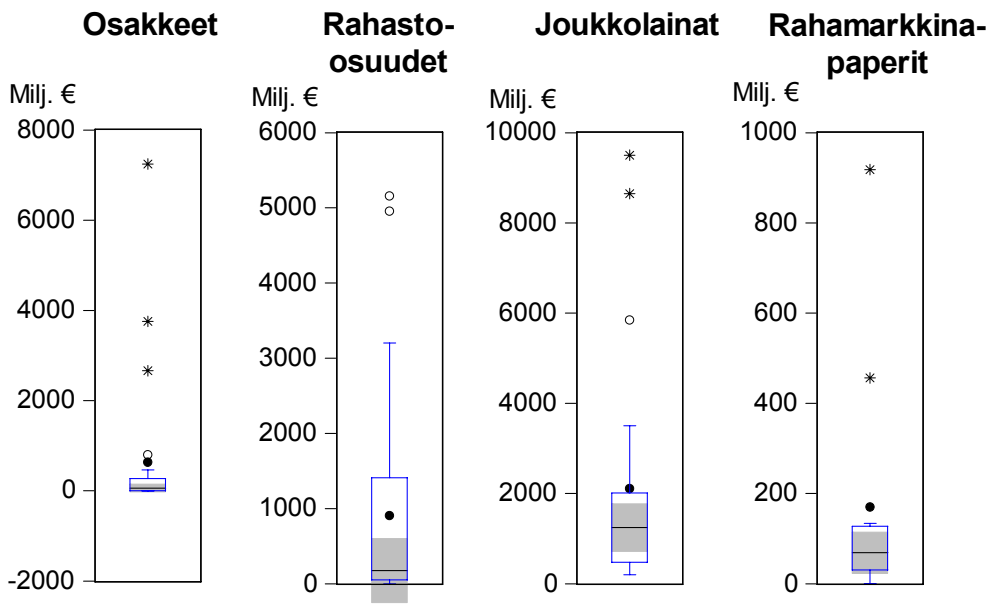
selyaineistojen tapauksessa. Kuvioiden perusteella voidaan esimerkiksi saada viitteelliset maksimiarvot yksittäisen vastaajan antamille luvuille kunkin muuttujan tapauksessa.

SV-aineiston kysymyksiä on kuvattu myös box-plot -kuvioilla (kuvio 2), joissa laatikon alaraja kuvaa 25 %:n kvartiilia (alakvartiili) ja yläraja 75 %:n kvartiilia (yläkvartiili); mediaani on 50 %:n kvartiili eli aineiston suuruusjärjestyksessä keskimmäinen havainto. Mediaanit on merkitty kuviin laatikon poikki kulkevalla vaakaviivalla ja keskiarvot mustalla pisteellä. Poikkeavat havainnot on merkitty kuviin valkoisilla pisteillä (near outliers) ja tähdillä (far outliers). Viivat eli viikset kuvaavat sitä, miten muut havainnot sijoittuvat. Vastaavia kuvaajia on piirretty myös aikasarja-aineistoille, jolloin kuvioista voidaan tehdä havaintoja esimerkiksi siitä, miten suuri hajonta on yksittäisen vastaajan tietyn muuttujan raportoiduissa luvuissa ja onko luvuissa selkeästi poikkeavia havaintoja (outlying observations, outliers).

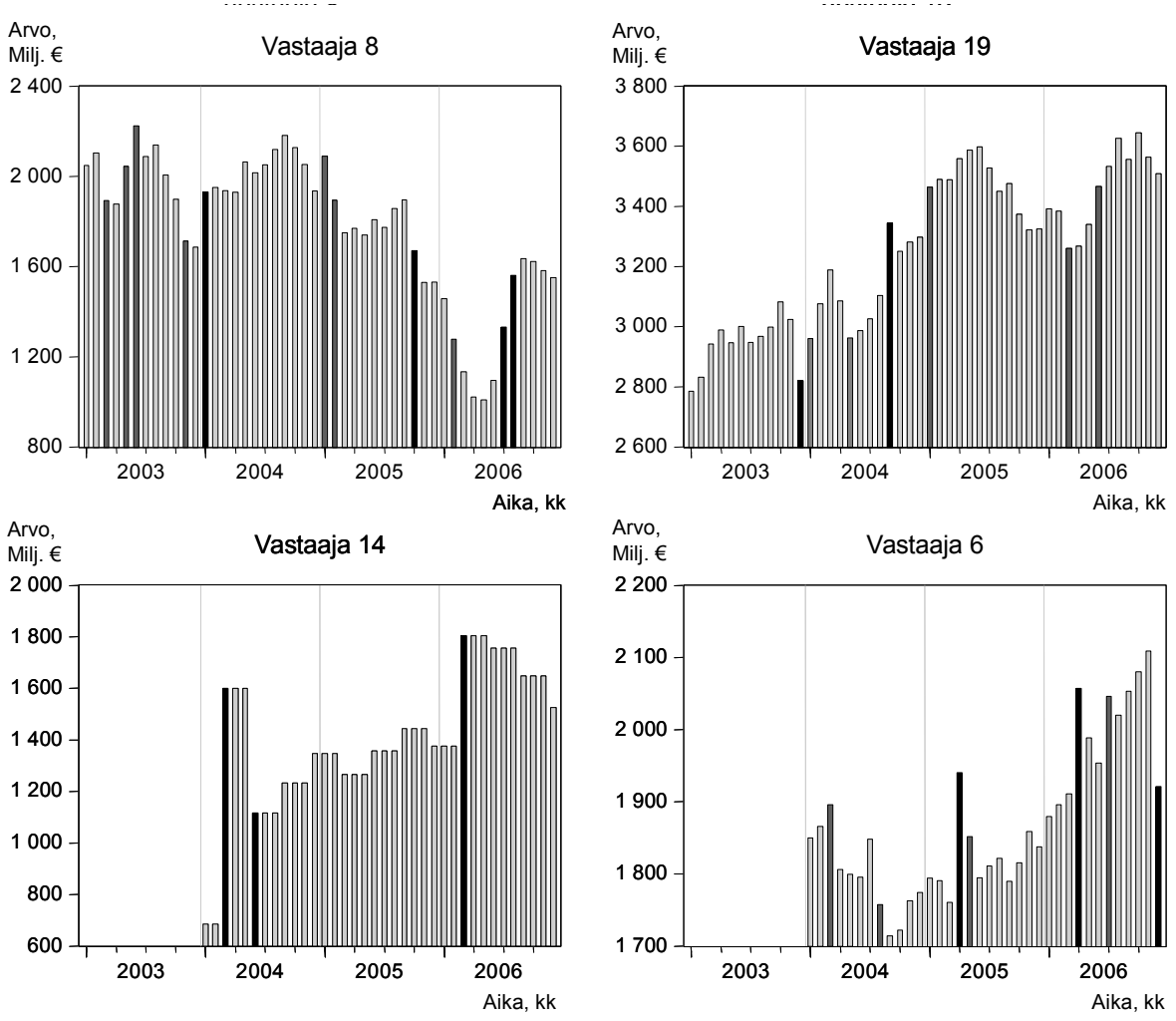
Kuvio 1. SV-aineiston muuttujien (havaintomatriisi) sarakesummat prosenttiosuuksineen suuruusjärjestyksessä 2006M12 kantatietojen perusteella



