Tilastokeskuksen liikevaihtoindeksien ennakkotietojen estimointimenetelmän kehittäminen

Heli Holtari

Tilastotieteen pro gradu -tutkielma

Jyväskylän yliopisto
Matematiikan ja tilastotieteen laitos
Kevät 2007

Tiivistelmä

Heli Holtari: *Tilastokeskuksen liikevaihtoindeksien ennakkotietojen estimointimenetelmän kehittäminen.*
Tilastoetieteen pro gradu -tutkielma, Jyväskylän yliopisto, kevät 2007.
Sivuja 54, liitteitä 3.

Tilastokeskus julkaisee Suomessa toimivien yritysten toiminnasta liikevaihtoindeksejä, jotka kuvaavat yritysten liikevaihdon kehitystä. Indeksit lasketaan teollisuuden, rakentamisen, kaupan ja muiden palveluiden toimialoilla. Indekseistä julkaistaan ennakkotiedot noin 25-45 päivää kohdekuukauden päättymisestä ja tarkentuneet tiedot noin 75 päivää kohdekuukauden päättymisestä. Ennakkotiedot perustuvat noin 2000 toimialansa merkittävimmän yrityksen otokseen.

Tutkielmassa selvitetään, onko nykyinen 2000 yrityksen otos paras mahdollinen liikevaihtoindeksien ennakkotietojen estimointiin, vai voisiko toisenlaisella otoksella saada tarkempia estimaatteja. Tutkielmassa testataan kolmea erilaista otanta-asetelmaa, joissa kussakin yrityspopulaatio ositetaan ennen varsinaista otoksen poimintaa. Otantamenetelminä ovat PPS-otanta, yksinkertainen satunnaisotanta sekä katkaistu suurten yritysten otanta. Empiirinen osa toteutetaan aineistolla, joka on laadittu Tilastokeskuksen yritysrekisteristä ja Verohallinnon maksuvalvontatiedoista. Tutkimuskohteena ovat neljän päätoimialan vuoden 2005 ennakkoindeksit.

Tutkimustulosten mukaan ositetusta aineistosta poimittu yksinkertainen satunnaisotanta antaa epätarkimmat estimaatit liikevaihtoindeksien ennakkotietoihin. Ositetun aineiston PPS-otanta toimii keskinkertaisesti; joinakin vuoden 2005 kuukausista se tuo tehokkuutta estimointiin verrattuna nykyiseen otokseen ja joinakin ei. Sen sijaan ositetun aineiston katkaistulla suurten yritysten otannalla saadaan selkeästi paremmat estimaatit kuin nykyisestä otoksesta lasketut ennakkotietojen estimaatit.

Tutkielma on tehty yhteistyössä Tilastokeskuksen kanssa.

Avainsanoja: Otanta-asetelmat, osittaminen, Tilastokeskuksen liikevaihtoindeksit, lisäinformaatio

Sisältö

1 Johdanto

Tilastokeskus julkaisee kuukausittain Suomessa toimivien yritysten toiminnasta liikevaihtoindeksejä, jotka kuvaavat yritysten liikevaihdon kehitystä kuukausittain. Indeksit lasketaan teollisuuden, rakentamisen, kaupan ja muiden palveluiden toimialoilla karkeasti toimialaluokituksen 2-numerotasoa vastaavalla toimialatarkkuudella. Yrityksen liikevaihto lasketaan ilman arvonlisäveroa, joten arvonlisäveron muutokset eivät vaikuta lasketun indeksin tasoon, ja näin ollen indeksejä voidaan pitää ajallisesti vertailukelpoisina. Indeksit sopivat hyvin eri toimialojen kehityksen vertailuun, mutta ne eivät kerro liikevaihdon euromääräisestä jakautumisesta toimialojen kesken. Tietoja käytetään Eurostatissa, Euroopan keskuspankissa, eräissä kansainvälisissä organisaatioissa ja Suomen julkisessa hallinnossa sekä aluekehittämisessä yritystalouden kehityksen seurantaan ja analysointiin. Elinkeinoelämä ja tutkimuslaitokset käyttävät tietoja markkinoiden ja kilpailijoiden kehityksen arviointiin.

Liikevaihtoindekseistä julkaistaan ennakkotiedot noin 25-45 päivää kohdekuukauden päättymisestä ja tarkentuneet tiedot noin 75 päivää kohdekuukauden päättymisestä. Ennakkotietojen laskenta perustuu Tilastokeskuksen suoraan tiedonkeruuseen, jossa on mukana noin 2000 toimialansa merkittävintä yritystä. Tarkentuneiden tietojen laskenta perustuu suoran tiedonkeruun ohella lähes koko Suomen yritystoiminnan kattavaan verohallinnon aineistoon. Tiedot ovat käytettävissä noin 60 päivää kohdekuukauden päättymisestä.

Tämän tutkielman tavoitteena on selvittää, onko nykyinen 2000 yrityksen otos paras mahdollinen liikevaihdon ennakkotietojen laskemiseen, vai voisiko toisenlaisella otoksella saada tarkempia estimaatteja. Aihe on Tilastokeskukselle erityisen kiinnostava, sillä kun ennakkotiedot on julkaistu, on kiusallista mikäli julkaistuja tuloksia joudutaan korjaamaan myöhemmin paljon varsinaisten tietojen saapuessa. Tällä hetkellä ennakkotietojen osuvuus ei ole kovinkaan tarkka ja vaihtelee suuresti toimialoittain. Tarkimmat ennakoindeksit ovat kaupan toimialalla, ja epätarkimmat rakentamisen toimialalla. Kaupan toimialalla revisio, eli se kuinka paljon ennakkoindeksi poikkeaa oikeasta indeksistä, on keskimäärin yhden prosentin luokkaa, kun taas rakentamisen toimialalla revisio on 2-3 prosenttiyksikköä.

Nykyisessä otoksessa on 2000 suurinta yritystä, joten otoksen koostu-

mus on varsin yksipuolinen, eikä siinä oteta huomioon esimerkiksi eri alueilla toimivien yritysten suhdannekehitystä ollenkaan. Ratkaisuvaihtoehtona kokeillaan yrityspopulaation osittamista ja ositettua otantaa siten, että mahdollisimman monentyyppisiä yrityksiä tulisi otokseen nykyisen 2000 suurimman sijaan. Muitakin ratkaisuvaihtoehtoja ennakkotietojen tarkentamiseen toki olisi, kuten esimerkiksi otoksen ulkopuolelle jäävien yritysten uusimman kuukauden liikevaihtotiedon ennustaminen edellisten kuukausien liikevaihtotietoja hyväksikäyttäen. Nykyinen laskentamenetelmä ei huomioi otoksen ulkopuolelle jääviä yrityksiä lainkaan, vaikka niistäkin on täydellinen liikevaihtoaineisto aiempien kuukausien osalta. Tässä tutkielmassa pitäydyttiin kuitenkin vain ensimmäisenä mainitussa ratkaisuvaihtoehdossa, koska silläkin päästiin hyviin tuloksiin, ja koska toisena mainittu ratkaisumenetelmä poikkeaa aihealueeltaan täysin ensin mainitusta.

Tutkielman toisessa luvussa tutustutaan tarkemmin Tilastokeskuksen liikevaihtoindekseihin, niiden laskemiseen, itse tutkimusongelmaan sekä sen taustaan ja tutkielman empiirisessä osassa (Luku 4) käytettävään aineistoon. Kolmas luku esittelee otantamenetelmiä ja ryhmittelyanalyysiä puhtaasti teoreettiselta kannalta. Neljännessä luvussa sovelletaan kolmannessa luvussa esiteltyjä otantamenetelmiä tutkimusongelmaan, ja vertaillaan eri otantamenetelmillä saatuja tuloksia. Viides luku koostuu tutkielman yhteenvedosta.

2 Tilastokeskuksen liikevaihtoindeksit

Tässä luvussa esitellään Tilastokeskuksen liikevaihtoindeksien laskemista. Ensimmäisessä kappaleessa kerrotaan indeksien taustasta. Kappaleessa luetellaan toimialat, joilla indeksit julkaistaan ja käydään läpi ennakollisiin indekseihin liittyviä asioita, kuten kuinka pitkällä viiveellä ne julkaistaan. Toisessa kappaleessa esitellään sekä ennakollisen indeksin että lopullisen indeksin nykyinen laskentamenetelmä, myös matemaattisin merkinnöin. Kolmas kappale käsittelee varsinaista tutkimusongelmaa, mitä tässä tutkielmassa tullaan tutkimaan ja miksi. Tämän luvun viimeisessä eli neljännessä kappaleessa käsitellään liikevaihtoindeksien laskemisessa käytetyt aineistot yksityiskohtaisesti. Lisäksi kappaleessa kuvataan tämän tutkimuksen empiirisessä osassa käytetyt aineistot.

2.1 Tausta

Tilastokeskus julkaisee Suomessa toimivien yritysten toiminnasta kuukausittain liikevaihtoindeksejä, jotka kuvaavat toimialojen liikevaihdon kehitystä. Indeksien laskennassa käytetään lähes koko Suomen yritystoiminnan kattavaa verohallinnon aineistoa sekä Tilastokeskuksen suoraa tiedonkeruuta ja yritys- ja toimipaikkarekisteriä. Tiedot julkaistaan teollisuuden, rakentamisen, kaupan ja muiden palvelujen toimialoilla karkeasti toimialaluokituksen 2-numerotasoa vastaavalla toimialatarkkuudella. Kaikki indeksisarjat alkavat vuodesta 1995 lukuunottamatta kauppaa, joka alkaa jo vuodesta 1985.

Liikevaihtoindekseistä lasketaan ennakkotiedot noin 25-45 päivää kohdekuukauden päättymisestä, sekä tarkentuneet tiedot noin 75 päivää kohdekuukauden päättymisestä. Ennakkotietojen laskenta perustuu Tilastokeskuksen suoraan tiedonkeruuseen, jossa on yhteensä noin 2000 suurinta yritystä. Tarkentuneet tiedot sen sijaan lasketaan suoran tiedonkeruun ohella verottajan arvonlisäveroaineiston kuukausittaisista yritysten liikevaihtotiedoista. Tällä hetkellä vain kaupan ennakkotiedot julkaistaan kotimaassa.

2.2 Nykyinen laskentamenetelmä

Koska kuukausittainen liikevaihtoaineisto on edellisen vuoden osalta aina täydellinen, liikevaihdon taso estimoidaan vertaamalla sitä edellisen vuoden lii-

kevaihdon tasoon nähden. Käytännössä indeksin laskenta perustuu muutoksen estimointiin. Tällä hetkellä indeksit lasketaan kaavalla

$$I_t^* = (\frac{\sum_{k=1}^{n} y_{k,t}}{\sum_{k=1}^{n} y_{k,t-12}} - 1) * I_{t-12}, \tag{2.1}$$

missä I_t^* on ennakollinen pisteluku, $\sum_{k=1}^{n} y_{k,t}$ on kyseessä olevan toimialan laskentahetkellä tunnettu, siis puutteellinen, liikevaihdon summatieto ja $\sum_{k=1}^{n} y_{k,t-12}$ on muodostettu summaamalla niiden yritysten edellisen vuoden liikevaihtotieto, joilta on saatu liikevaihtotieto $y_{k,t}$ laskentaan. Liikevaihtosummien osamäärästä vähennetään yksi, jotta saadaan liikevaihtosummien eri vuosien välinen muutos prosentteina. Lopullinen indeksi on vastaavasti

$$I_t = (\frac{\sum_{k=1}^{N} y_{k,t}}{\sum_{k=1}^{N} y_{k,t-12}} - 1) * I_{t-12}, \tag{2.2}$$

Kaavan (2.2) merkinnät ovat samat kuin edellä kaavassa (2.1), paitsi että liikevaihtotiedot ja siten indeksikin ovat lopullisia eli summat ovat yli koko toimialan perusjoukon N.

Tietoja ei voida laskea suoraan summaamalla, koska aineisto on uusimpien kuukausien osalta vajaa: koossa on vain osa yrityksistä. Menetelmässä vertaillaan tarkasteltavan kuukauden ja edellisen vuoden vastaavan kuukauden vertailukelpoisia havaintoja. Lisäksi aloittaneiden ja lopettaneiden yritysten sekä yritysjärjestelyjen vaikutus huomioidaan indeksin laskennassa. Indeksit lasketaan kuukausittain uudelleen kunnes aineisto on täysin kertynyt.

2.3 Tutkimusongelma

Tässä tutkielmassa on tarkoitus selvittää, voidaanko ennakollinen indeksi laskea tarkemmin kuin mihin edellä kuvatulla ratkaisulla päästään. Tällä hetkellä ennakkoindeksin osuvuus ei ole kovin tarkka ja vaihtelee toimialoittain. Tarkoituksena ei kuitenkaan ole kasvattaa kerättävän tiedon määrää, vaan ennakkoindeksin laskemisessa käytettäisiin edelleen samaa määrää yrityksiä, kuin mitä on suorassa tiedonkeruussa tällä hetkellä. Oleellisin menetelmä tässä tutkimuksessa tulee olemaan aineiston osittaminen. Tavoitteena on kehittää menetelmä ennakollisen indeksin (2.1) estimoimiseksi, kun käytössä on täydellinen (tai lähes täydellinen) informaatio liikevaihdon historiallisesta kehityksestä ja epätäydellinen informaatio tarkasteltavan ajankohdan kehityksestä.

Koska nykyinen ennakollisen indeksin laskentamenetelmä (2.1) ei hyödynnä muuta kuin edellisvuoden saman kuukauden historiainformaatiota, on tässä tutkielmassa tarkoitus selvittää, kuinka toimialakohtaisten yrityspopulaatioiden osittaminen ja erilaisten otantojen teko ositetussa aineistossa vaikuttaa ennakkoindeksin tarkkuuteen. Aineiston osittamisen esimerkiksi alueittain ja toimialoittain (onnistuessaan!) voidaan ajatella huomioivan jossain määrin eri alueiden ja toimialojen suhdannekehitystä. Kun otos poimitaan eri ositteista tulisi suhdannekehitys näin ollen huomioitua myös ennakkoindeksien laskennassa.

2.4 Aineistot

Liiketoiminnan kuukausikuvaajien perusjoukkona ovat liikevaihtoindeksejä laskettaessa kuukausivalvonnassa olevat arvonlisäverovelvolliset yritykset. Liiketoiminnan harjoittajista otetaan mukaan ne, jotka Tilastokeskuksen yritysrekisterin toimialakoodin mukaan tai uusimpien yritysten osalta suoraan verohallinnolta saatujen toimialatietojen perusteella kuuluvat tarkasteltaville toimialoille.

Liikevaihtoindeksien laskenta perustuu siis verohallinnon maksuvalvontatietoihin ja Tilastokeskuksen suuryritystiedusteluun. Lisäksi hyödynnetään Tilastokeskuksen yritys- ja toimipaikkarekisteriä, josta saadaan tarkempia perustietoja yrityksistä, ja alueellisia indeksejä tuotettaessa myös toimipaikkatietoja (Tilastokeskus, 2006).

Verohallinnon maksuvalvontatiedot koostuvat arvolisävero- ja työnantajasuoritustiedoista. Yleisperiaatteena on, että tavaroita ja palveluja myyvä elinkeinoharjoittaja on arvonlisäverovelvollinen yritysmuodosta riippumatta. Maksunvalvontailmoitus palautetaan verohallintoon kuukausittain siten, että arvonlisäverotiedot ilmoitetaan puolentoista kuukauden kuluttua kohdekuukauden päättymisestä. Tilastokeskuksella arvonlisäverotiedoista lasketut liikevaihtotiedot ovat valmiit indeksisarjojen laskentaa varten noin 70 päivän kuluttua kohdekuukauden päättymisestä. Tällöin aineiston kattavuus on keskimäärin 90 prosenttia yritysten liikevaihdolla mitattuna. Aineisto täydentyy kohdekuukauden osalta kuuden kuukauden ajan.

Tilastokeskuksen suoran tiedonkeruun piirissä on noin 2000 toimialojensa merkittävää yritystä. Valinnan kriteereinä ovat liikevaihdon suuruus sekä yrityksen henkilöstön määrä. Näiltä yrityksiltä kerätään kuukausittain liikevaihtotietoa. Suoran tiedonkeruun yrityksistä poistetaan säännöllisesti lopettaneet yritykset ja yritykset, joiden toiminta on vähentynyt huomatta-

vasti. Täydennystä tehdään uusilla ja toimintaansa laajentaneilla yrityksillä. Otoksen osuus kunkin päätoimialan yrityksien lukumäärästä on keskimäärin 1-2 prosenttia, mutta liikevaihtotiedoista kertyy näiden yrityksien perusteella toimialasta riippuen noin puolet kuukauden kokonaisliikevaihdosta.

