Elinkeinoverorekisteriaineiston editointi ja imputointi. Sovellettujen menetelmien vertailu.

Janne Ikäheimonen

Tilastotieteen pro gradu-tutkielma
30. heinäkuuta 2001

Jyväskylän yliopisto
matematiikan ja tilastotieteen laitos
tilastotoimen menetelmät

Tilastotoimen menetelmien maisteriohjelma TTMM

Tilastotoimen menetelmien maisteriohjelman (TTMM-ohjelma) tavoite on kouluttaa opiskelija tilastotiedon keruun, jatkojalostuksen ja käytön asiantuntijaksi nykyaikaisissa tilastojärjestelmäympäristöissä, joissa aineistot ovat survey-, koeasetelma- tai rekisteriperusteisia. Tilastotieteelliseltä kannalta kyse on survey-menetelmiin, biostatistiikkaan tai teollisuustilastotieteeseen erikoistuneiden tilastoasiantuntijoiden koulutuksesta. Ohjelman laajuus on 60 opintoviikkoa. Ohjelman kesto päätoimisella opiskelijalla on kaksi lukuvuotta. Maisterin tutkinnon voi suorittaa vaihtoehtoisesti joko matemaattis-luonnontieteellisessä tai yhteiskuntatieteellisessä tiedekunnassa

Tilastotoimen menetelmien maisteriohjelman ytimen muodostavat tilastotieteen syventävät opinnot, erityisesti tilastotoimen teorian ja menetelmien kurssit, sekä tietojenkäsittelyopin kurssit. Ohjelmaan voidaan sisällyttää koulutustavoitteita tukevia sovellusalojen opintoja (esim. talous-, yhteiskunta- ja viestintätieteistä) sekä yritystoimintaan perehdyttäviä kursseja. Tarjottava opetus koostuu tilastotieteen laitoksen opetusohjelmasta, erityisistä tilastotoimen teorian ja menetelmien kursseista sekä erikseen hankituista erikoiskursseista. Opettajista pääosa on kotimaasta. Vierailijoina on myös ulkomaisia asiantuntijoita, joten opetus on osin englanninkielistä.

Ohjelman tärkeä osa on yliopiston ulkopuolisessa yhteistyötoimipaikassa suoritettu harjoittelu ja siinä yhteydessä tehtävä pro gradu -tutkielma. Yhteistyötoimipaikat ovat virallisesta tilastotoimesta, suuryrityksistä ja tutkimuslaitoksista. Maisteriohjelmasta vastaa Jyväskylän yliopiston matematiikan ja tilastotieteen laitoksen tilastotieteen yksikkö.

Jyväskylän yliopiston matematiikan ja tilastotieteen laitos
Tilastotieteen yksikkö
Tilastotoimen maisteriohjelma TTMM
Ohjelman johtaja: Prof. Risto Lehtonen
Ohjelman varajohtaja: Yliassistentti Kari Nissinen
Toimisto: Amanuenssi Sari Eronen
www.stat.jyu.fi/ttmm

(Päivitetty 26.9.2001)

TIIVISTELMÄ

Janne Ikäheimonen: *Elinkeinoverorekisteriaineiston editointi ja imputointi. Sovellettujen menetelmien vertailu.*
Tilastotieteen pro gradu-tutkielma, Jyväskylän yliopisto. 30. heinäkuuta 2001.
Sivuja 79, liitteitä 5.

Elinkeinoverorekisterillä (EVR) on tärkeä osa Tilastokeskuksen tuottaman yritysten rakennetilastotietokannan muodostamisessa. EVR sisältää tilinpäätöstiedot kaikilta elinkeinoverotuslain mukaan verotettavilta elinkeinonharjoittajilta. Aineisto tilataan Tilastokeskukseen vuosittain Verohallinnolta. Aineistoa ei voida sen virheellisyyden vuoksi käyttää sellaisenaan, vaan siihen sovelletaan editointi- ja imputointimenetelmiä. Tässä tutkielmassa keskitytään kahdenkertaista kirjanpitoa pitävien yritysten tuloslaskelmamuuttujien korjaamiseen. Vuoden 1998 EVR-aineistossa näistä yrityksistä noin 20 prosentille joudutaan soveltamaan korjausmenetelmiä.

EVR-aineiston editoinnissa ja imputoinnissa on sovellettu kolmea eri menetelmää, jotka ovat kehittämisjärjestyksessä suhdeimputointi, rescaling ja sekamenetelmä. Tutkielman tarkoituksena on verrata menetelmien tehokkuutta EVR-aineistossa käyttäen generoitua koeaineistoa. Koeaineisto pohjautuu vuoden 1998 EVR-datan virheettömiin havaintoihin, joihin generoidaan virheitä. Virheiden rakenne perustuu noin 1000 EVR:ssä virheelliseen yritykseen, joiden virheettömät tiedot saadaan Tilastokeskuksen suorasta kyselystä.

Menetelmien vertailussa käytetään Mahalanobisin etäisyysmitan sovellutusta kahden havainnon väliselle etäisyydelle. Etäisyydet lasketaan yritysten korjattujen ja virheettömien arvojen välille. Lisäksi lasketaan etäisyydet korjaamattomalle datalle, minkä avulla selvitetään, parantavatko editointi- ja imputointimenetelmät varmasti aineistoa. Yrityskohtaiseen etäisyyteen vaikuttavat kaikki varsinaiset tuloslaskelman korjattavat muuttujat. Lopulliseen etäisyysmittaan summataan yrityskohtaiset etäisyydet. Tämän lisäksi menetelmiä verrataan laskemalla suhteellisia virheitä muuttujien toimialoittaisissa summissa.

Tulosten perusteella voidaan korjausmenetelmien käyttöä pitää perusteltuna ja aineiston laatua parantavana tekijänä. Rescaling osoittautuu yleisesti tehokkaimmaksi menetelmäksi, mutta ei kuitenkaan kaikkien yksittäisten muuttujien kohdalla. EVR-aineiston rakenteen vuoksi on virheellisen muuttujan paikallistaminen vaikeaa ja näin korjataan turhaan oikeitakin arvoja. Tätä ongelmaa vähentää sekamenetelmässä käytettävä outlier-metodi, mutta sillä voidaan korjata vain osa havainnoista. Tutkielman tulosten perusteella voidaan kehittää entistä tehokkaampi menetelmä, jossa yhdistetään outlier- ja rescaling-metodi.

Avainsanoja: editointi ja imputointi, koeaineiston generointi, Mahalanobisin etäisyysmitta.

ABSTRACT

Janne Ikäheimonen: *Editing and imputation of Return Tax File. Evaluation of methods.*
Master's thesis in Statistics, University of Jyväskylä. July 30, 2001.
Pages 79, Appendices 5.

Return Tax File (RTF) is an important data source in construction of the Structural Business Statistics database. RTF includes financial statements of every business that is taxed due to business taxation act. RTF is ordered annually to Statistics Finland from the National Board of Taxes. RTF needs to be edited and imputed because of its incorrectness. This study concentrates on editing and imputation of profit and loss account. About 20 percent of records in RTF of year 1998 are erroneous and have to be edited.

Three different methods have been used for editing of RTF in Statistics Finland. In order of development they are ratio imputation, rescaling-method and mixed method. The purpose of this study is to compare methods by generation of experimental data. Experimental data is based on correct records of RTF of year 1998, which are generated erroneous. Information of the error structure is gained from about 1000 erroneous businesses in RTF for which the correct values can be obtained from Statistics Finland's direct inquiry.

Mahalanobis distance between two records is used in evaluation of methods. Distance is measured between edited and correct values of a record. Every actual edited variable of profit and loss account affects the distance. Also the distance for erroneous experimental data is measured to solve if editing improves the quality of data. Final distance measure is summed up of record-specific distances. Also, proportional errors of aggregates of variables are used in comparison of methods.

Results show that editing and imputation methods improve the quality of data and should thereby be used. Rescaling proves to be the most efficient method overall but not for some particular variables. Due to structure of RTF accurate location of uncorrect variables in record is difficult and therefore also correct values are edited. Outlier-method that is part of the mixed method takes this problem into account. Method works very well but only a part of incorrect records can be edited with it. Based on this study's results a new editing and imputation method combining outlier-method and rescaling-method can be recommended.

Keywords: editing and imputation, generation of experimental data, Mahalanobis distance.

Sisältö

1 Johdanto

Elinkeinoverorekisteri on oleellinen osa Tilastokeskuksen tilastotuotantoa. Erityisesti sitä käytetään Yritysten rakenteet-yksikön tuottaman rakennetilastotietokannan luomisessa. EVR-aineiston yrityksistä kuitenkin lähes 20 prosenttia on virheellisiä tuloslaskelman osalta eli niiden tuloslaskelman muuttujat eivät summaudu oikein tilikauden tulokseen nähden. Tällöin tuloslaskelma ei mene umpeen, minkä takia EVR-aineistoon on sovellettu editointi- ja imputointimenetelmiä, joilla virheelliset yritykset on korjattu umpeenmeneviksi.

Menetelmät perustuvat täysin SAS-ohjelmistopohjaiseen koodiin, jossa käytetään enimmäkseen SAS:in data-lauseita sekä joitakin SAS-proseduureja. Tässä tutkielmassa ei tulla esittelemään menetelmien SAS-koodeja niiden laajuuden vuoksi.

Menetelmien vertailussa käytetään simulointilähestymistapaa eli luodaan keinotekoisesti virheellinen aineisto, johon sovelletaan eri editointimenetelmiä. Simuloinnissa käytettiin apuna SAS-ohjelmistoa, mutta tuotettuja koodeja ei tässä juurikaan esitetä. Tämä ei kuitenkaan vähennä niiden merkitystä tämän tutkielman tekemisessä, vaan niiden tekemiseen kohdistettu työpanos on mittava.

Luvussa 3 kuvataan elinkeinoverotusaineistoa. Aineistosta kuvataan editoitavat ja imputoitavat muuttujat ja niiden väliset suhteet sekä aineiston rooli yritysten rakennetilastojen tuotannossa Tilastokeskuksessa. Samassa luvussa kerrotaan myös Tilastokeskuksen omasta suorakyselystä, jonka tietoja käytetään hyväksi menetelmien vertailussa tarvittavaa koedataa luotaessa. Lisäksi kerrotaan yritysten toimialaluokituksesta. Toimialaluokitus on oleellinen osa rakennetilastojen tuottamisessa ja sitä luonnollisesti käytetään apuna elinkeinoverotusaineiston editoinnissa ja imputoinnissa.

Luvussa 4 kerrotaan ensin yleisesti editoinnista ja editointiprosessista. Lisäksi kerrotaan editointisäännöistä ja automatisoidun editoinnin ja imputoinnin kä-

sitteestä. Luvussa esitellään EVR-aineiston editoinnissa käytetyt menetelmät yksityiskohtaisesti. Lopuksi kerrotaan yleisesti imputoinnista. Varsinaiset vertailun kohteena olevat korjausmenetelmät esitellään luvuissa 5, 6 ja 7.

Menetelmien vertailuun tarvitaan kokeellinen data, jonka luonti kuvataan luvussa 8. Luvussa määritellään virheetön data, jonka perusteella vertailut tehdään. Lisäksi kerrotaan erilaisista virhetyypeistä. Tulosten luotettavuuden arvioimisen kannalta on tärkeätä tuntea virheellisen aineiston generointimenetelmä ja sen yhteys aitoon tilanteeseen, joten se kuvataan hyvin tarkasti.

Varsinaiset menetelmien vertailuun käytettävät menetelmät ja tulokset esitellään luvussa 9. Vertailussa käytetään apuna Mahalanobisin etäisyysmittaa. Lisäksi menetelmiä arvioidaan toimialoittain vertaillen eri muuttujien toimialatason summia virheettömästä datasta ja kullakin menetelmällä korjatusta datasta.

2 Tutkimusongelma

Elinkeinoverorekisteriaineiston editointimenetelmät ovat olleet jatkuvan muutoksen kohteena tilastovuodesta 1994 lähtien. Menetelmiä on pyritty jatkuvasti parantamaan luotettavamman ja parempilaatuisen aineiston aikaansaamiseksi. Editointimenetelmistä, tai paremminkin editointiin käytettävästä ohjelmakoodista, on löydetty jopa suoranaisia virheitä, jotka ovat saattaneet vaikuttaa editoidun aineiston laatuun. Menetelmien suorituskyvystä ei kuitenkaan ole ollut saatavilla tarkempaa tietoa.

2.1 Menetelmien kehitys Tilastokeskuksessa

EVR-aineiston editoinnissa on sovellettu kolmea eri korjausmenetelmää tilastovuosien 1994–99 aikana. Uudemmat menetelmät on kehitetty vanhan ohjelmakoodin pohjalta, joten menetelmissä on paljon yhteisiä osioita. Etenkin loogiset editointisäännöt ja havaintoyksiköiden virheellisyyssäännöt ovat kaikissa kolmessa menetelmässä samat. Menetelmät eroavatkin toisistaan lähinnä virheen sijainnin paikallistamisen ja virheellisiksi todettujen yritysten korjaamisen suhteen.

Vuosien 1994–97 aineistoihin on käytetty suhdeimputointimenetelmää, joka on puhtaasti imputointimenetelmä. Menetelmä ei paikallista virheen sijaintia, vaan se muuttaa virheellisten yritysten kaikkien editoitavien muuttujien arvot. Tämän vuoksi kehitettiin Rescaling-menetelmä, jolla saatiin korjatun datan osuus pienemmäksi. Siinä virhe paikallistetaan tiettyihin muuttujiin, jotka korjaamalla koko havainto saadaan virheettömäksi. Rescaling on kehitetty Tilastokeskuksessa eikä se ole varsinaisesti imputointimenetelmä, vaan siinä muutetaan olemassaolevia havaintoarvoja. Rescaling-menetelmää on käytetty tilastotuotannossa vuoden 1998 EVR-aineistoon. Vuoden 1999 aineistoon otettiin käyttöön menetelmä, jossa yhdistetään kolme eri korjaustapaa. Aluksi käytetään outlier-korjausta, toiseksi lähimmän naapurin imputointimenetelmää ja lopuksi suhdeimputointia. Tästä yh-

distetystä menetelmästä käytetään nimeä sekamenetelmä. Sekamenetelmään siirryttiin, koska rescaling-metodin uskottiin toimivan huonosti tapauksissa, joissa virhe on suuri. Lisäksi outlier-korjauksella pyrittiin pienentämään korjatun datan osuutta entisestään.

2.2 Mitä verrataan?

Tutkielmassa verrataan menetelmiä

- Suhdeimputointi
- Rescaling-metodi
- Sekamenetelmä,

jotka kuvataan tarkasti luvuissa 5, 6 ja 7. Kuitenkaan tarkoituksena ei ole pelkästään edellä mainittujen metodien vertailu, vaan samalla tutkitaan koko EVR-aineiston editointi- ja imputointiprosessin kehittymistä Tilastokeskuksessa. Tähän prosessiin vaikuttaa varsinaisen korjausmenetelmän lisäksi muutkin seikat, kuten virheen paikallistamisen tehokkuus. Menetelmiä ei ole siten standardoitu näiden muiden aineiston editointiin kuuluvien osien suhteen. Ainoastaan kaikissa menetelmissä täysin samanlaisina toistuvat osiot on jätetty pois simuloidun aineiston korjaamisessa. Täten esimerkiksi suhdeimputoinnissa ei sovelleta virheen paikallistamiseen käytettäviä tarkistusmenetelmiä, koska niitä ei ole myöskään tuotantokäytössä siihen sovellettu. Vaikka tässä tutkielmassa puhutaan edellä mainittujen kolmen metodin vertailusta, tarkoitetaan tällä koko editointiprosessien vertailua.

Elinkeinoverotusaineiston tärkeimmät tietokokonaisuudet ovat tuloslaskelma ja tase. Niihin molempiin on sovellettu Tilastokeskuksessa täysin samoja editointimenetelmiä. Tässä tutkielmassa kuvatut menetelmät koskevat tuloslaskelman muuttujia, mutta taseen editoinnissa ja imputoinnissa käytetään periaatteessa täysin samoja menetelmiä sovellettuna tasemuuttujiin. Tässä tutkielmassa verrataan

menetelmiä tuloslaskelman muuttujien osalta tutkimusjoukon ollessa kahdenkertaista kirjanpitoa pitävät yritykset. Tarkoituksena ei ole kehittää uutta editointi- ja imputointimenetelmää, vaan tutkia käytettyjen menetelmien tehokkuutta. Tämä tutkielma kuitenkin antaa vihjeitä menetelmien tulevan kehittämisen avuksi.

2.3 Simulointi ongelman ratkaisuna

Tutkimuksessa käytetään vuoden 1998 EVR-aineistoa (EVR-98). Lähtökohtana pidetään editoidun aineiston havaintokohtaista vertailua virheettömään aineistoon. Koska EVR-tietojen kaikkien havaintojen virheettömiä arvoja ei voi saada mitenkään selville, käytetään vertailussa simuloitua aineistoa. Simulointi on yleinen lähtökohta editointi- ja imputointimenetelmien paremmuuden selvittämiseksi. Aineiston luominen simuloimalla on yleensä melko ongelmatonta, mutta välttämättä tuotettu aineisto ei vastaa aitoa tilannetta. Tässä tutkielmassa esitetään keino virheiden generoimiseksi, jonka uskotaan tuottavan mahdollisimman aidonkaltaisen tilanteen EVR-aineistoon sovellettuna. Vastaavaa menetelmää ei tiettävästi ole ennen sovellettu.

Menetelmien vertailussa tarvitaan havaintoyksiköiltä muuttujakohtaiset oikeat, virheettömät arvot. Virheettöminä havaintoina pidetään niitä, jotka editointisäännöillä todetaan virheettömiksi. Virheiden generoinnissa käytetään apuna Tilastokeskuksen suoran kyselyn tietoja, jotka kerätään kaikilta suurimmilta yrityksiltä. Suorasta kyselystä saatavia tietoja pidetään oikeina, ja vertaamalla näitä EVR:n vastaaviin virheellisiin yrityksiin saadaan tietoa virheiden rakenteesta. Tämän tiedon avulla virheitä generoidaan virheettömään EVR-dataan. Kun simuloitu aineisto on saatu valmiiksi, sovelletaan kutakin vertailtavaa menetelmää siihen. Näin saadaan virheetön aineisto sekä kolme erilaista korjattua dataa, joissa kaikissa on samat yritykset. Näitä korjattuja datoja verrataan virheettömään aineistoon sekä yritystasolla että yleisemmällä toimialatasolla. Yritystasolla vertailussa käytetään Mahalanobisin etäisyysmittaa, joka ottaa huomioon erot muuttujatasolla. Etäisyy-

det lasketaan jokaiselle yritykselle ja nämä etäisyydet summataan menetelmittäin. Mitä pienempi on etäisyyksien summa, sitä parempi on menetelmä. Toimialatason vertailussa summataan muuttujia toimialoittain ja verrataan oikeaan summaan. Näin saadaan eräänlaiset harhaa kuvaavat arvot, joiden avulla saadaan parempi käsitys editoinnin vaikutuksista aineistoon.

Simulointi toistetaan 10 kertaa, jotta sattuman vaikutusta tuloksiin saadaan pienennettyä. Suurempikin määrä simulointeja olisi tietysti hyödyllinen, mutta se ei ole käytettävän ajan puitteissa mahdollista.

3 Aineisto

Tässä kappaleessa kuvataan tutkimuksen kohteena oleva elinkeinoverorekisteriaineisto, sen rooli tilastotuotannossa sekä muita editoinnissa ja imputoinnissa apuna käytettäviä aineistoja.

3.1 EVR-aineisto

Elinkeinoverorekisterissä on kunkin verovuoden aikana päättyneiden tilikausien tilinpäätöstiedot kaikilta niiltä elinkeinonharjoittajilta, joita verotetaan elinkeinoverotuslain mukaan. Mukana ovat kaikki eri yritysmuodot, liikkeen- ja ammatinharjoittajat, yhden- ja kahdenkertaista kirjanpitoa käyttävät kirjanpitovelvolliset. Yhdenkertaista kirjanpitoa käyttävät ammatinharjoittajat täyttävät tietosisällöltään suppeamman veroilmoituslomakkeen 5, jonka kiinnostavia tietokokonaisuuksia ovat tuloslaskelma, irtaimen käyttöomaisuuden hankintamenosta tehdyt poistot ja selvitys saaduista julkisista tuista. Kahdenkertaista kirjanpitoa käyttävät elinkeinonharjoittajat taas täyttävät veroilmoituksen liitelomakkeet 62: erittely varauksista ja kuluvan käyttöomaisuuden poistoista ja 63: lisätiedot. Liitelomakkeen 63 tietokokonaisuudet ovat tuloslaskelma, tase ja tuloslaskelman erittelyjä. *Tässä tutkielmassa keskitytään kahdenkertaisen kirjanpidon yritysten tuloslaskelman editointiin ja imputointiin.* Myös yhdenkertaista kirjanpitoa käyttävien ammatinharjoittajien tuloslaskelmatietoihin käytetään editointia, mutta sitä ei tässä tarkastella. Elinkeinoverotusaineisto saadaan Verohallinnolta vuosittain Tilastokeskukseen. Huomautettakoon tässä, että EVR ei ole otosaineisto, vaan siinä on siis kaikkien elinkeinoverotuslain mukaan verotettavien elinkeinonharjoittajien tiedot.

Tässä tutkielmassa käytettävä aineisto on vuoden 1998 elinkeinoverorekisterin tuloslaskelmatiedot kahdenkertaista kirjanpitoa pitäviltä yrityksiltä. Koko EVR-98-aineiston yritysten lukumäärä on 272983. Näistä kahdenkertaisen kirjanpidon yrityksiä on 235906. Kun edellisestä poistetaan yritykset, joiden kaikkien tulos-

laskelmamuuttujien arvot ovat nollia tai joille ei löydy toimialatietoa yritys- ja toimipaikkarekisteristä, jää jäljelle 219269 yritystä.