Kuvio 2. Box-plot -kuviot kysymyksen ulkomaiset arvopaperisaamiset muuttujista

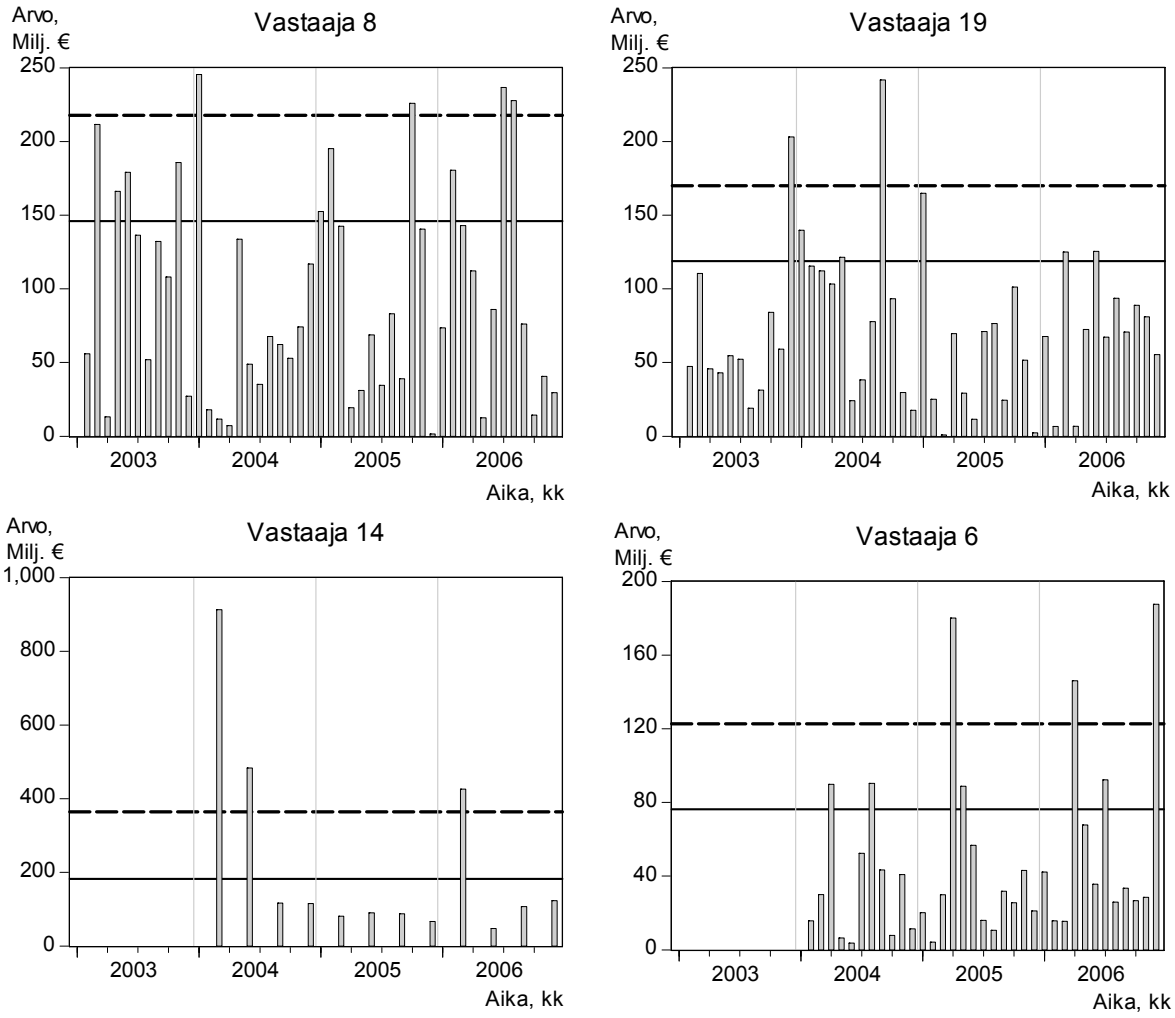


Kuvio 3. Aikasarja-aineiston pylväskuvioita, joihin on väritetty eritasoisen hälytyksen saaneet havainnot (tumma harmaa: alempi taso, musta: ylempi taso).



Aikasarja-aineistoa voidaan kuvata myös perinteisillä pylväskuvioilla esimerkiksi kuvion 3 tapaan. Siinä on esitetty pylväskuviot neljän vastaajan raportoimien lukujen aikasarjoista. Kuviossa 4 puolestaan on esitetty samojen vastaajien aikasarjoista kertaalleen differensoidut muutossarjat. Muutossarjojen kuvioihin on piirretty myös kaksi eritasoista ylähälytysrajaa, jotka ovat tässä tapauksessa mediaani + std. poikkeama ja mediaani + 2*sdt. poikkeama. Mustalla kuvioon 3 väritetyt havainnot ylittävät korkeamman hälytysrajan ja tummalla harmaalla väritetyt alemman. Luvun 6 taulukkoon 5 on kerätty kooste harjoituksesta, jossa liputettiin ylähälytysrajan ylittäviä vastauksia aikasarjassa kolmella eritasoisella hälytysrajalla.

Kuvio 4. Aikasarja-aineistosta differensoidut sarjat hälytysrajoinaan mediaani + standardipoikkeama (viiva) ja mediaani + 2*standardipoikkeama (katkoviiva).



3.3 Autokorrelaatio(AC)- ja osittaisautokorrelaatio(PAC)-estimaatit

Kun käsiteltävä aineisto on aikasarjatyyppiä, tilastollinen perusanalyysi voidaan ulottaa otoksen auto- ja osittaisautokorrelaatioiden laskentaan. Laadunvalvontamielessä havaintosarjan termien peräkkäinen korrelointivoimakkuus on luonnollisesti tärkeä mitata ja testata. Sarjan riittävä pituus (aikavälinä) on tietenkin tarpeellista. Kun aineiston jakaumasta ei ole tietoa (ainoastaan, että se ei ole normaalinen), voidaan kokeilla ei-parametrista estimaattoria, esimerkiksi laskea autokorrelaatio rank-muunnetulle sarjalle ja vertailla alkuperäiseen aineis-

toon perustuvaan estimointitulokseen. Kirjallisuudessa on käytettävissä myös autokorrelaation robusteja versioita (yleensä niitä ei toistaiseksi ole markkinoiden analyysiohjelmissa). Autokorrelaatio- ja osittaisautokorrelaatioiden laskemisen jälkeen voidaan estimoida vastaava aikasarjamalli[3].

Liitteen 1 taulukkoon 7 on kerätty kaikkien tarkastelussa mukana olleiden vastaajien autokorrelaatiot ja osittaisautokorrelaatiot joukkolainat-muuttujan tapauksessa. Taulukosta voidaan havaita, että lähes kaikkien vastaajien kohdalla autokorrelaatio yhdellä viiveellä on melko lähellä lukua 1. Tämä tarkoittaa, että näillä vastaajilla ajan hetkien *t* ja *t*-1 lukujen välillä on voimakas riippuvaisuus.

[3] SPC:ssä on autokorrelaatioita käytetty kahdella tavalla: pohjana on aikasarjamalli ja tarkastellaan alkuperäistä dataa hälytysrajoineen, mitkä sopeutetaan ottaen huomioon datan autokorreloituneisuus; toiseksi, käytetään aikasarjamallia parametriestimaatteineen ja analysoidaan mallin jäännöstermisarjaa. SPC:n ohella voidaan erottaa ns. automaattinen prosessin valvonta (APC, EPC), mitä käytetään etenkin insinöörisovelluksissa. Siinä aikasarjamallilla lasketaan yhden askeleen ennusteita ja tutkitaan prosessin sopeutumista kohti haluttua kehitysuraa (ks. Stoumbos et al., 2002). Hyvin tunnettu laadunvalvonnan aikasarjamalli on eksponentiaalisen tasoituksen liukuva keskiarvo (EWMA).

4 Kokeilussa tehtyjä tilastollisen laadunvalvonnan peruslaskelmia

Kynnysten muodostamista varten laskettiin poikkileikkausaineistojen kuvailevia tunnuslukuja. Keskeisiä ovat muuttujan %-jakauman perustunnusluvut sekä riveittäin (vastaajan eri muuttujat) että sarakkeittain (muuttujan jakauma vastaajittain). Koko kyselyn osalta oli myös mahdollista laskea muuttujakohtainen aggregaattijakauma.