Tässä tutkielmassa empiirinen osuus (Luku 4) toteutettiin neljällä toimialakohtaisella aineistolla, joka laadittiin Tilastokeskuksen yritysrekisteristä sekä Verohallinnon maksuvalvontatiedoista. Päätoimialat olivat teollisuuden, rakentamisen, kaupan ja muiden palveluiden toimialat. Kullekin toimialalle muodostettiin aluksi aineisto, jonka muuttujina olivat yritystunnus, liikevaihto (vuonna 2004), toimiala viiden numeron tarkkuudella (Tilastokeskus, 2002) ja postinumero. Lisäksi aineistoon laskettiin Verohallinnon maksuvalvontatiedoista vuoden 2004 keskimääräinen vuosimuutos verrattuna vuoteen 2003. Tällä aineistolla, joka koostui siis lähinnä yritysten "taustatiedoista", suoritettiin toimialan osittaminen toisensa poissulkeviin osajoukkoihin. Varsinaisessa indeksien- ja liikevaihdon vuosimuutosten laskemisessa käytettiin Verohallinnon maksuvalvonta-aineistoa, jossa on vuoden 2005 kuukausittaiset tiedot yrityksen liikevaihdosta sekä edellisen vuoden saman kuukauden liikevaihdosta. Aineistoista on karsittu pois kyseessä olevalla ajanjaksolla aloittaneet ja lopettaneet yritykset, fuusioituneet ja diffuusioituneet yritykset.

3 Otantamenetelmistä

Tutkielman kolmannessa luvussa esitellään otantamenetelmiä puhtaasti teoreettiselta kannalta. Otantamenetelmistä esitellään tutkimusongelman ratkaisun kannalta oleelliset menetelmät. Luku on jaettu neljään kappaleeseen, joista ensimmäinen kappale 3.1 käsittelee yksinkertaista satunnaisotantaa (*engl. Simple Random Sampling, SRS*). Alakappaleessa 3.1.1 käydään läpi kolme erilaista poimintamenettelyä, joilla yksinkertainen satunnaisotanta voidaan toteuttaa. Toinen alakappale 3.1.2 käsittelee esimointia. Siinä johdetaan perusjoukon kokonaismäärän, sen asetelmavarianssin ja keskivirheen estimaattorit palauttaen-tyyppisessä yksinkertaisessa satunnaisotannassa.

Kappale 3.2 esittelee pääpiirteissään suhteellisen otannan eli PPS-otannan. PPS on lyhenne, joka tulee englannin kielen sanoista "Probability Proportional to Size"ja tarkoittaa suomeksi otantaa otosyksikön koon mukaan. Myös PPS-otannasta käydään läpi eri poimintamenettelyjä sekä estimointia, joka tarkastellaan erikseen palauttaen-tyyppisessä ja palauttamatta-tyyppisessä otantatilanteissa. Kappaleen viimeisessä alakappaleessa tutkitaan PPS-otannan tehokkuutta suhteessa yksinkertaiseen satunnaisotantaan.

Luvun 3 kolmas kappale esittelee ryhmittelyanalyysiä ja havaintoaineiston osittamista toisensa poissulkeviin ryhmiin. Tässä kappaleessa on tarkoitus tutustua erilaisiin etäisyysmittoihin (kpl 3.3.1), joita käytetään ryhmiteltäessä havaintoaineistoa. Lisäksi tutustutaan ryhmittelymenetelmiin, jotka jaetaan tavallisesti hierarkkisiin ja epähierarkkisiin menetelmiin. Tutkielman empiirisessä osassa aineiston osittamisessa käytettiin erästä epähierarkkista ryhmittelymenetelmää, josta kerrotaan tarkemmin kappaleessa 4.1.1.

Luvun 3 viimeinen kappale 3.4 käsittelee ositetusta aineistosta tehtävää otantaa. Ensimmäisessä alakappaleessa 3.4.1 käydään läpi, miten perusjoukon kokonaismäärän ja sen asetelmavarianssin estimaattorit muodostetaan kun on kyseessä ositettu otanta. Lisäksi tutkitaan ositetun otannan tehokkuutta suhteessa yksinkertaiseen satunnaisotantaan asetelmavarianssin avulla. Toinen alakappale 3.4.2 esittelee kolme erilaista kiintiöintimenetelmää, eli sitä kuinka monta yksikköä kustakin ositteesta valitaan otokseen. Viimeisessä alakappaleessa käydään läpi ositetun otannan poimintamenettely.

3.1 Yksinkertainen satunnaisotanta

Yksinkertaista satunnaisotantaa (SRS) voidaan pitää todennäköisyyksiin perustuvien otantamenetelmien perusmuotona tilanteissa, joissa perusjoukosta ei ole saatavana ennakkotietoja. Yksinkertaisessa satunnaisotannassa jokaisella perusjoukon alkiolla on yhtäläinen todennäköisyys tulla valituksi otokseen, ja siksi valitun otoksen voidaan ajatella kuvaavan perusjoukkoa tasapuolisesti.

Yksinkertaista satunnaisotantaa käytetään periaatteessa kahdessa eri tarkoituksessa. Ensinnäkin se palvelee hyvänä vertailumenetelmänä kun halutaan verrata eri otantamenetelmien suhteellista tehokkuutta niin sanotun *asetelmakertoimen (Design Effect, DEFF)* avulla. Asetelmakertoimesta kerrotaan tarkemmin kappaleessa 3.4.1. Toiseksi yksinkertaista satunnaisotantaa voidaan käyttää lopullisena alkioiden valintametodina kehittyneemmissä otantamenetelmissä, kuten *ositetussa otannassa* (kpl 3.4), jolloin otanta-asetelmaan saadaan mukaan satunnaisuutta.

3.1.1 Poimintamenettely

Yksinkertainen satunnaisotanta voidaan toteuttaa kolmella eri valintatekniikalla; *Bernoulli otannalla (engl. Bernoulli sampling, SRSBE), palauttaen-tyyppisellä otannalla (engl. SRS with replacement, SRSWR)* sekä *palauttamatta -tyyppisellä otannalla (engl. SRS without replacement, SRSWOR)* (Lehtonen ja Pahkinen, 2004). Ensin mainitussa otoskokoa ei voida etukäteen määrittää, kahdessa jälkimmäisessä se on kiinteä. Bernoulli-otannassa ja palauttamatta-tyyppisessä otannassa otoksen valinta voidaan toteuttaa numeroimalla perusjoukon alkiot, mutta palauttaen-tyyppisessä otannassa jokainen alkio pitää arpoa perusjoukosta erikseen. Kaikissa kolmessa menetelmässä alkioiden otokseen sisältymistodennäköisyydet $\pi_k = \pi$ pysyvät vakioina kaikille perusjoukon alkioille.

Bernoulli-otannassa (SRSBE) asetetaan ensin samat sisältymistodennäköisyydet π kaikille perusjoukon alkioille. Vakion π arvo tulee määrittää siten, että odotettu otoskoko $E(n_s) = N\pi$, ja että $0 < \pi < 1$. Käytännössä valinta tehdään siten, että arvotaan kullekin perusjoukon alkiolle luku tasajakaumasta $(0, 1)$. Kaikki alkiot, joiden arvottu luku on pienempi kuin π tulevat otokseen. Metodi johtaa otoskokoon, jonka osotusarvo on siis $E(n_s) = N\pi$ ja varianssi $V(n_s) = N(1 - \pi)\pi$. Tästä seuraa ongelmia etenkin pienten otosten varianssin estimoinnissa, mutta vaihteleva otoskoko ei juurikaan haittaa

suurilla otoksilla. Bernoulli-otanta on palauttamatta-tyyppinen otantamenetelmä.

Yksinkertainen satunnaisotanta palauttaen (SRSWR) perustuu valintamenettelyyn, joka suoritetaan arpomalla perusjoukosta alkio, jonka jälkeen se palautetaan takaisin perusjoukkoon jokaisen valinnan jälkeen. Alkion valintatodennäköisyys pysyy muuttumattomana jokaisen valinnan jälkeen, ja samasta perusjoukosta erillään valitut otokset ovat toisistaan riippumattomia. Koska palauttamis-oletus yksinkertaistaa huomattavasti estimaattorien muotoa, eristyisesti varianssien estimaattoreita, otetaan se usein approksimaatioksi kun työskennellään monimutkaisempien otanta-asetelmien kanssa. SRSWR palvelee usein myös vertailumenetelmänä kun verrataan eri otantamenetelmien tehokkuutta asetelmakertoimien avulla.

Yksinkertainen satunnaisotanta palauttamatta (SRSWOR) on käytännössä yleisin yksinkertaisen satunnaisotannan metodeista. Kunkin alkion perusjoukkoon sisältymistodennäköisyys on vakio samalla poimintakerralla, mutta sisältymistodennäköisyys muuttuu riippuen siitä monesko alkio otoksesta valitaan: todennäköisyys kasvaa jokaisella valintakerralla, koska perusjoukossa on jäljellä vähemmän alkioita, joista otos valitaan. Tästä aiheutuu ongelmia laskettaessa varianssien estimaattoreita. SRSWR on tässä suhteessa helpompi. Myös palauttamatta-tyyppinen yksinkertainen satunnaisotanta voi olla vertailumenetelmänä asetelmakertoimien laskemisessa.

3.1.2 Estimointi

Tilastollisessa päättelyssä otos yleistetään kohdeperusjoukkoa koskevaksi laskemalla piste-estimaatit ja niiden luottamusvälit kiinnostaville parametreille. Usein kiinnostuksen kohteena on perusjoukon kokonaismäärä T. Ensin parametrille lasketaan piste-estimaatti, ja sen jälkeen piste-estimaatin luottamusvälit. Lisäksi voidaan suorittaa erilaisia tilastollisia testejä.

Johdetaan seuraavaksi kokonaismäärän T estimaattori $\hat{t}$ sekä sen asetelmavarianssi ja keskivirhe-estimaattori palauttamatta-tyyppisessä yksinkertaisessa satunnaisotannassa. Kokonaismäärän T estimaattori $\hat{t}$ voidaan kirjoittaa perusmuodossa

$$\hat{t} = N\bar{y} = N\sum_{k=1}^{n} y_k/n, \tag{3.1}$$

missä N on perusjoukon alkioiden lukumäärä, $\bar{y}$ on tutkittavan muuttujan y otoskeskiarvo, y_k on alkion k tutkittavan muuttujan arvo ja n on otoskoko.

Tutkielmassa tästä eteenpäin pienet kirjaimet (n ja y) viittaavat otokseen, ja isot kirjaimet (N ja Y) perusjoukkoon. Toinen tapa kirjoittaa estimaattori on

$$\hat{t} = \sum_{k=1}^{n} w_k y_k = (N/n) \sum_{k=1}^{n} y_k, \tag{3.2}$$

missä $w_k = N/n$ on otospaino. Vaihtoehtoisesti kokonaismäärän estimaattori voidaan kirjoittaa siten, että ensin määrätään sisältymistodennäköisyydet perusjoukon alkioille. Palauttamatta-tyyppisessä yksinkertaisessa satunnaisotannassa perusjoukon alkion k sisältymistodennäköisyys on $\pi_k = n/N$ tai sama vakio kaikille alkioille. Näin ollen kokonaismäärän estimaattori voidaan ilmaista yleisemmässä muodossa *Horvitz-Thompson*-tyyppisenä estimaattorina

$$\hat{t}_{HT} = \sum_{k=1}^{n} w_k y_k = \sum_{k=1}^{n} \frac{1}{\pi_k} y_k = \frac{N}{n} \sum_{k=1}^{n} y_k \tag{3.3}$$

(Lehtonen ja Pahkinen, 2004, s.25). Tässä tapauksessa estimaattorit $\hat{t}$ (3.2) ja $\hat{t}_{HT}$ (3.3) ovat samat, koska sisältymistodennäköisyydet $\pi_k = n/N$ ovat samat kaikille alkioille k. *Horvitz-Thompson*-tyyppistä estimaattoria käytetään usein esimerkiksi PPS-otannassa (kpl 3.2), jossa sisältymistodennäköisyydet vaihtelevat.

Kokonaismäärän estimaattorin $\hat{t}$ asetelmavarianssi on

$$V_{srs}(\hat{t}) = \frac{N^2}{n}(1 - \frac{n}{N}) \sum_{k=1}^{N} (Y_k - \bar{Y})^2/(N-1) = N^2(1 - \frac{n}{N})S^2/n, \tag{3.4}$$

missä $\bar{Y} = \sum_{k=1}^{N} Y_k/N$ on perusjoukon keskiarvo ja $S^2 = \sum_{k=1}^{N}(Y_k - \bar{Y})^2/(N-1)$ on perusjoukon varianssi. Harhaton estimaattori kokonaismäärän estimaattorin $\hat{t}$ asetelmavarianssille $V_{srs}(\hat{t})$ on

$$\hat{v}_{srs}(\hat{t}) = N^2(1 - \frac{n}{N}) \sum_{k=1}^{n} (y_k - \bar{y})^2/n(n-1) = N^2(1 - \frac{n}{N})\hat{s}^2/n, \tag{3.5}$$

missä $\bar{y} = \sum_{k=1}^{n} y_k/n$ on otoskeskiarvo ja $\hat{s}^2 = \sum_{k=1}^{n}(y_k - \bar{y})^2/(n-1)$ on estimaattori populaation varianssille S^2.

3.2 Suhteellinen otanta, PPS

Joskus tutkittava perusjoukko sisältää alkioita, joilla tutkittavan muuttujan arvo on huomattavasti suurempi kuin muilla alkioilla. Tällaisessa tapauksessa, etenkin estimoitaessa perusjoukon kokonaismäärää, kannattaa käyttää otantaa, jossa alkion sisältymistodennäköisyys riippuu sen koosta perusjoukossa. Tällöin suuret alkiot tulevat otokseen suuremmalla todennäköisyydellä kuin pienet.

PPS-otanta on lisäinformaatiota hyödyntävä menetelmä. Perusjoukosta tarvitaan siis tarkasteltavan muuttujan y lisäksi tieto alkioiden koosta, jota merkitään tutkielmassa tästä eteenpäin kirjaimella z (pieni kirjain z viitaa muuttujaan ja iso kirjain Z muuttujan arvoon). Oletuksena on, että lisäinformaatiomuuttujan arvo Z_k tunnetaan jokaiselle perusjoukon alkiolle k. Otannassa painotetaan suuria alkioita eli alkioita, joiden suhteellinen koko $p_k = Z_k/T_z$, missä $T_z = \sum_{k=1}^{N} Z_k$, on muihin alkioihin verrattuna suuri. Sisältymistodennäköisyys $\pi_k = n * p_k$ siis vaihtelee alkioittain.

Lisäinformaatiomuuttuja z tulisi valita siten, että sen ja tulosmuuttujan y suhde pysyisi mahdollisimman samana kaikille perusjoukon alkioille. PPS-otannan tehokkuus nimittäin riippuu siitä, onko suhde Y_k/Z_k vakio C koko perusjoukossa. Jos näin on, estimaattorin asetelmavarianssi on pieni.

PPS-otannan voi tehdä joko palauttamatta- tai palauttaen tyyppisesti. Sisältymistodennäköisyyksien laskeminen on helpompaa otannassa, joka on tehty palauttaen. Palauttamatta-tyyppisessä otannassa sisältymistodennäköisyyksien laskeminen on vaikeaa, koska perusjoukko pienenee yhdellä alkiolla jokaisen alkion valinnan jälkeen, joten jokaisen alkion valinnan jälkeen pitäisi laskea uudet sisältymistodennäköisyydet.

3.2.1 Poimintamenettely

PPS-otannan voi tehdä monella eri tavalla, joista esitellään nyt pääpiirteissään viisi; *Poisson-otanta, sequential Poisson-otanta, kumulatiivisen summan menetelmä palauttaen ja palauttamatta* sekä *systemaattinen otanta* (Lehtonen ja Pahkinen, 2004, s.50).

Poisson-otannassa lasketaan ensin sisältymistodennäköisyydet $\pi_k = n * Z_k/T_z$, jonka jälkeen arvotaan luvut $\epsilon_1, ..., \epsilon_k, ..., \epsilon_N$ väliltä $[0, 1]$. Jos $\epsilon_k < \pi_k$, alkio k tulee otokseen. Tämä proseduuri käydään läpi kaikille perusjoukon alkioille $k = 1, ..., N$. Nyt siis otoskokoa ei voida etukäteen määritellä, mutta

sen odotusarvo on $E(n_s) = \sum_{k=1}^{N} \pi_k$.

Toisinaan otoskoon satunnaisuus voi aiheuttaa ongelmia ja otoskoko pitää saada määrättyä tarkalleen etukäteen. Tällöin hyvä menetelmä on niin sanottu *vaiheittainen Poisson-otanta* (engl. *sequential Poisson-sampling*) (Teikari, 2001). Kuten Poisson-otannassakin, myös nyt ensin lasketaan sisältymistodennäköisyydet π_k ja arvotaan satunnaisluvut ϵ_k kaikille perusjoukon alkioille. Nyt kuitenkin lasketaan lisäksi "normeeratut satunnaisluvut"$\eta_k = \epsilon_k / Z_k$, missä Z_k on apumuuttujan z arvo alkiolle k. Alkio k tulee otokseen, jos $\eta_k \leq n/T_z$.