3.1.1 Muuttujien esittely

Tuloslaskelman muuttujat on esitelty liitteessä A. Liikevaihto, myyntikate, käyttökate ja tilikauden tulos eroavat muista muuttujista siinä, että ne ovat muiden muuttujien summia. Liikevaihto saadaan summana *liikevai = verm1 + verm2 + verm3 + myyntimu*. Vastaavasti myyntikate on *myyntika = liikevai + tuottomu - ostoyht - varastmu - ostopalv - palkkamu - kulumuut*. Käyttökate ja tilikauden tulos lasketaan vastaavasti. EVR-aineisto on tallennettu tiedostoon SAS-muodossa. Tässä tiedostossa vain muuttujilla *varastmu, myyntika, kayttoka, poivarmu, verovali ja tiliktul* voi olla negatiivisia arvoja. Kaikki muut muuttujat saavat arvoja, jotka ovat suurempia tai yhtä suuria kuin nolla, vaikka ne käsitteellisesti olisivatkin negatiivisia tuloslaskelman kaavassa, kuten esimerkiksi *ostoyht* (ostot tilikauden aikana). Tuloslaskelmassa käsitteellisesti positiivisia muuttujia ovat *verm1–3, myyntimu, liikevai, tuottomu, rahtuott, satunntu, poivarmu* ja *verovali*. Negatiivisia ovat *varastmu, ostopalv, palkkamu, kulumuut, palkkaki, vuokra, kulukiin, sumupois, korkokuk, rahkulum* ja *satunnku*. Muuttujat *poivarmu* ja *verovali* ovat käsitteellisesti positiivisia, vaikka ne voivatkin saada sekä negatiivisia että positiivisia arvoja. Muuttuja *varastmu* voi myöskin saada positiivisia ja negatiivisia arvoja, mutta käsitteellisesti se on negatiivinen. Täten esimerkiksi, jos SAS-tiedostossa *varastmu*$=-2500$, niin sen vaikutus tilikauden tulokseen on $-(-2500) = 2500$ eli se vaikuttaa positiivisesti. Tästä eteenpäin käytetään muuttujista $x_1, \ldots, x_{18}$ ja $y_1, \ldots, y_3$ (katso liite A) niiden käsitteellisiä etumerkkejä. Tämä yksinkertaistaa asioiden esittämistä, koska tällöin voidaan esimerkiksi tilikauden tulos ilmaista kaavalla $y_3 = \sum_{i=1}^{18} x_i$. Muuttujia *verm1, verm2, verm3* ja *myyntimu* ei ole merkitty liitteessä x-muuttujiksi, koska niitä ei käytetä tilastotuotannossa, ja täten niitä ei myöskään tarvitse editoida. Niitä kuitenkin käytetään editoinnissa apuna joissakin tapauksissa.

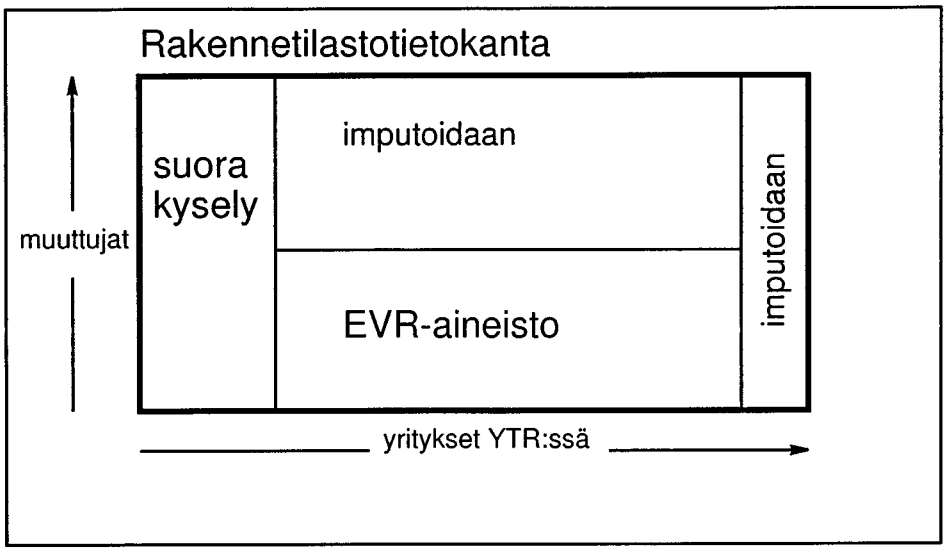
Kaikki tuloslaskelman editoitavat muuttujat ovat jatkuvia. Lisäksi kaikkien editoitavien muuttujien yksikkö on yksi Suomen markka. Imputointi ei ole perinteisessä mielessä puuttuvien arvojen imputointia, vaan pikemminkin editointisääntöjen perusteella virheellisiksi todettujen arvojen imputoimista tai muuttamista.

3.2 EVR-aineisto osana tilastotuotantoa

Yritysten rakennetilastotietokanta sisältää kaikkien Suomessa toimivien yritysten tilinpäätöstiedot. Tuloslaskelma- ja tasetietojen lisäksi se sisältää tuottojen ja kulujen erittelytietoja, tietoja henkilöstöstä ja käyttöomaisuuden lisäyksistä sekä erilaisia taustatietoja. Rakennetilastojen pääasiallinen käyttäjä on Tilastokeskuksen Yritysten rakenteet-yksikkö, mutta myös muut Tilastokeskuksen yksiköt sekä talon ulkopuoliset tutkijat käyttävät aineistoa. Rakennetilastotietokanta (kuva 1) muodostetaan seuraavista eri tietolähteistä:

- Suora kysely
- Elinkeinoverotusaineisto
- Yritys- ja toimipaikkarekisteri

Yritys- ja toimipaikkarekisteri (YTR) toimii eräänlaisena kehikkona, jonka sisältämille yrityksille pääsääntöisesti tuotetaan tiedot rakennetilastotietokantaan. Suoralla kyselyllä saadaan yritysten tilinpäätöstiedot sekä suuri määrä muita erittelytietoja. EVR-aineiston suppeampi tietosisältö yhdistetään suoran kyselyn tietoihin, ja niihin suoran kyselyn muuttujiin, jotka eivät EVR-aineistoon sisälly, imputoidaan arvot. Lisäksi niille YTR:n yrityksille, joille ei löydy tietoja kyselystä tai EVR:stä, imputoidaan suoran kyselyn muuttujia vastaavat arvot yritys- ja toimipaikkarekisteristä saatavan toimialan ja liikevaihdon perusteella. Näiden yritysten osuus toimialan liikevaihdosta vaihtelee välillä 0.6–5.0 prosenttia (Taulukko 1). Näihin edellä mainittuihin imputointijärjestelmiin ei tässä tutkielmassa



Kuva 1: EVR-aineiston rooli rakennetilastotietokannassa.

paneuduta lähemmin, vaan huomion kohteena on pelkästään EVR-aineistoon kuuluvan tietosisällön editointi ja imputointi. EVR-aineistossa on luonnollisesti myös suorassa kyselyssä olevien yritysten tiedot, mutta niitä ei käytetä, koska suoran kyselyn tiedot ovat tarkempia ja luotettavampia. Tätä päällekkäisyyttä käytetään hyväksi koeaineiston luomisessa.

3.3 Tilastokeskuksen suora kysely

Suora kysely on Tilastokeskuksen oma kysely, jolla kerätään tilinpäätöstiedot sekä erilaisia erittelytietoja. Se on tietosisällöltään huomattavasti laajempi kuin EVR-aineisto. Kysely kattaa EU:n rakennetilastoasetuksen mukaisilta toimialoilta sellaiset yritykset, jotka työllistävät yli 10, 20 tai 50 henkilöä riippuen toimialasta. Rakennetilastoasetuksen piiriin kuuluvat toimialat, joiden pääluokka on C, D, E, F, G, H, I tai K (katso kappale 3.4 ja taulukko 1). Näiden lisäksi suoralla kyselyllä kerätään tiedot myös osalta muiden toimialojen yrityksiä, jotka työllistävät yli 100 henkilöä. Suoran kyselyn avulla saadaan siis tiedot kaikilta suurimmilta ja tilastotuotannon kannalta tärkeimmiltä yrityksiltä. Taulukosta 1 näkyy, että ky-

Toimiala		Pros. yrityksistä EVR	SK	YTR	Pros. liikevaihdosta EVR	SK	YTR
CDE	Mineraalien kaivu, teollisuus, sähkö-, kaasu- ja vesihuolto	83.1	8.4	8.5	8.2	91.2	0.6
F	Rakentaminen	84.1	3.6	12.3	37.2	59.5	3.3
G	Tukku- ja vähittäiskauppa	86.1	4.2	9.7	24.8	73.8	1.4
H	Majoitus- ja ravitsemistoiminta	85.8	1.7	12.5	51.3	43.7	5.0
I	Liikenne	88.2	5.2	6.6	23.3	75.7	1.0
K	Liike-elämän palvelut	87.5	1.5	11.0	56.6	39.6	3.8

Taulukko 1: Rakennetilastojen tietolähteiden osuudet määrällisesti ja liikevaihdon mukaan.

sely kattaa vain 1.5–8.4 prosenttia kaikista toimialan yrityksistä, mutta kuitenkin 39.6–91.2 prosenttia toimialan kokonaisliikevaihdosta. Suoraan kyselyyn vastasi vuonna 1998 yhteensä 7289 yritystä. Suoran kyselyn muuttujat ovat aineistossa tuhansina Suomen markkoina.

3.4 Toimialaluokitus

Toimialaluokituksen avulla ryhmitellään samankaltaisia toimintoja luokkiin. Näitä luokkia kutsutaan lyhyesti toimialoiksi. Toiminnot ryhmitellään samankaltaisiksi tuottamiensa hyödykkeiden, tuotantopanostensa ja tuotantoprosessiensa perusteella. Toimialaluokitus on Tilastokeskuksen vahvistama luokitusstandardi, jonka tarkoituksena on edistää tilastojen käsitteellistä selkeyttä ja vertailukelpoisuutta (Tilastokeskus 1999).

Tässä tutkielmassa käytetään toimialaluokitus 1995:tä (TOL-95). Se perustuu Euroopan Yhteisön toimialaluokitusstandardi NACE:en vuodelta 1990. EY:n toimialaluokituksen käyttöä säätelee asetus, jonka mukaan kaikkien EU-maiden tulee käyttää yhdenmukaista toimialaluokitusta laatimissaan tilastoissa vuodesta

1993 lähtien. Suomessa Tilastokeskuksessa ja muissa valtion toimesta laadittavissa tilastoissa TOL-95 otettiin käyttöön tilastovuonna 1995.

Toimialaluokitusta käytetään yleisesti toimipaikkojen luokitteluun. Sitä voidaan käyttää myös muiden tilastoyksiköiden, kuten tässä tapauksessa yritysten, luokitteluun. Luokkien määrittelyssä on erityisesti pyritty siihen, että tarkimman tason luokat ovat mahdollisimman homogeenisia. Tarkimmalla tasollaan luokat koostuvat viisinumeroisesta nimikkeestä. Kaksinumerotasosta viisinumerotasoon luokitus on hierarkinen eli esimerkiksi viisinumeroisesta koodista 12345 voidaan päätellä, että luokka kuuluu kaksinumerotasolla luokkaan 12. Luokituksen pääluokkia on yhteensä 17 ja niitä merkitään kirjaimilla A-Q. Lisäksi on luokka "tuntematon", jota merkitään X. Yksinumerotason luokitusta, eli viisinumeroisen koodin ensimmäistä numeroa, ei yleisesti käytetä julkaisuissa, mutta suhdeimputoinnissa ja sekamenetelmässä sitä käytetään apuna.

4 Editointi ja imputointi

Tässä luvussa kerrotaan editoinnin määritelmästä, editointiprosessista, editointisäännöistä sekä automatisoidusta editoinnista ja imputoinnista. Lisäksi kuvataan EVR-aineistoon sovellettuja editointimenetelmiä. Lopuksi kerrotaan yleisesti imputoinnista ja imputointimenetelmien jaottelusta.

4.1 Editointi

Editoinnilla tarkoitetaan yleensä toimintaa, jolla pyritään havaitsemaan, paikallistamaan ja korjaamaan aineiston hankinnassa syntyneitä virheitä. Editoinnille ei ole olemassa yleisesti hyväksyttyä virallista määritelmää. Granquist (1995) mainitsee muutamia määritelmiä, jotka vaihtelevat editoinnin erilaisten päämäärien mukaan. Niistä ehkäpä sopivimman määritelmän on esittänyt Federal Committee on Statistical Methodology. Sen mukaan editoinnilla tarkoitetaan

> " ... proseduuria tai proseduureja, jotka on suunniteltu ja sovellettu virheellisen ja/tai epäilyttävän surveyaineiston selvittämiseksi tarkoituksena korjata mahdollisimman paljon virheellistä aineistoa (manuaalisesti ja/tai elektronisesti), yleensä ennen imputointia ja aineiston yhteenvetoproseduureja."

Joissain määritelmissä kuitenkin imputoinnin katsotaan olevan osa editointia. Imputoinnillahan voidaan tarkoittaa perinteisessä mielessä puuttuvan tiedon paikkaamista ja toisaalta editoinnissa virheelliseksi todettujen arvojen korjaamista. EVR-aineiston tapauksessa imputoinnissa on kyse jälkimmäisestä. Granquistin (1995) mukaan editointi voidaan jakaa kolmeen eri osaan: syöttöeditointi, koneellinen editointi ja makroeditointi.

Syöttöeditointi (*input editing*) käsittää toiminnot siihen asti, kunnes data on syötetty elektroniseen muotoon. Ennen datan syöttöä tapahtuvassa manuaalisessa

editoinnissa (*manual editing*) koodataan vastaukset ja tarkastetaan, että kyselylomake on täytetty asianmukaisesti. Nykyiset tietokoneavusteiset kyselymenetelmät mahdollistavat editoinnin jo datankeräysvaiheessa (*data entry editing, integrated editing*). Tällöin kyselyyn vastaaja tai haastattelija korjaa vastaukset jo haastattelutilanteessa, koska tietokoneohjelma tarkistaa vastausten oikeellisuuden sille ennalta syötettyjen editointisääntöjen perusteella.

Koneellisessa editoinnissa (*machine editing*) tarkistetaan elektronisessa muodossa oleva aineisto yksi havainto kerrallaan. Tästä käytetään usein myös nimitystä *mikroeditointi* silloin, kun tarkistamisessa ei käytetä apuna tietoa muista havaintoyksiköistä. Tarkistamisessa käytetään tilanteeseen sopivia editointisääntöjä, joilla varmistetaan havaintoyksikön kelpoisuus. Mikäli havainto ei noudata editointisääntöjä, niin se joko tarkistetaan manuaalisesti tai ohjataan imputoitavien havaintojen joukkoon. Manuaalinen tarkistus johtaa usein siihen, että vastaajaan joudutaan ottamaan yhteys ja näin selvittämään epäselvyydet.

Makroeditoinnissa (*macro editing, output editing*) havaintoja tarkastellaan yksikkötason sijasta aggregaattitasolla. Erilaisia makroeditointimenetelmiä on esitelty teoksessa Economic Commission for Europe (1994, sivut 111–147). Tämäntyylisistä menetelmistä käytetään myös nimitystä valikoiva editointi (*selective editing*) (Granquist & Kovar 1997). Sen sijaan, että tarkistettaisiin kaikki havaintoyksiköt, keskitytäänkin vain osaan havainnoista. Tarkistettavat havainnot valitaan sen perusteella, paljonko niiden vaikutus on kokonaisestimaatteihin. Menetelmissä painotetaan enemmän niitä havaintoja, joilla on suurempi vaikutus estimaatteihin. Täten eri havaintojen virheet eivät ole välttämättä yhtä tärkeitä. Näiden menetelmien tarkoituksena on vähentää editoinnin rahallisia ja ajallisia kustannuksia, välttää datan ylieditointia ja vähentää vastaajien vastausrasitetta. Ylieditoinnilla tarkoitetaan tilannetta, jolloin editointiin käytetyt resurssit eivät ole suhteessa aineiston laadun kanssa eli editoinnin lisääminen ei enää paranna datan laatua. Granquist (1994) vertaili eri menetelmiä ja totesi, että makroeditointimenetelmät vähensivät työtaakkaa 35–80 prosenttia verrattuna mikroeditointimenetelmiin.

EVR-aineiston editointia voidaan pitää koneellisena editointina, jota sovelletaan enimmäkseen mikrotasolla, mutta myös makrotasolla. EVR-aineiston editoinnissa ei käytetä manuaalista editointia eikä myöskään valikoivaa editointia. Rakennetilastotietokannan tärkeimmät eli suurimmat yritykset pyritään saamaan selville suoran kyselyn avulla. EVR-aineiston editoinnilla taas pyritään siihen, että kaikille sen yrityksille saataisiin umpeenmenevät tiedot.

4.2 Editointisäännöt

Editoinnissa käytetään apuna editointisääntöjä (*edit rules, edits*), joiden avulla aineiston oikeellisuutta voidaan tutkia. Säännöt voidaan jakaa luotettavuus-, johdonmukaisuus- ja tilastollisiin sääntöihin tai toisaalta vakavien ja epäilyttävien virheiden sääntöihin (Granquist 1995). Vakavien virheiden säännöillä (*fatal edits*) etsitään tilanteita, joissa havainnon tiedetään varmasti olevan virheellinen. Luotettavuus- ja johdonmukaisuussäännöillä yritetään yleensä löytää vakavia virheitä. Kuvitteellinen esimerkki tällaisesta säännöstä voisi olla "jos sukupuoli = mies, niin synnytetyt lapset = 0". Vakavat virheet tulisi aina pyrkiä korjaamaan, vaikka niillä ei välttämättä olisikaan suurta vaikutusta aineiston kokonaislaatuun.

Epäilyttävien virheiden säännöillä (*suspicious edits, query edits*) taas etsitään havaintoyksiköitä, joissa voidaan suurella todennäköisyydellä olettaa olevan virhe. Epäilyttävien virheiden säännöt perustuvat kahden tai useamman havaintoyksikön välisiin suhteisiin. Tällöin havainnoille määrätään etukäteen jokin hyväksymisalue, jonka ulkopuolelle jäätyään havainto merkitään epäilyttäväksi ja se ohjataan manuaaliseen tarkistukseen tai imputoitavaksi. Esimerkiksi suhdetarkistus $a < X/Y < b$ on tällainen editointisääntö. Siinä X ja Y ovat tarkasteltavia muuttujien arvoja ja a ja b ovat hyväksymisalueen ylä- ja alaraja. Asiantuntijat voivat määritellä hyväksymisalueet subjektiivisesti tai ne voidaan johtaa tilastollisen analyysin tuloksena, jolloin puhutaan tilastollisista säännöistä (*statistical edits*).

EVR-aineiston editoinnissa käytetyt alkueditointisäännöt ovat vakavien vir-

heiden sääntöjä, koska niillä saadaan selville periaatteessa varmoja virheitä. Virheen sijainnin selvittämiseen käytettävät säännöt ovat oikeastaan vakavien ja epäilyttävien virheiden sekoitus. Niillähän saadaan selville, että virhe on tietyllä välillä olevissa muuttujissa, mutta itse virheellistä tai virheellisiä muuttujia ei saada kuitenkaan selville.

4.3 Automatisoitu editointi ja imputointi

Tekniikan kehittymisen ansiosta tilastolliset surveyt alkoivat 1960-luvulla käyttää tietotekniikan suomia mahdollisuuksia. Samalla myös editoinnissa siirryttiin manuaalisesta tietokonepohjaiseen tapaan. Tämä ei tietenkään merkinnyt manuaalisen editoinnin loppumista, vaan yhä edelleen osa editoinnista tehdään käsin. Pierzchala (1995) mainitsee siirtymisen syitä: tarkistamisen nopeuttaminen, mahdollisuus useampien editointisääntöjen soveltamiseen, imputoinnin automatisoiminen, editoinnin objektiivisuus eli tietokone käsittelee kaikkia havaintoja samanarvoisesti, editoijien työn yksipuolisuuden vähentäminen ja heidän työpanoksensa suuntaaminen merkittävien virheiden korjaamiseen.

Merkittävä askel automatisoidun editoinnin ja imputoinnin kehittymisen kannalta on Fellegin & Holtin (1976) artikkeli "*A Systematic Approach to Automatic Edit and Imputation*". Siinä esitellään editointimetodologia, jonka kolme pääperiaatetta ovat:

1. Jokaisen havainnon täytyy noudattaa kaikkia editointisääntöjä siten, että muutettujen arvojen määrä on mahdollisimman pieni.

2. Imputointisääntöjen tulee olla seurausta editointisäännöistä.

3. Aineiston rakenteen tulee säilyä mahdollisimman alkuperäisenä.

Ensimmäisen periaatteen taustalla on ajatus siitä, että mahdollisimman paljon alkuperäisestä aineistosta jätetään ennalleen ja näin imputoidun datan osuus on

mahdollisimman pieni. Toinen periaate varmistaa sen, että imputoidut havainnot varmasti noudattavat kaikkia editointisääntöjä. Lisäksi se yksinkertaistaa editointi- ja imputointisääntöjen määrittelemistä sekä johtaa paremmin kontrolloituun editointi- ja imputointiprosessiin. Kolmannella kohdalla halutaan säilyttää korjatun datan rakenne samanlaisena kuin virheettömillä havainnoilla eli niillä, jotka noudattavat kaikkia editointisääntöjä.

4.4 Editointi EVR-aineistossa

Tässä kappaleessa kuvataan niitä editointivaiheita, jotka tehdään ennen varsinaisia vertailtavia menetelmiä. Käytetään näistä menetelmistä tässä nimeä *alkueditoinnit*. Nämä vaiheet ovat suurimmaksi osaksi kaikille kolmelle menetelmälle yhteisiä, joten kaikkia niitä ei oteta huomioon vertailussa. Virheiden generointi koeaineistoon suoritetaan siten, että tietyntyyppisiä alkueditoinneilla korjattavia virheitä ei aineistoon tuoteta, koska eri menetelmät käsittelisivät niitä kuitenkin täysin samalla tavalla. Tästä kerrotaan lisää luvussa 8.