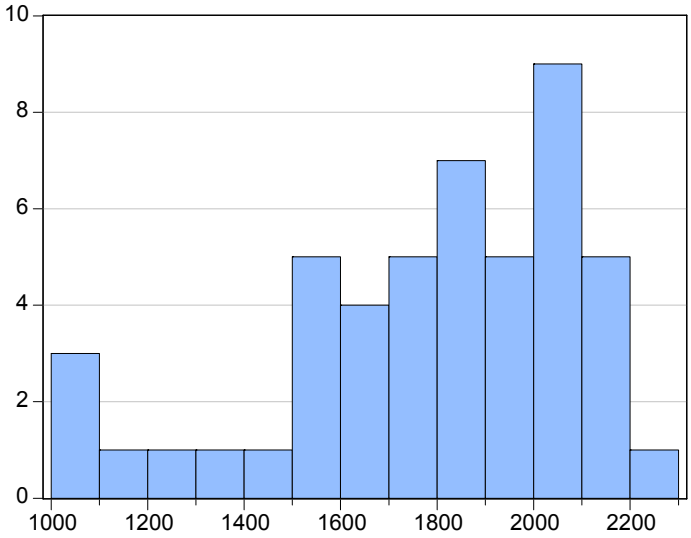
Liitteessä 1 on esitetty esimerkkinä matriisimuotoinen taulukko SV-kyselyn poikkileikkausaineistosta 2006M12 kantatietoihin perustuen. Taulukko 6 kuvaa vain ulkomaisia arvopaperisaamisia ja siinä on näkyvissä aineistolle laskettuja ominaisuuksia ja tunnuslukuja. Varsinainen matriisi on esimerkkitaulukkoa huomattavasti laajempi. SVA-kyselyn aineistosta on tarkasteluissa muodostettu täysin vastaava taulukko. Taulukosta 3 löytyvät molemmille kyselyille lasketut perustunnusluvut. Lisäksi liitteessä 1 on taulukko 8, johon on kerätty molempien aineistojen perustunnusluvut kysymyksittäin.

Aikasarja-aineistolle on ulkomaisten arvopaperisaamisten joukkolainat-muuttujan osalta laskettu kuvion 5 mukaiset perustunnusluvut ja piirretty vastausten jakaumaa kuvaavat histogrammit kaikille niille tarkastelujoukon vastaajille, joilla on vähintään yksi nollasta poikkeava vastaus. Joukkolainat-muuttujan kuukausisarjoille on laskettu myös autokorrelaatiot ja osittaisautokorrelaatiot tarkasteluissa mukana olleille vastaajille (liite 1 taulukko 7). Liitteessä 2 on autokorrelaatio- ja osittaisautokorrelaatioestimaatit korrelogrammeineen eräälle vastaajalle, jonka aikasarjassa autokorreloituneisuus on hyvin vahvaa. Autokorreloituneisuus vaikuttaa muun muassa uusien havaintojen ennustettavuuteen sekä sarjoille tehtävään eksponentiaaliseen tasoitukseen. Esitetyssä tapauksessa aikasarjaan voisi autokorrelaatioiden perusteella yrittää sovittaa autoregressiivista AR-mallia (ks. luku 3).

Eräistä sarjoista laskettiin kokeeksi eksponentiaalisen tasoituksen perusmalli sekä robustin filtterin 4253H avulla tasoitettu sarja (analyysiohjelmapaketti Statistica). Estimointiharjoitukset on esitetty liitteessä 2. Harjoituksista huomataan selkeä yhteys otosautokorrelaatiohin. Kun autokorrelaatiot ovat suuria, tulee eksponentiaalisen tasoituksen kertoimen alfa arvo hyvin lähelle yhtä ja tällöin edellinen havainto hetkellä *t*-1 estimoi seuraavan hetken *t* havaintoa melko hyvin.

Pilottiaineistoa muodostettaessa sovellettiin eräissä tapauksissa myös systemaattista otantaa (Statistica-ohjelmalla). Tämän otannan avulla on kätevä tutkia erityisesti keskisuuria ja pieniä vastaajia. Otantaväliä muuttamalla saadaan haluttu määrä vastaajia analyysin piiriin.

Kuvio 5. Aikasarja-aineiston perustunnusluvut ja histogrammi erään vastaajan arvopaperisaamisten joukkolainat-muuttujalle



Series: Vastaaja 8
Sample 2003M01 2006M12
Observations 48

Mean	1782.847
Median	1884.550
Maximum	2223.324
Minimum	1008.590
Std. Dev.	314.2172
Skewness	-0.922272
Kurtosis	3.125584
Jarque-Bera	6.836221
Probability	0.032774

5 Havaintojen tarkistus ja käsittely

5.1 Evaluoinnin kynnystarkastelut: poikittais- ja pitkittäisinfo

Lomaketietojen input-matriisin kunkin muuttujan X kynnysarvo kytkeytyy kyselyssä A olevan vastaajan Y lähettämän (tuottaman) tiedon laadun arviointiin. Saadun, yksittäisen tiedon arviointi perustuu seuraavassa esitettyihin tekijöihin.

Poikkileikkausaineisto

a) Nykyhetken *t* havaintoarvon vertaaminen edelliseen havaintoarvoon, $x_{i,j}(t-1)$, joka on valmiiksi näkyvissä vastauslomakkeen ao. kohdassa.

b) Vertaaminen tietoon, miten a-kohdan edellinen lukuarvo, $x_{i,j}(t-1)$, sijoittuu vastaajan kaikkien muuttujien vastausjakaumassa (%-osuus tms.) ajankohtana $t-1$.

c) Miten em. $x_{i,j}(t-1)$ sijoittuu kyselyn kaikkien vastaajien vastaavan muuttujan jakaumassa (esim. jakauman keskiarvon, mediaanin ja ääriarvojen suhteen) ajankohtana $t-1$.

Poikkileikkausinformaation osalta jakaumatunnusluvuista ääriarvojen ja IQR-arvojen (interquartale value, Q3-Q1) rooli korostuu. Tulokset ovat tietenkin hyvin muuttujakohtaisia.

Aikasarja-aineisto

Jos kyselystä A on riittävä määrä historiatietoa, voidaan kohtien a) – c) poikkileikkausanalyysin lisäksi käyttää aikasarja-analyysin välineitä laadun arvioimisessa. Tällöin voivat tulla kysymykseen autokorrelaatio ja osittaisautokorrelaatio, erilaisten tasoitusfunktioiden käyttö (klassinen ja/tai robusti), differensointi (jos aikasarjassa on selkeä trendikomponentti), kuvailevat tunnusluvut (tässä raportissa tarkemmin kohdassa 4) ja niin edelleen. Kynnys- ja mahdollinen hälytysrajaesitys voidaan sitten kehittää saatujen estimaattien ja lukuarvojen perusteella.

Korrelaatioestimaateilla voidaan mitata aikasarjan peräkkäisten havaintojen riippuvuuden voimakkuutta eri viiveillä (lag). Korreloinnin voimakkuutta voidaan sitten käyttää hyödyksi uuden havainnon laadun evaluoinnissa. Toinen, yksinkertainen menettely on soveltaa ai-

kasarjaan jotakin (mieluusti robustia) tasoitusta, esimerkiksi geometrisesti painotettua (eksponentiaalinen tasoitus eri analyysiohjelmissa) ja laskea mekaanisia ennusteita ajankohdalle *t*. Todettakoon vielä, että aikasarjatarkasteluja ei käytännön syistä voida tehdä kaikille kyselyn yksittäisille muuttujille.

Kanta-muodossa olevan muuttujan aikasarjasta voidaan hälytyskynnystä kokeilla myös seuraavasti: lasketaan sarjasta %-muutos, muunnetaan se itseisarvon muotoon, lasketaan itseisarvosarjalle n-termin liukuva mediaani ja lasketaan kynnysarvo = 3* liukuvan mediaanin tuorein arvo. Tätä hälytyskynnysmenettelyä voidaan vielä parantaa laskemalla edellä mainitulle itseisarvosarjalle robusti filtteri 4253H ja ottamalla filtterein tuorein arvo, z(rob), käyttöön, hälytyskynnys voisi silloin olla esimerkiksi muotoa 3*z(rob). Robustin filtterin tapauksessa tuorein havainto kohdistuu itseisarvosarjan tuoreimman havainnon kohdalle, kun taas liukuva mediaaniarvo jää sitä aiemmaksi aika-akselilla (sama ilmiö kuin liukuvan keskiarvon tapauksessa). Liitteessä 2 on kuvio harjoituksesta, joissa on laskettu eksponentiaalinen tasoitus sekä robusti filtteri 4253H eräälle vastaajan aikasarjalle.

Virta-muotoisen muuttujan tapauksessa voidaan menetellä, kuten edellä muodostamalla ensin virtatiedoista itseisarvo. Itseisarvosarjoista voidaan erikseen laskea kuvailevia tunnuslukuja erikseen vastaamaan alkuperäisiä + ja – tapauksia.

