Niina Ainiala

HELSINGIN OSA-ALUEIDEN TYÖVOIMATILASTOJEN ESTIMOINTI PIENALUETEKNIIKALLA VALTAKUNNALLISESTA TYÖVOIMATUTKIMUKSESTA

Tilastotieteen
pro gradu -tutkielma
6.6. 1997

Jyväskylän yliopisto
Tilastotieteen laitos
Informaatioteknologian maisteriohjelmat
Tilastotoimen menetelmät

Tilastotoimen menetelmien maisteriohjelma
(http://www.stat.jyu.fi/) (Päivitetty 16.5.1997)

Tilastotoimen menetelmien maisteriohjelman tavoite on kouluttaa opiskelija tilastotiedon keruun, jatkojalostuksen ja käytön asiantuntijaksi nykyaikaisessa tilastojärjestelmäympäristössä, jossa datat ovat survey-, koeasetelma- tai rekisteriperusteisia. Koulutus järjestetään yhteistyönä Jyväskylän yliopiston tilastotieteen laitoksen kanssa. Tilastotieteelliseltä kannalta kyse on survey-menetelmiin, biostatistiikkaan ja ekonometriaan erikoistuneiden tilastoasiantuntijoiden koulutuksesta. Ohjelman kesto päätoimisella opiskelijalla on kaksi lukuvuotta.

Tilastotoimen ytimenä ovat tilastotieteen, erityisesti tilastotoimen teorian ja tietojenkäsittelyn kurssit. Ohjelmaan voidaan sisällyttää myös koulutustavoitteita tukevia opintojaksoja lähitieteistä (esim. talous-, yhteiskunta- ja viestintätieteet) sekä yritystoimintaan perehdyttäviä kursseja. Tarjottava opetus koostuu osittain tilastotieteen laitoksen opetusohjelmasta, erityisesti suunnitelluista tilastotoimen teorian kursseista ja ostopalveluina hankituista muista erikoiskursseista. Opettajista pääosa on kotimaasta. Vierailijoina on myös ulkomaalaisia asiantuntijoita, joten opetus on osin englanninkielistä.

Ohjelman tärkeä osa on yliopiston ulkopuolisessa yhteistyötoimipaikassa suoritettu pro gradu -tutkielma ja siihen liittyvä harjoittelu. Yhteistyötoimipaikat ovat virallisesta tilastotoimesta, suuryrityksistä tai tutkimuslaitoksista. Ne osallistuvat osaltaan harjoittelun kustannuksiin. Perusajatus on, että pro gradu -tutkielma tai osa siitä toisi tutkimustulostensa osalta lisäarvoa yhteistyötoimipaikalle. Tätä tukenee se, että harjoittelu jaetaan kahteen osaan, joista ensimmäinen eli orientoiva osa on ensimmäisen opintovuoden lopussa. Sen jälkeen opiskelija palaa yliopisto-opiskeluihin yhden lukukauden ajaksi syventääkseen tietojaan siinä tilastotieteen osa-alueessa, joka on tarpeen aiotussa pro gradu -tutkielmassa. Maisteriohjelman viimeinen lukukausi muodostaa yhtenäisen tutkimusjakson, jonka aikana tehdään pro gradu -tutkielma.

Master's Programme in Statistical Systems
(http://www.stat.jyu.fi/) (Updated 16.5.1997)

The main traget of the Program is to educate high qualified professionals in the collection analysis, managing and dissemination of large data sets. The Program has been built mainly on the regular curriculum of the Department of Statistics in the University of Jyväskylä. In point of view of statistical sciences, the Program in Statistical Systems concentrate in the knowledge of survey methodology, biometry and econometry. Full-time students can graduate in two Academic years. The core courses cover the advanced theory in mathematical statistics, of statical systems expecially the survey methodology and some courses in information technology. Still the program is flexible that a moderate amount of different kinds of studies in the neighboring sciences can be included as studies in economics, social sciences and communication. Most of the instructors come from Finland but visiting professors from outboard are eventual using English.