EVR-aineiston editoitavat muuttujat on esitetty liitteessä A ja ne on merkitty $x_1, \ldots, x_{18}$ ja $y_1, \ldots, y_3$. Muuttujia *verm1–3* ja *myyntimu* ei editoida, koska niitä ei oteta mukaan rakennetilastotietokantaan. Nämä muuttujat sisältyvät luonnollisesti EVR-aineistoon, koska verottaja on niistä kiinnostunut, mutta tilastotuotannossa niistä ei olla kiinnostuneita.

4.4.1 Alkueditoinnit

Aluksi poistetaan tiedostosta kaikki sellaiset yritykset, joiden tiedot ovat pelkkiä nollia. Tämän jälkeen käytetään EVR-aineiston peruseditointisääntöä

$$|\sum_{i=1}^{18} x_i - y_3| < 1000, \tag{1}$$

jonka perusteella määrätään, onko havainto virheellinen vai ei. Mikäli tuloslaskelma menee umpeen alle 1000 markan erolla, havaintoyksikköä pidetään virheettö-

mänä. Editointisääntö (1) on luonnollinen tapa selvittää havainnon virheellisyys. Se sallii pienet alle 1000 markan virheet, mutta näillä ei ole merkitystä, koska lopulliseen aineistoon arvot pyöristetään tuhannen markan tarkkuudella. Toisaalta umpeenmenevässä havaintoyksikössä saattaa olla tämän editointisäännön jälkeenkin virheitä. Esimerkiksi siinä tapauksessa, että jossakin tuottomuuttujassa on yhtä suuri virhe kuin jossain toisessa kulumuuttujassa, nämä virheet ikäänkuin eliminoivat toisensa ja jäävät huomaamatta kaavalla (1). Tietysti voitaisiin tutkia havaintojen umpeenmenevyys myös välisummien osalta, mutta tällöin virheiden tulisi olla eri pätkissä, jotta ne huomattaisiin. Toisaalta myös välisummamuuttujissa *myyntika* ja *kayttoka* saattaa olla virheitä. Näillä ei kuitenkaan ole merkitystä umpeenmenevien havaintojen osalta, koska editoinnin lopuksi nämä välisummat lasketaan uudestaan tuloslaskelman muuttujista kaikille sekä virheellisille että virheettömille havaintoyksiköille.

Seuraavaksi käytetään virheellisiin havaintoihin erilaisia loogisia editointeja. Näillä etsitään sellaisia virheitä, jotka johtuvat yhdestä muuttujasta ja ovat käytännössä mahdollisia löytää. Ensimmäisenä näistä kokeillaan sijoittaa liikevaihdon tilalle myyntimuuttujien summa $x_1^* = verml + verm2 + verm3 + myyntimu$ ja testataan sitten umpeenmenevyys kaavalla

$$|x_1^* + \sum_{i=2}^{18} x_i - y_3| < 1000. \tag{2}$$

Mikäli ehto (2) toteutuu sijoitetaan aineistoon liikevaihdon tilalle $x_1 = x_1^*$ ja havaintoyksikkö merkitään korjatuksi.

Tämän jälkeen kokeillaan vaihtaa etumerkkiä sellaisilla muuttujilla, jotka voivat olla joko positiivisia tai negatiivisia. Näitä ovat *varastmu, poivarmu, verovali* ja *tiliktul*. Myyntikate ja käyttökate eivät ole tässä mukana, koska ne eivät vaikuta umpeenmenevyyteen. Esimerkiksi muuttujan *varastmu* tilanteessa sijoitetaan $x_4^* = -x_4$ ja testataan toteutuuko

$$|x_4^* + \sum_{i \neq 4} x_i - y_3| < 1000. \tag{3}$$

Ehdon toteutuessa sijoitetaan aineistoon $x_4 = x_4^*$. Vastaavasti tehdään muillekin edellä mainituille muuttujille. Testaukset tehdään yksi kerrallaan, joten sellaisia virheitä, joissa samassa havaintoyksikössä olisi useampi etumerkkivirhe, ei etsitä. Toisaalta silloin kasvaisi riski, että aineistoon tuotettaisiin näin lisää virheitä korjaamisen sijaan.

4.4.2 Virheen sijainti

EVR-aineisto on siinä mielessä hankala editoitava, että muuttujakohtaisia sääntöjä on vaikea edellä mainittujen loogisten sääntöjen lisäksi asettaa. Editoinnissa käytetäänkin hyväksi välisummatietoja, joiden avulla voidaan paikantaa virheen sijainti aineistossa tietylle välille. Tätä tietoa käytetään sitten hyväksi, kun korjausmenetelmiä sovelletaan. Tosin suhdeimputoinnissa tätä ei käytetä, vaan siinä imputoidaan arvot kaikille muuttujille. Rescaling-menetelmä ja sekamenetelmä käyttävät molemmat tätä tietoa, tosin sekamenetelmä vähän tarkemmin kuin rescaling. Voidaan tietysti ajatella, että menetelmien vertailun kannalta olisi oikein käyttää samaa tapaa molemmilla menetelmillä, mutta vertailussa halutaan saada selville menetelmien välisten erojen lisäksi se, kuinka editointi on Tilastokeskuksessa kehittynyt vuosien saatossa. Tämän takia menetelmiä sovelletaan sellaisina kuin ne ovat olleet tilastotuotannossa mukana, tosin joiltain osin niitä on muutettu. Näistä kerrotaan luvussa 8.

Editoitaville muuttujille pätevät tuloslaskelmasääntöjen perusteella seuraavat summautumissäännöt:

$$\sum_{i=1}^{7} x_i = y_1, \sum_{i=1}^{10} x_i = y_2 \text{ ja } \sum_{i=1}^{18} x_i = y_3. \tag{4}$$

Lisäksi liikevaihto on myyntimuuttujien summa. Näitä tietoja hyväksi käyttäen etsitään virheen sijainti aineistossa. Ensin sijoitetaan myyntikatteeseen vaikuttavien

muuttujien $x_1, \ldots, x_7$ tilalle muuttujan *myyntika* arvo y_1 ja testataan ehto

$$|y_1 + \sum_{i=8}^{18} x_i - y_3| < 1000. \tag{5}$$

Jos ehto (5) toteutuu, voidaan päätellä virheen sijaitsevan jossakin muuttujissa $x_1, \ldots, x_7$. Vastaavasti sijoitetaan käyttökate y_2 muuttujien $x_1, \ldots, x_{10}$ tilalle ja testataan ehtoa

$$|y_2 + \sum_{i=11}^{18} x_i - y_3| < 1000. \tag{6}$$

Ehdon (6) toteutuessa virhe ei sijaitse ainakaan muuttujissa $x_{11}, \ldots, x_{18}$. Samaan tapaan voidaan testata käyttökatteen jälkeisiä muuttujia $x_{11}, \ldots, x_{18}$ ehdolla

$$|\sum_{i=1}^{10} x_i - y_2| < 1000. \tag{7}$$

Mikäli ehto toteutuu, on virhe muuttujissa $x_{11}, \ldots, x_{18}$. Vastaava ehto voidaan muodostaa myös myyntikatteen jälkeisille muuttujille $x_8, \ldots, x_{18}$ kaavalla

$$|\sum_{i=1}^{8} x_i - y_1| < 1000. \tag{8}$$

Edellisten lisäksi käytetään ehtoa

$$|y_1 + \sum_{i=8}^{10} x_i - y_2| < 1000. \tag{9}$$

Tällä saadaan selville myyntikatteen ja käyttökatteen väliin jäävien muuttujien oikeellisuus. Edellä mainittuja ehtoja voidaan myös yhdistellä siten, että esimerkiksi ehtojen (6) ja (8) toteutuessa samanaikaisesti, päätellään virheen olevan myyntikatteen ja käyttökatteen väliin jäävissä muuttujissa $x_8, \ldots, x_{10}$. Edellä kuvatulla tavalla suoritetaan virheen sijainnin etsiminen rescaling-menetelmässä. Sekamenetelmässä käytetään edellä mainittujen ehtojen lisäksi tietoa $y_1 =$ *verm1 + verm2 + verm3 + myyntimu* eli testataan summautuvatko myyntimuuttujat liikevaihtoon ± 1000 markan erolla. Jos näin on, niin liikevaihtoa ei imputoida turhaan.

Mikäli mikään edellä mainituista ehdoista ei toteudu, täytyy korjata kaikki muuttujat. Virheen sijaintitietoa käytetään apuna korjausvaiheessa. Rescaling-menetelmää vastaava tapa jakaa virheelliset havainnot 7 eri luokkaan ja sekamenetelmän tarkempi tapa vastaavasti 11 luokkaan.

Niitä havaintoyksikköjä, jotka noudattavat perussääntöä (1) tai korjaantuvat aiemmin mainituilla alkueditointimenetelmillä, käytetään kaikissa kolmessa vertailumenetelmässä virheettöminä tietueina, joiden avulla virheellisiä korjataan. Kun kaikki yritykset on käyty läpi ja kaikille löydetty virheen sijainti, voidaan aloittaa varsinaiset korjausmenetelmät.

4.5 Virheellisten yritysten määrä EVR-98:ssa

Virheellisten yritysten osuuden määrittely EVR:ssä on hieman hankalaa, koska ei ole aivan selvää mistä yritysjoukosta se halutaan tietää. EVR-aineiston editointiproseduurissa käytetään myös sellaisia yrityksiä, joiden tiedot eivät tule rakennetilastotietokantaan tiettyjen rajoitusten takia. Yrityksen toimialatieto täytyy olla selvillä ja lisäksi yrityksen liikevaihdon tulee olla yli 50000 markkaa.

Koko EVR-98-aineiston tuloslaskelmatietojen osalta kaikkien yritysten lukumäärä on 272983. Näistä kahdenkertaisen kirjanpidon yrityksiä on 235906. Kun joukosta poistetaan sellaiset yritykset, joiden kaikkien tuloslaskelmamuuttujien arvot ovat nollia, jää jäljelle 235239 yritystä. Ehdon (1) täyttäviä eli virheettömiä yrityksiä on tässä joukossa 189114 (80.4 %). Liikevaihdon korvaaminen myyntimuuttujilla korjaa 939 (0.4 %) yritystä. Muuttujan *varastmu* etumerkin vaihtaminen korjaa 5025 (2.1 %) yritystä. Vastaavat luvut muille etumerkkikorjauksille ovat *poivarmu* 636 (0.3 %), *verovali* 4222 (1.8 %) ja *tiliktul* 905 (0.4 %). Yhteensä siis alkueditoinneilla saadaan korjattua 5.0 % yrityksistä. Täten jäljelle jäävä osuus (14.6 %) on virheellisiä yrityksiä.

Kun edellä kuvattua joukkoa rajoitetaan vielä siten, että otetaan mukaan yritykset, joiden toimiala tiedetään, jää jäljelle 219269 yritystä. Kun tästä otetaan

mukaan vain rakennetilastoasetuksen mukaisiin toimialojen pääluokkiin kuuluvat yritykset, jää jäljelle 179699 yritystä. Näistä 143506 (79.9 %) on virheettömiä. Liikevaihdon korvaaminen myyntimuuttujilla korjaa 677 (0.4 %) yritystä. Muuttujan *varastmu* etumerkin vaihtaminen korjaa 4169 (2.3 %) yritystä. Vastaavat määrät muilla etumerkkikorjauksilla ovat *poivarmu* 536 (0.3 %), *verovali* 3307 (1.8 %) ja *tiliktul* 658 (0.4 %). Yhteensä alkueditoinneilla korjaantuu 5.2 % virheellisistä yrityksistä, joten jäljelle jäävä 14.9 % yrityksistä on virheellisiä, jotka korjataan varsinaisilla imputointimenetelmillä.

4.6 Imputointi

Vaikka imputointimenetelmiä on olemassa useita ja monet niistä eroavat toisistaan paljonkin, voidaan lähes kaikki menetelmät esittää yleisen regressiomallin

$$\hat{y}_i = b_0 + \sum_j b_j z_{ij} + \hat{e}_i \tag{10}$$

avulla (Kalton & Kasprzyk 1986). Mallissa $\hat{y}_i$ on imputoitava arvo i:nnelle imputoitavalle havainnolle, z_{ij} kuvaa havainnon i apumuuttujien arvoja (indeksoitu j), b_0 ja b_j ovat regressiokertoimia muuttujan y ja apumuuttujien z_j väliselle regressiolle virheettömien yritysten osalta ja $\hat{e}_i$ on kullekin menetelmälle ennalta määrätty virhetermi. Esimerkiksi keskiarvoimputointi soluittain saadaan asettamalla $\hat{e}_i = 0$ ja määrittelemällä z_{ij} dummymuuttujaksi, joka ilmaisee imputointisolun h. Tällöin kaava (10) supistuu muotoon $\hat{y}_i = \bar{y}_h$ eli imputoidaan solun h keskiarvo.

Imputointimenetelmät voidaan jakaa Laaksosen (2000) mukaan neljään luokkaan:

1. Ei imputointia
2. Deduktiivinen tai looginen imputointi
3. Malliarvoimputointi (*model-donor imputation*)

4. Havaittuarvoimputointi (*real-donor imputation*).

Näistä ensimmäinen ei ole oikeastaan imputointimenetelmä, vaan ennemminkin vertailukohta menetelmille. Siinä ei imputoida ollenkaan puuttuvia tai virheellisiä havaintoja eli aineistosta käytetään vain niitä havaintoja, jotka eivät tarvitse imputointia (*available case method*).

Deduktiivinen imputointi perustuu loogiseen päättelyyn. Havaittujen ja puuttuvien arvojen välillä tiedetään olevan jokin tunnettu yhteys, jonka perusteella voidaan suurella varmuudella päätellä puuttuva tieto. Tällaista imputointia voidaan pitää ideaalisena imputointimenetelmänä.

Malliarvoimputoinnissa ajatellaan imputoitavan arvon luovuttajaksi (*donor*) jokin malli. Tällaisia menetelmiä ovat esimerkiksi keskiarvoimputointi, suhdeimputointi ja regressioimputointi. Malliin perustuvien menetelmien ominaisuutena on niiden kyky tuottaa myös sellaisia arvoja, joita ei oikeasti ole havaittu. Tästä saattaa tosin olla seurauksena epärealistisiakin arvoja.

Havaittuihin arvoihin perustuvissa menetelmissä imputoitavan arvon luovuttajana on oikeasti havaittu havaintoyksikkö, joten nämä menetelmät tuottavat realistisia oikeasti havaittuja arvoja. Esimerkkinä näistä menetelmistä mainittakoon hot deck-menetelmät, joita on kehitetty moniin eri tilanteisiin. Havaittuarvoimputoinnissa voi muodostua ongelmia, jos esimerkiksi havaitut arvot eivät kata kaikkia mahdollisia arvoja.

Edellisten menetelmien lisäksi voidaan puhua sekamenetelmistä, joissa käytetään sekä malliin että havaittuun arvoon perustuvaa imputointia. Joissain tilanteissa voidaan jopa osa havainnoista imputoida eri menetelmällä kuin toiset. Tällainen tapaus on vertailussa mukana oleva sekamenetelmä, jossa käytetään kolmea eri menetelmää.

Imputointimenetelmät voidaan luokitella myös toisella tavalla, deterministisiin ja stokastisiin menetelmiin (Kalton & Kasprzyk 1986, Kovar & Whitridge

1995, Schulte Nordholt 1998). Deterministisillä menetelmillä imputoitaessa määräytyy imputoitava arvo yksiselitteisesti yleensä johonkin malliin perustuen, kun taas stokastiset menetelmät sisältävät aina jonkinasteisen satunnaisuuden. Jos $\hat{y}_{mid}$ on jollakin deterministisellä menetelmällä imputoitu arvo, niin vastaava arvo stokastisella menetelmällä on $\hat{y}_{mis} = \hat{y}_{mid} + \hat{e}_{mi}$. Yleensä virhetermi $\hat{e}_{mi}$ valitaan siten, että sen odotusarvo $E(\hat{e}_{mi}) = 0$, mistä seuraa $E(\hat{y}_{mis}) = E(\hat{y}_{mid})$. Determinististen menetelmien ongelmana on usein se, että ne vääristävät imputoitavan muuttujan jakaumaa. Imputoitujen arvojen keskittyminen tiettyihin kohtiin aiheuttaa varianssin aliestimoitumisen. Tämän takia käytetään stokastisia menetelmiä, joiden avulla muuttujien jakaumat pyritään säilyttämään.

5 Suhdeimputointi

Suhdeimputointi (*ratio imputation*) on deterministinen eli ei-satunnainen menetelmä. Laaksosen (2000) jaottelun mukaan se on malliarvoimputointimenetelmä. Suhdeimputointi on yleisesti käytetty menetelmä (Shao 2000) ja sen periaate on melko yksinkertainen. Asettamalla $b_0 = 0$ ja $\hat{e}_i = 0$ suhdeimputointi voidaan esittää kaavan (10) avulla muodossa

$$\hat{y}_i = bz_i, \tag{11}$$

missä b on suhde, jonka avulla imputoidaan ja z_i on apumuuttuja, jonka arvo tiedetään kaikille virheellisillekin havainnoille. Merkitään virheettömien havaintojen joukkoa C. Suhde voi olla esimerkiksi kahden eri muuttujan y ja z summien suhde, jolloin imputoitava arvo saadaan

$$\hat{y}_i = \left(\sum_{i \in \mathrm{C}} y_i \Big/ \sum_{i \in \mathrm{C}} z_i \right) z_i.$$

Vastaavasti voidaan käyttää vaikkapa korjattavan havainnon aiempia muuttujien arvoja y_{t-1} ja z_{t-1} imputoitaessa ajanhetkellä t. Nämä tiedot voidaan saada esimerkiksi aiemmasta surveystä. Yleensä suhdeimputointi suoritetaan imputointisoluittain. Oletetaan aineiston jakautuvan imputointisoluihin $h = 1, \ldots, H$. Tällöin suhde b_h lasketaan virheettömistä havainnoista C_h.

5.1 Suhdeimputointi EVR-aineistossa

Suhdeimputointia on käytetty vuosien 1994–1997 EVR-aineiston editoinnissa. Imputoitaessa soluun h kuuluvan yrityksen arvoja käytetään imputoitavan muuttujan summaa joukossa C_h jaettuna saman joukon EVR:stä saatavalla liikevaihdon summalla. Tämä suhde kerrotaan apumuuttujalla, joka tässä tapauksessa on yritys- ja toimipaikkarekisteristä saatava liikevaihto. YTR:ssä on jokaiselle imputoitavalle yritykselle tieto liikevaihdosta. YTR:n liikevaihtotiedot on koottu eri

tietolähteistä ja ne vastaavat suurimmalta osalta EVR:n tietoja. Imputointisoluina käytetään toimialan ja liikevaihdon suuruuden mukaisia luokkia. Liikevaihdon suuruus on jaettu neljään eri luokkaan. Pienimpään luokkaan kuuluvat alle miljoonan markan liikevaihdon yritykset, seuraavaan alle 10 miljoonan, kolmanteen alle 50 miljoonan ja suurimpaan yritykset, joiden liikevaihto on yhtä suuri tai suurempi kuin 50 miljoonaa markkaa. Näin saadaan imputointisolut $h = 1, \ldots, H$, jotka jakautuvat toimialojen lisäksi vielä neljään eri luokkaan. Täten soluun h kuuluvan virheellisen havainnon i muuttujan x_j, $j = 2, \ldots, 18$, imputointi voidaan esittää

$$\hat{x}_{ij} = \left(\sum_{i \in \mathbf{C}_h} x_{ij} \Big/ \sum_{i \in \mathbf{C}_h} x_{i1} \right) x_{i1}^*, \quad j = 2, \ldots, 18, \tag{12}$$

missä x_{i1}^* on YTR:stä saatu liikevaihto. Liikevaihtomuuttujaa x_1 ei korjata suhdeimputoinnilla, vaan EVR:ssä olevan arvon paikalle sijoitetaan YTR:stä saatava liikevaihto.

EVR:n suhdeimputointi eroaa rescaling- ja sekamenetelmästä siinä, että suhdeimputoinnissa ei etsitä virheen sijaintia aineistossa, vaan kaikki tuloslaskelman virheellisten yritysten muuttujat $x_1, \ldots, x_{18}$ korjataan. Tosin liikevaihtoa x_1 ei suhdeimputoida, vaan sen tilalle sijoitetaan YTR:ssä oleva tieto.

6 Rescaling-metodi

Rescaling on Tilastokeskuksessa EVR-aineistoa varten kehitetty menetelmä virheiden korjaamiseen ja sitä on käytetty tilastotuotannossa vuoden 1998 EVR-aineiston editoinnissa. Se soveltuu jatkuvien muuttujien korjaamiseen tilanteissa, joissa aineistossa on muuttujia, joihin muut muuttujat summautuvat. EVR-aineistossa näitä ovat myyntikate, käyttökate ja tilikauden tulos. Liikevaihtoa ei tässä tarkoituksessa käytetä summamuuttujana, koska myyntimuuttujia ei käytetä tilastotuotannossa ja täten niitä ei tarvitse korjata. Menetelmän nimi tulee tavasta, jolla virheitä korjataan. Siinä virhe jaetaan korjattavien muuttujien kesken siten, että summautumissäännöt saadaan toteutumaan.

Rescaling ei ole varsinainen imputointimenetelmä, joten tässä on parempi puhua korjausmenetelmästä. Imputointia käytetään yleensä puuttuviin arvoihin tai sitten editointisäännöillä virheellisiksi todettuihin arvoihin, jolloin virheellinen arvo korvataan uudella imputoidulla arvolla. Imputoinnissa ei siis käytetä tietoa virheellisestä muuttujan arvosta. Rescaling-menetelmässä taas tämä tieto käytetään ja valmiita arvoja "skaalataan" kertoimilla, jotta ne saadaan summautumaan oikein. Tämä tuntuukin järkevältä, koska virheiden etsimissäännöillä saadaan selville joukko muuttujia, joissa yhdessä tai useammassa on virhe ja näin tiedetään monien korjattavien arvojen olevan oikeita. Rescaling yrittää säilyttää arvot mahdollisimman lähellä oikeita. Tosin tämä riippuu korjattavan havainnon virheen suuruudesta.