Huomattakoon, että kynnysarvon määrittäminen on tässä aina luonteeltaan empiirinen asia, useasti keskiarvoisiin ja ääriarvopoikkeamiin tukeutuen. Jos päädymme teoreettiseen lähestymistapaan, normaalijakaumaan pohjaavat analyysit tulee korvata muulla, mahdollisesti gamma- tai Pareto-jakaumaan perustuvilla hajontatunnusluvuilla. Jos voimme odottaa, että aineisto on edes likimääräisesti normaalijakaumaa noudattava, luottamusväli (±2*std. poikkeama) edustaa 5 %:n riskiä kaksisuuntaisessa testitapauksessa. Aineiston normaalijakautuneisuutta on eräissä kokeissa tutkittu soveltamalla Jarque-Bera -normaalisuustestiä (ks. luku 7) sekä qq-plot -kuvioita.

Jarque-Bera -testin tulokset arvopaperisaamisten joukkolainat-muuttujalle on esitetty taulukossa 2 ja tulosten mukaan vastaajaa 8 lukuun ottamatta nollahypoteesia ei hylätä. Tämä siis tarkoittaisi, että kaikkien muiden paitsi vastaajan 8 aikasarja on normaalisti jakautunut. Kuitenkaan qq-plot -kuvioiden (liite 2 kuvio 7) perusteella mikään sarjoista ei ole normaalisti jakautunut, sillä kaikissa kuvioissa ääripäiden havainnot poikkeavat selkeästi normaalijakaumaa kuvaavasta suorasta. Jarque-Bera -testin heikkous onkin, että siinä verrataan testattavan jakauman vinoutta ja huipukkuutta normaalijakauman vinouteen ja huipukkuuteen. Tällöin saattaa käydä niin, että nollahypoteesia (jakauma on normaalisti jakautunut) ei hylätä, vaikka vinous ja huipukkuus ovatkin vain sattumalta samat kuin normaalijakaumalla. Yleensä

Taulukko 2. Jarque-Bera -testin tulokset arvopaperisaamisten joukkolainat-muuttujalle

	Jarque-Bera	p-arvo
Vastaaja 1	0.927	0.629
Vastaaja 2	5.779	0.056
Vastaaja 3	4.110	0.128
Vastaaja 4	3.653	0.161
Vastaaja 6	3.538	0.171
Vastaaja 7	3.511	0.173
Vastaaja 8	6.836	0.033
Vastaaja 9	2.576	0.276
Vastaaja 11	1.504	0.472
Vastaaja 14	3.698	0.157

normaalisuuden testaus edellyttää varsin suuren otoskoon. Kirjallisuudessa on käytettävissä useita rakenteeltaan parametrisia ja ei-parametrisia normaalisuustestejä. Lisäksi tiedetään, että varsinkin huipukkuus lähestyy normaalijakaumaa hyvin hitaasti otoskoon kasvaessa.

5.2 Poikkeavat havainnot

Poikkeava havainto (outlying observation, outlier) määritellään väljästi havaintona, joka selvästi eroaa otoksen muista havainnoista (remainder, majority of observations). Yritysaineistoissa on yleensä paljon tämänlaisia havaintoja; tehtävänä on tässä virheellisten outlierien tunnistaminen ja korjaaminen. Yritysaineistoissa poikkeavien havaintojen suhteuttaminen esimerkiksi mediaaniin voi olla perusteltua varsinkin volatiilin aineiston (havaintomatriisin rivi- tai sarakedata) tapauksessa. Usein aineistojen graafiset tarkastelut kuten box-plot -kuviot auttavat paljon kynnysarvojen suuruusluokan määrittämisessä. Kyselyn substanssiasiantuntija voi myös ratkaisevasti auttaa kynnysarvon määritysongelmissa ja outlierien tulkinnassa. Todettakoon lisäksi, että aineistossa voi olla poikkeavien havaintojen rypäs, minkä selvittämiseen usein tarvitaan erityistuntemusta. Tällaisen ryppään voi aiheuttaa häiriö mittausvälineessä tai uuden mittaustekniikan soveltaminen, mikä sitten osoittautuu soveltumattomaksi.

Kynnysanalyysin ja poikkeavien havaintojen tunnistamisen tueksi erilaisten robustien suotimien (filters, smoothers) soveltaminen on erittäin suositeltavaa. Ne ovat useimmiten yksinkertaisia rakenteeltaan ja nopeita soveltaa. Kirjallisuudessa on näistä tarjolla monia algoritmeja. Robustien suotimien avulla poikkeavien havaintojen analyysiä voidaan terävöittää (= outlierit saadaan selkeämmin esiin aineistosta); liputettujen havaintojen listasta voidaan sitten täsmentää virheettömät ja virheelliset tapaukset.

6 Tulokset

6.1 SV-kysely

Molempien kyselyaineistojen osalta tarkasteluissa saatiin arvokasta tietoa aineiston rakenteesta. SV-aineisto koostuu yhteensä 92 vastaajasta, joilla vähintään yksi raportoitu arvo on suurempi kuin 10 000 euroa. Laskettaessa muuttujien vastaajakohtaiset summat (rivisummat) ja laskettaessa kunkin vastaajan osuus rivisummien kokonaissummasta huomattiin, että rivisummien mukaan viisi suurinta vastaajaa muodostaa yli 50 prosenttia kokonaissummasta ja 14 suurinta yhteensä yli 75 prosenttia. Kummallekin aineistoille on laskettu niin kysymys- kuin muuttujakohtaisia perustunnuslukuja, joiden avulla voidaan esimerkiksi vertailla aineisto- ja keskenään. Molempien aineistojen rivisummien tunnusluvut sekä edellä mainitut lukumäärät on kerätty taulukkoon 3.

Aikasarja-aineistolle tehdyt ylähälytysrajoja koskevat harjoitukset antoivat taulukon 5 mukaisia tuloksia. Joukkolainojen tapauksessa tiukin ylähälytysraja aiheutti tarkasteltujen kymmenen vastaajan joukossa 0-2 liputusta vastaajaa kohti, toiseksi tiukimmalla rajalla 2-5 liputusta ja löysimmällä rajalla 3-11 hälytystä. Myös kuviot 3 ja 4 havainnollistavat ylähälytysrajan tarkasteluita.

Myös autokorrelaatioiden ja osittaisautokorrelaatioiden sekä normaalisuuden tutkiminen antoi tietoa aikasarjojen rakenteesta. Autokorreloituneisuus oli lähes kaikissa tapauksissa melko voimakasta lyhyillä viiveillä. Tämä tarkoittaa, että uuden vastauksen arvo riippuu vahvasti korreloivien tapauksessa selkeästi ainakin muutamasta edellisestä vastauksesta. Tällöin on siis perusteltua käyttää vastaajan aiemmin raportoimia lukuja uusimman luvun oikeellisuuden tarkistuksessa.

Taulukko 3. SV- ja SVA-pilottiaineistolle laskettuja tunnuslukuja 2006M12 (SV) ja 2006 (SVA) kantatietojen perusteella

	Vastaajien lkm	50 % muod. lkm	75 % muod. lkm	Rivisummat (milj. euroa) Keskiarvo	Mediaani	Maksimi	Keskihajonta	Rivisummien Summa
SV	92	5	14	1264.0	275.5	18525.5	3149.4	116287.9
SVA	302	19	47	44.5	8.9	995.3	104.7	13429.2

Taulukko 4. Kolmen eritasoisen ylähälytysrajan antamien hälytysten lukumäärä joukkolainat-muuttujalle