Kumulatiivisen summan PPS-otanta palauttaen on kuin yksinkertainen satunaisotanta palauttaen (SRSWR). Ero näiden kahden metodin välillä on se, että SRSWR-otannassa perusjoukon alkioille määrätään numerot väliltä [1,N], ja otanta arvotaan näistä numeroista (kullakin alkiolla on sama sisältymistodennäköisyys), mutta nyt alkioille määrätäänkin valintavälit kumulatiivisten summien $1, ..., G_k, ..., G_n$ mukaan. Kumulatiivinen summa alkiolle k on

$$G_k = \sum_{j=1}^{k} Z_j, k = 1, ..., N, G_N = T_z.$$

Seuraavaksi väliltä $[1, G_N]$ arvotaan luku. Perusjoukon alkio, jonka valintaväli sisältää tämän satunnaisluvun, valitaan otokseen. Todennäköisyys valita alkio otokseen yhdellä poiminnalla on $p_k = Z_k/T_z$. Tätä proseduuria jatketaan, kunnes otoksen n kaikki alkiot on poimittu. Otokseen sisältymistodennäköisyys alkiolle k on nyt $\pi_k = n * p_k$. Sama perusjoukon alkio voi nyt tulla valituksi otokseen useita kertoja, koska suurilla alkioilla myös niiden sisältymistodennäköisyydet ovat suuria.

Kumulatiivisen summan palauttamatta- tyyppinen PPS-otanta sisältää ongelman liittyen sisältymistodennäköisyyksien laskentaan. Valittaessa otokseen ensimmäistä alkiota perusjoukosta, sisältymistodennäköisyys on $\pi_k = p_k = Z_k/T_z$, mutta sen jälkeen tilanne hankaloituu, koska T_z pienenee ja sisältymistodennäköisyys muuttuu koko ajan. Erityisesti suurille otoksille sisältymistodennäköisyyksien laskeminen on vaikeaa, ja tämän takia on kehitetty useita vaihtoehtoisia palauttamatta-tyyppisiä otantatekniikoita.

Systemaattinen PPS-otanta on ehkä helpointa toteuttaa palauttamatta-tyyppisellä otannalla. Tässä metodissa yhdistyvät PPS-otanta ja systemaattinen otanta yhdeksi otanta-asetelmaksi. Tavallisessa systemaattisessa otan-

nassa otantaväli on $q = N/n$, mutta nyt se on $q = T_z/n$. Valitaan satunnaisluku g_0 väliltä $[1, q]$. Otoksessa tarvittavat n satunnaislukua ovat

$$g_0, g_0 + q, g_0 + 2q, ..., g_0 + (n-1)q.$$

Otokseen tuleva perusjoukon alkio on kullakin valinnalla ensimmäinen alkio listalla, jonka kumulatiivinen summa G_k on sama tai suurempi kuin kyseessä oleva satunnaisluku. Alkion k sisältymistodennäköisyys on taas $\pi_k = n * p_k$.

3.2.2 Estimointi

PPS-otannan estimointi pitää tarkastella erikseen palauttamatta-tyyppisessä ja palauttaen-tyyppisessä otanta-tilanteissa. Palauttaen-tyyppisessä otannassa todennäköisyys, että valitaan alkio yhdellä poiminnalla, on vakio. Sen sijaan palauttamatta-tyyppisessä otannassa nämä todennäköisyydet vaihtelevat alkioittain, joka aiheuttaa ongelmia etenkin varianssin estimointiin.

Alkion k suhteellinen koko p_k on palauttamatta-tyyppisessä otantatilanteessa, käyttäen perusjoukon alkioiden kokoa Z_k, on

$$p_k = \frac{Z_k}{\sum_{k=1}^{N} Z_k} = \frac{Z_k}{T_z}.$$

Suhde p_k on myös todennäköisyys, että valitaan alkio k yhdellä poiminnalla. Sisältymistodennäköisyys π_k n:n alkion otoksessa on

$$\pi_k = n * p_k = n * \frac{Z_k}{T_z}.$$

Sisältymistodennäköisyyksien π_k pitäisi täyttää ehto $\pi_k \leq 1$. Näin ei kuitenkaan aina ole, kun $n > 1$ ja jotkut Z_k:t ovat todella suuria. Tällöin $\pi_k = n * (Z_k/T_z) > 1$. Tämä ongelma on kuitenkin ratkaistavissa. Yksi keino on asettaa sisältymistodennäköisyys $\pi_k = 1$ kaikille alkioille, joille $nZ_k > \sum_{k=1}^{N} Z_k$ eli nämä alkiot otetaan otokseen varmasti. Jäljelle jääneille alkioille sisältymistodennäköisyydet π_k on suhteutettu niiden kokoon. Esimerkiksi mikäli perusjoukon alkioista vain yksi on todella suuri, alkio k', ja sille $\pi_{k*} = 1$, sisältymistodennäköisyys lopuille $N - 1$ alkioille on

$$\pi_k = (n-1)\frac{Z_k}{\sum_{k=1}^{N} Z_k - Z_{k'}}, k \neq k',$$

joka varmistaa, että tilanne $\pi_k \leq 1$ pätee.

Kaksi tunnettua PPS-otannassa käytettyä estimaattoria, jotka perustuvat todennäköisyyssuhteisiin, ovat *Horvitz-Thompson* eli HT-estimaattori sekä *Hansen-Hurwitz* eli HH-estimaattori (Lehtonen ja Pahkinen, 2004, s.53). Käyttäen PPS-otantaa palauttamatta, harhaton HT-estimaattori perusjoukon kokonaismäärälle on

$$\hat{t}_{HT} = \sum_{k=1}^{n} \frac{y_k}{\pi_k}, \tag{3.6}$$

missä π_k on sisältymistodennäköisyys. Vastaava PPS-otannan palauttaentyyppinen HH-estimaattori on

$$\hat{t}_{HH} = \frac{1}{n} \sum_{k=1}^{n} \frac{y_k}{p_k} = \frac{1}{n}(\hat{t}_1 + ... + \hat{t}_k + ... + \hat{t}_n), \tag{3.7}$$

missä kukin $\hat{t}_k = y_k/p_k$ estimoi perusjoukon kokonaismäärää T.

PPS-otannassa palauttaminen yksinkertaistaa myös asetelmavariansseja. Estimaattorille $\hat{t}_{HH}$ asetelmavarianssi on

$$V_{ppswr}(\hat{t}_{HH}) = \frac{N^2}{n} \sum_{k=1}^{N} p_k(\frac{Y_k}{Np_k} - \bar{Y})^2 = \frac{1}{n} \sum_{k=1}^{N} p_k(T_k - T)^2, \tag{3.8}$$

missä $T_k = Y_k/p_k$. Huomataan, että jos Y_k on suoraan suhteessa muuttujaan Z_k, siten että suhde Y_k/Z_k pysyy mahdollisimman samana kaikille alkioille k, asetelmavarianssi olisi nolla. Tämä olisi ideaalitilanne, mutta käytännössä ei yleinen. Harhaton estimaattori varianssille on

$$\hat{v}_{ppswr}(\hat{t}_{HH}) = \frac{N^2}{n(n-1)} \sum_{k=1}^{n} (\frac{y_k}{Np_k} - \bar{y})^2 = \frac{1}{n(n-1)} \sum_{K=1}^{n} (\hat{t}_k - \hat{t}_{HH})^2, \tag{3.9}$$

missä $\bar{y}$ on otoskeskiarvo.

3.2.3 PPS-otannan tehokkuus

PPS-otannan asetelmavarianssi $V_{ppswr}(\hat{t}_{HT})$ estimaattorille $\hat{t}_{HT}$ (kaava (3.6)) on verrannollinen kokoa mittaavan muuttujan z ja tulosmuuttujan y regressioon (Lehtonen ja Pahkinen, 2004, s.56)

$$Y_k = A + BZ_k + E_k,$$

missä $E_k, k = 1, ..., N$ on jäännöstermi. Jäännösten neliösumman ja perusjoukon varianssin suhde on

$$\frac{1}{N-1} * \sum_{K=1}^{N} (Y_k - A - BZ_k)^2 \approx S^2(1-\rho_{yz}^2),$$

missä S^2 on perusjoukon varianssi ja ρ_{yz}^2 on muuttujien y ja z neliöity korrelaatio. Jäännösten varianssi on pieni, jos korrelaatio on lähellä ± 1. PPS-otannan tehokkuutta pitäisi tutkia ylläolevan regression avulla, mutta vahva korrelaatio ρ_{yz} ei yksinään takaa tehokasta estimointia.

Yksinkertainen esimerkki, jossa voidaan tutkia PPS-otannan tehokkuutta, on yksinkertaisen satunnaisotannan palauttaen (SRSWR) ja PPS-otannan palauttaen (PPSWR) varianssien vertaaminen

$$V_{srswr}(\hat{t}) - V_{ppswr}(\hat{t}_{HT}) = N^2 Cov(z, y^2/z)/n.$$

Jos korrelaatio muuttujaparin $(z, Y^2/z)$ välillä on positiivinen, on PPS-otanta SRS-otantaa tehokkaampi. Toisaalta, kuten aiemmin todettiin, tehokkain PPS-otanta vaatii myös että suhde Y_k/Z_k pysyy vakiona kaikille alkioille. Silloin asetelmavarianssi minimoituu eli saa arvon 0.

3.3 Ryhmittelyanalyysi ja osittaminen

Ryhmittelyanalyysi pyrkii sijoittamaan havaintoyksiköt toisensa poissulkeviin ryhmiin siten, että havaintoyksiköt ryhmän sisällä ovat mahdollisimman samankaltaisia keskenään ja erilaisia toisten ryhmien havaintoyksiköiden kesken. Tässä tutkielmassa perusjoukosta eli yrityspopulaatiosta yritetään muodostaa ryhmiä eli ositteita, jotka sisältäisivät mahdollisimman samankaltaisia yrityksiä. Otos poimitaan muodostuneista ositteista, jolloin monenlaisia yrityksiä pääsee otokseen.

Ryhmittelyanalyysissä tarkasteltava havaintoaineisto X voidaan esittää $n * p - matriisina$ seuraavasti

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \dots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$$

jossa X:n rivit edustavat havaintoyksiköitä ja sarakkeet muuttujia.

Ryhmittelyanalyysin tavoitteena on tiivistää ja jäsentää havaintoaineiston sisältämää informaatiota, joten sitä käytetään data-analyyttisenä kuvausmenetelmänä. Toinen ryhmittelyanalyysin käyttötapa on asettaa lähtökohdaksi oletus, että havaintoyksiköt ovat peräisin erillisistä aliperusjoukoista. Tällöin ryhmittelyanalyysin tehtävänä on tunnistaa nämä perusjoukot havaintoaineistoon liittyvän informaation avulla ja ryhmitellä yhteen havaintoyksiköt, jotka kuuluvat samaan perusjoukkoon.

Ryhmittelyanalyysin menetelmät perustuvat johonkin havaintojen samanlaisuuden tai etäisyyden mittaan. Päämääränä on yhdistää ne havainnot, jotka ovat käytetyn etäisyysmitan suhteen riittävän lähellä toisiaan. Tavallisesti menetelmät jaetaan kahteen ryhmään; hierarkkisiin ryhmittelymenetelmiin ja epähierarkkisiin (tilastollisiin) ryhmittelymenetelmiin (Mellin, 2004).

3.3.1 Etäisyysmitat

Etäisyyteen perustuvissa ryhmittelymenetelmissä perustana on jokin metriikka. Matriisin X pisteparien muodostaman numeerisen etäisyysfunktion $d(x_i, x_j)$ sanotaan olevan metrinen, jos se toteuttaa seuraavat ehdot:

- $d(x_i, x_j) \geq 0$
- $d(x_i, x_j) = 0 \qquad jos \qquad x_i = x_j$
- $d(x_i, x_j) = d(x_j, x_i) \qquad (symmetrisyys)$
- $d(x_i, x_k) + d(x_j, x_k) \geq d(x_i, x_j)$.

Merkitään pisteparin (x_i, x_j) etäisyyttä $d(x_i, x_j)$ d_{ij}:llä. Silloin kaikkien pisteparien etäisyydet muodostavat symmetrisen matriisin

$$\mathbf{D} = \begin{pmatrix} d_{11} & d_{12} & \dots & d_{1p} \\ d_{21} & d_{22} & \dots & d_{2p} \\ \vdots & \vdots & \dots & \vdots \\ d_{n1} & d_{n2} & \dots & d_{np} \end{pmatrix}.$$

Yleisin etäisyysmitta on *Minkowskin metriikka*. Olkoot

$$x_i' = (x_{i1}, x_{i2}, ..., x_{ip})$$

ja

$$x'_j = (x_{j1}, x_{j2}, ..., x_{jp})$$

kaksi havaintoyksikköä avaruudessa R^p. Havaintoyksiköiden x'_i ja x'_j Minkowskin etäisyys voidaan määritellä muodossa

$$d_{ij} = [\sum_{l=1}^{p}(x_{il} - x_{jl})^k]^{\frac{1}{k}}, \tag{3.10}$$

jossa $k = 1, 2, ..., \infty$, ja jossa potenssi k painottaa erotuksen $(x_{il} - x_{jl})$ arvot (Mellin, 2004, s.8). Yksi Minkowskin metriikan erikoistapaus on *painotettu Euklidinen metriikka*

$$d_{ij} = [\sum_{l=1}^{p} w_l(x_{il} - x_{jl})^2]^{\frac{1}{2}}. \tag{3.11}$$

Tämän tutkielman empiirisessä osassa kappaleessa 4.1.1 käytetty SAS-ohjelmiston FASTCLUS-proseduuri perustuu euklidiseen metriikkaan. Kun $w_l = 1/s_l^2$, jossa s_l on muuttujan x_l keskihajonta, ja valitaan $k = 2$, saadaan *Pearsonin etäisyysmitta*

$$d_{ij} = [\sum_{l=1}^{p}(x_{il} - x_{jl})^2/s_l^2]^{\frac{1}{2}}. \tag{3.12}$$

Mahalanobiksen etäisyysmitta saadaan kaavasta

$$d_{ij} = [(x'_i - x'_j)'S^{-1}(x'_i - x'_j)]^{\frac{1}{2}}, \tag{3.13}$$

jossa S^{-1} on muuttujien $x_1, x_2, .., x_p$ havaituista arvoista lasketun otoskovarianssimatriisin käänteismatriisi (Mellin, 2004, s.9).

Vastaavasti voidaan muodostaa *similariteettimatriisi* C. Sen alkioille tehdään oletukset

- $c_{ij} \geq 0$
- $c_{ij} = 1 \qquad kun \qquad i = j$
- $c_{ij} = c_{ji} \qquad (symmetrisyys)$.

Similariteettikertoimet saavat arvoja välillä $[0, 1]$. Suuri etäisyys tapausten välillä merkitsee pientä similariteettiä, ja sama pätee myös kääntäen eli pieniä etäisyyksiä vastaavat suuret similariteetit. Similariteetin ja etäisyyden välinen yhteys voidaan ilmaista kaavalla

$$c_{ij} = (1 + d_{ij})^{-1}. \tag{3.14}$$

3.3.2 Hierarkkiset ryhmittelymenetelmät

Hierarkkiset ryhmittelymenetelmät perustuvat *sisäkkäisiin ryhmityksiin.* Jos tehtävänä on ryhmitellä n havaintoyksikköä H ryhmään ja ryhmien lukumäärä voi olla mikä tahansa luvuista $1, 2, ..., n$, niin

$$C_1, C_2, ..., C_H$$

on jono sisäkkäisiä ryhmityksiä. Ryhmitystä C_1, jossa on n ryhmää, ja jossa ryhmät muodostuvat yksittäisistä havainnoista, kutsutaan *heikoksi.* Ryhmitystä C_H, jossa on H ryhmää, kutsutaan *vahvaksi.* Hierarkkisessa ryhmittelyssä lähdetään liikkeelle heikosta ryhmittelystä C_1, jossa jokainen havainto on siis oma ryhmänsä. Heikosta ryhmityksestä C_1 edetään $(H-1)$ askeleella vahvaan ryhmitykseen C_H siten, että ryhmitykset $C_i, i = 1, 2, ..., H$ ovat *sisäkkäisiä* seuraavassa mielessä: Jokainen ryhmityksen $C_i, i = 2, 3, ..., H$ ryhmä on *sama* kuin ryhmityksessä C_{i-1} tai saadaan *yhdistämällä* kaksi ryhmitykset C_{i-1} ryhmistä. Askeleessa $i = 1, 2, ..., H$ valitaan yhdistettäväksi ne kaksi ryhmää, joiden *etäisyys* (Etäisyysmitat kpl 3.3.1) on käytetyn metriikan mukaan pienin. Hierarkkisen ryhmittelyn tulos esitetään tavallisesti kaaviona, *dendrogrammina*, joka on ns. puurakenne havaintoaineiston yhdistämisestä.

Hierarkkinen ryhmittely voidaan esittää seuraavana algoritmina:

1. Valitaan aloitusryhmitykseksi heikko ryhmitys C_1, jossa jokainen havainto muodostaa oman ryhmänsä. Muodostetaan ryhmitystä C_1 vastaava etäisyysmatriisi D.

2. Muodostetaan ryhmityksestä C_1 uusi ryhmitys C_{i+1} yhdistämällä ne kaksi ryhmää, joiden etäisyys on käytetyssä metriikassa pienin ja pitämällä muut ryhmät muuttumattomina.