Aluksi virheen sijainti selvitetään kappaleessa 4.4.2 kerrotulla tavalla. Merkitään tätä virheellisten muuttujien joukkoa $\mathbf{E}$. Tämän jälkeen lasketaan virheen suuruus ε käyttäen apuna summamuuttujia, joiden oletetaan olevan virheettömiä. Lisäksi lasketaan summa

$$S = \sum_{i \in \mathbf{E}} |x_i| \tag{13}$$

eli virhevälin muuttujien absoluuttinen summa. Tässä tulee huomata, että indeksi

i ei viittaa havaintoon, vaan muuttujaan. Esityksen selkeyttämiseksi havaintoja ei ole indeksoitu. Skaalauskerroin k saadaan virheen suhteena absoluuttisesta summasta $k = \varepsilon/S$. Tämän jälkeen virhevälin positiiviset muuttujat kerrotaan luvulla $(1-k)$ ja negatiiviset luvulla $(1+k)$. Lisäksi aineistossa on muuttujat *varastmu*, *poivarmu*, *verovali* ja *tiliktul*, joissa voi olla joko positiivisia tai negatiivisia arvoja. Näissä tapauksissa käytetään vastaavia kertoimia arvon etumerkin mukaan. Täten skaalaus voidaan ilmaista

$$x_i^* = \begin{cases} (1-k)\cdot x_i, & \text{kun } x_i \geq 0 \\ (1+k)\cdot x_i, & \text{kun } x_i < 0, \end{cases} \tag{14}$$

missä x_i^* on korjattu arvo muuttujalle $i \in \mathbf{E}$.

Seuraavaksi kuvataan rescaling-menetelmää esimerkin avulla. Oletetaan kaavan (1) mukaan virheellisen tietueen toteuttavan ehdon (5). Tällöin päätellään virheen sijaitsevan muuttujissa $x_1, \ldots, x_7$ ja oletetaan muuttujat y_1, y_2, y_3 sekä $x_8, \ldots, x_{18}$ virheettömiksi. Ensin lasketaan virhe kaavalla

$$\varepsilon = \sum_{i=1}^{7} x_i - y_1. \tag{15}$$

Seuraavaksi lasketaan virhevälin muuttujien absoluuttinen summa

$$S = \sum_{i=1}^{7} |x_i|. \tag{16}$$

Skaalauskerroin on siten

$$k = \frac{\varepsilon}{S} = \frac{\sum_{i=1}^{7} x_i - y_1}{\sum_{i=1}^{7} |x_i|}. \tag{17}$$

Tällöin virhevälin eri muuttujien korjaaminen tapahtuu seuraavilla säännöillä:

- $x_i^* = (1-k)\cdot x_i$, kun $i = 1, 2$
- $x_i^* = (1+k)\cdot x_i$, kun $i = 3, 5, 6, 7$
- $x_4^* = (1-k)\cdot x_4$, kun $x_4 \geq 0$

- $x_4^* = (1 + k) \cdot x_4$, kun $x_4 < 0$.

Muuttujat *liikevai* ja *tuottomu* (x_1, x_2) ovat positiivisia, joten niiden kertoimeksi tulee $(1 - k)$. Tällöin virheen ε ollessa positiivinen eli muuttujien $x_1, \ldots, x_7$ summan ollessa suurempi kuin myyntikate y_1, pienennetään muuttujien x_1 ja x_2 arvoja kertomalla ne luvulla $(1 - k)$, missä k on positiivinen. Vastaavasti arvoja kasvatetaan, jos k on negatiivinen eli jos muuttujien $x_1, \ldots, x_7$ summa on pienempi kuin myyntikate y_1. Täten muuttujat "skaalataan" summautumaan oikein.

Edellä kuvatut säännöt pätevät mikäli $-1 \leq k \leq 1$. Jos $k > 1$ tai $k < -1$, eivät edellä mainitut säännöt toimi, koska tällöin skaalaus vaihtaisi muuttujien etumerkit. Tällöin toimitaan seuraavasti. Mikäli $k > 1$, niin lasketaan summaan S vain negatiiviset muuttujan arvot eli esimerkkitapauksessa muuttujat x_3, x_5, x_6 ja x_7 sekä x_4, jos se on negatiivinen. Tämän jälkeen lasketaan k käyttäen edellä mainittua arvoa S ja kerrotaan ainoastaan negatiiviset muuttujat luvulla $(1 + k)$. Vastaavasti tapauksessa $k < -1$ lasketaan summaan S vain positiiviset muuttujat eli x_1 ja x_2 sekä x_4, jos se on positiivinen. Tällöin skaalataan ainoastaan positiiviset muuttujat.

Rescaling-menetelmä korjaa virheellisen välin muuttujat siten, että ne summautuvat skaalaamisen jälkeen oikein. Todistetaan tämä seuraavaksi käyttäen edellä kuvattua esimerkkitilannetta. Asian yksinkertaistamiseksi oletetaan $x_4 < 0$. Sijoitetaan korjatut muuttujien arvot $x_1^*, \ldots, x_7^*$ käytettävään editointisääntöön ja

saadaan

$$\sum_{i=1}^{7} x_i^* - y_1 =$$
$$\sum_{i=1}^{2} x_i(1-k) + \sum_{i=3}^{7} x_i(1+k) - y_1 =$$
$$\sum_{i=1}^{2} x_i(1-\varepsilon/S) + \sum_{i=3}^{7} x_i(1+\varepsilon/S) - y_1 =$$
$$\sum_{i=1}^{2} x_i + \sum_{i=3}^{7} x_i - y_1 - (\varepsilon/S)(\sum_{i=1}^{2} x_i - \sum_{i=3}^{7} x_i) =$$

$$\varepsilon - (\varepsilon/S)S = 0,$$

koska

$$\sum_{i=1}^{2} x_i + \sum_{i=3}^{7} x_i - y_1 = \sum_{i=1}^{7} x_i - y_1 = \varepsilon$$

ja

$$\sum_{i=1}^{2} x_i - \sum_{i=3}^{7} x_i = \sum_{i=1}^{7} |x_i| = S.$$

Vastaavasti voidaan todistaa tilanne $x_4 > 0$. Eli rescaling-menetelmä siis jakaa virheen muuttujiin siten, että ne summautuvat täysin oikein käytettävään apusummaan nähden. Tämän seurauksena editoitava tietue noudattaa varmasti perussääntöä (1). Tosin ero ei välttämättä ole nolla, koska muualla kuin korjatulla välillä saattaa olla alle 1000 markan ero. Mikäli virhettä ei pystytä paikallistamaan millekään välille, käytetään rescaling-menetelmää kaikille muuttujille lukuun ottamatta liikevaihtoa x_1.

7 Sekamenetelmä

Uusinta tilastotuotannossa käytettyä menetelmää on sovellettu vuoden 1999 EVR-aineistoon. Käytetään siitä nimitystä *sekamenetelmä*, koska siinä sovelletaan kolmea eri korjaustapaa. Menetelmät ovat soveltamisjärjestyksessä

- Outlier-menetelmä
- Lähimmän naapurin imputointi
- Suhdeimputointi.

Ensin käytetään siis outlier-lähestymistapaa, toiseksi lähimmän naapurin menetelmää niihin, joita ei outlier-metodilla saatu korjattua. Lopuksi suhdeimputointia sovelletaan niihin yrityksiin, joille ei voitu päätellä virheen sijaintia. Suhdeimputoinnilla korjataan siis kaikki muuttujat.

Sekä outlier- että lähimmän naapurin menetelmässä käytetään apuna toimialaluokitusta. Tarkimmillaan toimialaluokitus on viisinumerotasolla ja pudottamalla aina viimeinen numero päästään hierarkisesti yksinumerotasolle asti. Näistä käytetään apuna 5-, 3- ja 1-numerotasoja. Jokaisella virheellisellä ja virheettömällä yrityksellä on siis toimialatieto, joka saadaan YTR:stä. Virheettömästä datasta lasketaan kuhunkin viisinumerotason luokkaan h_5 kuuluvien yritysten lukumäärä $\#h_5$. Mikäli $\#h_5 \geq 50$, käytetään korjausmenetelmissä luokkaan h_5 kuuluville virheellisille yrityksille apuna virheettömän datan vastaavan luokan viisinumerotason tietoja. Jos $\#h_5 < 50$, mutta $\#h_3 \geq 50$, käytetään luokkaan h kuuluvalle virheelliselle yritykselle kolminumerotason tietoja virheettömistä yrityksistä. Jos $\#h_3 < 50$, käytetään yksinumerotason tietoja.

7.1 Outlier-menetelmä

Outlier-menetelmällä etsitään yksittäisiä muuttujia, jotka korjaamalla tietue saadaan hyväksytyksi. Tällä pyritään välttämään tietoista oikeiden arvojen muuttamista, mitä tapahtuu lähes kaikissa muissa EVR-aineiston korjausmenetelmissä. Apuna käytetään virheettömästä datasta laskettuja muuttujakohtaisia fraktiileja, joiden avulla yritetään löytää ja korjata virhe ainoastaan yhtä tai kahta muuttujaa korjaamalla.

Jokaiselle virheelliselle yritykselle lasketaan virhevälin muuttujille niiden prosentuaalinen osuus liikevaihdosta

$$s_i = \frac{x_i}{x_1} \cdot 100, \quad i \in \mathbf{E}. \tag{18}$$

Vastaavasti lasketaan osuudet myös virheettömille yrityksille. Näille virheettömien yritysten osuuksille lasketaan 1. ja 9. desiilit. Merkitään näitä $D_1(s_i)$ ja $D_9(s_i)$. 10 % havainnoista on pienempiä kuin 1. desiili ja vastaavasti 10 % on suurempia kuin 9. desiili. Täten näiden desiilien väliin jää 80 % aineistosta. Lisäksi tarvitaan virheen $\varepsilon = \sum_{i=1}^{18} -y_3$ osuus

$$s_\varepsilon = \frac{\varepsilon}{x_1} \cdot 100. \tag{19}$$

Periaatteena on siirtää virhe ε johonkin yksittäiseen muuttujaan x_i, $i \in \mathbf{E}$, seuraavien ehtojen avulla. Eron ollessa positiivinen, $\varepsilon > 0$, käytetään negatiivisille muuttujille ehtoja

$$s_i < D_1(s_i) \quad \text{ja} \quad D_1 \leq s_i + s_\varepsilon \leq D_9. \tag{20}$$

Molempien ehtojen (20) toteutuessa korjataan arvo $x_i^* = x_i + \varepsilon$. Vastaavasti positiivisille muuttujille käytetään ehtoja

$$s_i > D_9(s_i) \quad \text{ja} \quad D_1 \leq s_i - s_\varepsilon \leq D_9. \tag{21}$$

Molempien ehtojen (21) toteutuessa korjataan $x_i^* = x_i - \varepsilon$. Eron ollessa negatiivinen, $\varepsilon < 0$, muuttuu negatiivisten muuttujien ensimmäinen ehto muotoon

$s_i > D_9(s_i)$ ja vastaavasti positiivisilla $s_i < D_1(s_i)$. Toinen ehto ja korjauskaava pysyvät samoina. Menetelmää sovelletaan muuttuja kerrallaan siinä järjestyksessä, jossa muuttujat ovat tuloslaskelmassa. Täten menetelmällä voi yleensä korjaantua vain yksi kunkin virheellisen yrityksen muuttujan arvo. Poikkeuksen tekee tapaus, jossa virhe on sekä välillä liikevaihdosta myyntikatteeseen että käyttökatteesta tilikauden tulokseen. Tällöin molemmat välit tutkitaan erikseen, joten yhdestä tietueesta voidaan korjata kaksi muuttujan arvoa. Siinä tapauksessa virheen ε sijasta lasketaan erikseen virheet molemmille väleille käyttäen apuna myyntikate-, käyttökate- ja tilikauden tulos-tietoja.

On huomattava, että vaikka molemmat edellä mainitut ehdot toteutuvat ja muuttujan x_i arvo korjataan, ei havainto välttämättä korjaannu, jos sekä välillä liikevaihdosta myyntikatteeseen että käyttökatteesta tilikauden tulokseen on virhe. Tämän takia lopuksi vielä testataan kaavalla (1) umpeenmenevyys. Mikäli ehto ei toteudu, jää havainto lähimmän naapurin menetelmällä korjattavaksi.

7.2 Lähimmän naapurin imputointi

Lähimmän naapurin imputointi (*nearest neighbor imputation, distance function matching*) luetaan yleensä kuuluvaksi hot deck-menetelmiin. Hot deck-menetelmille ominaista on, että imputoitava arvo on jokin oikeasti havaittu arvo eli Laaksosen (2000) luokituksen mukaan kyseessä on havaittuarvoimputointi. Lähimmän naapurin imputointi on deterministinen menetelmä, koska luovuttajahavainto määräytyy yksiselitteisesti etäisyysmitan avulla. Hot deck-menetelmä voi olla myös stokastinen, kuten satunnainen hot deck (*random hot deck*), jossa luovuttaja valitaan satunnaisesti joko kaikista tai samaan imputointisoluun kuuluvista virheettömistä havainnoista.

Lähimmän naapurin menetelmässä etsitään kullekin virheelliselle havaintoyksikölle i ominaisuuksien puolesta mahdollisimman samankaltainen yksikkö virheettömistä j (Chen & Shao 2000). Tämän mahdollisimman samanlaisen havain-

toyksikön eli lähimmän naapurin etsiminen tapahtuu käyttäen apuna etäisyysmittaa. Lessler & Kalsbeek (1992, sivut 218–219) esittävät erilaisia etäisyysmittoja. Eräänä vaihtoehtona mainitaan myös kaavan (32) Mahalanobisin etäisyys. EVR-aineistossa havaintojen i ja j välisenä etäisyysmittana käytetään

$$D_{ij} = \sum_{k \in \mathbf{F}} |\log(x_{ik}) - \log(x_{jk})|, \tag{22}$$

missä $\mathbf{F}$ on käytettävien muuttujien x_k joukko. Käytettäviä muuttujia on 4–10 kappaletta. Tämä joukko vaihtelee sen mukaan, missä välissä virhe on aineistossa. Käytetyt muuttujat ovat *liikevai, ostoyht, palkkamu, kulumuut, palkkaki, vuokra, kulukiin, rahtuott, korkokuk* ja *poivarmu*. Luotettavinta olisi tietysti käyttää ainoastaan niitä muuttujia, joiden tiedetään olevan oikein eli virhevälin ulkopuolisia muuttujia. Joissain tilanteissa, kuten virheettömän välin ollessa käyttökatteesta tilikauden tulokseen, käytetään kuitenkin tietoisesti havainnon i virhevälin muuttujia. Näin toimitaan, koska virheettömän välin muuttujat ovat luonteeltaan sellaisia, että ne voivat "samanlaisillakin" yrityksillä vaihdella paljon. Havainnolle i imputoidaan sen virheelliseen osaan havainnon j, jonka kanssa etäisyys saa pienimmän arvonsa min D_{ij}, tiedot.

Lähimmän naapurin imputointi suoritetaan EVR:ssä imputointisoluittain. Soluina toimivat toimialaluokat. Kalton & Kasprzyk (1986) mainitsevat eräänä imputointimenetelmänä hierarkisen hot deckin (*hierarchical hot-deck imputation*). Siinä imputoinnissa käytetään imputointisoluja, joita voidaan tarpeen tullen yhdistää laajemmiksi soluiksi. Näin tehdään mikäli tarkimmalla solutasolla ei ole mahdollista löytää arvon luovuttajaa eli esimerkiksi samassa solussa ei ole yhtään virheetöntä havaintoa. EVR-aineiston imputoinnissa käytetään hyväksi toimialaluokitusta, joka on hierarkinen. Sääntönä pidetään sitä, että solussa täytyy olla vähintään 50 virheetöntä yritystä, jotta imputointi suoritetaan kyseisellä tasolla. Tästä on kerrottu tarkemmin sivulla 35. Tällä pyritään välttämään saman luovuttajan käyttämistä useampaan kertaan. Kuitenkin on mahdollista, että sama yritys toimii luovuttajana useamman kerran.

Imputoimalla saadut arvot joudutaan lopuksi muuttamaan rescaling-menetelmällä, jotta ne summautuvat oikein. Tämän takia ei ehkä ole oikein puhua puhtaasta hot deck-menetelmästä.

Sekamenetelmän lopuksi ne yritykset, joiden virheen sijaintia ei pystytä löytämään, imputoidaan käyttäen luvussa 5 kuvattua suhdeimputointia.

8 Koeaineisto

Tässä kappaleessa kuvataan tarkasti se prosessi, jolla keinotekoinen virheellinen aineisto tehdään. Aluksi määritellään käytetty virheetön aineisto, minkä jälkeen kuvataan virhekertoimien muodostaminen sekä virheiden vieminen virheettömään aineistoon.

8.1 Virheetön aineisto

Jotta voidaan luoda keinotekoinen virheellinen data ja verrata menetelmien kykyä aineiston korjaamiseen, tarvitaan tietysti jokin virheetön aineisto. Nimetään tämä virheetön aineisto tässä OK-dataksi. Barcaroli & D'Aurizio (1997) mainitsevat useita eri keinoja OK-datan luomiseksi. Yksi mahdollisuus on uudelleenkysely, joka suoritetaan erittäin huolellisesti, ja näin saadaan parempilaatuinen aineisto. Myös aineiston syöttövaiheen editointia voidaan tehostaa ja näin vähentää syöttövaiheessa syntyviä virheitä. Tällaiset menetelmät eivät useinkaan ole käytännössä toimivia ratkaisuja. Ensinnäkin ne ovat yleensä kalliita toteuttaa ja toisekseen dataan jää kuitenkin virheitä, tosin virheitä on luultavasti vähemmän kuin ennen. Näiden syiden takia on usein parempi hankkia OK-data jollakin seuraavista tavoista. Jos on käytössä muita tietolähteitä, kuten jokin toinen kysely tai rekisteri, jossa on samoja muuttujia samoilta havaintoyksiköiltä samalta ajanjaksolta, voidaan tietoja yhdistää. Tämä tietysti vaatii, että muiden tietolähteiden aineisto on korkealaatuista. Yksi tapa on luoda aineisto täysin keinotekoisesti johonkin malliin perustuen. Tämä on tietysti melko vaivaton keino, mutta menetelmien vertailun luotettavuus riippuu voimakkaasti siitä, kuinka hyvin valittu malli kuvaa todellista tilannetta. Yleisesti käytetty keino on käyttää tutkittavaa editointimenetelmää raakaan aineistoon ja määritellä OK-dataksi ne havaintoyksiköt, joissa ei ollut ollenkaan virheitä (Garcia Rubio & Peirats 1994). Se on helppo toteuttaa, mutta saatava aineisto ei tietenkään täysin vastaa aitoa dataa. Jos näin olisi, niin virheethän olisivat silloin harmittomia.

Tässä käytetään viimeksi mainittua menetelmää eli sovelletaan kaavan (1) editointisääntöä raakaan EVR-98-aineistoon ja näin saadaan selville virheettömät havainnot eli OK-data. Havainnon oikeellisuutena pidetään siis sitä, että sen tuloslaskelma menee umpeen ± 1000 markan erolla. Tässä on huomioitava, että alkueditoinneilla (kappale 4.4.1) korjatut yritykset eivät tule mukaan tähän OK-dataan. Lisäksi OK-dataan tulevat yritykset rajataan siten, että mukaan otetaan rakennetilastoasetuksen piiriin kuuluvat yritykset eli ne joiden toimialan pääluokka on C, D, E, F, G, H, I tai K. Lisäksi OK-datasta on poistettu sellaiset yritykset, joille ei löydy toimialatietoa viisinumerotasolla tai joilla kaikkien tuloslaskelman muuttujien arvot ovat nollia tai joiden tiedot saadaan suoralla kyselyllä. Näiden putsausten jälkeen OK-dataan jää 116146 yritystä.

8.2 Virhetyypit

Jotta virheiden generoiminen olisi mahdollisimman aidonmukaista, täytyy virheiden luonteesta ja rakenteesta olla jonkinlaista etukäteistietoa. Artikkelissaan Barcaroli & D'Aurizio (1997) olettavat virheiden syntyvän kahdessa eri vaiheessa. Ensinnäkin vastausten antamisessa syntyvät virheet ja toisaalta datan syöttämisessä syntyvät virheet. He olettavat syntyvän sekä stokastisia että systemaattisia virheitä. Vastausten antamisvaiheessa syntyviä virheitä ovat muun muassa eräkato eli puuttuvat arvot, vastausten vaihtuminen eli annetaan vahingossa vastaukseksi jonkin toisen kysymyksen vastaus, väärän mittayksikön käyttäminen (esimerkiksi annetaan tieto tuhansina markkoina markkojen sijasta) tai joidenkin vastausten vähättely tai liioittelu. Datan syöttövaiheessa mahdollisia virheitä ovat muun muassa näppäilyvirheet. Näiden virheiden generointi eroaa siinä, että virheet ajatellaan merkkikohtaisiksi, kun ensimmäisen vaiheen virheitä generoidaan muuttujatasolla. Merkkitasolla virheiden generointi suoritetaan siten, että esimerkiksi vaihdetaan vastatun jatkuvan muuttujan kaksi peräkkäistä merkkiä keskenään.

Jotta generoitavat virheet olisivat mahdollisimman samankaltaisia kuin aidot

virheet, edellä mainittu virheiden generointitapa vaatii tarkkaa tietoa eri virhetyyppien tapahtumisen todennäköisyydestä sekä niiden syntymekanismista. EVR-aineiston tapauksessa tällaisen tiedon hankkiminen voisi olla joiltain osin mahdollista, mutta pääsääntöisesti tällaista tietoa ei voida saada selville. Tämän takia virheiden generointi toteutetaan seuraavalla tavalla.

8.3 Suora kysely aputietona

Virheiden generoinnissa käytetään apuna Tilastokeskuksen suoraa kyselyä yrityksiltä. Suoran kyselyn tietoja voidaan pitää oikeina. Kyselyssä on $R = 1093$ sellaista vertailukelpoista yritystä, jotka ovat virheellisiä EVR:ssä. Näistä yrityksistä tiedetään siis sekä oikeat että virheelliset tiedot, minkä avulla virheet luodaan koeaineistoon. Suoran kyselyn tiedoista saadaan kaikki EVR:n editoitavien muuttujien arvot lukuun ottamatta myyntikatetietoa. Palkkamuuttujille (x_6, x_8) ja kulumuuttujille (x_7, x_{10}) saadaan kyselystä ainoastaan summatiedot, mikä on otettava huomioon virheiden generoinnissa. Osa EVR-aineistoa vastaavista muuttujista lasketaan suorassa kyselyssä useamman muuttujan summana.