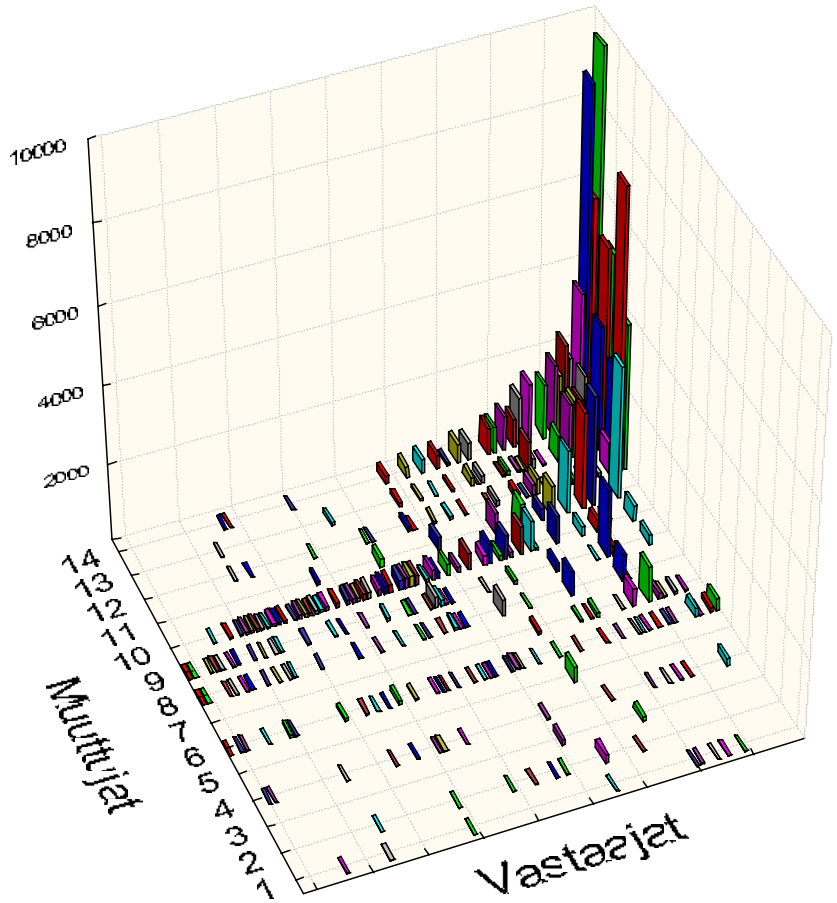
	Median	StDev	Median + 3*stdev	Hälytykset	Median + 2*stdev	Hälytykset	Median + stdev	Hälytykset
Vastaaja 1	98.0	141.1	521.2	2	380.1	4	239.1	8
Vastaaja 2	85.1	84.9	339.7	0	254.8	5	170.0	9
Vastaaja 3	125.2	155.5	591.6	1	436.1	4	280.6	9
Vastaaja 4	67.7	51.2	221.2	1	170.0	2	118.8	7
Vastaaja 6	29.9	46.3	169.0	2	122.6	3	76.3	7
Vastaaja 7	34.8	52.3	191.8	2	139.5	3	87.2	6
Vastaaja 8	74.2	71.8	289.5	0	217.7	4	146.0	11
Vastaaja 9	70.6	92.2	347.0	2	254.9	2	162.7	6
Vastaaja 11	60.6	59.9	240.3	0	180.4	3	120.5	7
Vastaaja 14	0.0	182.3	546.8	1	364.6	3	182.3	3
				11		**33**		**73**

Taulukko 5. Vastausten lukumääriä muuttujittain ja vastaajittain

		Vastausten lukumäärä muuttujittain				**Vastausten lukumäärä per vastaaja**			
	Kaikki vastaukset	**Osakkeet**	**Rahasto-osuudet**	**Joukko-lainat**	**Raha-markkina-paperit**	**Keskiarvo**	**Mediaani**	**Maksimi**	**Keski-hajonta**
SV	279	27 29 %	25 27 %	25 27 %	11 12 %	3	3	7	2
SVA	647	*60* 20 %	*74* 25 %	*55* 18 %	*18* 6 %	2	2	7	1

Tarkastelujen ohessa saatiin tietoa myös siitä, minkälaisia tyhjiä alueita havaintomatriisissa on. Poikkileikkausaineiston tapauksessa tämä tuli esiin erityisesti siinä, että käytännössä systemaattista otantaa ei voitu tehdä aineiston aukollisuudesta johtuen. Tietoa SV-poikkileikkausaineiston aukollisuudesta on kerätty taulukkoon 5. Taulukko kaikkiin muuttujiin saatujen vastausten lukumääristä löytyy liitteestä 1 (taulukko 9). Vastausintensiteetit ovat tarkasteltavien muuttujien osalta huomattavan alhaisia. Keskimäärin SV-aineiston vastaajalla oli lukuja kolmessa muuttujassa. Myös kuvio 6 havainnollistaa SV-aineiston aukollisuutta. Siitä voidaan nähdä, että erityisesti muuttujien 1-5 vastausintensiteetti on muita muuttujia pienempi.

Kuvio 6. SV-aineiston (2006M12) lukujen jakautuminen vastaajakohtaisten rivisummien mukaisessa suuruusjärjestyksessä



1 Muut saamiset, muut saamiset	**5** Muut velat, vuokrausluotot	**9** Muut saamiset, kauppaluotot	**13** Arvopaperisaamiset, rahasto-osuudet
2 Muut saamiset, lainat	**6** Muut saamiset, talletukset	**10** Muut velat, lainat	**14** Arvopaperisaamiset, joukkolainat
3 Arvopaperivelat, rahamarkkinapaperit	**7** Arvopaperisaamiset, rahamarkkinapaperit	**11** Arvopaperivelat, joukkolainat	
4 Muut velat, muut velat	**8** Muut velat, kauppaluotot	**12** Arvopaperivelat, osakkeet	

6.2 SVA-kysely

SVA-aineisto koostuu yhteensä 302 vastaajasta, joilla vähintään yksi raportoitu arvo on suurempi kuin 10 000 euroa. 19 vastaajaa muodostaa 50 prosenttia kokonaissummasta ja 75 prosenttia kokonaissummasta koostuu 47 vastaajan luvuista. Taulukosta 5 nähdään, että

SVA-aineisto on ainakin tarkastelussa olleiden muuttujien osalta SV-aineistoa aukollisempaa. Keskimäärin SVA-aineiston vastaajalla oli lukuja kahdessa muuttujassa.

Raportoidun arvon $x_{i,j}(t)$ validoinnissa voidaan SVA-kyselyssä käyttää hyväksi taulukossa 8 esitettyjä tunnuslukuja. Esimerkkitilanteessa (2) voidaan hetken *t* odotusarvoa (1) approksimoida taulukon 9 SVA-kyselyn kysymyksen hetken *t*-1 maksimiarvolla (Max). Tämä voidaan kuvata siten, että jos raportoitu arvo $x_{Arvopaperisaamiset,j}(t)$ on suurempi kuin 834,4 miljoonaa euroa, niin silloin syntyy tarve arvon $x_{Arvopaperisaamiset,j}(t)$ validoinnille ehdollisena indeksille *j* eli vastaajalle.

7 Yhteenveto ja jatkoselvityksiä

7.1 Yhteenveto

Suomen Pankille toimitettavan maksutasetilaston perusaineiston laadun arvioimiseksi kyselylomakkeisiin on implementoitu muuttujien erilaisia tarkistuksia, jotka pyrkivät eliminoimaan ristiriitaiset arvot muiden muuttujien suhteen. Kyselylomakkeisiin on mahdollista implementoida tarkistuksia sekä muuttujien, vastaajien että ajan suhteen (aikasarja-analyysi). Kehikkoperusjoukon ja tilaston estimointiin liittyvät tilastovirheet on rajattu tarkastelun ulkopuolelle.

Työn yksi tavoite oli löytää approksimaatioita odotusarvolle (1) (ks. s. 9). Pilottiraportin taulukossa 8 raportoituja arvoja voidaan käyttää odotusarvon (1) approksimointiin epäyhtälössä (2) sivulla 9 kuvatun ongelman arvioimiseksi. Jatkossa on tarve keskittyä liian pienien arvojen raportoimiseen (ongelma kuvattu epäyhtälönä (3)) suhteessa muihin vastaajiin ja muihin kysymyksiin yli ajan, sekä ongelmaan, jossa raportoitu arvo $x_{i,j}(t)$ on raportoitu virheellisesti väärän muuttujan kohdalle.

Pilottikokeisiin valittujen SV- ja SVA-kyselyn tiettyjen kysymysten ja muuttujien aineistotarkastelut osoittavat:

1. jako suuriin ja muuhun osaan vastaajista ("laaja massa") on varsin selvä
2. vastaajien jakaumat (muuttujat ja niiden aggregaatit, rivi- ja sarakesummat) ovat siten hyvin vinoja
3. systemaattinen otanta on käyttökelpoinen väline, mutta sen soveltamiselle tarvitaan aina tarpeeksi havaintoja
4. jako poikkileikkaus- ja aikasarjatarkasteluihin on tarpeellinen: tulokset täydentävät toisiaan
5. graafiset esitykset, erityisesti box-plot - ja qq-plot -kuviot ovat nopeita ja selkeitä rakenteiden ilmaisimia kynnys- ja hälytysrajojen muodostamista varten
6. valittu aineisto osoittaa, että on suositeltavaa käyttää tilastollisten menetelmien sekä klassisia että robusteja versioita; myös ei-parametrisia menetelmiä tulisi käyttää, koska aineistojen tilastollisesta jakaumasta ei tässäkään tapauksessa ole tietoa.