3. Muodostetaan uusi, ryhmitystä C_{i+1} vastaava etäisyysmatriisi D.

4. Toistetaan vaiheita (2) ja (3) kunnes ollaan päästy vahvaan ryhmitykseen C_H.

Hierarkkisia ryhmittelymenetelmiä on lukuisia, joista käsitellään tarkemmin neljää yleisintä (Mellin, 2004, s.16); *lähimmän naapurin menetelmää (engl. nearest-neighbour, single-linkage method), kaukaisimman naapurin menetelmää (engl. farthest-neighbour, complete-linkage method), keskiarvomenetelmää (engl. average-linkage method)* sekä *Wardin menetelmää.*

Lähimmän naapurin menetelmässä kahden ryhmän välinen etäisyys saadaan määräämällä kaikki parittaiset etäisyydet kahden eri ryhmän havaintojen välillä, ja valitsemalla etäisyyksistä lyhin. Kaukaisimman naapurin menetelmässä sen sijaan ryhmien välinen etäisyys saadaan valitsemalla etäisyyksistä pisin, jolloin se erottelee havainnot usein hyvin erottuviksi ryhmiksi.

Keskiarvomenetelmässä kahden ryhmän välinen etäisyys saadaan määräämällä kaikki parittaiset etäisyydet kahden eri ryhmän havaintojen välillä, mutta nyt valitaan etäisyysmitaksi keskiarvo. Keskiarvomenetelmä on yleisimmin käytetty ryhmittelymenetelmä.

Wardin menetelmässä päämääränä on ryhmien homogeenisuuden maksimointi. Yhdistettäviksi valitaan kullakin askeleella ne kaksi ryhmää, joiden yhdistämisestä muodostuneen uuden ryhmän sisäinen vaihtelu kasvaa vähiten. Heikossa ryhmiteyksessä jokaisessa ryhmässä on vain yksi havainto, jolloin ryhmien sisäistä vaihtelua ei ole. Jokainen ryhmien yhdistäminen kasvattaa ryhmien sisäistä vaihtelua.

3.3.3 Epähierarkkiset ryhmittelymenetelmät

Tehtävänä on edelleen ryhmitellä n havaintoa H:hon ryhmään. Ryhmien lukumäärästä H tarvitaan jonkinlaista priori-tietoa. Epähierarkkisessa ryhmittelyssä on seuraavat vaiheet:

1. Valitaan ryhmien lukumäärä H.
2. Muodostetaan aloitusryhmitys allokoimalla jokainen havainto johonkin ryhmään.
3. Pyritään parantamaan ryhmitystä siirtämällä havaintoja ryhmästä toiseen.
4. Jatketaan parantamisyrityksiä, kunnes parantaminen ei ole enää mahdollista.

Monissa epähierarkkisen ryhmittelyn algoritmeissa aloitusryhmitys muodostetaan valitsemalla ensin H ryhmäkeskipistettä ja liittämällä sitten jokainen havainto siihen ryhmään, jonka keskipistettä lähimpänä se on. Aloitusryhmityksen ryhmäkeskipisteet voidaan valita esimerkiksi siten, että valitaan H ($H < n$) ensimmäistä havaintoa ryhmäkeskipisteiksi tai valitaan satunnaisesti H havaintoa ryhmäkeskipisteiksi. Ryhmityksen parantamiseen voidaan käyttää esimierkiksi seuraavia menetelmiä:

- Määrätään ryhmäkeskipisteet ja allokoidaan jokainen havainto siihen ryhmään, jonka keskipistettä lähimpänä se on. Kun kaikki havainnot on saatu allokoiduksi, päivitetään ryhmäkeskipisteet ja suoritetaan uusi allokointi. Havaintojen allokoinnin ja ryhmäkeskipisteiden päivittämisen vuorottelua jatketaan, kunnes uusien ja vanhojen ryhmäkeskusten etäisyydet muuttuvat vähemmän kuin etukäteen asetettu kynnysarvo.

- Määrätään ryhmäkeskipisteet ja allokoidaan jokainen havainto siihen ryhmään, jonka keskipistettä lähimpänä se on. Ryhmäkeskipisteet päivitetään välittömästi jokaisen havainnon jälkeen. Havaintojen allokoinnin ja ryhmäkeskipisteiden päivittämisen vuorottelua jatketaan, kunnes uusien ja vanhojen ryhmäkeskusten etäisyydet muuttuvat vähemmän kuin etukäteen asetettu kynnysarvo.

- Siirretään havainto ryhmästä toiseen jos siirto parantaa ryhmitystä jollakin tilastollisella kriteerillä mitattuna.

Kun ryhmityksen parantamisessa sovelletaan kahta ensin mainittua menetelmää, valitaan ryhmäkeskipisteiksi tavallisesti aloitusryhmää lukuunottamatta havaintojen ryhmäkohtaisten aritmeettisten keskiarvojen määräämät pisteet. Kahta ensimmäistä menetelmää kutsutaan usein yhteisnimityksellä *k:n keskiarvomenetelmä*. Tämän tutkielman empiirisessä osassa kappaleessa 4.1.1 käytetty SAS-ohjelmiston FASTCLUS-proseduuri perustuu k:n keskiarvomenetelmään. Kolmantena mainittussa menetelmässä tilastollisena kriteerinä sovelletaan tavallisesti jotakin ryhmien sisäistä homogeenisuutta kuvaavaa mittaa. Näitä menetelmiä kutsutaan usein *tilastollisiksi ryhmittelymenetelmiksi* (Mellin, 2004, s.23).

3.4 Otanta ositetusta aineistosta

Edellisessä kappaleessa 3.3 keskityttiin aineiston ryhmittelyn teoreettiseen tarkasteluun, mutta tämän kappaleen tarkoituksena on syventyä itse otannan tekemiseen ositetusta aineistosta. Kun yrityspopulaatio on ositettu, kuinka otos poimitaan siitä ja miten suoritetaan estimointi? Tässä kappaleessa saadaan vastaus näihin kysymyksiin.

Aineiston osittaminen tarkoittaa sitä, että perusjoukko jaetaan toisensa poissulkeviin ryhmiin eli osiitteisiin. Osittamiseen tarvitaan lisäinformaatiota perusjoukosta, ja sillä pyritään lisäämään estimoinnin tehokkuutta. Lisäinformaatiota on usein helposti saatavilla eri rekistereistä tai tietokannoista. Tyypillisiä muuttujia, joita osittamisessa käytetään, ovat sijainnilliset (läänit), väestötieteeseen liittyvät (kuten sukupuoli ja ikäryhmä) ja taloudelliset

(kuten tuloluokka) muuttujat. Kun perusjoukko on ositettu, otanta tehdään riippumattomasti kustakin ositteesta. Se, kuinka monta yksikköä kustakin ositteesta otokseen valitaan, riippuu mitä *kiintiöintimenetelmää* käytetään (kpl 3.4.2).

Ositetun otannan käytölle on olemassa kolme perussyytä. Ensinnäkin osittamisella voidaan lisätä estimoinnin tehokkuutta olennaisesti, koska kun otos poimitaan ositetusta aineistosta, saadaan otokseen mukaan kattavammin erilaisia alkioita. Toiseksi, se mahdollistaa eri otantamenetelmien käytön otannan eri osajoukoissa eli ositteissa. Joskus on hyvä tehdä tietyissä ositteissa kokonaistutkimus, kun taas toisissa ositteissa käyttää otantaa. Kolmanneksi, osittamalla perusjoukko sopivasti voidaan varmistaa, että perusjoukon kaikki halutut osajoukot saavat riittävän otoskoon. Perusjoukko saattaa nimittäin olla vinosti jakautunut eli alkiot ovat hyvin eri kokoisia, ja osittamalla perusjoukko voidaan varmistaa, että otokseen tulee valituksi sekä suuria että pieniä alkioita.

3.4.1 Estimointi ja asetelmakerroin (Design effect, DEFF)

Koska kutakin ositetta pidetään riippumattomana aliperusjoukkona, jossa otanta tehdään erikseen, estimaattorit ovat ositekohtaisten estimaattoreiden painotettuja summia, joissa painoina ovat ositepainot $W_h = N_h/N$. Esimerkiksi perusjoukon kokonaismäärän T estimaattori $\hat{t}$ on

$$\hat{t} = N * \sum_{h=1}^{H} W_h \bar{y}_h = \sum_{h=1}^{H} \hat{t}_h = \hat{t}_1 + ... + \hat{t}_h + ... + \hat{t}_H, \tag{3.15}$$

missä $W_h = N_h/N$ on ositepaino ja $\bar{y}_h = \sum_{k=1}^{n_h} y_k/n_h$ on ositteen h keskiarvoestimaattori, jonka laskeminen perustuu siihen miten alkioiden poiminta on tehty.

Koska otokset on poimittu riippumattomasti kustakin ositteesta, estimaattorin $\hat{t}$ asetelmavarianssi on ositekohtaisten varianssiestimaattoreiden summa. Esimerkiksi jos ositteissa on käytetty yksinkertaista satunnaisotantaa palauttamatta (engl. lyhenne = SRSWOR), varianssiestimaattori on

$$V_{str}(\hat{t}) = \sum_{h=1}^{H} V_{srs}(\hat{t}_h), \tag{3.16}$$

jonka harhaton estimaattori on

$$\hat{V}_{str}(\hat{t}) = \sum_{h=1}^{H} \hat{V}_{srs}(\hat{t}_h) \tag{3.17}$$

(Lehtonen ja Pahkinen, 2004, s.62).

Tarkastellaan asetelmakertoimen (DEFF) laskemista kokonaismäärän T estimoinnissa käyttäen suhteellista kiintiöintiä, jossa ositteen otoskoko on $n_h = n * W_h$ ja $n = \sum_{h=1}^{H} n_h$. Jos alkiot valitaan ositteista käyttäen yksinkertaista satunnaisotantaa palauttamatta (SRSWOR), $\hat{t}$ (3.15) on kokonaismäärän T harhaton estimaattori ja

$$V_{str}(\hat{t}) = N^2(1 - n/N)\sum_{h=1}^{H} W_h S_h^2/n \tag{3.18}$$

on $\hat{t}$:n asetelmavarianssi, missä S_h^2 on y:n varianssi ositteessa h. Vaihtoehtoisesti estimaattorin $\hat{t}$ SRSWOR-varianssi voidaan kirjoittaa ositetun otannan termein seuraavasti (kun oletetaan otoksen n olevan suuri)

$$V_{str}(\hat{t}) \doteq N^2(1 - n/N)[\sum_{h=1}^{H} W_h S_h^2 + \sum_{h=1}^{H} W_h(\bar{Y}_h - \bar{Y})^2]/n, \tag{3.19}$$

missä $\bar{Y}_h$ on perusjoukon keskiarvo ositteessa h ja $\sum_{h=1}^{H} W_h S_h^2$ on ositteiden sisäisen vaihtelun termi ja $\sum_{h=1}^{H} W_h(\bar{Y}_h - \bar{Y})^2$ on ositteiden välisen vaihtelun termi (Lehtonen ja Pahkinen, 2004, s.63). Kokonaisvarianssi on siis jaettu kahteen osaan; ositteiden sisäiseen varianssiin ja ositteiden väliseen varianssiin, joten estimaattorin $\hat{t}$ astelmakerroin DEFF on nyt

$$DEFF_{str}(\hat{t}) \doteq \frac{\sum_{h=1}^{H} W_h S_h^2}{\sum_{h=1}^{H} W_h[S_h^2 + (\bar{Y}_h - \bar{Y})^2]} \tag{3.20}$$

(Lehtonen ja Pahkinen, 2004, s.63). Asetelmakerroin on siis ositteiden sisäisen varianssin suhde kokonaisvarianssiin, ja kokonaisvarianssi on ositteiden sisäisen ja välisen varianssin summa.

Ositetun otannan tehokkuus riippuu nyt siis siitä, miten perusjoukon kokonaisvaihtelu jakautuu ositteiden kesken. Jos osittaminen jakaa perusjoukon tutkittavan ominaisuuden kannalta homogeenisiin ryhmiin, tulevat ositekohtaiset varianssit pieniksi, jolloin myös niiden painotettu summa on pieni. Silloin $0 < DEFF < 1$ eli perusjoukon ositus on tuonut tehokkuutta estimointiin verrattuna yksinkertaiseen satunnaisotantaan (SRS).

3.4.2 Otoksen kiintiöinti

Otoksen kiintiöinti määrittelee, kuinka monta yksikköä kustakin ositteesta poimitaan otokseen eli miten otos jaetaan ositteiden kesken. Tavoitteena on jakaa otos ositteiden kesken siten, että koko perusjoukkoa koskevan parametrin estimointi tulisi mahdollisimman tehokkaaksi.

Kiintiöintimenetelmän valinta vaikuttaa olennaisesti otanta-asetelman tehokkuuteen, ja siihen milloin ositettu otanta on yksinkertaista satunnaisotantaa (SRS) tehokkaampi. Yleisemmin tämä ongelma tarkoittaa sitä, kuinka suuria ositteet ovat, miten tulosmuutujan kokonaisvaihtelu jakautuu ositteiden sisäiseen ja ositteiden väliseen vaihteluun, ja kuinka tämä tulee huomioitua eri kiintiöintimenetelmissä.

On olemassa kolme kiintiöimisperiaatetta (Lehtonen ja Pahkinen, 2004, s.64):

1. *Tasakiintiöinti*

   Tasakiintiöinnissä jokaisesta ositteesta poimitaan yhtä paljon otosalkioita. Otoskoko ositteessa h on tällöin $n_{h.eq} = n/H$, jossa n on koko otos ja H ositteiden lukumäärä. Alkion k otokseen sisältymistodennäköisyys π_k vaihtelee ositteittain. Pienestä otoksesta valituksi tuleminen on todennäköisempää kuin suuresta ositteesta. Tasakiintiöinti hyödyntää vain ositteisiin jakoa. Ositteiden sisäinen informaatio tutkittavasta muuttujasta jää käyttämättä.

2. *Suhteellinen kiintiöinti*

   Suhteellinen kiintiöinti on yksinkertaisin kiintiöintimenetelmä ja käytännössä yleinen. Siinä otetaan huomioon ositteen koko N_h. Suuremmasta ositteesta poimitaan otokseen enemmän alkioita kuin pienemmästä, joten alkion k sisältymistodennäköisyys $\pi_k = n_h/N_h$ on vakio kaikissa ositteissa. Otoskoko n_h ositteessa h on nyt

   $$n_{h.pro} = n * \frac{N_h}{N} = n * W_h,$$

   missä W_h on ositepaino.

   Suhteellinen kiintiöinti jakaa otoksen ositteisiin suhteessa niiden kokoon, mutta ei välttämättä tuota parhampia estimaatteja. Koska sisältymistodennäköisyys π_k on sama kaikille alkioille, niin tilanne vastaa osite-tasolla palauttamatta-tyyppistä yksinkertaista satunnaisotantaa

(SRSWOR). Tämä ominaisuus yksinkertaistaa estimointia, eikä ositteiden sisäisiä keskiarvoja tarvitse laskea, joten suhteellista kiintiöintiä kutsutaankin *itsestään painottuvaksi*. Muissa kiintiöintimenetelmissä ei ole tätä ominaisuutta, koska sisältymistodennäköisyydet vaihtelevat ositteiden välillä.

3. *Optimaalinen kiintiöinti*

   Optimaalisessa kiintiöinnissä lähdetään liikkeelle siitä otantaa ohjailevasta yleisperiaatteesta, että tehokkaalla estimaattorilla on mahdollisimman pieni varianssi. Tätä kiintiöintimenetelmää käytetään, mikäli tutkittavan muuttujan ositekohtaiset keskihajonnat S_h (tai niiden arviot) ovat tiedossa. Nyt siis minimoidaan estimaattorin varianssi, kun otoskoko oletetaan vakioksi. Otoskoko ositteessa h on

$$n_{h.opt} = n * \frac{N_h S_h}{\sum_{h=1}^{H} N_h S_h}.$$

   Optimaalisessa kiintiöinnissä otosalkioita poimitaan enemmän ositteista, jotka ovat kooltaan muita suurempia, tai joissa on muita suurempi varianssi. Alkion k sisältymistodennäköisyys π_k voi vaihdella ositteittain.

3.4.3 Poimintamenettely

Ositetussa otannassa poiminta tehdään riippumattomasti kustakin ositteesta, joten olisi mahdollista, että niissä käytettäisiin eri otantamenetelmiä. Yleensä kuitenkin käytetään samaa menetelmää kaikissa ositteissa. Poiminta tehdään kahdessa vaiheessa. Ensin lisäinformaatiota käytetään jakamaan perusjoukko N erillisiin aliperusjoukkoihin $N_1, ..., N_h, ..., N_H$, joiden summa on N. Sen jälkeen kustakin aliperusjoukosta poimitaan kokoa $n_1, ..., n_h, ..., n_H$ oleva otos yksinkertaisella satunnaiotannalla (SRS), systemaattisella otannalla (SYS) tai PPS-otannalla.