Kuten taulukosta 1 nähdään, ovat suoran kyselyn yritykset suurempia kuin tässä koedatana käytettävät EVR-yritykset. Tässä siis generoidaan suurten yritysten virheitä pienempiin yrityksiin. Sitä on mahdoton selvittää, poikkeavatko eri kokoluokkien virhetyypit toisistaan, koska käytettävissä on vain suurten yritysten oikeat tiedot. Tässä kuitenkin oletetaan, että ne eivät merkittävästi eroa toisistaan. Virheiden suuruusluokka on tietysti kyselyn yrityksillä suurempi, koska muuttujien arvotkin ovat keskimäärin suurempia. Tämä on otettu huomioon generoimalla suhteellisia virheitä. Lisäksi virheiden generointi pyritään suorittamaan mahdollisimman huolellisesti, jotta kokoeron vaikutus minimoituisi. Esimerkiksi pienillä yrityksillä on enemmän nolla-arvoja kuin suurilla. Näihin virheen vieminen saattaa epäonnistua, koska nollan kertominen toisella luvulla ei muuta muuttujan arvoa. Tämän takia generoinnissa valvotaan, että jokainen virhe tulee koeaineistoon

sellaisena kuin on tarkoitus.

8.4 Virhevektorit

Virheiden generoinnissa käytetään eräänlaisia virhevektoreita, joita viemällä virheettömän aineiston tietueisiin luodaan virheitä. Nämä virhevektorit saadaan periaatteessa jakamalla kunkin EVR-muuttujan arvo vastaavan yrityksen kyselytiedolla. Tällöin siis oikeiden tietojen kertoimina on ykkösiä ja virheellisten tietojen kertoimina jotain eri suurta kuin 1.

Merkitään suoran kyselyn tietoja x_i^*. Ensin etsitään virheelliset yritykset, joissa on ainoastaan yhdessä muuttujassa virheellinen arvo. Ne saadaan selville sijoittamalla kyselyn oikea tieto kunkin EVR-muuttujan arvon paikalle yksi kerrallaan ja testaamalla umpeenmenevyys. Virheen sijainti EVR-tietueessa tiedetään kappaleessa 4.4.2 esitellyllä tavalla. Esimerkiksi virheen ollessa välillä $x_{11}, \ldots, x_{18}$ käytetään testiä

$$|y_2 + x_i^* + \sum_{i' \neq i} x_{i'} - y_3| < 1500,\ i = 11, \ldots, 18. \tag{23}$$

Tässä käytetään lievempää rajaa ± 1500 markkaa, koska osa suoran kyselyn tiedoista on yhdistetty useista muuttujista, mikä tuo lukuihin pientä heittoa verrattuna EVR-lukuihin. Lisäksi täytyy muistaa, että kyselyn tiedot on ilmoitettu tuhansina markkoina, kun EVR:ssä tiedot on markan tarkkuudella. Tämänkin takia luvut poikkeavat toisistaan. Jos ehto (23) toteutuu, niin voidaan päätellä virheen olevan muuttujassa x_i, jonka arvo korvattiin oikealla tiedolla x_i^*. Soveltamalla vastaavaa testiä muillekin virheväleille saadaan selville virheen sijainti havainnoissa, joissa virhe on yhdessä muuttujassa.

Niissä havainnoissa, jotka eivät toteuta edellä olevaa ehtoa millään muuttujalla x_i^*, on useampi kuin yksi virhe. Nämä saadaan selville ehdolla

$$|x_i - x_i^*| < 1000,\ i = 1, \ldots, 18. \tag{24}$$

Jos ehto ei toteudu, niin silloin kyseessä olevassa muuttujassa x_i on virhe. Ehdon toteutuessa tulee virhevektoriin vastaavaan kohtaan kerroin 1.

Kun kaikki virheelliset arvot on saatu selville, voidaan generoida virheitä OK-dataan eli EVR:n oikeisiin tietoihin. Tämä tehdään virhevektoreiden, jotka koostuvat eräänlaisista kertoimista, avulla. Kertoimet ovat muotoa

$$c_{x_i} = \frac{x_i}{x_i^*},\ i = 1, \ldots, 18. \tag{25}$$

Muuttujien $x_1, \ldots, x_{18}$ lisäksi kertoimet lasketaan myös käyttökatteelle y_2 ja tilikauden tulokselle y_3. Myyntikatetta vastaavaa tietoa ei suoran kyselyn tiedoista saada muodostettua, koska palkkamuuttujista *palkkamu* ja *palkkaki* sekä kulumuuttujista *kulumuut* ja *kulukiin*, jotka ovat tuloslaskelmassa myyntikatteen eri puolilla, tiedetään vain niiden summat. Virhevektori r muodostuu siis kertoimista

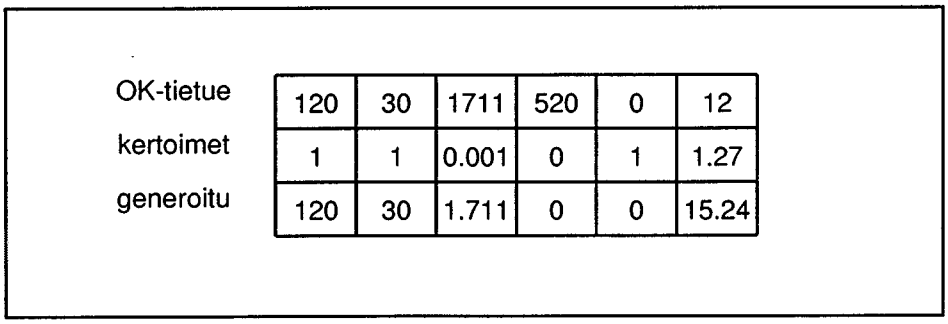
$$\mathbf{c_r} = \begin{pmatrix} c_{rx_1} \\ c_{rx_2} \\ \vdots \\ c_{rx_{18}} \\ c_{ry_2} \\ c_{ry_3} \end{pmatrix}. \tag{26}$$

Edellä kuvattu ei sellaisenaan kuitenkaan sovellu kaikille virhetyypeille. Jos virhe on sellainen, että EVR:ssä on arvona jokin nollasta poikkeava luku, mutta kyselyn oikea tieto on nolla, on kyseessä niin sanottu ylimääräinen arvo. Tällaisessa tapauksessa käytetään koedatan generoimisessa hyväksi väärän arvon suhdetta oikeaan liikevaihtoon

$$s_i = \frac{x_i}{x_1^*}. \tag{27}$$

Käytännössä tämä tapahtuu siten, että kerroinvektoriin laitetaan kertoimen tilalle piste ja erilliseen muuttujaan laitetaan suhteen (27) arvo.

Mikäli virhetyyppi on eräkatoa eli puuttuva tieto, tulee kertoimeksi nolla eli virhevektorimenetelmä toimii tässä tapauksessa hyvin. Jos virhe on sellainen, että on vahingossa annettu tieto tuhansina markkoina, tulee kertoimeksi 0.001 ja niin

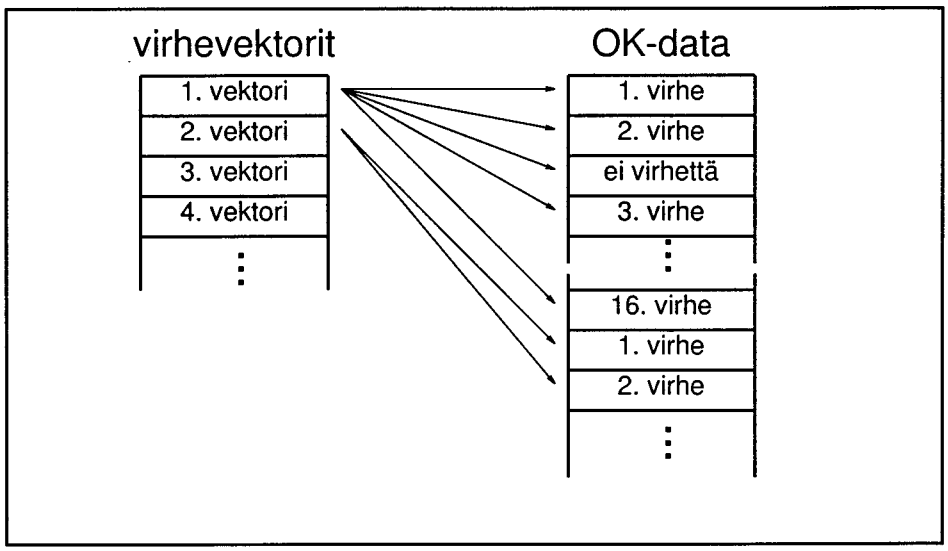


Kuva 2: Esimerkki virheiden generoitumisesta.

edelleen. Tämä virhevektorimenetelmä ei tietenkään osaa luoda esimerkiksi sellaisia virheitä, joissa kahden peräkkäisen muuttujan arvot ovat vaihtaneet paikkaa. Toisaalta tällaisessa tapauksessa se kuitenkin luo virheen molempiin muuttujiin, mutta kertoimiin tulee suhteelliset virheet molempiin muuttujiin erikseen. Mikäli oltaisiin testaamassa sellaisia editointimenetelmiä, jotka yrittävät korjata esimerkiksi edellä mainitun arvojen vaihtumisen kaltaisia virheitä, tulisi myös virheiden generoinnissa ottaa se huomioon.

8.5 Virheiden generointi

Virheitä pyritään viemään OK-dataan siten, että virheellisten osuus aineistossa on mahdollisimman lähellä oikeata tilannetta. Aidossa EVR-aineistossa on 179699 rakennetilastoasetuksen mukaisten toimialapääluokkien yritystä, joista 9347 eli 5.2 % korjaantuu alkueditoinnilla. Alkueditoinnin jälkeen aineistossa on 26846 virheellistä eli 14.9 %. Tässä kokeellisessa OK-datassa on 116146 yritystä, joten virheellisten havaintojen osuuden tulisi olla $0.149 * 116146 = 17352$. Täten kutakin virhevektoria viedään OK-dataan $17352/1093 \approx 16$ kertaa eli virheitä generoidaan $M = 16 * 1093 = 17488$ kappaletta. Tulokseksi saadaan siis koedata, jossa on 116146 havaintoa, joista 17488 eli 15.1 % on virheellisiä. Koedataan ei tuoteta alkueditoinneilla korjaantuvia virheitä, koska kaikissa vertailtavissa menetelmissä niitä käsitellään samalla tavalla.



Kuva 3: Virheiden generointi.

Virheiden vieminen OK-dataan tapahtuu seuraavasti. Sekä virhevektorit että OK-datan tietueet järjestetään satunnaiseen järjestykseen. Tämän jälkeen ensimmäisenä oleva virhevektori kerrotaan ensimmäisellä OK-datan havainnolla. Tämä kertominen ei ole normaalia matriisilaskennan vektorien kertomista vaan kukin OK-tietueen muuttuja siis kerrotaan virhevektorin vastaavalla arvolla. Poikkeuksena on ylimääräisen arvon vieminen, jolloin muuttujan i virhe generoidaan kaavalla

$$x_i^{gen} = s_i x_1,\ i \neq 1, \tag{28}$$

jossa s_i on kaavan (27) suhde ja x_1 on generoitavan OK-datan yrityksen liikevaihto. Tällä tavalla ei pystytä luomaan 'ylimääräinen arvo'-tyyppisiä virheitä itse liikevaihtomuuttujaan, mutta toisaalta jokaisella OK-yrityksellä yleensä on liikevaihtotiedossa jokin arvo, joten tällä ei ole merkitystä.

Kun ensimmäinen virhevektori on viety, testataan syntyikö havaintoon tarpeeksi suuri virhe peruseditointisäännöllä (1). Tätä jatketaan siten, että otetaan seuraava OK-havainto ja viedään siihen ensimmäisenä oleva virhevektori ja jatketaan tätä niin kauan kunnes ensimmäistä virhevektoria on viety OK-aineistoon

16 kertaa. Tämän jälkeen toisena olevaa virhevektoria viedään seuraaviin OK-tietueisiin, kunnes virheitä on saatu syntymään 16 kappaletta (katso kuva 3). Näin edetään niin kauan kunnes kaikkia vektoreita on toistettu onnistuneesti 16 kertaa eli aineistoon on tuotettu yhteensä $M = 16 * 1093 = 17488$ virhettä. Mikäli OK-data käydään kokonaan läpi ennen kuin kaikki virheet on generoitu, jatketaan generointia aloittaen OK-datan alusta siten, että samaan yritykseen ei viedä enää uutta virhettä.

Virheiden generoinnissa otetaan huomioon myös muutettavan OK-tietueen rakenne. Jos generoidaan esimerkiksi puuttuva arvo eli kerroin on nolla, tarkastetaan onko OK-tietueessa vastaavassa kentässä jokin arvo. Kentän ollessa valmiiksi jo nolla, ei virhettä viedä tähän OK-tietueeseen, vaan siirrytään järjestyksessä seuraavaan OK-tietueeseen. Jos taas generoitava virhe on tyypiltään ylimääräinen arvo, tarkastetaan että vastaavassa OK-tietueessa on kyseessä olevassa muuttujassa arvo nolla. Mikäli virhevektorissa on useampi kuin yksi virhe eli ykkösestä poikkeava arvo, tarkastetaan että kaikki virheet saadaan vietyä OK-tietueeseen. Näin varmistetaan generoitujen virheiden rakenteen säilyminen mahdollisimman aidonmukaisena.

8.6 Koeaineiston editointi ja imputointi

Generoidun aineiston koko on sama kuin OK-datan eli 116146 yritystä. Näistä 17488 yritystä eli 15.1 % on virheellisiä. Näiden lisäksi koedataan otetaan mukaan EVR-tiedot 4491 suurelta yritykseltä, joiden tiedot rakennetilastotietokantaan saadaan suoralla kyselyllä. Nämä yrityksethän eivät ole mukana virheettömässä aineistossa, johon generoitu koedata perustuu. Nämä havainnot ovat virheettömiä ja toimivat aputietona virheellisten korjaamisessa. Näiden suurten yritysten tiedoilla on merkittävä osa esimerkiksi suhdeimputoinnissa, joten on tärkeää ottaa ne mukaan koeaineistoon. Täten koedatan lopullinen koko on 120637 yritystä. Koedatan editoinnissa käytetään apuna $116146 + 4491 - 17488 = 103149$ yrityksen

virheettömiä tietoja.

Virheiden generoinnissa varmistetaan, että kaikkiin virheellisiin yrityksiin luodaan vähintään 1000 markan virhe. Täten myös kullakin editointimenetelmällä saadaan selville yhtä monta virheellistä yritystä. Koedatan editointi poikkeaa aidosta tilanteesta siinä, että siihen ei sovelleta alkueditointeja, jotka ovat kaikissa vertailtavissa menetelmissä samat. Koeaineistoon ei myöskään luoda virheitä, jotka korjaantuvat alkueditoinneilla. Tästä syntyy myös pieni ero virheettömien havaintojen määrittelyssä aidon ja simuloidun tilanteen välillä. Aidossa tilanteessa alkueditoinneilla korjattavat yritykset yhdistetään täysin virheettömiin yrityksiin ja niitä käytetään yhdessä aputietona varsinaisissa korjausmenetelmissä poislukien rescaling, joka ei käytä hyväkseen virheettömien havaintojen tietoja. Simuloidussa tilanteessa käytetään aputietona ainoastaan täysin virheettömiä yrityksiä. Täysin virheettömällä tarkoitetaan tässä sitä, että havainto täyttää ehdon (1). Tämä ero aidon ja simuloidun tilanteen välillä ei ole kuitenkaan merkittävä, koska se ei vaikuta vertailutuloksiin.

Koedatan editointi poikkeaa aidosta tilanteesta suhdeimputoinnin osalta. Suhdeimputointia käytetään myös sekamenetelmässä osaan havainnoista. Näissä imputointisoluina käytetään toimialan ja liikevaihdon suuruuden mukaisia luokkia. Aidossa tilanteessa liikevaihto jaetaan neljään luokkaan rajojen ollessa miljoona, 10 miljoonaa ja 50 miljoonaa markkaa eli suurimman liikevaihtoluokan alaraja on 50 miljoonaa markkaa. Koska virheitä ei ole generoitu suuriin yrityksiin, joiden tiedot saadaan suorasta kyselystä, käytetään koedatan suhdeimputointiosioissa suurimman luokan alarajana 25 miljoonaa markkaa.

Sekamenetelmän outlier-osiolla korjattujen yritysten osuus koedatan virhehavainnoista on noin 27 %, lähimmän naapurin 67 % ja suhdeimputoinnin 6 %. Luvut eivät juurikaan vaihtele eri simulointikertojen välillä. Vastaavat luvut ovat aidossa aineistossa 20 %, 73 % ja 7 % eli luvut ovat melko lähellä toisiaan.

9 Menetelmien vertailu

Tarkoituksena on vertailla, kuinka lähelle oikeita havaintoarvoja kunkin eri editointi- ja imputointimenetelmän tuottamat arvot sijoittuvat. Koska oikeat muuttujien arvot tiedetään havaintoyksikkötasolla, voidaan menetelmien tehokkuutta arvioida hyvinkin yksityiskohtaisesti. Tällaisessa usean kvantitatiivisen muuttujan tilanteessa luonnollinen tapa vertailuun voisi olla kahden havainnon r ja s välinen euklidinen etäisyys

$$||\mathbf{x}_r - \mathbf{x}_s||^2 = (\mathbf{x}_r - \mathbf{x}_s)'(\mathbf{x}_r - \mathbf{x}_s). \tag{29}$$

Euklidinen etäisyys siis kuvaa kahden pisteen, havaintojen r ja s, välistä etäisyyttä j-ulotteisessa avaruudessa muuttujien toimiessa koordinaatteina. Tämä etäisyys ei kuitenkaan ota huomioon eroja koordinaattien skaaloissa, eli tässä tapauksessa muuttujien variansseissa, vaan kaikki muuttujat vaikuttavat yhtä paljon euklidisen etäisyyden laskemiseen. Mielekkäämpää onkin käyttää sellaista mittaa, joka painottaa vähemmän sellaisia muuttujia, joiden vaihtelu on suurempaa ja päinvastoin. Seuraavassa kappaleessa esitellään tämän seikan huomioonottava Mahalanobisin etäisyys.

9.1 Mahalanobisin etäisyysmitta

Mahalanobis kehitti etäisyysmittansa antropometristen tutkimustensa yhteydessä. Antropometria tutkii ja vertailee ihmisen kehon mittoja. Mahalanobis julkaisi mitan lopullisessa muodossaan 1930-luvulla. Tämän jälkeen mittaa on käytetty menestyksellisesti mitä moninaisemmilla tieteenaloilla.

Mahalanobisin etäisyysmitan perusmuoto kahden populaation 1 ja 2 väliselle etäisyydelle on

$$D^2 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \boldsymbol{\Sigma}^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2), \tag{30}$$

missä $\bar{\mathbf{x}}_1$ ja $\bar{\mathbf{x}}_2$ ovat populaatioiden 1 ja 2 keskiarvovektoreita ja $\boldsymbol{\Sigma}$ on kovarianssimatriisi (Mardia, Kent & Bibby 1979). Mitta ottaa kovarianssimatriisin käänteismatriisin $\boldsymbol{\Sigma}^{-1}$ avulla huomioon muuttujakohtaiset varianssit sekä muuttujien väliset kovarianssit. Tätä perusmuodossa olevaa mittaa voitaisiin käyttää periaatteessa menetelmien vertailuun, mutta sen antama tieto ei ole tarpeeksi tarkkaa. Sehän vain kuvaisi eri menetelmien kykyä kunkin muuttujan keskiarvon kannalta. Koska tiedossa on yrityskohtaiset muuttujien oikeat arvot, ei kaavan (30) etäisyysmittaa käytetä. Mahalanobisin etäisyyttä on sovellettu myös tapaukseen, jossa halutaan verrata yksittäisten havaintojen etäisyyksiä oman populaationsa keskiarvoon. Tällöin mitta on

$$D^2 = (\mathbf{x}_i - \bar{\mathbf{x}})' \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}), \tag{31}$$

missä $\mathbf{x}_i$ on i. havainto ja $\bar{\mathbf{x}}$ on populaation keskiarvovektori. Tätä mittaa voidaan käyttää etsittäessä outliereita eli poikkeavia havaintoja (Barnett & Lewis 1994). Kolmas sovellutus mitasta on kahden eri havainnon r ja s välinen etäisyys

$$D^2 = (\mathbf{x}_r - \mathbf{x}_s)' \boldsymbol{\Sigma}^{-1} (\mathbf{x}_r - \mathbf{x}_s), \tag{32}$$

joka soveltuu EVR-aineiston havaintojen vertailuun.