7.2 Jatkoselvityskohteita

Apuinformaation (-muuttujien) käyttö yksinkertaisten regressiolaskelmien avulla on erillinen kokonaisuus, joka jätettiin myöhemmän tarkastelun kohteeksi. Aineistojen substanssiasiantuntijaa tarvitaan aina tukena tässä analyysissä. On luonnollisesti suotavaa, että apumuuttuja on teoreettisesti "sukua" päämuuttujalle, vähintäänkin sen pitää olla tilastollisesti merkitsevästi korreloitunut päämuuttujan kanssa.

Jatkoselvityksistä (-piloteista) keskeinen on erilaisen apuinformaation etsiminen tukemaan etenkin poikkileikkaustapausta (*t, t*-1). Ulkopuolinen apuinformaatio voi liittyä myös tapaukseen, jossa vastaajat pyritään jakamaan kokoluokkiin, joille kullekin määritellään omat kynnys- tai hälytysrajat. Ihannetapauksessa kullekin kokoluokalle olisi käytettävissä oma apumuuttuja. Toinen selvityskohde voisi olla havaintomatriisin aukollisuuden, siis vastausintensiteetin analysointi, minkä tulokset ohjaavat systemaattisen otannan käyttöä ja vaikuttavat kynnysten laskemiseen. Esimerkiksi vastausintensiteettiä tutkittaessa havaintomatriisi on kätevää muuntaa riveittäin ja sarakkeittain suuruusjärjestysmuotoon (rearranged form). Muunnetusta matriisista näkyvät silloin informaation tihentymät, harventumat ja aukot.

Yhteistä kaikille analyyseille on tarve yksinkertaisiin, selkeisiin ja "bulkkitavaratyyppisiin" ratkaisuihin. Eri maksutasekyselyjen yhteiset rakenneominaisuudet ja suurimmat eroavuudet olisi myös hyödyllistä kartoittaa, ainakin karkealla tasolla ja taulukoida tulokset. Metadatan ohella eri kyselyprosessien paradataa (esim. miten vastaus-% ja virheiden määrä kasvaa kyselyn aikana) tulisi jatkossa kerätä systemaattisesti hyödyntämistä ajatellen.

Metodologiapuolelta yksinkertaiset aikasarjatarkastelut (suotimet ja mallit) on syytä vakioida ja sijoittaa teknisesti linkkien päässä tehtäviksi ja rajoittaa vain hyvin keskeisten vastaajien dataongelmiin. Graafisten analyysien (lähinnä uudentyyppiset ja 3D-kuviot) yhdistelmien kokeilut, myös uudelleenjärjestetyn havaintomatriisin tai sen osien tapauksessa, voivat osoittautua hyödyllisiksi.

Kirjallisuutta

Alwan, L. C. and Roberts, H. V. (1988), Time-series modeling for statistical process control. Journal of Business and Economic Statistics, 6, 87-95.

Betlehem, J. (2005), Internet Survey Developments at Statistics Netherlands. Discussion Paper 05003, Statistics Netherlands.

Granquist, L. and Kovar, J. G. (1997), Editing of survey data: how much is enough? Chapter 18 in *Survey Measurement and Process Quality*. Edited by Lyberg, L., Biemer, P., Collins, M., de Leeuw, E., Dippo, C., Schwarz, N. and Trewin, D. John Wiley & Sons, Inc.

Hella, H. and Huhtaniska, M. (2006), , Matrix procedure in business surveying, Bank of Finland, Department of the Financial Markets and Statistics, Statistics Paper 2/2006.

Hella, H. and Viertola, H. (1999), On quality control methods in business surveys, the pilot project of BOP surveys in the Bank of Finland. Bank of Finland, Report No. 1. Statistics department, working papers, 3/99.

Kume, H. (1998), *Laadun parantamisen tilastolliset menetelmät*. MET-julkaisu, toinen, korjattu painos.

Laatua tilastoissa, Tilastokeskus 2002, Käsikirjoja 43.

Lee, H. (1995), Outliers in business surveys, Chapter 26 in *Business Survey Methods* edited by Cox, B., Binder, D., Chinnappa, B., Christianson, A., Colledge, M. and Kott, P. John Wiley & Sons.

Liedes, M. ja Manninen, P. (1975), *Otantamenetelmät*, toinen korjattu painos. Gaudeamus.

Maronna, R. A., Martin, R. D. and Yohai, V. J. (2006), *Robust Statistics, Theory and Methods*. John Wiley & Sons, Ltd. Chichester, England.

Maravall, A. (2005), Data quality control with ARIMA filtering. Irving Fisher Committee (IFC) Bulletin 22, 48-52, September 2005.

Montgomery, D. C. (2001), *Introduction to Statistical Quality Control*, 4th Edition. John Wiley & Sons, Inc.

Niemi, H. , Tourunen, K. (toim.), (1996), *Tilastoista tiedoiksi*, Tilastokeskus.

Olson, J. E. (2003), *Data Quality, The Accuracy Dimension*. Morgan Kaufmann Publishers, An Imprint of Elsevier.

Raj, D. (1968), *Sampling Theory*. McGraw-Hill Book Company.

Rocke, D. M. (1989), Robust control charts. Technometrics, 31, 173-184.

Ryan, T. P. (2000), *Statistical Methods for Quality Improvement*, 2nd Edition. Wiley.

Stoumbos, Z. G., Reynolds, M. R., Jr., Ryan, T. P., and Woodall, W. H. (2002), The state of statistical process control as we proceed into the 21st century. *In Statistics in the 21st Century*, 243-255, edited by A. E. Raftery, M. A. Tanner and M. T. Wells. Chapman & Hall/CRC.

Tilastollinen laatusanasto, Glossary of statistical terms of quality control. SFS-KÄSIKIRJA 178, 1. painos, toukokuu 2007. Suomen Standardisoimisliitto

Xiao, P., Kalhok, P., and March, M. (1998), Detecting outliers in large time series databases. *Proceedings of the Survey Research Methods Section*. American Statistical Association.

Liite 1. Taulukot

Taulukko 6. Matriisitaulukko SV-poikkileikkausaineistosta (näkymä ulkomaisten arvopaperisaamisten osalta) 2006M12 kantatietojen perusteella

Vastaajat	Muuttuja a	Muuttuja b	Muuttuja c	Muuttuja d	…	**Rivisummat[3]**	**Osuus**	**Kumul. osuus**	**Kvantiilit**	**Lkm**
Vastaaja 1	3759.72	4958.08	9503.09		…	**18525.54**	15.9 %	15.9 %		
Vastaaja 2	7251.23	5160.49	5845.99		…	**18381.88**	15.8 %	31.7 %	> 25 %	2
Vastaaja 3	2668.46	3204.79	8649.16		…	**14522.41**	12.5 %	44.2 %		
Vastaaja 4	804.57	379.76	3507.83	29.86	…	**4723.51**	4.1 %	48.3 %		
Vastaaja 5					…	**4497.07**	3.9 %	52.2 %	> 50 %	5
Vastaaja 6	92.25	2041.8	1921.03	69.03	…	**4124.11**	3.5 %	55.7 %		
Vastaaja 7	43.79	1450.02	2296.83		…	**3790.64**	3.3 %	59.0 %		
Vastaaja 8	465.58	1464.75	1550.61	33.68	…	**3514.62**	3.0 %	62.0 %		
Vastaaja 9	23.67	1398.71	1829.79	134.51	…	**3390.35**	2.9 %	64.9 %		
Vastaaja 10					…	**3104.24**	2.7 %	67.6 %		
Vastaaja 11	110.66	659.91	1397.32	918.27	…	**3088.19**	2.7 %	70.2 %		
Vastaaja 12	1.56				…	**3032.44**	2.6 %	72.8 %		
Vastaaja 13					…	**2402.73**	2.1 %	74.9 %		
Vastaaja 14		24.85	1525.23	456.5	…	**2033.97**	1.7 %	76.6 %	> 75 %	14
Vastaaja 15					…	**1873.41**	1.6 %	78.3 %		
Vastaaja 16	3.56	254.57	1246.94	0.39	…	**1508.23**	1.3 %	79.6 %		
Vastaaja 17		729.57	718.29		…	**1447.86**	1.2 %	80.8 %		
Vastaaja 18	406.21	25.51			…	**1247.38**	1.1 %	81.9 %		
Vastaaja 19	279.64	69.03	868.86		…	**1236.99**	1.1 %	82.9 %		
:	:	:	:	:	…	**:**	:	:		
:	:	:	:	:	…	**:**	:	:		
SARAKESUMMA	*16516.46*	*22728.26*	*44424.73*	*1870.60*	…	**116287.89**				
KESKIARVO	*611.72*	*909.13*	*1776.99*	*170.05*	…	**1264.00**				
MEDIAANI	*64.58*	*178.13*	*718.29*	*69.03*	…	**275.49**				
MAKSIMI	*7251.23*	*5160.49*	*9503.09*	*918.27*	…	**18525.54**				
KESKIHAJONTA	*1580.33*	*1478.95*	*2551.11*	*278.77*	…	**3149.39**				
OSUUS YHTEISSUMMMASTA	14 %	20 %	38 %	2 %	…	**100 %**				