Sisältymistodennäköisyydet riippuvat ositteessa käytetystä otantamenetelmästä. Esimerkiksi jos käytetään yksinkertaista satunnaisotantaa palauttamatta kussakin ositteessa, on sisältymistodennäköisyys vakio kunkin ositteen h sisällä oleville alkioille k. Sisältymistodennäköisyys voidaan merkitä $\pi_{hk} = n_h/N_h$, missä n_h on otoskoko ositteessa h ja N_h on ositteen h ositekoko. Jos taas ositteissa käytetään PPS-otantaa, sisältymistodennäköisyydet ovat $\pi_{hk} = n_h * (Z_{hk}/T_{hz})$, missä $T_{hz} = \sum_{k=1}^{N_h} Z_{hk}$ on apumuuttujan z kokonaismäärä ositteessa h. Tässä tapauksessa sisältymistodennäköisyys vaihtelee kullakin alkiolla.

4 Sovellukset

Luku neljä on tämän tutkielman empiirinen osa. Ensimmäisessä kappaleessa 4.1 käydään läpi aineiston osittamiseen liittyvät seikat. Osittaminen on tutkielman "punainen lanka", koska juuri sen vaikutusta haluttiin tutkia liikevaihdon ennakkoindeksien laskemisessa. Saisiko ositetulla otannalla tarkemmat ennakkoindeksit? Perusjoukkoja on empiirisessä osassa neljä; päätoimialat, ja kukin perusjoukko ositettiin erikseen. Kappaleessa 4.1 esitellään myös mitä ositusmuuttujia on käytetty, ja miksi juuri niitä. Osittaminen suoritettiin SAS-ohjelmiston FASTCLUS-proseduurilla, joka esitellään tarkemmin kappaleessa 4.1.1.

Kappaleessa 4.2 esitellään yksityiskohtaisesti kolme eri otanta-asetelmaa, joiden sopivuutta ennakkoindeksien laskemisessa on testattu. Ensimmäinen otanta-asetelma on ositetun aineiston PPS-otanta (teoriaa kappaleessa 3.2). Se esitellään kappaleessa 4.2.1. Kappaleessa on kuvailtu algoritmi, jota käytettiin PPS-otannan simuloinnin toteutuksessa. Lisäksi kappaleessa on esitelty kaavoina tilastolliset tunnusluvut, jotka kertovat otantamenetelmän onnistumisesta, sekä teollisuuden toimialan estimaatit kyseisille tunnusluvuille; Taulukko 4.1 Muiden toimialojen osalta estimaatit ovat liitteessä 1: Taulukot L.1-L.3. Kappaleen lopussa on esitelty toinen tapa, revisiointi, jolla tuloksia voidaan analysoida ja verrata suoran tiedonkeruun antamiin tuloksiin.

Kappaleessa 4.2.2 esitellään ositetun aineiston yksinkertainen satunnaisotanta. Aluksi on esitelty otanta-asetelman lähtötilanne ja simulointialgoritmi, jota sovellettiin SAS-ohjelmistolla. Teollisuuden toimialalta on estimaatit Taulukossa 4.3 ja revisiot Taulukossa 4.4. Muiden toimialojen osalta tulokset ovat liitteessä 2: Taulukot L.1-L.3. Estimaattorit ovat samat kuin PPS-otannan tulosten tarkastelussa, joten niitä ei ollut aihetta kirjoittaa tähän kappaleeseen uudelleen.

Kolmas testattu otanta-asetelma esitellään kappaleessa 4.2.3. Siinä ositetusta aineistosta poimitaan katkaistu suurten yritysten otos. Se poikkesi muista otanta-asetelmista siinä mielessä, että siinä otosta ei kannattanut simuloida perusjoukosta, koska tulos olisi ollut joka kerralla sama. Tämä siksi, että otannassa ei ole mukana satunnaisuutta, vaan otokseen poimitaan ositteiden suurimmat yritykset. Tilastollisia tunnuslukuja ei siis ollut keskiarvoa lukuunottamatta järkevää laskea, ja tuloksia tarkasteltiinkin nyt vain revisioiden kautta. Teollisuuden toimialan revisiot ovat Taulukossa 4.5 ja muiden

toimialojen vastaavat taulukot ovat liitteessä 3: Taulukot L.1-L.3.

4.1 Aineiston osittaminen

Tässä tutkielmassa perusjoukkoja on neljä; kunkin päätoimialan yrityspopulaatio. Kaupan toimialalla se koostui 35 200 yrityksestä, palveluissa 91 067 yrityksestä, teollisuudessa 28 684 yrityksestä ja rakentamisessa 27 591 yrityksestä. Yhteensä kaikilla toimialoilla yrityksiä oli 180 184.

Ennakkoindeksit lasketaan erikseen kullekin toimialalle, joten toimialat kannatti pitää omina perusjoukkoinaan, joille ositus tehtäisiin erikseen. Osittamisessa käytettiin perusjoukkojen "taustatietoaineistoa"kun taas ennakkoindeksien laskemisessa, tai pikemminkin ennakkoindeksien tarkkuusvertailussa käytössä oli aineisto, jossa on kullekin yritykselle kuukausittaiset liikevaihtotiedot. Koska "taustatietoaineisto"oli vuodelta 2004, päätettiin testattavilla otanta-asetelmilla laskettavat ennakkoindeksien tarkkuusvertailut laskea vuoden 2005 kuukausille. Tämä siksi, että myös todellisessa tilanteessa, jos uusi menetelmä otettaisiin käyttöön,ositukseen käytettävä taustatietoaineisto olisi noin vuoden vanhempaa kuin laskettavat indeksit.

Luokittelumuuttujina osittamisessa käytettiin tarkempaa toimiala-muuttujaa (Tilastokeskus, 2002) sekä postinumeroa (Tilastokeskus, 2006), joka kuvaa yrityksen fyysistä sijaintia. Voidaan ajatella, että kun otokseen valikoituu yrityksiä maantieteellisesti eri alueilta, saadaan kattavampaa tietoa toimialan kokonaisliikevaihdon kehityksestä. Aluksi osittamiseen ajateltiin käyttää myös yrityksen kokoa mittaavaa muuttujaa, esimerkiksi yrityksen vuoden 2004 liikevaihtosummaa

$$y_{k,year04} = \sum_{t=1}^{12} y_{k,t}$$

ja/tai muuttujaa, joka kuvaisi yrityksen liikevaihdon kehitystä, kuten vuoden 2004 liikevaihtosumman muutosta verrattuna vuoteen 2003

$$m_k = \frac{y_{k,year04}}{y_{k,year03}} - 1,$$

missä m_k on yrityksen k kokonaisliikevaihdon vuosimuutos vuodesta 2003 vuoteen 2004, $y_{k,year04}$ on yrityksen k kokonaisliikevaihto vuonna 2004 ja $y_{k,year03}$ vastaava vuonna 2003.

Empiiriset kokeilut kuitenkin osoittivat, että kumpaakaan muuttujaa m_k ja $y_{k,year04}$ ja ei kannata käyttää osittamiseen, koska molemmissa muuttujissa

joillakin yrityksillä muuttujien arvot ovat todella poikkeavat verrattuna koko yrityspopulaation arvoihin, ja näin ollen ne huonontaisivat ositusta, ja sen myötä otosta. Esimerkiksi jos perusjoukossa on muutamia pieniä yrityksiä, joiden liikevaihto on kaksinkertaistunut vuodesta 2003 vuoteen 2004 välisenä aikana, niin tällöin muuttujan m_k arvo on näillä yrityksillä poikkeavan suuri verrattuna muiden yritysten muuttujan m_k arvoon. Kun perusjoukko ositetaan, nämä poikkevasti kehittyneet yritykset muodostavat oman ositteensa. Kun otos sitten valitaan perusjoukon kustakin ositteesta, nämä yritykset pääsevät otokseen mukaan suuremmalla todennäköisyydellä kuin muut omissa ositteissaan, ja täten vääristävät ennakkoindeksin tarkkuutta, koska koko toimiala ei suinkaan ole kaksinkertaistanut liikevaihtoaan tänä aikana.

Koska muuttujilla oli eri mitta-asteikot, ne piti standardoida käyttäen SAS-ohjelmiston STANDARD-proseduuria. Se skaalaa muuttujat siten, että keskiarvona on nolla ja varianssina yksi. Ellei standardointia olisi tehty, olisivat muuttujat, joiden arvot vaihtelevat paljon, vaikuttaneet enemmän ositteiden muodostukseen kuin sellaiset muuttujat, joiden arvot vaihtelevat vain vähän.

Aineiston osittamiseen käytettiin SAS-ohjelmiston FASTCLUS-proseduuria (tarkempi kuvaus kappaleessa 4.1.1), jota varten pitää määritellä ositteiden maksimäärä. Muutamien empiiristen kokeilujen jälkeen päätettiin ositteiden maksimimääräksi valita 10, koska se tuntui toimivan kaikilla toimialoille siten, että yritykset jakautuivat ositteisiin suhteellisen tasaisesti eikä pieniä muutaman yrityksen ositteita päässyt syntymään. Jonkinlaisena "hyvyysmittarina"voitiin pitää FASTCLUS-proseduurin automaattisesti laskemaa selitysastetta $0 < R^2 < 1$. Mitä suurempi selitysaste, sen onnistuneempi ositus. Selitysaste ei kuitenkaan suoraan kerro sitä, tulevatko ennakkoindeksit tarkentumaan, vaan se kuvaa lähinnä sitä, kuinka etäällä havaintoaineiston yksiköt ovat toisistaan standardoiduilla ositusmuuttujilla mitattuna.

4.1.1 FASTCLUS-proseduuri

Tässä tutkielmassa käytettiin aineiston osittamiseen SAS-ohjelmiston FASTCLUS-proseduuria, sillä se tarjoaa tehokkaan metodin toisensa poissulkevien ryhmien löytämiseen ja soveltuu hyvin keskisuurien ja suurehkojen aineistojen ryhmittelyyn. Luokittelumuuttujia voi olla yksi tai useampia. Proseduuri käyttää *euklidista etäisyysmittaa* (3.11, s.19) ja ryhmien keskipisteiden etsiminen tapahtuu *pienimmän neliösumman menetelmällä*. FASTCLUS-proseduuri perustuu *iteratiiviseen uudelleen sijoitukseen* (*iterative relocation*)

ja menetelmänä on *k:n keskiarvoalgoritmi* (kappale 3.3.3) yhdistettynä *lähimmän keskipisteen lajitteluun* (*nearest centroid sorting*). Proseduuri toimii seuraavissa vaiheissa:

1. Valitaan havainnot ryhmien alkupisteiksi (cluster seeds).
2. Muodostetaan väliaikaiset ryhmät siten, että kukin havainto sijoitetaan lähintä alkupistettä vastaavaan ryhmään. Samalla päivitetään alkupiste. Tämä vaihe on valinnainen.
3. Muodostetaan ryhmät osoittamalla jokainen havainto lähimpään alkupisteeseen. Kun kaikki havainnot on sijoitettu, alkupisteet korvataan ryhmän keskiarvolla. Tämä vaihe on valinnainen ja se voidaan toistaa iteroiden.
4. Lopulliset ryhmät muodostetaan osoittamalla kukin havainto lähimpään alkupisteeseen.

4.2 Eri otannat ositetusta aineistosta

Tutkielman tarkoituksena on siis selvittää ovatko nyt suorassa tiedonkeruussa olevat yritykset optimaaliset kuvaamaan kaikkien yritysten uusimman kuukauden t liikevaihdon muutosta edellisen vuoden saman kuukauden liikevaihtoon nähden. Voidaan ajatella, että uusimman kuukauden liikevaihtoindeksin laskenta perustuu nimenomaan tämän vuosimuutoksen estimointiin. Tämän hetkinen ennakollisen indeksin laskentakaava voidaan siis kirjoittaa muodossa

$$I_t^* = g_t^* * I_{t-12},$$

missä

$$g_t^* = \frac{\sum_{k=1}^{n} y_{k,t}}{\sum_{k=1}^{n} y_{k,t-12}} - 1$$

on kuukauden t liikevaihdon vuosimuutos ja estimoidaan suoran tiedonkeruun yritysten antamien liikevaihtotietojen perusteella. Koska nyt suoran tiedonkeruun yritykset ovat kunkin toimialan suurimmat yritykset eli kyseessä on "katkaistu otanta", jossa ei ole minkäänlaista satunnaisuutta mukana, päätettiin tässä tutkielmassa kokeilla, tarkentaako otanta-astelmassa käytetty satunnaisuus osittamisen ohella ennakoindeksien tarkkuutta. Päätettiin testata kolmea erilaista otanta-asetelmaa, joissa kaikissa aineiston osittaminen on merkittävässä asemassa. Ensimmäinen kokeiltava otanta-asetelma on

ositetusta aineistosta tehtävä PPS-otanta, toinen otanta-asetelma perustuu ositetusta aineitosta tehtävään hieman muunneltuun yksinkertaiseen satunnaisotantaan ja kolmannessa otanta-asetelmassa otokseen tulevat ositetun aineiston suurimmat yritykset. Seuraavissa kappaleissa on tarkemmat kuvaukset eri otanta-asetelmista ja niiden tuottamista tuloksista.

4.2.1 Ositetun aineiston PPS-otanta

Ensimäinen testattava otanta-asetelma oli siis ositetusta aineistosta tehtävä PPS-otanta. Ensin kunkin toimialan yrityspopulaatio ositettiin SAS:n FASTCLUS-proseduurilla. Ositusmuuttujina käytettiin tarkempaa toimialaa sekä postinumeroa, jotta ositteet jakautuisivat alueittain, ja jotta kaikilta toimialoilta saataisiin yrityksiä otantaan. Yrityksen kokoa mittaavaa muuttujaa kuten liikevaihtotietoa ei PPS-otannassa kannata käyttää osittamiseen, koska suuria yrityksiä tulisi olla mahdollisimman tasaisesti kaikissa ositteissa. PPS-otannassa suuret yritykset tulevat otokseen todennäköisemmin kuin pienet. Kaikilla toimialoilla ositteiden lukumääräksi valittiin 10.

Poimintamenettelyksi valittiin *Poisson-otannan* muunnos *vaiheittainen Poisson-otanta* (engl. *sequential Poisson-sampling*) (kpl 3.2.1). Se on palautta–matta-tyyppinen poimintamenettely, jossa otoskoko on kiinteä, etukäteen määritelty luku. Otos jaettiin ositteiden kesken soveltaen suhteellista kiintiöintimenetelmää eli liikevaihdoltaan suuremmista ositteista valittiin otokseen enemmän yrityksiä kuin pienistä. Yhteensä otokseen tulleita yrityksiä oli saman verran kuin mitä on suorassa tiedonkeruussa yrityksiä, koska tässä tutkielmassa ei ollut tarkoitus kasvattaa kerättävän tiedon määrää vaan kokeilla, voiko ennakkoindeksejä laskea nykyistä tarkemmin muiden kuin suoran tiedonkeruun yritysten perusteella. Ositekohtainen otoskoko voidaan siis kirjoittaa muodossa

$$n_h = \frac{T_{hz}}{T_z} * n,$$

missä n on koko otoskoko, T_{hz} on ositteen h liikevaihtosumma ja T_z on koko toimialan liikevaihtosumma.

Ensin kaikille yrityksille laskettiin sisältymistodennäköisyydet $\pi_k = n_h * Z_{hk}/T_{hz}$. Yrityspopulaatiossa oli toimialasta riippuen jonkin verran yrityksiä, joiden sisältymistodennäköisyys oli suurempi kuin yksi. Näin voi tapahtua, jos yrityksen koko Z_{hk} on todella suuri suhteessa ositteen liikevaihtoon T_{hz}. Näille yrityksille sisältysmistodennäkösyys pakotettiin ykköseksi, ja ne

tulivat otokseen varmasti. Ne olivat siis mukana ikäänkuin omina ositteinaan.

Otos simuloitiin otoskehikosta 1000 kertaa, ja otanta-asetelman tehokkuutta tutkittiin tilastollisin tunnusluvuin. Simuloinnissa käytettiin SAS-ohjelmistoa, ja tutkielman kirjoittaja ohjelmoi itse simulointiohjelman, jonka algoritmi oli pääpiirteissään seuraava:

1. Kaikille perusjoukon alkioille arvotaan satunnaisluvut $\epsilon_1, ..., \epsilon_k, ..., \epsilon_N$ väliltä $[0, 1]$

2. Lasketaan kaikille alkioille "normeeratut satunnaisluvut"$\eta_k = \epsilon_k * T_{hz}/Z_{hk}$, missä Z_{hk} on alkion k liikevaihtotieto ja T_{hz} on ositteen h liikevaihtosumma

3. Järjestetään alkiot "normeeratun satunnaisluvun"mukaan pienimmästä suurimpaan. Kustakin ositteesta h valitaan otokseen n_h ensimmäistä alkiota.