9.1.1 Etäisyysmitan sovellutus EVR-aineistoon

Simulointien avulla saadaan oikeat ja eri menetelmillä korjatut tiedot $M = 17488$ yritykselle. Mahalanobisin etäisyydet lasketaan käyttäen muuttujia $x_1, \ldots, x_{18}$. Summamuuttujia y_1, y_2, y_3 ei oteta mukaan etäisyysmitan laskemiseen. Niitähän ei varsinaisesti editoida, vaan ne lasketaan kussakin menetelmässä lopuksi summaten x-muuttujia. Merkitään eri menetelmillä editoituja havaintoja

$$\mathbf{x}_i^{ed} = \begin{pmatrix} x_{i1}^{ed} \\ x_{i2}^{ed} \\ \vdots \\ x_{i,18}^{ed} \end{pmatrix}, \quad i = 1, \ldots, M,$$

missä *ed*=suhdeimputointi, rescaling tai sekamenetelmä. Näiden lisäksi tarvitaan oikeat havaintojen arvot, joita merkitään

$$\mathbf{x}_i^{ok} = \begin{pmatrix} x_{i1}^{ok} \\ x_{i2}^{ok} \\ \vdots \\ x_{i,18}^{ok} \end{pmatrix}, \quad i = 1, \ldots, M.$$

EVR-aineiston tapauksessa havaintokohtainen Mahalanobisin etäisyysmitta on

$$D_i^2 = (\mathbf{x}_i^{ed} - \mathbf{x}_i^{ok})' \, \Sigma_{ok}^{-1} (\mathbf{x}_i^{ed} - \mathbf{x}_i^{ok}), \quad i = 1, \ldots, M, \tag{33}$$

missä Σ_{ok} on OK-datasta laskettu kovarianssimatriisi. Se on siis laskettu kaikista OK-havainnoista, ei pelkästään niistä, joihin on generoitu virheitä. Kovarianssimatriisin laskemisessa käytetty SAS-koodi esitetään liitteessä B. Yleensä otostilanteissa kovarianssimatriisin tilalla käytetään otoskovarianssimatriisia S, mutta koska nyt tiedetään koko perusjoukon oikea kovarianssimatriisi, käytetään luonnollisesti sitä. Koko aineistoa kuvaavana etäisyysmittana käytetään havaintokohtaisten etäisyyksien summaa

$$D = \sum_{i=1}^{M} D_i^2. \tag{34}$$

Mahalanobisin etäisyyksien laskemisessa käytetty SAS-ohjelmakoodi on liitteessä C.

9.1.2 Tulokset

Kappaleessa 8.5 kuvattu koedatan generointi toistetaan 10 kertaa eli saadaan 10 koedataa, joissa jokaisessa on 120637 yritystä, joista 17488 on virheellisiä. Kullekin koedatalle suoritetaan editointi ja imputointi kolmella eri menetelmällä. Näin saatujen kolmen erilaisen editoidun datan lisäksi tiedetään havaintojen oikeat arvot sekä generoidut virheelliset arvot. Tuloksissa esitetään myös virheellinen data sen tutkimiseksi, saavutetaanko editoinnilla parempilaatuinen aineisto kuin ilman

editointia. Virheellinen aineisto on muuttujien $x_1, \ldots, x_{18}$ osalta muuttamaton, mutta summamuuttujat myyntikate, käyttökate ja tilikauden tulos on laskettu uudelleen eli aineisto on umpeenmenevä. Tämä voisi olla yksi keino aineiston editoimiseksi, mikäli todettaisiin, että editointi ei paranna aineistoa.

Mahalanobisin etäisyydet 10 eri simuloinnista esitetään liitteessä D sivulla 68. Taulukossa 2 on kaavan (34) mukaiset etäisyyksien summat virheelliseltä eli korjaamattomalta datalta sekä eri menetelmin korjatuilta. Taulukoissa 3, 4 ja 5 ovat vastaavat luvut jaettuna sen mukaan, millä menetelmällä havainto korjaantuu sekamenetelmässä. Tällä jaottelulla saadaan tarkempaa tietoa sekamenetelmän eri osioiden tehokkuudesta. Taulukoissa on kunkin menetelmän etäisyyksien 95 ja 5 prosentin kvantiilit, mediaani ja kolme suurinta yksittäistä etäisyyttä (max). Samalla rivillä olevat max-arvot eivät yleensä ole saman yrityksen etäisyyksiä, vaan eri menetelmillä voi tietysti tulla eri suuruisia etäisyyksiä. Esimerkiksi korjaamattoman aineiston suurimman virheen eli maksimiarvon omaavan yrityksen saattaa jokin menetelmä korjata niin hyvin, että se ei ole sen menetelmän kolmen suurimman virheen joukossa. Pienimmät eli parhaat Mahalanobisin etäisyyksien summat on lihavoitu.

Taulukon 2 korjaamaton-sarakkeessa on generoidun koedatan virheellisistä havainnoista lasketut etäisyydet. Niistä nähdään, että virheellinen aineisto on selkeästi huonompi kuin yksikään korjatuista aineistoista. Tämän perusteella voidaan todeta, että editointi- ja imputointimenetelmien käyttäminen on hyvin perusteltua EVR-aineistossa.

Kuten huomataan, on rescaling-menetelmä kaikissa simuloinneissa toiminut parhaiten. Sekamenetelmä on toiseksi paras ja suhdeimputointi heikoin. Suhdeimputointimenetelmän heikkoudelle on olemassa selkeä selitys. Menetelmähän korjaa kaikkien muuttujien arvot, kun taas rescaling ja sekamenetelmä korjaavat vain virheellisen välin muuttujat, mikäli virheen sijainti voidaan paikallistaa. Tämä ero menetelmien välillä vaikuttaa ratkaisevasti tulokseen varsinkin, kun etäisyysmitta

ottaa huomioon kaikkien muuttujien arvot.

Taulukossa 2 esitetään kunkin simulaation 95 ja 5 prosentin kvantiilit, ala- ja yläkvartiilit sekä mediaani. Jakaumienkin mukaan rescaling toimii yleisesti parhaiten. 5 prosentin kvantiililukujen perusteella sekamenetelmä korjaa kuitenkin osan virheellisistä havainnoista rescaling-menetelmää paremmin. Syy tähän näkyy taulukossa 3 sekamenetelmän nollaetäisyyksinä 5 prosentin kvantiilipisteessä. Sekamenetelmän outlier-osuus korjaa osan havainnoista täysin oikein, mikä on tietysti menetelmän tarkoituskin. Sekamenetelmän outlier-metodilla täysin oikein korjattujen yritysten osuus kaikista outlier-korjatuista vaihtelee 5 ja 10 prosentin välillä. Taulukon 3 perusteella outlier-menetelmä toimii selvästi paremmin kuin suhdeimputointi ja rescaling. Maksimiarvoista kuitenkin nähdään, että aineistoon jää outlier-korjauksen jäljiltä suuriakin virheitä, joten kaikilta osin menetelmä ei toimi kuten pitäisi.

Taulukossa 4 on Mahalanobisin etäisyydet yrityksiltä, jotka korjaantuvat sekamenetelmässä lähimmän naapurin imputoinnilla. Tämä onkin mielenkiintoinen vertailu, koska sekamenetelmässä suurin osa havainnoista korjataan juuri lähimmän naapurin imputoinnilla. Tulosten perusteella lähimmän naapurin menetelmä ei ole erityisen toimiva. Rescaling toimii kaikissa simuloinneissa parhaiten ja lähimmän naapurin menetelmä yleensä toiseksi parhaiten. Simuloinneissa 1, 6 ja 9 suhdeimputointi toimii kuitenkin paremmin kuin lähimmän naapurin imputointi, mikä johtuu naapurimenetelmän suurista yksittäisistä maksimiarvoista. Jakaumien perusteella suhdeimputointi on selkeästi huonoin, mutta toisaalta naapurimenetelmällä saattaa aineistoon jäädä suuria yksittäisiä virheitä. Rescaling-menetelmä poistaa suuria virheitä aineistosta selvästi tehokkaammin kuin muut menetelmät.

Taulukossa 5 on etäisyystiedot niiltä yrityksiltä, jotka sekamenetelmässä korjaantuvat suhdeimputoinnilla. Näiden virheen sijaintia ei onnistuta paikallistamaan sekamenetelmässä käytetyllä virheen etsimismenetelmällä, vaan kaikkien

muuttujien arvot korjataan. Suhdeimputoinnin ja sekamenetelmän tulokset ovat tietysti täysin samat, koska menetelmät ovat samat, ja täten vertailtavana onkin oikeastaan vain kaksi eri menetelmää: rescaling ja suhdeimputointi. Rescaling-menetelmä toimii paremmin kahdeksassa simuloinnissa kymmenestä. Simuloinneissa 1 ja 8 suhdeimputointi toimii paremmin. Koska virheen sijaintia tietueessa ei voida paikallistaa, on jokaisella välillä periaatteessa vähintään yksi virhe, ja täten tässä joukossa virheitä on yleensä enemmän kuin muissa. Tämä saattaa aiheuttaa sen, että rescaling ei aina toimi parhaiten. Koska rescaling perustuu valmiiden arvojen, oikeiden ja virheellisten, muuttamiseen kertoimien avulla, voidaan sen olettaa toimivan sitä huonommin mitä enemmän tietueessa on virheitä.

9.2 Muuttujakohtaiset suhteelliset virheet

Mahalanobisin etäisyyden lisäksi tutkitaan menetelmiä yleisemmällä tasolla verraten korjattujen muuttujien toimialakohtaisia summia oikeisiin summiin, jotka tiedetään. Aineistosta lasketaan muuttujakohtaiset suhteelliset virheet

$$e_{hj} = \left| \frac{\sum_{i \in h} x_{ij}^{ed} - \sum_{i \in h} x_{ij}^{ok}}{\sum_{i \in h} x_{ij}^{ok}} \right|, \tag{35}$$

missä h on toimiala, i on havainto ja j on muuttuja, jolle suhteellinen virhe lasketaan. Näin saadaan eräänlaiset suhteelliset harhat, jotka kuvaavat menetelmien kykyä muuttujittain summatasolla. On huomattava, että nämä suhteelliset virheet lasketaan ainoastaan siitä yritysjoukosta, johon on tuotettu virheitä, joten näitä suhteellisia virheitä ei pidä suoraan tulkita koko EVR-aineiston suhteellisiksi virheiksi. Toimialaluokituksena käytetään yksinumeroista luokitusta muille kuin luokalle 5, joka on jaettu kaksinumerotason luokkiin.

9.2.1 Tulokset

Suhteellisten virheiden keskiarvot 10 simuloinnista esitetään liitteessä E. Siinä ovat muuttujien *liikevai, ostoyht, varastmu, verovali, rahtuott, kayttoka* ja *tiliktul*

tulokset. Käyttökate (*kayttoka*) ja tilikauden tulos (*tiliktul*) eroavat muista muuttujista siinä, että niitä ei ole suoraan imputoitu, vaan ne on summattu muista muuttujista. Täten nämä kaksi muuttujaa kuvaavat laajemmin koko editointi- ja imputointiprosessia, varsinkin *tiliktul*-muuttuja, joka on kaikkien muuttujien summa. Pienimmät suhteelliset virheet on lihavoitu. Alimmalla rivillä on menetelmittäin toimialaluokkien virheiden keskiarvo painotettuna eri simulointien yritysten lukumäärän keskiarvolla. Nämä frekvenssien keskiarvot ovat taulukossa 6. Muihin tauluihin näitä ei ole merkitty, koska ne ovat kaikissa samat.

Taulukossa 6 on liikevaihtomuuttujan tulokset. Liikevaihto on tärkeä muuttuja, ja sen käsittely poikkeaa muiden muuttujien käsittelystä suhde- ja sekamenetelmissä. Suhdeimputoinnissa käytetään apuna yritys- ja toimipaikkarekisterin liikevaihtotietoa, jonka avulla muut muuttujat imputoidaan. Kaikille virheellisille yrityksille tulee EVR-liikevaihdon tilalle YTR:n tieto. Tämä varmasti vaikuttaa siihen, että suhdeimputointi toimii parhaiten liikevaihdon tapauksessa. Sekamenetelmä toimii yllättäen lähes yhtä hyvin, vaikka siinä käytetään suhdemenetelmän lisäksi rescalingia. Sekamenetelmän lähimmän naapurin osiossa ei tuoda korjattaville yrityksille luovuttajayritysten liikevaihtotietoja, mutta sen sijaan näille yrityksille käytettävällä rescaling-metodilla muutetaan myös liikevaihtomuuttuja. Itse rescaling-menetelmä toimii selvästi huonommin kuin suhde- ja sekamenetelmä, mihin vaikuttaa varmasti suhdeimputointiin kuuluva YTR:n liikevaihtotietojen hyväksikäyttö. Rescaling-menetelmässä ei virheen sijainnin etsimisessä tutkita liikevaihdon oikeellisuutta, toisin kuin sekamenetelmässä, mikä aiheuttaa liikevaihtomuuttujan oikeidenkin arvojen muuttamisen. Tämä voitaisiin toki muuttaa, mikä varmastikin parantaisi rescalingin tulosta.

Muuttujan *ostoyht* tulokset esitetään taulukossa 7. Keskimäärin parhaiten toimii sekamenetelmä. Suhdeimputointi toimii yllättävän hyvin ottaen huomioon, että siinä korjataan paljon oikeitakin arvoja. Ilman toimialaluokkien 6 ja 7 suuria virheitä se olisi keskimäärin rescalingia parempi, koska lähes kaikissa muissa luokissa se on sitä parempi. Rescaling on huonompi kuin sekamenetelmä kaikissa

muissa luokissa paitsi luokassa 7. Rescaling ei osaa muuttaa nolla-arvoja, mikä voi vaikuttaa tällaisen vähän aitoja nollia sisältävän muuttujan tapauksessa. Jos virheenä on puuttuva arvo, ei rescaling muuta nolla-arvoa, kun taas suhdeimputointi ja sekamenetelmä voivat tuottaa nollasta poikkeavan arvon.

Muuttujat *varastmu* ja *verovali* taulukoissa 8 ja 9 poikkeavat muista siinä, että ne voivat saada sekä positiivisia että negatiivisia arvoja. Tämä vaikuttaa eri menetelmien tehokkuuteen varsinkin virheissä, joissa oikean ja virheellisen arvon etumerkit eivät ole samat. *Varastmu*-muuttujan kohdalla rescaling on keskimäärin paras, kun taas *verovali*-muuttujalla rescaling on huonoin. Kun katsotaan tarkemmin taulukon 8 luokkakohtaisia keskiarvoja, huomataan rescalingin olevan kuudessa luokassa kymmenestä huonoin menetelmä. Taulukossa 9 rescaling on huonoin menetelmä kaikissa luokissa. Täten voidaan sanoa, että rescaling ei toimi muuttujilla, jotka saavat positiivisia sekä negatiivisia arvoja. Tämä johtuu siitä, että rescaling-menetelmä ei koskaan vaihda korjattavan arvon etumerkkiä. Rescalingin menestys *varastmu*-muuttujan kohdalla johtuu pitkälti luokan 3 tuloksesta. Siinä yksittäisten simulointien suuret virheet nostavat keskiarvon korkeaksi. Myös luokissa 6 ja 7 on suuria virheitä. Näistä huomataan, että positiivisia ja negatiivisia arvoja saavien muuttujien editointi on ongelmallista ja vaatii lähempää tarkastelua. Samat menetelmät eivät sovellu kaikille muuttujille yhtä hyvin. *Verovali*-muuttujan kohdalla tuloksissa ei ole yhtä suurta vaihtelua kuin edellä. Tähän syynä on se, että yleensä veroja maksetaan enemmän kuin saadaan takaisin, jolloin etumerkki on usein negatiivinen ja suuria positiivisia arvoja ei ole, toisin kuin *varastmu*-muuttujalla.

Taulukossa 10 ovat *rahtuott*-muuttujan suhteelliset virheet. Tälle muuttujalle on tyypillistä nolla-arvojen suuri määrä. Muita samanlaisia muuttujia ovat *rahkulum*, *satunntu* ja *satunnku*. Kuten tuloksista nähdään, on rescaling selvästi paras menetelmä tämäntyyppisille muuttujille. Rescalingin etuna on se, että se ei muuta valmiita nolla-arvoja, vaan säilyttää ne nollina. Suhdeimputointi puolestaan tuottaa muuttujaan lähes aina jonkin nollasta poikkeavan arvon, mikä aiheuttaa virhet-

tä useissa tapauksissa. Toisaalta tämä on toivottavaa, mikäli nolla-arvo on puuttuva arvo. Tällöin rescaling ei toimi oikein, vaan jättää arvon nollaksi. Tämän tuloksen perusteella rescalingin ominaisuudet soveltuvat kuitenkin parhaiten paljon oikeita nolla-arvoja sisältäville muuttujille. Huomattavaa on toimialaluokan 7 erittäin suuri generoitu virhe. Virhe johtuu yksittäisestä suuresta yrityksestä, jonka *rahtuott*-muuttujan virhekerroin on yli 3000. Kaikki menetelmät kuitenkin korjaavat tämän virheen hyvin.

Käyttökatemuuttujan tuloksissa (taulukko 11) rescaling ja sekamenetelmä ovat melko tasavahvoja jälkimmäisen ollessa niukasti parempi. Näissä menetelmissä käyttökatteen virheet ovat pienempiä kuin yksittäisissä muuttujissa, mikä johtuu korjaustavasta. Rescaling-menetelmässä ja sekamenetelmän naapuriosiossa tuotetaan virheellisiin muuttujiin arvot siten, että ne summautuvat oikeiksi tiedettyjen myyntikate- ja käyttökatetietojen mukaan. Tämän takia nämä menetelmät toimivat käyttökatteen tapauksessa selvästi paremmin kuin suhdeimputointi, jossa summatietoja ei käytetä hyväksi.

Tilikauden tulos summataan kaikista editoiduista ja imputoiduista muuttujista. Tulosten (taulukko 12) mukaan rescaling on selvästi tehokkain menetelmä. Käyttökatteen ja tilikauden tuloksen tuloksia vertaamalla voidaan päätellä, että rescaling toimii tuloslaskelman käyttökatteen jälkeisille muuttujille paremmin kuin sekamenetelmä. Tähän voi vaikuttaa se, että joukossa on useita paljon oikeita nolla-arvoja saavia muuttujia, joiden korjaamisessa rescaling osoittautuu parhaaksi.

10 Yhteenveto ja johtopäätökset

Tämän tilastotieteen pro gradu-tutkielman tarkoituksena on tutkia elinkeinoverorekisteriaineiston editointi- ja imputointimenetelmiä Tilastokeskuksessa ja verrata niitä käyttäen simuloitua aineistoa. Seuraavassa kerrotaan tiivistetysti käytetystä aineistosta, menetelmistä, tuloksista ja johtopäätöksistä.

EVR-aineistossa on kunkin verovuoden tilinpäätöstiedot kaikilta elinkeinoverotuslain mukaan verotettavilta elinkeinonharjoittajilta. Aineisto saadaan vuosittain Tilastokeskukseen Verohallinnolta. Tässä tutkielmassa keskitytään kahdenkertaista kirjanpitoa pitävien yritysten tuloslaskelman editointiin ja imputointiin. Aineistossa noin 20 prosenttia yrityksistä on virheellisiä, kun virheenä pidetään vähintään 1000 markan eroa tuloslaskelman muuttujien summan ja tilikauden tuloksen välillä. EVR-aineisto on osa yritysten rakennetilastotietokantaa (kuva 1). Lukumäärällisesti valtaosa tietokannan yritystiedoista saadaan EVR-aineistosta, mutta suurimpien yritysten tiedot kerätään Tilastokeskuksen suoralla kyselyllä (taulukko 1).

Kaikki kolme vertailtavaa menetelmää ovat olleet käytössä tilastotuotannossa. Menetelmät ovat kehittämisjärjestyksessä suhdeimputointi, rescaling-metodi ja sekamenetelmä. Menetelmien vertailussa käytetään simuloitua koeaineistoa, johon on generoitu virheitä. Virheettömänä aineistona käytetään vuoden 1998 EVR-datan niitä havaintoja, joissa ei ole virhettä peruseditointisäännön (kaava 1) mukaan. Mukaan ei oteta suuria yrityksiä, joiden tiedot rakennetilastotietokantaan saadaan suoralla kyselyllä. Tähän virheettömään aineistoon generoidaan virheitä käyttäen apuna EVR:n virheellisiä havaintoja, joiden vastaavat virheettömät tiedot saadaan suorasta kyselystä. Näistä yrityksistä, joita on yhteensä 1093, muodostetaan yhtä monta virhevektoria, joita viedään toistaen virheettömiin yrityksiin.

Virhevektorien lukumäärä on valitettavan pieni aineiston kokoon nähden. Vaikka koedatassa onkin $M = 17488$ virheelliseksi generoitua havaintoa, on siinä kui-

tenkin vain 1093 erilaista virhetyyppiä. Toisena ongelmana voidaan pitää sitä, että virhevektorit lasketaan suurilta yrityksiltä, mutta niitä generoidaan pääasiassa pienempiin yrityksiin. Tällöin oletetaan virheiden olevan samantyyppisiä sekä suurissa että pienissä yrityksissä, mikä ei välttämättä pidä paikkaansa. Kun kaikkia menetelmiä on sovellettu generoituun dataan, tiedetään jokaisen havainnon virheettömät EVR-datan arvot sekä eri menetelmillä korjatut arvot. Näiden avulla voidaan menetelmiä verrata havaintokohtaisesti. Generointi suoritetaan 10 kertaa, jolloin saadaan 10 yhtä suurta koedataa, joissa kaikissa on yhtä monta virheellistä yritystä. Virheelliset yritykset vaihtelevat generointiin sisältyvästä satunnaistamisesta johtuen.

Menetelmien vertailuun sovelletaan Mahalanobisin kahden havainnon etäisyysmitan muotoa

$$D_i^2 = (\mathbf{x}_i^{ed} - \mathbf{x}_i^{ok})' \, \Sigma_{ok}^{-1} (\mathbf{x}_i^{ed} - \mathbf{x}_i^{ok}), \quad i = 1, \ldots, M,$$

missä $\mathbf{x}_i^{ed}$ on korjattu havainto, $\mathbf{x}_i^{ok}$ on saman havainnon virheettömät arvot ja Σ_{ok} on OK-datasta laskettu kovarianssimatriisi. M on virheellisten yritysten lukumäärä koedatassa. Lopullinen etäisyysmitta lasketaan summana

$$D = \sum_{i=1}^{M} D_i^2.$$

Etäisyyksien summat lasketaan kaikilta yrityksiltä sekä jaettuna kolmeen osaan sen mukaan, millä metodilla yritys korjaantuu sekamenetelmässä. Jaottelun avulla voidaan verrata sekamenetelmän eri osioiden tehokkuutta. Mahalanobisin etäisyysmitta reagoi herkästi suuriin virheisiin, ja täten muutamassa yrityksessä olevat suuret virheet vaikuttavat erittäin paljon lopulliseen etäisyyksien summaan D. Toisaalta tämä on toivottavaa, koska aineistoon ei haluta jäävän suuria virheitä. Tulokset ovat kuitenkin samansuuntaisia kaikissa simuloinneissa, joten etäisyysmittaa voidaan pitää soveltuvana menetelmien vertailuun.