[3] Sisältää koko havaintomatriisin rivisummat.

Taulukko 7. Aikasarja-aineiston autokorrelaatiot ja osittaisautokorrelaatiot joukkolainat-muuttujalle

		Viive															
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Vastaaja 1	AC	0.82	0.67	0.54	0.41	0.23	0.09	-0.02	-0.11	-0.22	-0.30	-0.40	-0.53	-0.53	-0.50	-0.48	-0.38
	PAC	0.82	0.00	-0.02	-0.10	-0.22	-0.05	-0.02	-0.06	-0.12	-0.10	-0.19	-0.30	0.18	0.05	-0.07	0.20
Vastaaja 2	AC	0.94	0.88	0.81	0.74	0.68	0.62	0.57	0.53	0.48	0.43	0.37	0.29	0.20	0.13	0.07	0.00
	PAC	0.94	-0.09	-0.12	0.01	0.05	-0.03	0.04	-0.01	-0.08	-0.06	-0.08	-0.21	-0.09	0.06	0.01	-0.18
Vastaaja 3	AC	0.94	0.88	0.83	0.78	0.72	0.67	0.61	0.54	0.47	0.39	0.33	0.27	0.21	0.17	0.10	0.04
	PAC	0.94	-0.08	0.04	-0.05	-0.03	-0.02	-0.10	-0.09	-0.11	-0.05	0.00	0.07	-0.09	0.07	-0.17	-0.03
Vastaaja 4	AC	0.89	0.78	0.69	0.64	0.57	0.49	0.41	0.34	0.28	0.27	0.25	0.18	0.14	0.12	0.14	0.11
	PAC	0.89	-0.10	0.06	0.14	-0.14	-0.03	-0.06	-0.05	0.04	0.18	-0.11	-0.15	0.11	0.02	0.10	-0.16
Vastaaja 6	AC	0.81	0.68	0.63	0.53	0.41	0.33	0.22	0.10	0.08	0.00	-0.04	-0.04	-0.11	-0.14	-0.14	-0.18
	PAC	0.81	0.09	0.17	-0.10	-0.10	-0.04	-0.12	-0.12	0.15	-0.13	0.12	0.03	-0.16	0.01	-0.02	-0.14
Vastaaja 7	AC	0.91	0.82	0.70	0.62	0.53	0.47	0.42	0.39	0.38	0.38	0.36	0.33	0.28	0.20	0.13	0.02
	PAC	0.91	-0.05	-0.23	0.17	-0.04	-0.03	0.10	0.07	0.02	0.08	-0.09	-0.08	-0.13	-0.13	0.01	-0.26
Vastaaja 8	AC	0.92	0.79	0.69	0.60	0.49	0.40	0.33	0.26	0.17	0.12	0.13	0.12	0.07	0.01	-0.03	-0.08
	PAC	0.92	-0.28	0.11	-0.05	-0.13	0.02	0.09	-0.21	-0.03	0.22	0.15	-0.20	-0.17	-0.05	0.01	-0.06
Vastaaja 9	AC	0.85	0.66	0.48	0.29	0.10	-0.08	-0.24	-0.36	-0.43	-0.46	-0.40	-0.27	-0.11	0.06	0.15	0.24
	PAC	0.85	-0.24	-0.05	-0.17	-0.14	-0.15	-0.10	-0.07	-0.05	-0.03	0.20	0.16	0.11	0.09	-0.24	0.11
Vastaaja 11	AC	0.85	0.68	0.49	0.31	0.15	0.08	0.05	0.00	-0.02	-0.05	-0.09	-0.15	-0.14	-0.14	-0.15	-0.15
	PAC	0.85	-0.14	-0.18	-0.07	-0.08	0.19	0.02	-0.17	0.04	-0.10	0.00	-0.10	0.14	-0.04	-0.11	-0.03
Vastaaja 14	AC	0.64	0.28	0.17	0.26	0.35	0.30	0.24	0.18	0.15	0.05	-0.02	-0.06	-0.04	-0.03	-0.04	-0.07
	PAC	0.64	-0.23	0.17	0.21	0.11	0.00	0.10	-0.04	0.00	-0.17	-0.05	-0.11	0.00	-0.05	0.03	-0.03

Taulukko 8. Kysymyskohtaisia tunnuslukuja kyselyaineistoille, 2006M12

		Arvopaperisaamiset	Muut saamiset	Arvopaperivelat	Muut velat
SV	**Sarakesumma**	85540.1	4685.5	13560.8	12501.6
	Keskiarvo	2376.1	79.4	1130.1	181.2
	Mediaani	475.5	11.9	630.4	107.9
	Max	18257.7	1736.7	3425.2	790.3
	Keskihajonta	4701.0	238.0	1173.5	209.3
SVA	**Sarakesumma**	5981.1	2221.6	779.6	4447.0
	Keskiarvo	57.0	12.7	129.9	25.0
	Mediaani	11.2	1.9	44.3	4.6
	Max	834.4	497.7	570.4	497.6
	Keskihajonta	118.6	44.0	218.4	61.3