4. Lisätään otokseen suuret yksiköt, joiden sisältymistodennäköisyys $\pi_k = 1$.

5. Lasketaan otoksesta *Horvitz-Thompson-* eli *HT-estimaatti* liikevaihdon kokonaismäärälle. PPS-otannan HT-estimaattori voidaan ositetussa aineistossa kirjoittaa muodossa

$$\hat{t}_{t,HT} = \sum_{k=1}^{n} \frac{y_{k,t}}{\pi_k} = \sum_{k=1}^{n} \frac{y_{k,t}}{n_h * (Z_{hk}/T_{hz})} = \frac{T_{hz}}{n_h} \sum_{k=1}^{n} \frac{y_{k,t}}{Z_{hk}},$$

   missä $y_{k,t}$ on yrityksen k uusimman kuukauden t liikevaihto eli tutkittavan muuttujan arvo alkiolle k, $\pi_k = n_h * (Z_{hk}/T_{hz})$ on yrityksen k otokseen sisältymistodennäköisyys, ja missä n_h on ositteen h otoskoko, Z_{hk} on yrityksen k vuoden 2004 kokonaisliikevaihto eli apumuuttujan arvo alkiolle k, ja T_{hz} on sen kokonaismäärä ositteittain.

6. Otetaan *HT-estimaatti* talteen ja aloitetaan alusta.

Tämänkaltainen liikevaihtoindeksien laskentamenetelmä poikkeaa nykyisestä menetelmästä siinä mielessä, että nyt otos "korotetaan toimialatasolle"jo ennen indeksin laskemista, kun otoksesta estimoidaan uusimman kuukauden liikevaihdon kokonaismäärä, ja vuosimuutosprosentti

$$\hat{g_{i,t}} = \frac{\hat{t}_{t,HT}}{T_{y,t-12}} - 1$$

estimoidaan jakamalla tämä estimoitu kokonaismäärä edellisvuoden vastinkuun liikevaihdon summalla ja vähentämällä siitä yksi. Muutosprosentin alaindeksi i viittaa yhteen simulointikertaan ja alaindeksi t vuoden 2005 kuukausiin $1, ..., 12$. Koska ennakkoindeksin tarkkuus riippuu vain tästä muutosprosentista, kannattaa menetelmän antamia tuloksia tarkastella muutosprosentin kautta.

Simulointituloksista estimoiduista liikevaihdon vuosimuutosprosenteista $\hat{g}_i$ voidaan tutkia esimerkiksi *Monte Carlo-hajontaa ja -harhaa* (Särndal, Svensson and Wretman, 2002, s.277). Muutosprosentin *varianssi* voidaan laskea kaavalla

$$Var(\hat{g}_t) = \frac{1}{1000}\sum_{i=1}^{1000}(\hat{g}_{i,t} - \frac{1}{1000}\sum_{i=1}^{1000}\hat{g}_{i,t})^2, i = 1, ..., 1000 \tag{4.1}$$

missä alaindeksi i viittaa simulointituloksiin. Alaindeksi t viittaa vuoden 2005 kuukausiin $1, ..., 12$. Muutosprosentin *keskihajonta* saadaan varianssin neliöjuuresta

$$s.e(\hat{g}_t) = \sqrt{Var(\hat{g}_t)}. \tag{4.2}$$

Estimoidun muutosprosentin *harhan* (*engl. bias*) laskemiseen on käytetty tiedossa olevaa liikevaihdon oikeaa muutosprosenttia

$$g_t = \frac{\sum_{k=1}^{N} y_{k,t}}{\sum_{k=1}^{N} y_{k,t-12}} - 1, k = 1, ..., N \tag{4.3}$$

missä N viittaa yrityspopulaatioon. Harha saadaan keskimääräisen simulointituloksen ja oikean vuosimuutosprentin erotuksena

$$BIAS(\hat{g}_t) = (\frac{1}{1000}\sum_{i=1}^{1000}\hat{g}_{i,t}) - g_t. \tag{4.4}$$

Tärkeä tunnusluku on myös estimoitujen vuosimuutosprosenttien *keskineliöpoikkeama* (*engl. mean square error, MSE*)

$$MSE(\hat{g}_t) = \frac{1}{1000}\sum_{i=1}^{1000}(\hat{g}_{i,t} - g_t)^2 = Var(\hat{g}_t) + BIAS(\hat{g}_t)^2 \tag{4.5}$$

ja sen neliöjuuri

$$\sqrt{MSE(\hat{g}_t)} = \sqrt{Var(\hat{g}_t) + BIAS(\hat{g}_t)^2}. \tag{4.6}$$

Simulointituloksista lasketaan myös keskineliöpoikkeaman *variaatiokerroin* $CV(MSE(\hat{g}_t))$, joka on hajontaluku, joka suhteuttaa estimoidun muutosprosentin keskineliöpoikkeaman neliöjuuren estimoituun muutosprosenttiarvoon

$$CV(MSE(\hat{g}_t)) = \frac{\sqrt{MSE(\hat{g}_t)}}{\hat{g}_t} * 100. \tag{4.7}$$

Keskineliöpoikkeaman variaatiokerroin kuvaa estimoidun muutosprosentin suhteellista hajontaa, ja sen avulla on mielekkäämpää vertailla kahta eri otanta-astelmaa. Mitä suurempi luku itseisarvoltaan on, sitä suurempi epävarmuus otanta-asetelmaan liittyy, eli kun variaatiokerroin on suuri, voi yhdellä simulointikerralla tulla todennäköisemmin otokseen valituksi sellaisia yrityksiä, joiden liikevaihdon vuosimuutosprosentti poikkeaa paljon oikeasta vuosimuutoksesta.

Seuraavassa taulukossa on teollisuuden toimialan simulointituloksista estimoitujen muutosprosenttien tuloksia. Tulokset on laskettu kuukausittain vuodelle 2005, koska aineiston osittamisessa ja otannassa käytetty lisäinformaatio on vuoden 2004 yritystiedoista. Teollisuuden toimialan perusjoukkona on 26326 yritystä ja otoskoko n on 550. Suuria yrityksiä, joiden sisältymistodennäköisyys $\pi_k = 1$, on 62 kappaletta. Kolmen muun päätoimialan vastaavat taulukot ovat liitteessä 1: Taulukot L.1-L.3.

Teollisuus, PPS-otanta						
Kuukausi	g_t	$\hat{g}_t$	$s.e.(\hat{g}_t)$	$BIAS(\hat{g}_t)$	$\sqrt{MSE(\hat{g}_t)}$	$CV(MSE(\hat{g}_t))$
01	0.1068	0.1062	0.00941	-0.00062	0.00942	8.8737
02	0.0840	0.0826	0.00890	-0.00140	0.00906	10.9775
03	0.0344	0.0342	0.01011	-0.00016	0.01010	29.5470
04	0.0767	0.0774	0.00984	-0.00071	0.00986	12.7386
05	0.0453	0.0453	0.01088	-0.00003	0.01087	24.0094
06	-0.0756	-0.0749	0.01054	0.00069	0.01055	-14.0966
07	0.0053	0.0048	0.01095	-0.00048	0.01096	228.5807
08	0.0766	0.0762	0.01056	-0.00048	0.01056	13.8653
09	0.0860	0.0836	0.01002	-0.00242	0.01031	12.3425
10	0.0065	0.0064	0.00981	-0.00006	0.00980	153.2877
11	0.0480	0.0470	0.01039	-0.00096	0.01043	22.1809
12	0.0122	0.0118	0.01257	-0.00041	0.01257	106.9843

Taulukko 4.1: Teollisuuden toimialan tunnuslukuja PPS-otannalla. Liikevaihdon oikea muutosprosentti kyseessä olevana kuukautena verrattuna edellisvuoteen, estimoitu muutosprosentti, sen keskihajonta, harha ja keskineliöpoikkeaman neliöjuuri, sekä keskineliöpoikkeaman variaatiokerroin.

Oikeat liikevaihdon vuosimuutosprosentit (g_t) ja estimoidut muutosprosentit ($\hat{g}_t$) ovat hyvin lähellä toisiaan. Keskihajonnat ($s.e.(\hat{g}_t)$) vaikuttavat kovin pieniltä, mutta pitää huomioida, että estimoidut muutosprosentit ovat alle ykkösen, ja niille lasketut hajontaluvut ovat suhteessa niiden kokoon. Harhat ($BIAS(\hat{g}_t)$) ovat lähes olemattomia. Itse *HT-estimaattorihan* on harhaton, joten nämä minimaaliset harhat johtuvat siitä että simulointikertoja oli vain 1000. Mikäli simulointikertoja olisi lisätty, olisi harha todennäköisesti miltei kadonnut. Estimoitujen muutosprosenttien keskineliöpoikkeaman neliöjuuret ($\sqrt{MSE(\hat{g}_t)}$) ovat lähes samoja kuin keskihajonnat, johtuen tämäkin minimaalisesta harhasta ja g_t:n pienistä arvoista.

Keskineliöpoikkeaman variaatiokertoimet ($CV(MSE(\hat{g}_t))$) sen sijaan vaihtelevat paljon. Arvoja, joiden itseisarvo on suurempi kuin 20, voidaan pitää huonoina, sillä tällöin otanta-asetelmaan liittyy suuri epävarmuus. Yhden poimitun PPS-otannan tulos saattaa siis antaa hyvinkin vääristyneitä muutosprosentteja. Huomattavasti 20:tä suurempia kuukausia ovat heinäkuu (07), lokakuu (10) ja joulukuu (12), joilla kaikilla variaatiokerroin on sataa suurempi. Negatiivinen variaatiokerroin on vain silloin kun liikevaihdon oikea muutosprosentti $\hat{g}$ on negatiivinen kuten nyt kesäkuu (06).

Toinen tapa tutkia PPS-otannan onnistumista, on sen tulosten vertaaminen suoran tiedonkeruun antamiin tuloksiin. Tällöin lasketaan ensin kun-

kin simulointikerran tuottamista vuosimuutoksista *revisiot* kuukausittain eli kuinka paljon estimoidut kuukausittaiset vuosimuutokset poikkeavat oikeista vuosimuutosarvoista, ja lasketaan näistä sitten keskiarvo

$$r_{t,pps} = \frac{\sum_{i=1}^{1000} |g_t - \hat{g_{i,t}}|}{1000} \tag{4.8}$$

missä alaindeksi i viittaa simulointituloksiin. Alaindeksi t viittaa taas vuoden 2005 kuukausiin $1, ..., 12$. Yksinkertaistettuna revisio $r_{t,pps}$ kertoo, kuinka paljon yhden simulointikerran estimoitu muutosprosentti eroaa oikeasta muutosprosentista keskimäärin. Samat revisiot lasketaan suoran tiedonkeruun yrityksistä lasketuille vuosimuutosprosenteille

$$r_t^* = |g_t - g_t^*|, t = 1, ..., 12. \tag{4.9}$$

Nyt ei tietenkään pystytä laskemaan *keskiarvoa*, koska tätä otosta ei ole simuloitu, vaan sitä käytetään vertailukohteena.

Seuraavassa taulukossa on teollisuuden toimialan suoran tiedonkeruun yrityksistä lasketun liikevaihdon vuosimuutosprosenttien revisiot oikeisiin vuosimuutoksiin nähden r_t^*, sekä pps-otannan tuottamien muutosprosenttien revisiot oikeisiin arvoihin nähden $r_{t,pps}$. Ne on jälleen laskettu vuoden 2005 kuukausille. Muiden toimialojen vastaavat taulukot löytyvät liitteestä 3: Taulukot L.1-L.3.

Teollisuus, PPS-otanta		
Kuukausi	r_t^*	$r_{t,pps}$
01	0.0233	0.0073
02	0.0024	0.0072
03	0.0258	0.0074
04	0.0028	0.0078
05	0.0199	0.0080
06	0.0124	0.0086
07	0.0016	0.0081
08	0.0217	0.0082
09	0.0050	0.0082
10	0.0434	0.0075
11	0.0297	0.0080
12	0.0319	0.0098

Taulukko 4.2: Teollisuuden toimialan revisiot PPS-otannalla

Taulukosta nähdään, ovatko PPS-otannalla saadut liikevaihdon vuosimuutokset lähempänä todellisia vuosimuutosarvoja kuin suoran tiedonke-

ruun yrityksistä saadut vuosimuutosarvot. Mitä pienempi revisio, sitä lähempänä todellista vuosimuutosprosenttia otoksesta laskettu vuosimuutosprosentti on. Näin ollen kahdeksana kuukautena kahdestatoista (8/12) PPS-otanta olisi tarkempi ennakkoindeksejä laskettaessa, kuin suoran tiedonkeruun yritykset.

4.2.2 Ositetun aineiston yksinkertainen satunnaisotanta

Toinen otanta-asetelma, jota testattiin, oli *palauttamatta-* tyyppinen *yksinkertainen satunnaisotanta* (*engl. Simle Random Sampling Without Replacement, SRS WOR*). Otanta-asetelmaa päätettiin muokata siten, että kaikkein suurimmat yritykset tulisivat otokseen varmasti todennäköisyydellä 1. Suurimpia yrityksiä ovat ne, joiden henkilöstömäärä vuonna 2004 oli suurempi tai yhtäsuuri kuin 250. Nämä suuret yritykset poistettiin perusjoukosta, ja jäljelle jäänyt yrityspopulaatio ositettiin samalla tavalla kuin PPS-otannassakin, jotta tuloksia voitaisiin vertailla myöhemmin. Ositus tehtiin siis SAS:n FASTCLUS-proseduurilla, ja ositusmuuttujina olivat tarkempi toimiala-muuttuja ja postinumero. Ositteita muodostettiin 10. Kuten PPS-otannassakin, suuria yrityksiä käsiteltiin omina ositteinaan.

Ennen varsinaista otoksen valintaa otos piti vielä kiintiöidä eri ositteiden kesken. Kuten PPS-otannassakin, nytkin käytettiin hieman muunneltua suhteellista kiintiöintimenetelmää, eli kukin osite saa niin paljon otosalkioita, kuin siinä ositteessa on liikevaihtoa suhteessa koko toimialaan

$$n_h = \frac{T_{hz}}{T_z} * n,$$

missä n_h on ositteen h otoskoko, T_{hz} on sen ositteen yritysten kokonaisliikevaihto vuonna 2004, T_z on koko toimialan yritysten kokonaisliikevaihto vuonna 2004 ja n on koko otoskoko. Koko otoskoosta on vähennetty pois suurten yritysten lukumäärä, koska niitä ei kiintiöidä, vaan ne tulevat otokseen joka tapauksessa ja omina ositteinaan.

Otos simuloitiin perusjoukosta, josta oli aiemmin poistettu suuret yritykset 1000 kertaa. Simuloinnissa käytettiin SAS-ohjelmistoa, ja tutkielman kirjoittaja ohjelmoi itse simulointiohjelman yksinkertaiselle satunnaisotannalle, jonka algoritmi on pääpiirteissään seuraava:

1. Kaikille perusjoukon alkioille arvotaan satunnaisluvut $\epsilon_1, ..., \epsilon_k, ..., \epsilon_N$ väliltä $[0, 1]$

2. Järjestetään satunnaisluvut pienimmästä suurimpaan. Kustakin ositteesta h valitaan otokseen n_h ensimmäistä alkiota.

3. Lisätään otokseen suuret yritykset.

4. Lasketaan otoksesta *Horvitz-Thompson-* eli *HT-estimaatti* liikevaihdon kokonaismäärälle. Yksinkertaisen satunnaisotoksen HT-estimaattori voidaan ositetussa aineistossa kirjoittaa muodossa

$$\hat{t}_{t,HT} = \sum_{k=1}^{n} \frac{N_h}{n_h} * y_{k,t},$$

missä y_k on yrityksen k uusimman kuukauden t liikevaihto eli tutkittavan muuttujan arvo alkiolle k, N_h on alkioiden lukumäärä ositteessa h, ja n_h on ositteen h otoskoko. Suurille yrityksille $N_h = n_h = 1$, joten niiden uusimman kuukauden liikevaihtotieto tulee "sellaisenaan"ilman kertoimia.

5. Otetaan *HT-estimaatti* talteen ja aloitetaan alusta.

Yksinkertaisen satunnaisotannan tuloksia voidaan vertailla ja analysoida kuten aiemmin kuvatussa PPS-otannassakin (kappale 4.2.2), eli lasketaan *Monte Carlo-hajontaluvut* (kaava (4.2)) ja -*harhat* (kaava (4.4)) sekä *keskineliöpoikkeaman neliö* (kaava (4.6)) ja sen *variaatiokerroin* (kaava (4.7)). Seuraavassa taulukossa on esitetty nämä tunnusluvut laskettuna teollisuuden toimialalle, joka koostuu siis (kuten PPS-otannassakin) 26 326 yrityksestä, ja jossa suuria yrityksiä on 217. Koko otoskoko n on 550. Tunnusluvut on jälleen laskettu vuoden 2005 kuukausille. Muiden toimialojen vastaavat taulukot ovat liitteessä 2: Taulukot L.1-L.3.