Mahalanobisin mitan lisäksi simuloiduista datoista lasketaan muuttujakohtaiset suhteelliset virheet (kaava 35), joiden keskiarvot esitetään liitteessä E. Näiden

avulla saadaan selville eri menetelmien tehokkuus eri muuttujien editoinnissa.

Suhdeimputointia on käytetty EVR-aineiston korjaamiseen verovuosina 1994-97. Se perustuu imputoitavan muuttujan ja liikevaihdon summien suhteeseen imputointisoluissa. Imputointisoluina käytetään toimialaluokitusta ja sen lisäksi vielä yrityksen suuruusluokkaa liikevaihdon perusteella. Suhdeimputoinnissa ei etsitä virheen sijaintia, vaan kaikki muuttujat, lukuun ottamatta liikevaihtoa, korjataan suhdemenetelmällä. Tämä onkin menetelmän suurin heikkous, mikä näkyy sekä Mahalanobisin etäisyyksissä että suhteellisissa virheissä. Menetelmällä korjataan ehdottomasti liikaa arvoja, mitä voidaan pitää ylieditointina. Huomattakoon että kyseessä ei ole itse menetelmän heikkous, vaan tavan, jolla sitä on sovellettu EVR-aineistoon. Menetelmää voitaisiin kehittää sovellettavaksi ainoastaan virheellisen välin muuttujiin, mikä selvästi vähentäisi turhien korjausten määrää. Suhdemenetelmässä EVR:n virheellisten yritysten liikevaihtomuuttujan tilalle tulee YTR:n tieto. Liikevaihdon korjaamisessa tämä menetelmä toimiikin hyvin (taulukko 6).

Rescaling-menetelmä on kehitetty Tilastokeskuksessa EVR-aineiston korjausmenetelmien kehittämisen yhteydessä. Sitä on sovellettu tilastotuotannossa verovuoden 1998 EVR-aineistoon. Rescalingin yhteydessä etsitään virheen sijainti muuttujien joukossa käyttäen apuna summamuuttujia. Korjaukset tehdään ainoastaan virhevälin muuttujille, mikä vähentää ylieditointia. Korjausmenetelmässä virhe jaetaan virhevälin kaikkiin muuttujiin samassa suhteessa. Menetelmä eroaa kahdesta muusta siinä, että se ei ole aito imputointimenetelmä, vaan se ottaa huomioon valmiina olevat muuttujien arvot, jotka skaalataan summautumaan oikein. Menetelmä korjaa varsinkin pienten virheiden tapauksessa muuttujien arvot erittäin lähelle oikeita arvoja, koska yleensä suurin osa virhevälin muuttujista on oikeita. Tulosten perusteella rescaling on yleisesti tehokkain menetelmä (taulukot 2 ja 12). Rescalingilla korjattuun aineistoon kuitenkin jää suuria yksittäisiä virheitä. Rescaling-metodin ja suhdeimputoinnin puhdas metodien välinen ero nähdään taulukosta 5, jossa on etäisyydet havainnoista, joille ei löydetä virheväliä. Tällöin rescaling-menetelmällä ei ole etua virheen sijainnin etsimisestä. Tulokset osoit-

tavat rescaling-metodin olevan tehokkaampi myös tässä tapauksessa. Yksittäisten muuttujien kohdalla rescaling ei kuitenkaan aina toimi parhaiten. Ongelmana ovat rescalingin kyvyttömyydet puuttuvien arvojen imputoinnissa sekä muuttujan etumerkin vaihtamisessa.

Sekamenetelmä koostuu kolmesta erillisestä metodista, joita sovelletaan eri havaintoyksiköille. Menetelmää on käytetty verovuoden 1999 EVR-aineiston editoinnissa ja imputoinnissa. Aluksi korjataan outlier-menetelmällä kaikki sillä korjattavissa olevat havainnot, joiden virheen sijainti tiedetään. Seuraavaksi käytetään lähimmän naapurin imputointimenetelmää niille, joita ei onnistuta outlier-menetelmällä korjaamaan. Lopuksi käytetään suhdeimputointia niille yrityksille, joiden virheen sijaintia ei saada selville.

Outlier-menetelmässä lasketaan muuttujille suhteessa liikevaihtoon desiilirajat, joiden ulkopuolelle jäävien arvojen korjaamisella yritetään nämä havainnot saada korjattua. Outlier-menetelmä toimiikin varsin hyvin verrattuna rescaling- ja suhdemenetelmiin (taulukko 3). Menetelmä pienentää ylieditointia entisestään, koska siinä ei tarvitse muuttaa kuin yhden tai kahden muuttujan arvoja riippuen virhevälistä. Menetelmä ei kuitenkaan aina toimi siten kuin halutaan, koska aineistoon jää suuriakin yksittäisiä virheitä. Menetelmällä saadaan korjattua vain osa virheellisistä havainnoista, mutta korjaus kannattaa tehdä, koska menetelmä toimii hyvin tässä joukossa. Menetelmää kannattaa kehittää, jotta aineistoon ei jäisi suuria virheitä. Esimerkiksi desiilirajojen tiukentaminen voisi olla yksi keino.

Lähimmän naapurin menetelmässä etsitään virheettömien havaintojen joukosta eniten imputoitavaa yritystä vastaava yritys. Tämän lähimmän naapurin eli luovuttajan tiedot kopioidaan virheellisen yrityksen virhevälin muuttujille ja muutetaan rescaling-menetelmällä summautumaan oikein. Havaintojen läheisyyteen vaikuttavat toimiala ja tärkeimmät tuloslaskelman muuttujat. Havaintojen i ja j

välinen etäisyys lasketaan kaavalla

$$D_{ij} = \sum_{k \in \mathbf{F}} |\log(x_{ik}) - \log(x_{jk})|, \tag{36}$$

missä $\mathbf{F}$ on tulosmuuttujien x_k joukko, jota käytetään laskemisessa. Muuttujat vaihtelevat sen mukaan, millä välillä virhe on. Joissain tapauksissa käytetään etäisyyden laskemisessa virhevälinkin muuttujia, mikä on ongelma tässä menetelmässä. Luovuttajaksi valitaan se saman toimialan yritys j, jonka kanssa korjattavan yrityksen i etäisyys on pienin eli min D_{ij}. Toimialajaon hierarkisuutta käytetään hyväksi siten, että mikäli tarkimmalla viisinumerotasolla ei samassa luokassa ole vähintään 50 luovuttajayritystä, käytetään kolminumerotason luokitusta. Vastaavasti siirrytään yksinumeroiseen luokitteluun, jos kolminumeroinen ei toimi. Tässä kannattaisi ottaa mukaan myös neli- ja kaksinumeroiset luokat, joilla saataisiin pientä lisätarkkuutta imputointiin. Naapurimenetelmä ei toimi erityisen hyvin verrattuna rescaling-menetelmään (taulukko 4).

Tämän tutkielman tulosten perusteella voidaan rescaling-menetelmää pitää käytetyistä menetelmistä parhaana. Se ei kuitenkaan toimi yhtä hyvin kaikkien muuttujien kohdalla. EVR-aineiston editoinnissa ja imputoinnissa on ongelmana karkea virheen paikallistamisen tarkkuus, minkä johdosta korjataan paljon oikeitakin arvoja. Sekamenetelmässä käytetty outlier-metodi auttaa tässä ongelmassa, mutta sitäkin kannattaa kehittää. Näiden tulosten perusteella kannattaa EVR-aineistoon jatkossa soveltaa outlier-metodin ja rescalingin yhdistelmää siten, että ensin käytetään outlieria ja lopuille korjaantumattomille rescalingia.

Viitteet

[1] Barcaroli, G. & D'Aurizio, L. (1997): *Evaluating editing procedures: the simulation approach.* Conference of European Statisticians. Work Session on Statistical Data Editing, Praha 1997. Working Paper No. 17.

[2] Barnett, V. & Lewis, T. (1994): *Outliers in Statistical Data.* 3. painos. Chichester: John Wiley & Sons.

[3] Chen, J. & Shao, J. (2000): Nearest Neighbor Imputation for Survey Data. *Journal of Official Statistics* **16**, 113–131.

[4] Economic Commission for Europe (1994): *Statistical Data Editing, Volume No 1, Methods and Techniques.* Conference of European Statisticians, Statistical Standards and Studies, No. 44. Geneve: United Nations.

[5] Fellegi, I. P. & Holt, D. (1976): A Systematic Approach to Automatic Edit and Imputation. *Journal of the American Statistical Association* **71**, 17–35.

[6] Garcia Rubio, E. & Peirats, V. (1994): Evaluation of Data Editing Procedures: Results of a Simulation Approach. Teoksessa Economic Commission for Europe: *Statistical Data Editing, Volume No 1, Methods and Techniques.* Geneve: United Nations.

[7] Granquist, L. (1994): Macro-Editing — A Review of Some Methods for Rationalizing the Editing of Survey Data. Teoksessa Economic Commission for Europe: *Statistical Data Editing, Volume No 1, Methods and Techniques.* Geneve: United Nations.

[8] Granquist, L. (1995): Improving the Traditional Editing Process. Teoksessa Cox, Binder, Chinnappa, Christianson, Colledge & Kott (toim.) *Business Survey Methods*, sivut 385–401. New York: John Wiley & Sons.

[9] Granquist, L. & Kovar, J. G. (1997): Editing of Survey Data: How Much Is Enough? Teoksessa Lyberg, Biemer, Collins, de Leeuw, Dippo, Schwarz & Trewin (toim.) *Survey Measurement and Process Quality*, sivut 415–435. New York: John Wiley & Sons.

[10] Kalton, G. & Kasprzyk, D. (1986): The Treatment of Missing Survey Data. *Survey Methodology* **12**, 1–16.

[11] Kovar J. G. & Whitridge P. J. (1995): Imputation of Business Survey Data. Teoksessa Cox, Binder, Chinnappa, Christianson, Colledge & Kott (toim.) *Business Survey Methods*, sivut 403–423. New York: John Wiley & Sons.

[12] Laaksonen, S. (2000): Regression-Based Nearest Neighbour Hot Decking. *Computational Statistics* **15**, 65–71.

[13] Lessler, J. T. & Kalsbeek, W. D. (1992): *Nonsampling Error in Surveys*. New York: John Wiley & Sons.

[14] Mardia, K. V., Kent, J. T. & Bibby J. M. (1979): *Multivariate Analysis*. London: Academic Press.

[15] Pierzchala, M. (1995): Editing Systems and Software. Teoksessa Cox, Binder, Chinnappa, Christianson, Colledge & Kott (toim.) *Business Survey Methods*, sivut 425–441. New York: John Wiley & Sons.

[16] Schulte Nordholt, E. (1998): Imputation: Methods, Simulation Experiments and Practical Examples. *International Statistical Review* **66**, 157–180.

[17] Shao, J. (2000): Cold Deck and Ratio Imputation. *Survey Methodology* **26**, 79–85.

[18] Tilastokeskus (1999): *Toimialaluokitus 1995*. Toinen tarkistettu painos. Käsikirjoja 4. Helsinki: Tilastokeskus.

A Tuloslaskelman muuttujat

Muuttuja		Kuvaus
verm1		Verollinen myynti (22%, ilman arvonlisäveroa).
verm2		Verollinen myynti (17%, ilman arvonlisäveroa).
verm3		Verollinen myynti (12% / 6%, ilman arvonlisäveroa).
myyntimu		Veroton myynti (sisältyy myynti ulkomaille).
liikevai	x_1	Liikevaihto yhteensä, ilman alv-osuutta.
tuottomu	x_2	Liiketoiminnan muut tuotot.
ostoyht	x_3	Ostot tilikauden aikana.
varastmu	x_4	Varastojen muutos. Sisältää valmiste- ja raaka-aine-varaston muutoksen. Lisäys +, vähennys –.
ostopalv	x_5	Ulkopuolisten palvelujen ostot tilikauden aikana.
palkkamu	x_6	Muuttuvat palkat ja palkkiot.
kulumuut	x_7	Muut muuttuvat kulut.
myyntika	y_1	Myyntikate.
palkkaki	x_8	Kiinteät palkat ja palkkiot.
vuokra	x_9	Vuokrat.
kulukiin	x_{10}	Muut kiinteät kulut.
kayttoka	y_2	Käyttökate.
sumupois	x_{11}	Suunnitelman mukaiset poistot.
rahtuott	x_{12}	Rahoitustuotot.
korkokuk	x_{13}	Korkokulut.
rahkulum	x_{14}	Muut rahoituskulut kuin korkokulut.
satunntu	x_{15}	Satunnaiset tuotot eli epäsäännölliset tuotot.
satunnku	x_{16}	Satunnaiset kulut eli epäsäännölliset kulut.
poivarmu	x_{17}	Poistoeron ja varausten muutos yhteensä. Lisäys –, vähennys +.
verovali	x_{18}	Välittömät verot, etumerkki +/–.
tiliktul	y_3	Tilikauden tulos. Voitto +, tappio –.

B Kovarianssimatriisi SAS-koodina

```
/* OIKEAN KOV.MATRIISIN LUONTI */

proc iml;

/* Tuodaan oikeat havainnot.     */
/* HUOM!Kyseessä on kaikki oikeat havainnot. */

use data.ookoot;
read all var _num_ into koko;

/* Lasketaan kov.matriisin kääntmatriisi is */

n=nrow(koko);
one=j(n,1,1);
mt=one`*koko/n;
z=koko-one*mt;
s=z`*z/(n-1);
is=inv(s);

/* Luodaan SAS-datasetti 'data.invs', */
/* jonne viedään kääntmatriisi.      */

create data.invs from is;
append from is;

quit;
```

C Mahalanobisin etäisyys SAS-koodina

```
proc iml;

/* Tuodaan kov.matriisin käänt.matriisi */

use data.invs;
read all var _num_ into is;

/* Tuodaan havaintojen oikeat arvot */

use data.oikeat;
read all var _num_ into x;

/* Vaihda korj_i:n tilalle korjatun datan nimi */

use data.korj_i;
read all var _num_ into korj;

/* Lasketaan mahalanobisin etäisyydet */
/* mahvekt[i] on i:nnen hav. etäisyys */
/* MAHD on hav.koht. etäisyyksien summa */

erot=korj-x;
M=nrow(erot);
mahvekt=j(M,1,0);
DO i=1 to M;
  mahvekt[i]=erot[i,]*is*erot[i,]`;
END;
MAHD=sum(mahvekt);
print MAHD;

/* Luodaan SAS-datasetti 'mahvektx', */
/* jossa hav.kohtaiset etäisyydet.   */

create mahvektx from mahvekt;
append from mahvekt;

quit;
```

D Mahalanobisin etäisyydet

Taulukko 2: Mahalanobisin etäisyydet 10 eri simuloinnista.

Sim.		Korjaamaton	Suhdeimp.	Rescaling	Sekamen.
1	Mahal.sum	44574764	857696,9	**317748,3**	815637,5
	95 %	33,2672	21,4658	8,8715	11,2672
	75 %	0,5554	1,7356	0,2884	0,5927
	med	0,0272	0,3404	0,0182	0,0631
	25 %	0,0009	0,0751	0,0007	0,0028
	5 %	9,88E-06	0,0088	8,64E-06	3,43E-12
2	Mahal.sum	43206488,9	1336549,2	**458302,3**	619360,7
	95 %	32,3580	22,6478	9,7997	12,0027
	75 %	0,5311	1,7401	0,2909	0,5747
	med	0,0290	0,3572	0,0201	0,0629
	25 %	0,0010	0,0760	0,0008	0,0027
	5 %	9,87E-06	0,0091	8,85E-06	2,95E-12
3	Mahal.sum	30291502,6	638394,5	**373586,9**	417596,7
	95 %	33,2673	21,9093	8,9576	11,2197
	75 %	0,5224	1,7074	0,2765	0,5700
	med	0,0276	0,3331	0,0187	0,0597
	25 %	0,0009	0,0704	0,0007	0,0027
	5 %	9,95E-06	0,0091	8,95E-06	2,95E-12
4	Mahal.sum	1,01E+10	1370065,1	**358685,5**	494344,6
	95 %	31,2078	22,3862	9,7600	11,7022
	75 %	0,5262	1,7052	0,2859	0,5473
	med	0,0263	0,3417	0,0180	0,0607
	25 %	0,0009	0,0720	0,0007	0,0028
	5 %	1,00E-05	0,0090	7,97E-06	3,43E-12
5	Mahal.sum	35954998,4	1145849,9	**221881,9**	373434,0
	95 %	34,8263	22,5639	9,3845	10,7620
	75 %	0,5408	1,7957	0,2871	0,5714
	med	0,0278	0,3423	0,0191	0,0626
	25 %	0,0009	0,0728	0,0008	0,0030
	5 %	1,06E-05	0,0090	8,94E-06	7,34E-12

Taulukko 2: (jatkuu)

Sim.		Korjaamaton	Suhdeimp.	Rescaling	Sekamen.
6	Mahal.sum	458619617	535890,9	**222419,5**	492122,7
	95 %	31,0366	21,9771	8,2601	10,9158
	75 %	0,5481	1,6584	0,2832	0,5579
	med	0,0265	0,3242	0,0184	0,0563
	25 %	0,0009	0,0691	0,0007	0,0025
	5 %	1,03E-05	0,0090	8,87E-06	4,43E-12
7	Mahal.sum	45929455	803510,8	**235561,4**	409376,0
	95 %	31,4131	21,3288	9,6321	11,0771
	75 %	0,5593	1,6893	0,2970	0,5711
	med	0,0289	0,3393	0,0196	0,0601
	25 %	0,0009	0,0710	0,0007	0,0028
	5 %	1,00E-05	0,0092	8,25E-06	9,54E-12
8	Mahal.sum	75473678,9	746226	**456001,6**	544214,4
	95 %	29,6443	22,7714	8,7146	11,3446
	75 %	0,5223	1,7327	0,2881	0,5399
	med	0,0260	0,3399	0,0178	0,0567
	25 %	0,0009	0,0720	0,0007	0,0024
	5 %	1,02E-05	0,0089	9,10E-06	9,37E-12
9	Mahal.sum	68124298,4	918637,3	**357682,7**	812359,0
	95 %	33,5282	21,6416	8,7286	11,0489
	75 %	0,5272	1,6856	0,2755	0,5297
	med	0,0261	0,3296	0,0174	0,0590
	25 %	0,0009	0,0692	0,0007	0,0027
	5 %	1,03E-05	0,0085	8,53E-06	2,95E-12
10	Mahal.sum	162189333	621577,7	**193952,1**	414941,8
	95 %	34,1597	21,6416	9,9030	11,9879
	75 %	0,5307	1,7210	0,2850	0,5415
	med	0,0274	0,3383	0,0181	0,0561
	25 %	0,0009	0,0701	0,0007	0,0024
	5 %	1,07E-05	0,0090	9,71E-06	5,45E-12

Taulukko 3: Mahalanobisin etäisyydet sekamenetelmän outlier-metodilla korjatuille. 10 simulointia.

Sim.		Korjaamaton	Suhdeimp.	Rescaling	Outlier
1	Mahal.sum	6826821,8	76079,4	73758,4	**10868,7**
	max 1.	4925989,5	24155,6	50509,9	4068,9
	max 2.	1172124,5	12198,3	4582,9	1078,8
	max 3.	200689,7	5790,8	1925,2	512,9
	95 %	35,8965	14,8092	8,4494	2,1965
	med	0,0955	0,2345	0,0467	8,77E-05
	5 %	0,0001	0,0067	6,98E-05	0
2	Mahal.sum	13017610,9	35999,4	35230,4	**7023,2**
	max 1.	10627745	10096,4	4402,7	625,8
	max 2.	941525	1721,9	4335,3	424,7
	max 3.	311340	960,5	2455,5	344,9
	95 %	49,3469	12,9264	10,5437	2,1953
	med	0,103	0,2552	0,052	7,75E-05
	5 %	0,0001	0,0068	6,36E-05	0
3	Mahal.sum	2408895,6	73733,5	147030,9	**21712,8**
	max 1.	1403331,4	12198,25	99965,4	14568,4
	max 2.	356939,4	9436	12119,3	1499,6
	max 3.	155075	5190,6	4640,9	346,3
	95 %	44,0477	13,8261	9,5427	2,2411
	med	0,0949	0,2302	0,0426	7,38E-05
	5 %	0,0002	0,0071	7,85E-05	0
4	Mahal.sum	5082576,6	85233,8	60061,6	**21500,2**
	max 1.	1676069	29976,2	7265,4	8059,9
	max 2.	1435206	17946,2	7045,4	5645,9
	max 3.	1241157	5544,2	6873,7	1190
	95 %	42,6894	14,4451	10,407	2,1481
	med	0,0997	0,2387	0,0476	7,85E-05
	5 %	0,0001	0,0067	5,91E-05	0
5	Mahal.sum	5290024,4	41323,8	47509,4	**16142**
	max 1.	2909811	5630,1	9424,5	6494,6
	max 2.	848390	3242,5	4167,2	1390,3
	max 3.	477378	2759,6	3074,4	1237,5
	95 %	52,8936	16,7942	11,366	2,2296
	med	0,1056	0,2462	0,0544	0,0001
	5 %	0,0001	0,0065	7,37E-05	0

Taulukko 3: (jatkuu)

Sim.		Korjaamaton	Suhdeimp.	Rescaling	Outlier
6	Mahal.sum	9877288,7	84695,7	41871,3	**10775,7**
	max 1.	8631132	40893,6	15118	4915,9
	max 2.	289737	18566,2	5357,5	609,1
	max 3.	205204	2509	2257,5	428,9
	95 %	39,7641	13,2287	7,5563	2,0081
	med	0,0996	0,234	0,0485	0,0001
	5 %	0,0001	0,0065	7,50E-05	0
7	Mahal.sum	2090742,1	98586,7	41996,2	**30634,1**
	max 1.	1272749,7	30841,3	19988,7	24154,4
	max 2.	227307,4	30423,8	1151	840,3
	max 3.	96540,6	4769,6	1037,9	414,5
	95 %	40,9045	14,3197	8,8507	1,9215
	med	0,0991	0,2466	0,054	0,0001
	5 %	9,99E-05	0,0069	5,64E-05	0
8	Mahal.sum	35898795,8	67228,8	155028,7	**8193,5**
	max 1.	35049767	28026,1	124975,1	4017,9
	max 2.	170011,5	4433,2	5061,6	305,1
	max 3.	167707,5	3394,9	4661,5	263,7
	95 %	40,5759	13,3277	8,6479	1,8889
	med	0,089	0,2461	0,0492	9,53E-05
	5 %	0,0001	0,0069	8,77E-05	0
9	Mahal.sum	1275331,1	182925,9	**21265,7**	21878,4
	max 1.	206825,6	100789,2	1685,1	15565,4
	max 2.	178206,9	40889,3	1340,7	978,4
	max 3.	120153,2	5736,3	1243,6	951,5
	95 %	40,7778	13,2372	8,3729	1,7257
	med	0,0862	0,2255	0,0412	9,33E-05
	5 %	0,0001	0,0063	6,77E-05	0
10	Mahal.sum	1883006,9	68352,6	21206,7	**6443,2**
	max 1.	1031798,3	28026,1	2102,3	775,4
	max 2.	535996,1	5745,3	2040,8	389,8
	max 3.	67430,5	5190,6	1007,2	277,4
	95 %	39,0751	13,6692	9,7807	1,992
	med	0,0976	0,2576	0,0448	9,45E-05
	5 %	0,0001	0,0075	7,26E-05	0

Taulukko 4: Mahalanobisin etäisyydet sekamenetelmän lähimmän naapurin menetelmällä korjatuille. 10 simulointia.