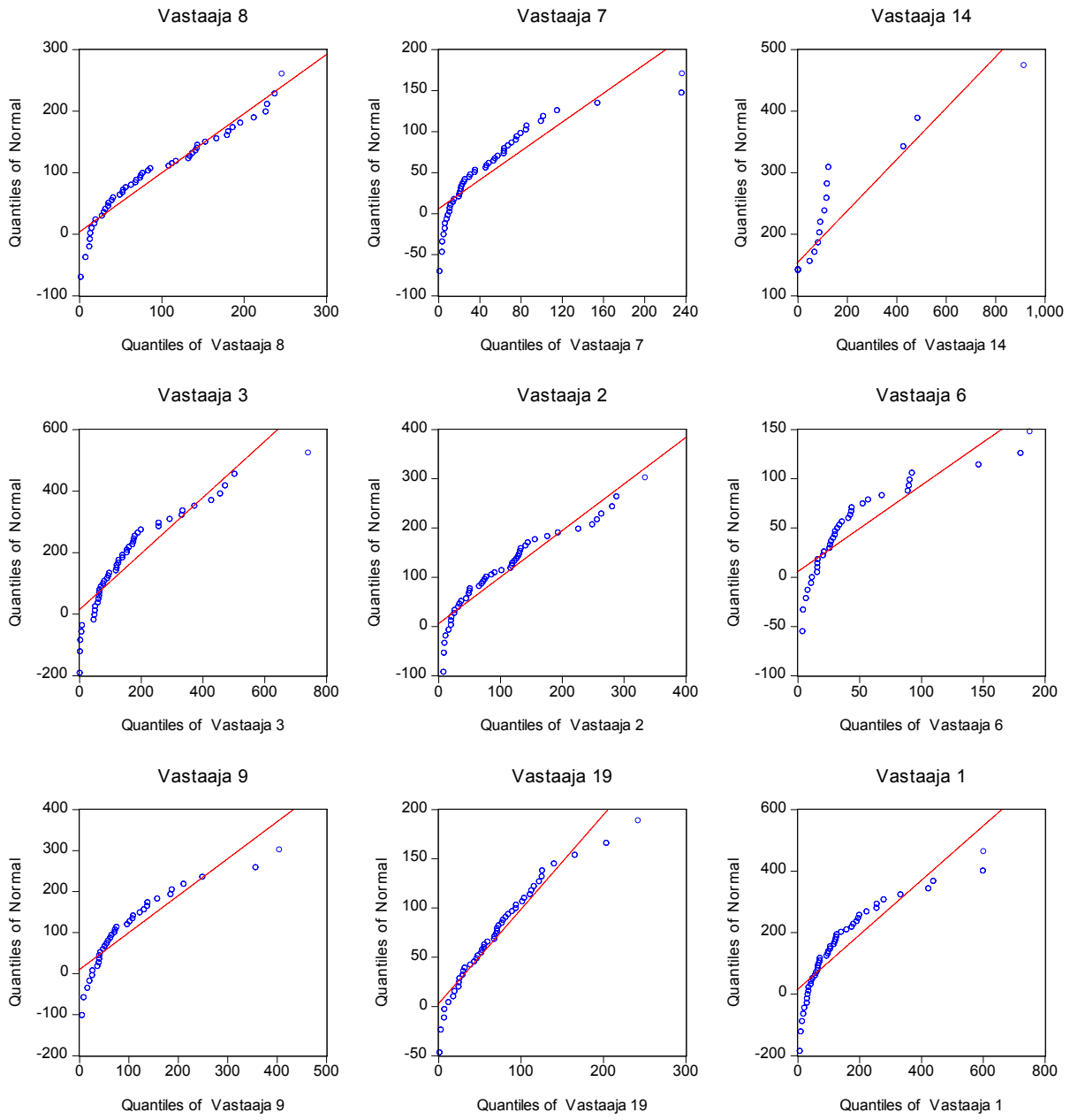
Taulukko 9. Muuttujakohtaiset vastausten lukumäärät, 2006M12

	Arvopaperisaamiset				Muut saamiset				
	Osakkeet	Rahasto-osuudet	Joukko-lainat	Raha-markkina-paperit	Lainat	Vuokraus-luotot	Talletuk-set	Kauppa-luotot	Muut saamiset
SV	27	25	25	11	8	0	38	30	9
SVA	*60*	*74*	*55*	*18*	*46*	*2*	*66*	*0*	*109*

	Arvopaperivelat		Muut velat			
	Joukko-lainat	Raha-markkina-paperit	Lainat	Vuokraus-luotot	Kauppa-luotot	Muut velat
SV	11	2	51	1	26	15
SVA	*5*	*1*	*96*	*7*	*0*	*108*

Liite 2. Kuviot

Kuvio 7. Aikasarja-aineiston normaalisuutta kuvaavat qq-plot -kuviot eräiden muutossarjojen tapauksissa

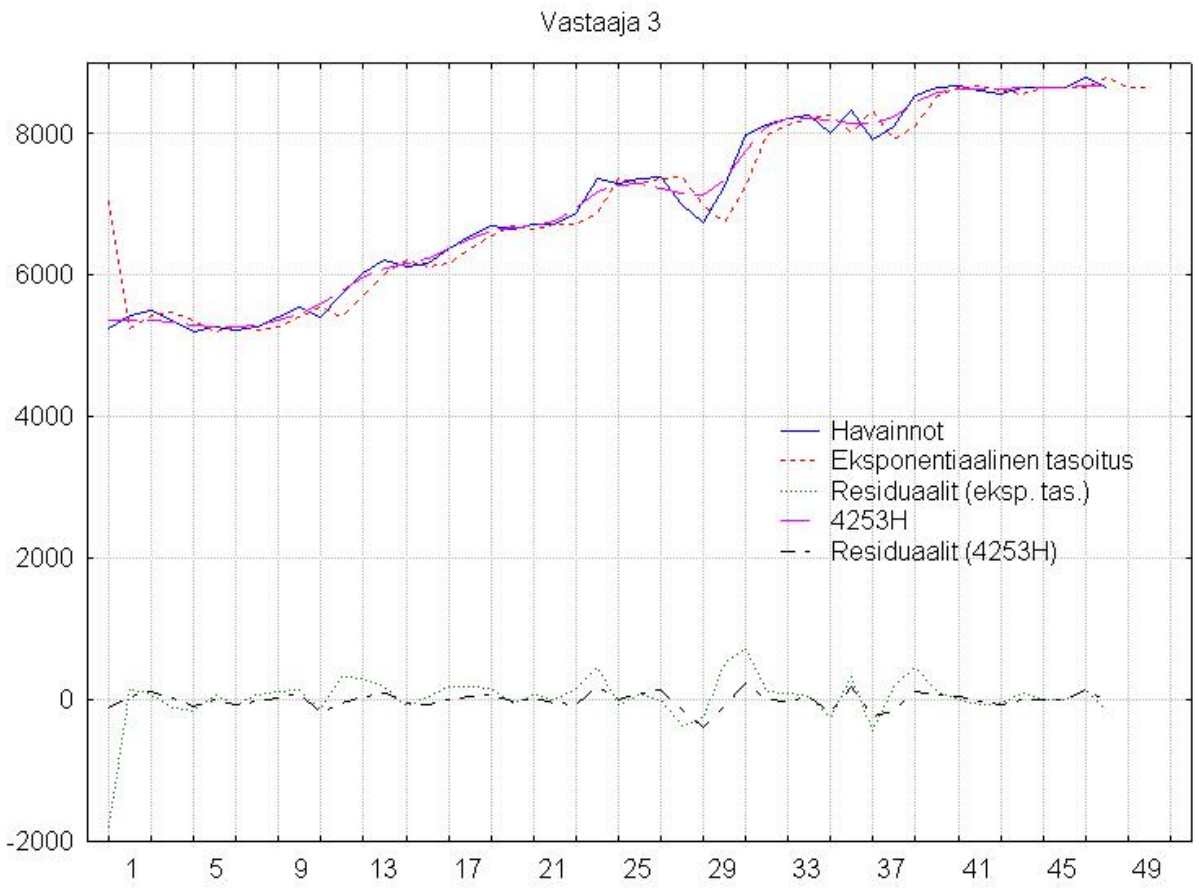


Kuvio 8. Vastaajan 3 autokorrelaatio- ja osittaisautokorrelaatioestimaatit

Date: 08/21/07 Time: 12:37
Sample: 2003M01 2006M12
Included observations: 48

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.944	0.944	45.512	0.000
		2	0.883	-0.077	86.187	0.000
		3	0.829	0.040	122.88	0.000
		4	0.775	-0.046	155.63	0.000
		5	0.720	-0.028	184.55	0.000
		6	0.667	-0.016	209.95	0.000
		7	0.606	-0.103	231.46	0.000
		8	0.539	-0.093	248.87	0.000
		9	0.465	-0.107	262.17	0.000
		10	0.392	-0.049	271.86	0.000
		11	0.325	0.004	278.73	0.000
		12	0.273	0.072	283.68	0.000
		13	0.214	-0.094	286.82	0.000
		14	0.165	0.067	288.75	0.000
		15	0.104	-0.170	289.53	0.000
		16	0.041	-0.028	289.66	0.000
		17	-0.014	-0.015	289.67	0.000
		18	-0.066	-0.045	290.02	0.000
		19	-0.101	0.110	290.87	0.000
		20	-0.126	0.013	292.23	0.000

Kuvio 9. Vastaajan 3 aikasarja sekä sen eksponentiaalisesti ja robustilla filtterillä 4253H tasoitettu sarja residuaaleineen mukaan lukien eksponentiaalisen tasoituksen ennusteita



Liite 3. Laskelmissa käytetyt toimisto- ja analyysiohjelmat

ProClarity: Aineisto haetaan kuutiosta, aineisto siirretään Exceliin.

Excel: Laskentaa, hälytystestauksia, tunnuslukuja ja kuvaajia

EViews: Kuvaajia, normaalisuuden testaus (Jarque-Bera), eksponentiaalinen tasoitus, autokorrelaatiot, qq-plot, box-plot

Statistica: Systemaattinen otanta, kuvaajia (3D), eksponentiaalinen tasoitus, autokorrelaatiot, robusti filtteri 4253H