Teollisuus, Yksinkertainen satunnaisotanta						
Kuukausi	g_t	$\hat{g}_t$	$s.e.(\hat{g}_t)$	$BIAS(\hat{g}_t)$	$\sqrt{MSE(\hat{g}_t)}$	$CV(MSE(\hat{g}_t))$
01	0.1068	0.1082	0.1599	0.0014	0.1598	147.6401
02	0.0840	0.0850	0.1632	0.0011	0.1632	191.9059
03	0.0344	0.0334	0.1641	-0.0007	0.1641	487.1089
04	0.0767	0.0767	0.1542	-0.0000	0.1542	200.9827
05	0.0453	0.0464	0.1517	0.0011	0.1516	326.9734
06	-0.0756	-0.0746	0.1225	0.0010	0.1224	-164.1292
07	0.0053	0.0068	0.1759	0.0015	0.1758	2592.5799
08	0.0766	0.0798	0.1965	0.0032	0.1964	246.0554
09	0.0860	0.0870	0.1667	0.0010	0.1666	191.5231
10	0.0065	0.0134	0.2211	0.0069	0.2211	1653.7455
11	0.0480	0.0497	0.1663	0.0017	0.1662	334.5206
12	0.0122	0.0172	0.1899	0.0050	0.1899	1104.2991

Taulukko 4.3: Teollisuuden toimialan tunnuslukuja SRS-otannalla. Liikevaihdon oikea muutosprosentti kyseessä olevana kuukautena verrattuna edellisvuoteen, estimoitu muutosprosentti, sen keskihajonta, harha ja keskineliöpoikkeaman neliöjuuri, sekä keskineliöpoikkeaman variaatiokerroin.

Oikeat liikevaihdon muutosprosentit (g_t) ja estimoidut muutosprosentit ($\hat{g}_t$) näyttäisivät olevan yhtä lähellä toisiaan kuin PPS-otannassakin. Keskihajonnat ($s.e.(\hat{g}_t)$) sen sijaan on aivan eri luokkaa kuin PPS-otannassa, nyt arvot ovat huomattavasti suurempia. Yksinkertaisessa satunnaisotannassa, vaikka mukana onkin perusjoukosta katkaistu suurten yritysten osa, tulee otokseen paljon pieniä yrityksiä, jotka voivat suuria helpommin käyttäytyä "oudosti"eli liikevaihto saattaa vaihdella paljonkin vuoden sisällä, ja tällöin otos antaa väärän kuvan koko toimialan kehityksestä.

Harhat ($BIAS(\hat{g}_t)$) ovat jonkin verran suurempia kuin PPS-otannassa, mutta eivät merkittävästi. Harhat olisivat myös nyt kadonneet miltei kokonaan, jos simulointikertoja olisi lisätty, koska tämäkin menetelmä on harhaton. Keskineliöpoikkeaman neliöjuuret ($\sqrt{MSE(\hat{g}_t)}$) ovat nytkin lähes samat kuin keskihajonnat, koska harhat ovat niin pieniä. Keskineliöpoikkeaman variaatiokertoimet sen sijaan ($CV(MSE(\hat{g}_t))$) ovat todella suuria, kaikki ovat reilusti $>$20. Yksi yksinkertainen satunnaisotos antaa siis suurella todennäköisyydellä vuosimuutosprosentit, jotka ovat kaukana todellisista arvoista.

Kuten PPS-otannassakin, myös yksinkertaisessa satunnaisotannassa voi ja kannattaa estimoituja vuosimuutosprosentteja verrata suoran tiedonkeruun yrityksistä saatuihin muutosprosenttehin. Tässäkin tapauksessa se tapahtuu *revisioita* tarkastelemalla, jotka voidaan laskea kuten PPS-otannan

tapauksessa (kaava (4.8)). Seuraavassa taulukossa on teollisuuden toimialan vuoden 2005 kuukausille suoran tiedonkeruun yrityksistä laskettujen vuosimuutosprosenttejen revisiot r_t^* (kaava (4.9)) ja yksinkertaisen satunnaisotannan tuottamien vuosimuutosprosenttien revisiot r_{srs}. Muiden toimialojen vastaavat taulukot löytyvät liitteestä 3: Taulukot L.1-L.3.

Teollisuus, Yksinkertainen satunnaisotanta		
Kuukausi	r_t^*	$r_{t,srs}$
01	0.0233	0.1078
02	0.0024	0.1140
03	0.0258	0.1160
04	0.0028	0.1094
05	0.0199	0.1080
06	0.0124	0.0903
07	0.0016	0.1155
08	0.0217	0.1250
09	0.0050	0.1146
10	0.0434	0.1255
11	0.0297	0.1163
12	0.0319	0.1216

Taulukko 4.4: Teollisuuden toimialan revisiot SRS-otannalla

Taulukosta voidaan helposti huomata, että yksinkertaisen satunnaisotannan revisiot ovat moninkertaiset verrattuna suoran tiedonkeruun revisioihin tai PPS-otannan revisioihin (vrt. Taulukko 4.2). Yksinkertaista satunnaisotantaa käyttämällä revisio oikeisiin arvoihin nähden olisi siis $9-12$ prosenttiyksikön välillä, kun suoran tiedonkeruun yrityksillä revisio olisi vain $0-4$ prosentin luokkaa.

4.2.3 Ositetun aineiston katkaistu suurten yritysten otanta

Kolmas eli viimeinen testattava otanta-asetelma poikkeaa jonkin verran kahdesta edellisestä. Siinäkin lähdettiin liikkeelle tilanteesta, jossa suuret yritykset otetaan erilleen perusjoukosta, ja ne tulisivat otokseen varmasti. Suuriksi yrityksiksi luokiteltiin taas ne yritykset, joiden henkilöstömäärä oli vuonna 2004 250 tai suurempi. Jäljelle jäänyt yrityspopulaatio ositettiin taas täsmälleen kuten muissakin otanta-asetelmissa käyttäen SAS:n FASTCLUS-proseduuria, ja ositusmuuttujina olivat tarkempi toimiala ja postinumero. Ositteita määrättiin olevan kuten aiemminkin 10 kappaletta. Kiintiöintimenetelmänä oli taas hieman muunneltu versio suhteellisesta kiintiöintimenetelmästä eli otos jaettiin ositteiden kesken siten, että kustakin ositteesta otettiin

yrityksiä otokseen suhteessa ositteen kokoon kokonaisliikevaihdolla mitattuna. Erilleen otettuja suuria yrityksiä kohdeltiin "omina ositteinaan"ja ne tulivat otokseen varmasti.

Erilaisen otanta-asetelmasta tekee se, että nyt otokseen poimittiin kustakin ositteesta järjestelmällisesti n_h liikevaihdoltaan *suurinta* yritystä, eikä otoksen valinnassa ole mukana minkäänlaista satunnaisuutta. Ja kun otokseen tulleet yritykset oli valittu, laskettiin niiden perusteella liikevaihdon vuosimuutosprosentit. Nyt ei siis pystytä estimoimaan kokonaismääriä, koska katkaistulle suurten yrityksen otannalle ei ole olemassa kokonaismäärän estimaattoria. Liikevaihdon vuosimuutokset lasketaan siis suoraan otoksen yritysten perusteella

$$\hat{g}_t = \frac{\sum_{k=1}^{n} y_{k,t}}{\sum_{k=1}^{n} y_{k,t-12}},$$

missä $y_{k,t}$ on yrityksen k kuukauden t liikevaihto ja $y_{k,t-12}$ on saman yrityksen edellisvuoden vastinkuun liikevaihto. Summataan siis otoksen yritysten liikevaihdot.

Koska nyt kyseessä oli katkaistu otos ilman satunnaisuutta, otosta ei kannattanut simuloida perusjoukosta, koska samat yritykset tulisivat otokseen jokaisella simulointikerralla. Näin ollen ei myöskään pystytä laskemaan tilastollisia tunnuslukuja keskiarvoa lukuunottamatta. Ainoa keino vertailla otanta-asetelmaa muihin on laskea liikevaihtojen vuosimuutosprosenttien *revisiot* eli se kuinka paljon suoran tiedonkeruun ja nyt testatun katkaistun otoksen liikevaihdon vuosimuutokset poikkeavat oikeista vuosimuutosarvoista. Katkaistun otoksen revisio voidaan laskea kaavalla

$$r_{t,suuret} = |g_t - \hat{g}_t|. \tag{4.10}$$

Seuraavassa taulukossa on teollisuuden toimialan vuoden 2005 kuukausille suoran tiedonkeruun yrityksistä laskettujen vuosimuutosprosenttejen revisiot r_t^* (kaava (4.9)) ja katkaistun otoksen tuottamien vuosimuutosprosenttien revisiot $r_{t,suuret}$. Muiden toimialojen vastaavat taulukot löytyvät liitteestä 3: Taulukot L.1-L.3.

Teollisuus, Katkaistu otos		
Kuukausi	r_t^*	$r_{t,suuret}$
01	0.0233	0.0109
02	0.0024	0.0021
03	0.0258	0.0053
04	0.0028	0.0034
05	0.0199	0.0212
06	0.0124	0.0431
07	0.0016	0.0047
08	0.0217	0.0049
09	0.0050	0.0010
10	0.0434	0.0173
11	0.0297	0.0045
12	0.0319	0.0101

Taulukko 4.5: Teollisuuden toimialan revisiot katkaistulla otannalla

Taulukosta huomataan selkeästi, että revisiot ovat pienemmät katkaistulla suurten yritysten otannalla kuin suoran tiedonkeruun yritysten otannalla. Kahdeksassa kuukaudessa kahdestatoista revisio on pienempi katkaistulla suurten yritysten otannalla. Molemmissa otoksissahan olivat yrityspopulaation suurimmat yritykset, mutta nyt testatussa otanta-asetelmassa suurimmat yritykset oli poimittu *ositetusta* yrityspopulaatiosta. Voidaan siis ajatella että otokseen tulleet yritykset ovat maantieteellisesti enemmän hajallaan ja tarkemmat toimialat paremmin edustettuina otoksessa, ja näin ollen nyt testattu suurten yritysten katkaistu otos kuvastaisi suoran tiedonkeruun yrityksiä paremmin koko perusjoukkoa.

5 Tulosten analysointi ja yhteenveto

Tämän pro gradu -tutkielman tarkoituksena on ollut tutkia, voidaanko Tilastokeskuksen liikevaihtoindeksien ennakkotietojen 2000 suurimman yrityksen otokseen perustuvia estimaatteja laskea tarkemmin toisenlaisella otoksella. Aihe on Tilastokeskukselle erityisen kiinnostava, sillä kun ennakkotiedot on julkaistu, on kiusallista mikäli julkaistuja tuloksia joudutaan korjaamaan myöhemmin paljon varsinaisten tietojen saapuessa. Tällä hetkellä ennakkotietojen osuvuus ei ole kovinkaan tarkka ja vaihtelee suuresti toimialoittain. Tarkimmat ennakoindeksit ovat kaupan toimialalla, ja epätarkimmat rakentamisen toimialalla. Kaupan toimialalla revisio, eli se kuinka paljon ennakkoindeksi poikkeaa oikeasta indeksistä, on keskimäärin yhden prosentin luokkaa, kun taas rakentamisen toimialalla revisio on 2-3 prosenttiyksikköä.

Tutkimusaineistona on ollut Verohallinnon maksuvalvonta-aineisto vuodelle 2005, jossa on kyseisen vuoden kuukausittaiset tiedot kaikkien yritysten liikevaihdosta sekä edellisen vuoden saman kuukauden liikevaihdosta. Lisäksi yrityksille saatiin taustatietoja osittamista varten Tilastokeskuksen yritysrekisteristä. Tutkimuskohteena on ollut neljän päätoimialan; teollisuuden, rakentamisen, kaupan ja muiden palveluiden vuoden 2005 kuukausittaiset ennakkoindeksit.

Tutkielman empiirisessä osassa (Luku 4) testattiin neljälle päätoimialalle kolmea erilaista otanta-asetelmaa. Teollisuuden toimialan tulokset käsiteltiin tekstissä (Taulukot 4.1-4.5), muiden toimialojen tutkimustulokset ovat taulukoituna liitteissä 1-3. Kaikilla toimialoilla PPS-otanta oli SRS-otantaa selkeästi tehokkaampi. Tämän voi todeta tarkastelemalla talukoiden L1-L3 lukuja liitteessä 1 ja vertaamalla niitä talukoiden L1-L3 lukuihin liitteessä 2. SRS-otannan keskihajonta ja keskineliöpoikkeaman neliöjuuri ovat kaikille vuoden 2005 kuukausille kymmenkertaiset verrattuna PPS-otantaan. Harhat ovat molemmissa otanta-asetelmissa kutakuinkin samaa luokkaa, koska molemmissa otanta-asetelmissa käytetty HT-estimaattori on harhaton, ja täten simulointikertojen lisääntyessä harha lähestyisi nollaa molemmissa otanta-asetelmissa. Myös keskineliöpoikkeaman variaatiokerroin on SRS-otannassa moninkertainen verrattuna PPS-otantaan.

Tutkielman kannalta ehkä kiinnostavampaa on tarkastella eri otanta–asetelmilla laskettujen liikevaihdon vuosimuutosprosenttien estimaattien revisioita verratuna todelliseen vuosimuutosprosenttiin. Revisio kertoo paljon-

ko liikevaihdon estimoitu vuosimuutosprosentti poikkeaa todellisesta vuosimuutosprosentista. Revisioitakin tarkastelemalla on täysin selvää, että SRS-otannalla liikevaihdon vuosimuutosprosenttien estimaatit ovat kauempana todellisista vuosimuutosprosenteista kuin suoran tiedonkeruun yritysten otannalla (taulukot L1-L3 liitteessä 3, $r_{t,srs}$ vrt. r_t^*). SRS-otannan revisiot ovat kaikilla toimialoilla 20 prosenttiyksikön luokkaa, kun suoran tiedonkeruun vastaavat luvut ovat muutaman prosenttiyksikön toimialasta riippuen.

PPS-otannan revisiot $r_{t,pps}$ olivat kaikilla muilla kuin teollisuuden toimialalla keskimäärin isommat kuin suoran tiedonkeruun revisiot r_t^*. Kaupan ja rakentamisen toimialalla PPS-otannan revisiot olivat pienemmät neljässä kahdestatoista kuukaudesta ja palveluiden toimialalla PPS-otannan revisiot olivat pienemmät kolmessa kahdestatoista kuukaudesta. Näillä eri otanta-asetelmilla ero ei kuitenkaan ollut kovin suuri.

Katkaistulla suurten yritysten otannalla lasketut revisiot todellisiin vuosimuutosprosentteihin $r_{t,suuret}$ ovat suoran tiedonkeruun vastaavia lukuja pienemmät kaikilla toimialoilla kauppaa lukuunottamatta. Teollisuuden ja rakentamisen toimialoilla revisiot ovat pienemmät katkaistulla suurten yritysten otannalla kahdeksassa kuukaudessa kahdestatoista, palveluissa seitsemässä ja kaupassakin puolessa tarkastelluista kuukausista. Näin ollen ositetun aineiston katkaistu suurten yritysten otanta olisi testatuista otanta-asetelmista paras, ja tarkentaisi liikevaihdon ennakkotietojen estimaatteja verrattuna nykyiseen suoran tiedonkeruun yrityksistä laskettuihin ennakkotietojen estimaatteihin.

Yksinkertainen satunnaisotanta ositetusta aineistosta ei tuonut liikevaihtoindeksien ennakkotietojen estimaatteihin tehokkuutta verrattuna nykyiseen. Tämä johtuu ehkä siitä, että yrityspopulaatio jakautuu muutamiin satoihin liikevaihdoltaan hyvin suuriin yrityksiin ja liikevaihdoltaan pienempiin yrityksiin, toisin sanoen suuret yritykset ovat muita yrityksiä *huomattavasti* suurempia. Pienten yritysten liikevaihto saattaa kehittyä hyvin eri tavalla kuin koko toimiala tai se voi heittelehtiä paljon kuukaudesta riippuen, suurissa yrityksissä kehitys on usein koko toimialan suuntainen. SRS-otannassa otokseen tulee todennäköisemmin pieniä yrityksiä kuin suuria, koska suuri osa yrityspopulaatiosta on verrattain pieniä yrityksiä, ja näin ollen otoksen perusteella estimoidut liikevaihdon ennakkoindeksit saattavat vaihdella paljonkin.

PPS-otannan keskinkertaiset tulokset verrattuna nykyiseen suoran tiedonkeruun yrityksistä laskettuihin ennakkotietojen estimaatteihin selittyvät

osittain samalla asialla kuin SRS-otannankin. Myös PPS-otannassa otokseen tulee pieniä, "epästabiilimpia"yrityksiä, jotka heikentävät liikevaihdon ennakkoindeksejä. Sen sijaan ositetun aineiston katkaistulla suurten yritysten otannalla saatiin selkeästi paremmat estimaatit kuin nykyiset suoran tiedonkeruun yrityksistä lasketut ennakkotietojen estimaatit. Molemmissa otoksissa olivat kunkin toimialan suurimmat yritykset, erona vain ennen ennen otantaa tehty yrityspopulaation osittaminen. Voidaan siis ajatella että ositetun aineiston katkaistulla suurten yritysten otannalla otokseen tulleet yritykset ovat maantieteellisesti enemmän hajallaan ja tarkemmat toimialat paremmin edustettuina otoksessa, ja näin ollen nyt testattu suurten yritysten katkaistu otos kuvastaisi suoran tiedonkeruun yrityksiä paremmin koko perusjoukkoa.