Sim.		Korjaamaton	Suhdeimp.	Rescaling	Naapurim.
1	Mahal.sum	36340723,3	702281,1	**130724**	725432,4
	max 1.	7187170	121560,5	31347,8	310800,1
	max 2.	4566225	100789,2	21140,6	188482,7
	max 3.	4213483	64516,2	15777	36220,6
	95 %	21,9604	21,2138	6,6703	12,3623
	med	0,0095	0,3603	0,0076	0,1155
	5 %	5,76E-06	0,0098	5,35E-06	0,0004
2	Mahal.sum	30056294,3	1174153	**382757,7**	485940,7
	max 1.	19563178	316023	275949	111876,5
	max 2.	2456637	118548,1	19976,7	69823
	max 3.	1750413	105345,4	11417,1	66016,7
	95 %	20,0846	24,1987	6,5892	13,1786
	med	0,0102	0,366	0,0083	0,1124
	5 %	5,76E-06	0,0099	5,30E-06	0,0004
3	Mahal.sum	27533882,5	412668,9	**112656,6**	243891,8
	max 1.	7229729	53087,4	11415	37632
	max 2.	5220197	33881,5	4916,6	28290,1
	max 3.	4954320	31190,2	4850,3	19961,3
	95 %	21,4586	23,1449	6,1412	12,814
	med	0,0088	0,3468	0,0074	0,1093
	5 %	5,85E-06	0,0097	5,40E-06	0,0003
4	Mahal.sum	1,01E+10	1215625,5	**274044,6**	403638,6
	max 1.	1,00E+10	818995,2	87181,8	147429,8
	max 2.	26154636	66373,6	38261,1	40893
	max 3.	12378846	37389,2	31741,8	36360
	95 %	21,6249	23,348	6,5173	13,4203
	med	0,0087	0,3493	0,0072	0,107
	5 %	5,82E-06	0,0103	5,31E-06	0,0003
5	Mahal.sum	27904608,7	1024016	**126332,9**	276781,8
	max 1.	14996001	245369,3	22966,6	44397,7
	max 2.	2428281	215930,5	19106,7	37632
	max 3.	2100048	70760,8	14718,3	36220,6
	95 %	19,1814	22,287	6,0076	11,2456
	med	0,0093	0,3487	0,0075	0,1101
	5 %	6,20E-06	0,0097	5,92E-06	0,0004

Taulukko 4: (jatkuu)

Sim.		Korjaamaton	Suhdeimp.	Rescaling	Naapurim.
6	Mahal.sum	448116734	332649,4	**78078,1**	362801,2
	max 1.	363632402	37292,6	5462,1	146986,3
	max 2.	60609186	29590,3	3854,9	33668
	max 3.	13714062	20326,5	3449,4	19961,3
	95 %	21,5738	24,1784	6,1557	12,7576
	med	0,0092	0,3363	0,0073	0,1027
	5 %	6,03E-06	0,0096	5,30E-06	0,0003
7	Mahal.sum	43761611,9	668627,6	**170249,6**	342445,3
	max 1.	26320260	116676,7	36639,7	188482,7
	max 2.	6506033	100789,2	25667,4	36547,8
	max 3.	1777499	95269,3	24845,4	8688,5
	95 %	22,1171	23,058	7,268	13,2174
	med	0,0092	0,3487	0,0076	0,1053
	5 %	5,47E-06	0,0097	5,11E-06	0,0004
8	Mahal.sum	38408884,8	647079,7	**180437,6**	504103,4
	max 1.	12666928	97209,6	66318,9	106607,5
	max 2.	9220791	92228,6	47463,6	66511,3
	max 3.	6890536	65200,9	5134,6	50685,4
	95 %	20,4413	24,8819	6,6916	13,0942
	med	0,0086	0,3465	0,0073	0,1069
	5 %	5,80E-06	0,0096	4,82E-06	0,0003
9	Mahal.sum	66687500,1	651735,3	**271339,8**	706504,5
	max 1.	33640385	105536	42169,4	257244,6
	max 2.	20729253	60336,5	40191,1	36393,5
	max 3.	2555234	52939,3	36686,4	36288,7
	95 %	22,3212	23,4373	6,5635	13,2215
	med	0,0083	0,3388	0,0072	0,107
	5 %	5,33E-06	0,0093	5,23E-06	0,0004
10	Mahal.sum	159932416	487139,6	**154303,9**	342413,1
	max 1.	142288346	51402,4	49102,7	63886,1
	max 2.	7307657	40884,8	23671,3	38866,8
	max 3.	4111324	37453,4	11808,9	36227,2
	95 %	23,6753	24,0161	7,5255	13,821
	med	0,009	0,3385	0,0072	0,1066
	5 %	5,63E-06	0,0095	5,86E-06	0,0003

Taulukko 5: Mahalanobisin etäisyydet sekamenetelmän suhdeimputoinnilla korjatuille. 10 simulointia.

Sim.		Korjaamaton	Suhdeimp.	Rescaling	Suhdeimp.
1	Mahal.sum	1407219,1	**79336,4**	113265,9	**79336,4**
	max 1.	820789,5	40892,5	53597,8	40892,5
	max 2.	382672,4	5256,7	18070,1	5256,7
	max 3.	39605,2	5190,6	15082,2	5190,6
	95 %	123,072	81,4489	54,7599	81,4489
	med	0,5914	0,9454	0,3786	0,9454
	5 %	0,0016	0,0226	0,0015	0,0226
2	Mahal.sum	132583,7	126396,8	**40314,2**	126396,8
	max 1.	46161,9	45477,8	10838,3	45477,8
	max 2.	13578	35228,4	4187,7	35228,4
	max 3.	10770,1	5719,5	3337,1	5719,5
	95 %	118,407	57,0555	44,4011	57,0555
	med	0,5489	1,0213	0,3934	1,0213
	5 %	0,0016	0,0313	0,0011	0,0313
3	Mahal.sum	348724,6	151992,1	**113899,4**	151992,1
	max 1.	202765,2	97703,6	52494,7	97703,6
	max 2.	45751,4	19239,1	20650,6	19239,1
	max 3.	28280	10272,9	11066,4	10272,9
	95 %	156,737	51,738	60,274	51,738
	med	0,5623	1,0139	0,3655	1,0139
	5 %	0,0015	0,0269	0,0013	0,0269
4	Mahal.sum	147513,9	69205,8	**24579,3**	69205,8
	max 1.	94319,1	40905,6	2971,2	40905,6
	max 2.	4737,6	5143,3	1866,6	5143,3
	max 3.	4447,1	3394,9	1603,8	3394,9
	95 %	118,116	56,64	48,287	56,64
	med	0,6017	1,0851	0,4047	1,0851
	5 %	0,0013	0,017	0,001	0,017
5	Mahal.sum	2760365,4	80510,1	**48039,7**	80510,1
	max 1.	2207981,8	41228,8	28314,5	41228,8
	max 2.	420281,9	12198,3	2463,2	12198,3
	max 3.	33545,3	5264,2	2085,5	5264,2
	95 %	168,506	75,9819	47,9617	75,9819
	med	0,6265	1,0608	0,4187	1,0608
	5 %	0,0012	0,0268	0,001	0,0268

Taulukko 5: (jatkuu)

Sim.		Korjaamaton	Suhdeimp.	Rescaling	Suhdeimp.
6	Mahal.sum	625594,7	118545,8	**102470**	118545,8
	max 1.	235314,5	43921,9	57474,3	43921,9
	max 2.	103735,9	35228,4	8890,2	35228,4
	max 3.	96148,8	5755,4	6681,3	5755,4
	95 %	117,654	60,1448	44,1989	60,1448
	med	0,6525	1,0083	0,4291	1,0083
	5 %	0,0014	0,0296	0,0014	0,0296
7	Mahal.sum	77101,1	36296,5	**23315,6**	36296,5
	max 1.	24156,1	5190,6	5482,9	5190,6
	max 2.	12027,5	5143,3	2037,6	5143,3
	max 3.	7941	3594,1	1820,4	3594,1
	95 %	91,5312	44,3875	46,0039	44,3875
	med	0,6507	0,9873	0,4486	0,9873
	5 %	0,0018	0,0307	0,0011	0,0307
8	Mahal.sum	1165998,3	**31917,5**	120535,3	**31917,5**
	max 1.	767819,1	5607,8	57474,3	5607,8
	max 2.	169985,6	4871	20485,1	4871
	max 3.	96148,8	2759,6	18751,3	2759,6
	95 %	100,272	61,9965	53,8119	61,9965
	med	0,5712	0,9362	0,4066	0,9362
	5 %	0,0026	0,0247	0,0021	0,0247
9	Mahal.sum	161467,2	83976,1	**65077,2**	83976,1
	max 1.	70129,2	30423,8	31914,2	30423,8
	max 2.	28076,3	11761	10789,8	11761
	max 3.	25786,4	7353,3	5423,9	7353,3
	95 %	118,528	72,7096	53,9714	72,7096
	med	0,4423	0,8883	0,3022	0,8883
	5 %	0,0014	0,0251	0,0022	0,0251
10	Mahal.sum	373910,3	66085,5	**18441,5**	66085,5
	max 1.	255497,3	40751,2	1903,6	40751,2
	max 2.	62130,6	3593,7	1529,1	3593,7
	max 3.	8363,9	2880,4	1266,3	2880,4
	95 %	116,885	48,1748	48,1816	48,1748
	med	0,6802	1,2056	0,4516	1,2056
	5 %	0,0018	0,0227	0,0015	0,0227

E Suhteelliset virheet summissa toimialoittain

TOL-95	Frekv.	Korjaamaton	Suhdeimp.	Rescaling	Sekamen.
1	666	0,0367	**0,0284**	0,0472	0,0335
2	2226,6	0,0460	**0,0051**	0,0388	0,0166
3	648,6	0,0438	**0,0123**	0,0434	0,0156
4	3060	0,0605	**0,0117**	0,0321	0,0223
50	959,6	0,0372	**0,0102**	0,0525	0,0133
51	1575,1	0,0510	0,0270	0,0307	**0,0164**
52	2476,5	0,0200	**0,0023**	0,0240	0,0064
55	1026	0,0467	0,0059	0,0402	**0,0057**
6	1817,2	0,0554	0,0184	0,0461	**0,0150**
7	3032,4	0,0447	0,0353	0,0306	**0,0206**
Painotettu ka		0,0452	**0,0159**	0,0355	0,0164

Taulukko 6: *Liikevai*-muuttujan suhteellisten virheiden keskiarvot.

TOL-95	Korjaamaton	Suhdeimp.	Rescaling	Sekamen.
1	0,0877	**0,0268**	0,1097	0,0652
2	**0,0549**	0,0638	0,0746	0,0565
3	0,0650	0,0436	0,0830	**0,0401**
4	**0,0430**	0,0913	0,0647	0,0548
50	0,0656	**0,0120**	0,0845	0,0230
51	0,0377	0,0392	0,0446	**0,0282**
52	0,0196	**0,0067**	0,0279	0,0092
55	0,1225	0,0357	0,0672	**0,0130**
6	0,2202	0,3070	0,1619	**0,0409**
7	0,1062	0,2532	**0,0725**	0,1351
Painotettu ka	0,0785	0,1098	0,0740	**0,0543**

Taulukko 7: *Ostoyht*-muuttujan suhteellisten virheiden keskiarvot.

TOL-95	Korjaamaton	Suhdeimp.	Rescaling	Sekamen.
1	3,5761	**0,5484**	1,1111	0,5574
2	1,4934	0,5058	**0,3792**	0,6830
3	**5,8623**	21,2032	6,0184	15,0629
4	5,0679	**0,2251**	0,5762	0,5640
50	3,2830	**0,3151**	0,6281	0,3605
51	8,8463	0,4282	0,5870	**0,3963**
52	1,4812	**0,2494**	0,3827	0,2604
55	2,7300	0,6331	0,9559	**0,5664**
6	0,9528	2,4623	**0,7347**	2,5297
7	4,8512	2,4802	**1,3402**	2,1949
Painotettu ka	3,7175	1,7253	**0,9210**	1,5346

Taulukko 8: *Varastmu*-muuttujan suhteellisten virheiden keskiarvot.

TOL-95	Korjaamaton	Suhdeimp.	Rescaling	Sekamen.
1	0,2891	**0,0916**	0,2336	0,1119
2	0,2291	**0,0489**	0,2118	0,0947
3	0,2930	0,0871	0,2243	**0,0577**
4	0,2205	0,0698	0,2190	**0,0307**
50	0,6358	**0,0719**	0,2112	0,1180
51	0,3799	0,1229	0,1712	**0,1025**
52	0,2327	0,0742	0,1858	**0,0607**
55	0,2617	0,1932	0,2162	**0,1473**
6	0,3028	**0,0804**	0,2423	0,1410
7	0,1848	**0,1692**	0,1900	0,1815
Painotettu ka	0,2705	**0,0997**	0,2066	0,1029

Taulukko 9: *Verovali*-muuttujan suhteellisten virheiden keskiarvot.

TOL-95	Korjaamaton	Suhdeimp.	Rescaling	Sekamen.
1	4,4563	1,2365	0,3454	**0,2849**
2	2,3715	0,5406	**0,1577**	0,3271
3	1,9039	0,9803	**0,1679**	0,3938
4	2,4540	0,7381	**0,1020**	0,1316
50	1,9316	0,3614	**0,1129**	0,1775
51	4,1639	0,4490	**0,0960**	0,3310
52	0,8539	0,1852	**0,0725**	0,1064
55	0,5721	0,3272	**0,1890**	0,2955
6	0,4806	0,4064	**0,1101**	0,3503
7	33,9579	0,3376	**0,1643**	0,3190
Painotettu ka	7,5454	0,4878	**0,1334**	0,2538

Taulukko 10: *Rahtuott*-muuttujan suhteellisten virheiden keskiarvot.

TOL-95	Korjaamaton	Suhdeimp.	Rescaling	Sekamen.
1	0,7219	0,0899	**0,0383**	0,0417
2	0,5018	0,0230	0,0220	**0,0099**
3	1,4586	0,1527	**0,0280**	0,0681
4	0,9247	0,0588	**0,0089**	0,0097
50	2,4651	0,0514	0,0585	**0,0312**
51	1,8302	0,1286	0,0326	**0,0155**
52	0,5105	0,0400	0,0073	**0,0048**
55	1,0896	0,0628	0,0078	**0,0063**
6	1,3224	0,0987	**0,0139**	0,0235
7	0,4141	0,1754	**0,0234**	0,0277
Painotettu ka	0,9528	0,0867	0,0200	**0,0185**

Taulukko 11: Käyttökatteen suhteellisten virheiden keskiarvot.

TOL-95	Korjaamaton	Suhdeimp.	Rescaling	Sekamen.
1	2,1009	0,1905	0,0637	**0,0299**
2	1,0431	0,0670	**0,0136**	0,0371
3	2,9444	0,1533	**0,0302**	0,0908
4	1,6358	0,0674	**0,0101**	0,0587
50	4,7625	**0,0596**	0,0598	0,0648
51	3,5071	0,1107	**0,0363**	0,0806
52	0,8751	0,0481	**0,0125**	0,0259
55	2,3638	0,1738	**0,0305**	0,0952
6	3,4939	0,1721	**0,0248**	0,0971
7	137,5294	23,5593	**1,4724**	25,9007
Painotettu ka	25,6586	4,1666	**0,2750**	4,5408

Taulukko 12: Tilikauden tuloksen suhteellisten virheiden keskiarvot.

Tilastotoimen menetelmien maisteriohjelman pro gradu -tutkielma sarja

1. Salmikuukka, J. (1997)	Aikasarjojen perusrakennemalleista ja niiden soveltaminen Jyväskylän kaukolämmön kulutuksen analysointiin ja ennustamiseen. (76 s., 1 liite)	Jyväskylän Energia Oy, Jyväskylä
2. Yrjölä, T. (1997)	Lasten päivähoidon tuottavuusvertailu suurissa kaupungeissa DEA-menetelmällä. (72 s., 2 liitettä)	Jyväskylän kaupungin terveystoimi, Jyväskylä
3. Ainiala, N. (1997)	Helsingin osa-alueiden työvoimatilastojen estimointi pienaluetekniikalla valtakunnallisesta työvoimatutkimuksesta. (73 s., 3 liitettä)	Helsingin kaupungin tietokeskus, Helsinki
4. Puhakka, E. (1997)	Kiintiöpoiminnan tilastolliset ominaisuudet pk-yritysbarometritutkimuksessa. Sovelluksena 2/1996 aineisto (82 s., 2 liitettä) (salainen)	Tietoykkönen, Jyväskylä
5. Salonen, R. (1997)	Muutoksen ja tason estimointi rotatoivassa paneeliaineistossa eri estimaattoreiden avulla. Sovellus työvoimatutkimuksen aineistoon. (39 s, 4 liitettä)	Tilastokeskus, Helsinki
6. Kunttu, S. (1997)	Alueellisen teollisuustuotannon volyymi-indeksin estimointi Etelä-Pohjanmaalle. (103 s.)	Tilastokeskus, Seinäjoki
7. Koponen, M. (1997)	Epätäydelliseen otantakehikkoon perustuvan yritysaineiston estimointi. Käsiteanalyysiä ja soveluksia Tilastokeskuskuksen järjestäytymättömiä yrityksiä koskevaan palkkatiedusteluun. (56 s., 5 liitettä)	Tilastokeskus, Helsinki
8. Kainulainen, P. (1997)	Toimialatilastot sanomalehdessä: Monimuuttujaisen rekisteriaineiston havainnollistaminen ja tuotteistaminen (62 s., 6 liitettä)	Sanomalehtien liitto, Helsinki
9. Storfors, S. (1998)	Toistojen määrän arviointi kokeellisessa tutkimuksessa. (34 s., 3 liitettä)	Cultor Oy, Helsinki
10. Peltoniemi-Laajanen, L. (1998)	Pääteiden kehittämistarpeet: kuljettajahaastattelun monimuuttuja-analyysi (78 s., 8 liitettä)	VTT Yhdyskuntatekniikka, Helsinki
11. Hirvonen, M. (1999)	Vastauskadon käsittely yritysten innovaatiokyselyssä 1996 (81s., 9 liitettä)	Tilastokeskus, Helsinki
12. Nevalainen, J. (1999)	Välttöfunktion estimointi kliinisessä toistomittauskokeessa (56 s., 5 liitettä)	Leiras, Helsinki
13. Vierinkari, K. (1999)	Aaltojuotoskoneen parametrien optimointi Taguchi-menetelmällä.(92 s, salainen)	Essex Communication EMS Oy, Äänekoski

14. Pylvänäinen, I. (1999)	Kliinisen toistomittauskokeen analysointi toisistaan riippuvien binääri- ja ordinaali-vasteiden tapauksessa. (82 s., 3 liitettä)	Orion Oyj, Turku
15. Saarnio, K. (1999)	Työvoimaa kuvaavien tunnuslukujen estimointi pienaluetasolla kuukausittaisesta työvoimatutkimuksesta; hierarkkisten mallien sovellus. (107 s., 4 liitettä)	Tilastokeskus, Helsinki
16. Lamberg, P. (1999)	Pienalue-estimaatteja pitkäaikaissairastavuudelle seutukuntatasolla. (76 s., 4 liitettä)	Tilastokeskus, Helsinki
17. Piela, P, (1999)	Massaimputointi kaupan lyhyen aikavälin tilastotuotannossa. (120 s., 12 liitettä, salainen)	Tilastokeskus Helsinki
18. Tuominen, T. (1999)	Vastausrasite ja sen mittaaminen Tilastokeskuksen yrityssurveyssä 1998. (173 s)	Tilastokeskus, Helsinki
19. Pohjanjousi, P. (2000)	Kiinteiden ja satunnaisvaikutusten meta-analyysi (65 s.+ 29 s. liitteitä)	Orion Helsinki
20. Partanen, A. (2000)	Parametriset ja epäparametriset menetelmät vaihtovuoroasetelmassa (70 s.)	Orion, Turku
21. Isohanni, P. (2000)	Ammatin pääluokkien keskipalkkojen estimointi järjestäytymättömiä yrityksiä koskevasta Tilastokeskuksen palkkatiedustelusta	Tilastokeskus, Helsinki
22. Ahlgren, M. (2001)	Tiedon louhinta myynnin, markkinoinnin ja asiakkuuksien hallinnassa.. (49 s. + 3 liitettä)	TietoEntra, Espoo
23. Luoma, S. (2001)	Tekes-rahoituksen saaneiden pk-yritysten ryhmittelyä eri yritystyyppeihin. (74s. + 3 liitettä)	Tekes, Helsinki
24. Ikäheimonen, J. (2001)	Elinkeinoverorekisteriaineiston editointi ja imputointi. Sovellettujen menetelmien vertailu. (79 s. + 5 liitettä)	Tilastokeskus, Helsinki