Kysymykseksi jää vielä, millaisia tutkimustuloksia olisi saatu tästä tutkielmasta poisjätetyillä menetelmillä, kuten otoksen ulkopuolelle jäävien yritysten uusimman kuukauden liikevaihtotiedon ennustamisella edellisten kuukausien liikevaihtotietoja hyväksikäyttäen. Tapoja tähän on lukuisia. Ensinnäkin ennustamisen voisi tehdä kaikille otoksen ulkopuolelle jääville yrityksille yhteensä tai kullekin yritykselle erikseen, tai sitten jotain siltä väliltä. Toiseksi ennustustapoja on lukuisa määrä. Luultavasti myös tällä menettelyllä olisi saatu hyviä tuloksia liikevaihtojen ennakkoindeksien tarkentumiseen, mutta se jääköön jonkin toisen tutkielman selvitettäväksi.

Lähdeluettelo

Lehtonen, R. and Pahkinen, E. 2004. *Practical Methods for Design and Analysis of Complex Surveys*. Wiley, New York.

Pahkinen, E. ja Lehtonen, R. 1989. *Otanta-asetelmat ja tilastollinen analyysi*. 0y Gaudeamus Ab, Helsinki.

Ohlsson, E. 1990. *Sequential Poisson Sampling from a Business Register and its Application to the Swedish Consumer PriceIndex*. Department of Enterprice Statistics, Statistics Sweden.

Hidiroglou, M.A., Särndal, C.-E., Binder, D.A. 1995. *Weighting and Estimation in Business Surveys*. Wiley, New York.

Jajuga, K., Sokolowski, A., Bock, H.-H. 2002. *Classification, Clustering and Data Analysis*. Springer, Berlin.

Teikari, I. 2001. *Poisson Mixture Sampling in Controlling the Distribution of Response Burden in Longitudinal and Cross section Business surveys*. Statistics Finland.

Särndal, C.-E., Svensson, B., Wretman, J. 2002 *Model Assisted Survey Sampling*. Springer, New York.

Tilastokeskus 2002. *Toimialaluokitus TOL 2002*. Käsikirjoja 4. Helsinki.

Tilastokeskus 2006. *Suomen yritykset 2005*. Helsinki.

Tilastokeskus. Yritys- ja toimipaikkarekisteri.

Tilastokeskus. Maksuvalvonta-aineisto MAVA

Verkkolähteet

Mellin, I. 2004. *Ryhmittelyanalyysi*. TKK/SAL 2004
http://www.sal.tkk.fi/Opinnot/Mat-2.112/pdf/CLUST10.pdf

Kaleva, O. *Tilastolliset monimuuttujamenetelmät*. TTY/ Matematiikan laitos
http://butler.cc.tut.fi/ kaleva/Tmm.pdf

Liite 1

PPS-otannan tulostaulukot kaupan, rakentamisen ja palveluiden toimialoille.

Kauppa, PPS-otanta						
Kuukausi	g_t	$\hat{g}_t$	$S.e.(\hat{g}_t)$	$BIAS(\hat{g}_t)$	$\sqrt{MSE(\hat{g}_t)}$	$CV(MSE(\hat{g}_t))$
01	0.0196	0.0201	0.01204	0.00041	0.01204	60.0292
02	0.0639	0.0639	0.01146	-0.00005	0.01145	17.9218
03	0.0271	0.0288	0.02612	0.00167	0.02616	90.9480
04	0.0696	0.0716	0.01381	0.00194	0.01394	19.4779
05	0.0932	0.0957	0.02424	0.00254	0.02436	25.4450
06	0.0561	0.0587	0.02805	0.00262	0.02815	47.9322
07	0.0208	0.0210	0.01978	0.00014	0.01978	94.3586
08	0.0822	0.0823	0.02783	0.00010	0.02782	33.7943
09	0.0748	0.0765	0.03642	0.00169	0.03644	47.6136
10	0.0602	0.0595	0.01474	-0.00067	0.01474	24.7530
11	0.0591	0.0594	0.05125	0.00036	0.05122	86.2097
12	0.0544	0.0569	0.07307	0.00251	0.07308	128.3267

Taulukko L.1: Kaupan toimialan tunnuslukuja PPS-otannalla vuoden 2005 kuukausille. Liikevaihdon oikea muutosprosentti kyseessä olevana kuukautena verrattuna edellisvuoteen, estimoitu muutosprosentti, sen keskihajonta, harha ja keskineliöpoikkeaman neliöjuuri, sekä keskineliöpoikkeaman variaatiokerroin.

Rakentaminen, PPS-otanta						
Kuukausi	g_t	$\hat{g}_t$	$S.e.(\hat{g}_t)$	$BIAS(\hat{g}_t)$	$\sqrt{MSE(\hat{g}_t)}$	$CV(MSE(\hat{g}_t))$
01	0.1336	0.1338	0.04874	0.00023	0.04871	36.4112
02	0.0428	0.0411	0.03512	-0.00166	0.03514	85.4302
03	-0.0188	-0.0208	0.03328	-0.00204	0.03333	-159.8798
04	0.0388	0.0374	0.04750	-0.00146	0.04750	127.0837
05	0.0941	0.0928	0.03517	-0.00129	0.03517	37.8828
06	0.0655	0.0633	0.03344	-0.00222	0.03349	52.9417
07	0.0223	0.0231	0.04307	0.00076	0.04306	186.4299
08	0.1288	0.1270	0.03633	-0.00186	0.03636	28.6336
09	0.1079	0.1070	0.03967	-0.00096	0.03966	37.0813
10	0.0929	0.0928	0.03616	-0.00012	0.03614	38.9460
11	0.0759	0.0732	0.03637	-0.00270	0.03645	49.7863
12	0.0423	0.0413	0.03959	-0.00096	0.03958	95.8201

Taulukko L.2: Rakentamisen toimialan tunnuslukuja PPS-otannalla vuoden 2005 kuukausille. Liikevaihdon oikea muutosprosentti kyseessä olevana kuukautena verrattuna edellisvuoteen, estimoitu muutosprosentti, sen keskihajonta, harha ja keskineliöpoikkeaman neliöjuuri, sekä keskineliöpoikkeaman variaatiokerroin.

Palvelut, PPS-otanta						
Kuukausi	g_t	$\hat{g}_t$	$S.e.(\hat{g}_t)$	$BIAS(\hat{g}_t)$	$\sqrt{MSE(\hat{g}_t)}$	$CV(MSE(\hat{g}_t))$
01	0.0477	0.0476	0.02142	-0.00006	0.02141	44.9370
02	0.0516	0.0510	0.01882	-0.00056	0.01882	36.9083
03	0.0135	0.0138	0.02172	0.00037	0.02171	157.1070
04	0.0587	0.0581	0.02247	-0.00061	0.02246	38.6761
05	0.0556	0.0560	0.02370	0.00037	0.02369	42.3409
06	0.0459	0.0461	0.02339	0.00021	0.02338	50.7046
07	0.0186	0.0180	0.02865	-0.00058	0.02864	159.3646
08	0.0700	0.0706	0.02855	0.00059	0.02854	40.4421
09	0.0646	0.0662	0.02695	0.00153	0.02698	40.7702
10	0.0593	0.0599	0.03014	0.00059	0.03013	50.3091
11	0.0741	0.0756	0.02983	0.00153	0.02985	39.4607
12	0.0403	0.0429	0.03904	0.00264	0.03911	91.1407

Taulukko L.3: Palveluiden toimialan tunnuslukuja PPS-otannalla vuoden 2005 kuukausille. Liikevaihdon oikea muutosprosentti kyseessä olevana kuukautena verrattuna edellisvuoteen, estimoitu muutosprosentti, sen keskihajonta, harha ja keskineliöpoikkeaman neliöjuuri, sekä keskineliöpoikkeaman variaatiokerroin.

Liite 2

SRS-otannan tulostaulukot kaupan, rakentamisen ja palveluiden toimialoille.

Kauppa, Yksinkertainen satunnaisotanta						
Kuukausi	g_t	$\hat{g}_t$	$S.e.(\hat{g}_t)$	$BIAS(\hat{g}_t)$	$\sqrt{MSE(\hat{g}_t)}$	$CV(MSE(\hat{g}_t))$
01	0.0196	0.0203	0.26914	0.00065	0.26900	1369.9674
02	0.0639	0.0622	0.29299	-0.00171	0.29285	458.2018
03	0.0271	0.0246	0.29349	-0.00223	0.29335	1083.6441
04	0.0696	0.0653	0.27941	-0.00436	0.27931	401.2645
05	0.0932	0.0910	0.28266	-0.00220	0.28253	303.3021
06	0.0561	0.0553	0.26986	-0.00080	0.26973	480.9271
07	0.0208	0.0188	0.28276	-0.00202	0.28263	1358.6040
08	0.0822	0.0802	0.31667	-0.00205	0.31652	358.1503
09	0.0748	0.0731	0.32334	-0.00170	0.32319	432.0368
10	0.0602	0.0594	0.31847	-0.00083	0.31832	528.6596
11	0.0591	0.0581	0.31641	-0.00096	0.31625	535.7669
12	0.0544	0.0524	0.31203	-0.00205	0.31188	573.2349

Taulukko L.1: Kaupan toimialan tunnuslukuja SRS-otannalla vuoden 2005 kuukausille. Liikevaihdon oikea muutosprosentti kyseessä olevana kuukautena verrattuna edellisvuoteen, estimoitu muutosprosentti, sen keskihajonta, harha ja keskineliöpoikkeaman neliöjuuri, sekä keskineliöpoikkeaman variaatiokerroin.

Rakentaminen, Yksinkertainen satunnaisotanta						
Kuukausi	g_t	$\hat{g}_t$	$S.e.(\hat{g}_t)$	$BIAS(\hat{g}_t)$	$\sqrt{MSE(\hat{g}_t)}$	$CV(MSE(\hat{g}_t))$
01	0.1336	0.1414	0.38802	0.00787	0.38791	274.28152
02	0.0428	0.0501	0.33952	0.00728	0.33943	677.8632
03	-0.0188	-0.0097	0.36298	0.00906	0.36291	-3724.3586
04	0.0388	0.0424	0.32723	0.00354	0.32714	772.0453
05	0.0941	0.0930	0.32108	-0.00110	0.32092	344.9282
06	0.0655	0.0661	0.31927	0.00061	0.31911	482.8569
07	0.0223	0.0196	0.25993	-0.00273	0.25982	1324.7669
08	0.1288	0.1247	0.32578	-0.00411	0.32565	261.1085
09	0.1079	0.1105	0.29167	0.00263	0.29153	263.7190
10	0.0929	0.0859	0.31550	-0.00706	0.31542	367.3882
11	0.0759	0.0732	0.29483	-0.00272	0.29470	402.5922
12	0.0423	0.0442	0.28479	0.00197	0.28465	643.4051

Taulukko L.2: Rakentamisen toimialan tunnuslukuja SRS-otannalla vuoden 2005 kuukausille. Liikevaihdon oikea muutosprosentti kyseessä olevana kuukautena verrattuna edellisvuoteen, estimoitu muutosprosentti, sen keskihajonta, harha ja keskineliöpoikkeaman neliöjuuri, sekä keskineliöpoikkeaman variaatiokerroin.

Palvelut, Yksinkertainen satunnaisotanta						
Kuukausi	g_t	$\hat{g}_t$	$S.e.(\hat{g}_t)$	$BIAS(\hat{g}_t)$	$\sqrt{MSE(\hat{g}_t)}$	$CV(MSE(\hat{g}_t))$
01	0.0477	0.0470	0.34062	-0.00070	0.34045	724.3087
02	0.0516	0.0546	0.32502	0.00301	0.32487	595.2662
03	0.0135	0.0159	0.30353	0.00246	0.30338	1907.0025
04	0.0587	0.0612	0.33688	0.00253	0.33672	549.9391
05	0.0556	0.0578	0.33261	0.00226	0.33246	574.7473
06	0.0459	0.0456	0.32238	-0.000341	0.32222	707.1808
07	0.0186	0.0188	0.34018	0.00023	0.34001	1809.9340
08	0.0700	0.0701	0.32788	0.00010	0.32771	467.4876
09	0.0646	0.0681	0.31604	0.00337	0.31590	464.4754
10	0.0593	0.0584	0.32999	-0.00092	0.32983	564.9061
11	0.0741	0.0744	0.33011	0.00032	0.32994	443.2420
12	0.0403	0.0466	0.33474	0.00635	0.33463	717.8678

Taulukko L.3: Palveluiden toimialan tunnuslukuja SRS-otannalla vuoden 2005 kuukausille. Liikevaihdon oikea muutosprosentti kyseessä olevana kuukautena verrattuna edellisvuoteen, estimoitu muutosprosentti, sen keskihajonta, harha ja keskineliöpoikkeaman neliöjuuri, sekä keskineliöpoikkeaman variaatiokerroin.

Liite 3

Katkaistun otannan tulostaulukot kaupan, rakentamisen ja palveluiden toimialoille.

Revisiot: Kauppa				
Kuukausit	r_t^*	$r_{t,pps}$	$r_{t,srs}$	$r_{t,suuret}$
01	0.0002	0.0085	0.1894	0.0041
02	0.0034	0.0091	0.2072	0.0074
03	0.0122	0.0090	0.2071	0.0127
04	0.0030	0.0106	0.1994	0.0012
05	0.0140	0.0117	0.1981	0.0036
06	0.0239	0.0135	0.1900	0.0001
07	0.0044	0.0117	0.1977	0.0075
08	0.0249	0.0122	0.2159	0.0022
09	0.0128	0.0143	0.2211	0.0054
10	0.0022	0.0113	0.2227	0.0026
11	0.0031	0.0145	0.2219	0.0043
12	0.0087	0.0189	0.2169	0.0080

Taulukko L.1: Kaupan toimialan revisiot vuoden 2005 kuukausille. Ensimmäisessä sarakkeessa r_t^* on suoran tiedonkeruun yrityksistä laskettujen vuosimuutosprosenttejen revisiot (kaava 4.9). Toisessa sarakkeessa $r_{t,pps}$ on PPS-otannan tuottamien vuosimuutosprosenttejen revisiot (kaava 4.8) ja kolmannessa sarakkeessa $r_{t,srs}$ vastaavat SRS-otannalle. Neljännessä sarakkeessa $r_{t,suuret}$ on katkaistun otannan tuottamien vuosimuutosprosenttejen revisiot (kaava 4.10).

Revisiot: Rakentaminen				
Kuukausi	r_t^*	$r_{t,pps}$	$r_{t,srs}$	$r_{t,suuret}$
01	0.0116	0.0381	0.2537	0.0115
02	0.0533	0.0277	0.2172	0.0366
03	0.0066	0.0262	0.2131	0.0021
04	0.0861	0.0313	0.2106	0.0667
05	0.0335	0.0274	0.2174	0.0175
06	0.0041	0.0256	0.2200	0.0011
07	0.0379	0.0296	0.1903	0.0341
08	0.0185	0.0287	0.2176	0.0185
09	0.0017	0.0298	0.2131	0.0023
10	0.0032	0.0283	0.2178	0.0063
11	0.0281	0.0292	0.2174	0.0279
12	0.0253	0.0304	0.1956	0.0330

Taulukko L.2: Rakentamisen toimialan revisiot vuoden 2005 kuukausille. Ensimmäisessä sarakkeessa r_t^* on suoran tiedonkeruun yrityksistä laskettujen vuosimuutosprosenttejen revisiot (kaava 4.9). Toisessa sarakkeessa $r_{t,pps}$ on PPS-otannan tuottamien vuosimuutosprosenttejen revisiot (kaava 4.8) ja kolmannessa sarakkeessa $r_{t,srs}$ vastaavat SRS-otannalle. Neljännessä sarakkeessa $r_{t,suuret}$ on katkaistun otannan tuottamien vuosimuutosprosenttejen revisiot (kaava 4.10).

Revisiot: Palvelut				
Kuukausi	r_t^*	$r_{t,pps}$	$r_{t,srs}$	$r_{t,suuret}$
01	0.0213	0.0167	0.2203	0.0134
02	0.0114	0.0143	0.2171	0.0056
03	0.0047	0.0151	0.2046	0.0091
04	0.0213	0.0158	0.2264	0.0175
05	0.0229	0.0154	0.2229	0.0224
06	0.0064	0.0172	0.2137	0.0081
07	0.0214	0.0214	0.2189	0.0220
08	0.0130	0.0211	0.2177	0.0099
09	0.0078	0.0201	0.2164	0.0094
10	0.0189	0.0219	0.2224	0.0150
11	0.0035	0.0217	0.2232	0.0013
12	0.0032	0.0281	0.2242	0.0048

Taulukko L.3: Palveluiden toimialan revisiot vuoden 2005 kuukausille. Ensimmäisessä sarakkeessa r_t^* on suoran tiedonkeruun yrityksistä laskettujen vuosimuutosprosenttejen revisiot (kaava 4.9). Toisessa sarakkeessa $r_{t,pps}$ on PPS-otannan tuottamien vuosimuutosprosenttejen revisiot (kaava 4.8) ja kolmannessa sarakkeessa $r_{t,srs}$ vastaavat SRS-otannalle. Neljännessä sarakkeessa $r_{t,suuret}$ on katkaistun otannan tuottamien vuosimuutosprosenttejen revisiot (kaava 4.10).