An important part of the Program crows from the cooperational research work created between the Department of Statistics and the research units located outside the University of Jyväskylä. Those units are research and development departments in the bodies of official statistics, big business firms and research institutes. They share the costs of practices. Basic idea is that the MS.c. Thesis wil be written from the topics given by the cooperational research units and so the results could be of some contributed value to them. As an appropriate mode of policy may be here thus that the time of practice skills for the Thesis. This is proceeded in 6 final months of the Program, in collaboration with the same cooperating research unit.

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 ratatoivassa 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

Tiivistelmä

Niina Ainiala: ***Helsingin osa-alueiden työvoimatilastojen estimointi pienaluetekniikalla valtakunnallisesta työvoimatutkimuksesta***
Tilastotieteen pro gradu-tutkielma, Jyväskylän yliopisto, 6. kesäkuuta 1997. Sivuja 73, liitteitä 3.

Tässä työssä on tutkittu millaiset ovat mahdollisuudet tuottaa pienalue-estimaattoreiden avulla Tilastokeskuksen työvoimatiedustelusta neljännesvuosittaisia otospohjaisia alueellisia tilastotietoja Helsingin työvoimasta. Päämääränä on, että Helsingin kaupungin tietokeskus pystyisi tulevaisuudessa tuottamaan alueellista tietoa Helsingin työvoimasta nopeammassa tahdissa kuin se tähän asti on tehnyt.

Pienaluetta on tässä tutkimuksessa nimitetty soluksi. Helsingin työvoima on jaettu soluiksi seitsemän suurpiirin, kolmen ikäryhmän ja sukupuolen mukaan. Otosaineistona on käytetty työvoimatiedustelun yksilötasoista neljännesvuosiaineistoa Helsingin osalta ja rekisteriaineistona Tilastokeskuksen tuottamaa rekisteripohjaista työssäkäyntitilastoa. Otosaineiston tiedot ovat peräisin vuoden 1994 kolmelta ensimmäiseltä ja työssäkäyntitilasto kuvaa tilannetta vuoden 1993 lopussa. Tulosmuuttujina ovat soluittaiset työllisten ja työttömien kokonaismäärät sekä työttömyysasteet.

Pienalue-estimaattorit voidaan jakaa ominaisuuksien perusteella kolmeen ryhmään: (1) suorat estimaattorit, (2) synteettiset estimaattorit ja (3) komposiittiestimaattorit. Tutkimuksessa on mukana estimaattoreita kaikista kolmesta pääryhmästä. Estimaattoreita on verrattu toisiinsa soluittaisten keskivirheiden ja variaatiokertoimien sekä kullekin estimaattorille laskettavien keskineliövirheen määrää kuvaavien mittareiden avulla.

Tulosten perusteella parhaimpaan tulokseen päästään komposiittiestimaattorilla. Tietojen julkaisukelpoisuuden parantamiseksi, täytyy tutkimusta hiukan vielä jatkaa. Avainasemassa on sopivimman mallin löytäminen ja sen kehittäminen niin, että myös ajassa tapahtuvat muutokset tulisi huomioiduksi.

Avainsanoja: pienalue, pienalue-estimointi, synteettinen estimaattori, komposiittiestimaattori, työvoima, työlliset, työttömät

Abstract

Niina Ainiala: ***Estimation of labour statistics for Helsinki's sub-districts, applying a small area technique on a nationwide labour force survey***

Pro gradu thesis in Statistics, University of Jyväskylä, 6. June 1997. Pages 73, appendixes 3.

In my work, I have evaluated the chances of applying small area estimators to produce sample-based local statistics on Helsinki's labour force. This would happen four times a year, and the material would be based on Statistics Finland's labour force survey. The purpose of my study is to make it possible for City of Helsinki Urban Facts to produce local data on Helsinki's labour force faster than today.

In my work, I have given the small areas another denomination, namely cell, and divided Helsinki's labour force cells according to 1) the seven major districts of the city, 2) three age groups and 3) the two sexes. My sample material consisted of quarterly individual data from the labour force survey for Helsinki, and my register data were employment statistics from Statistics Finland. The data of the sample material covers the three first months of 1994, and the employment statistics cover the situation at the end of 1993. The study variables are the total nmbers of employed and unemployed plus the unemployment rates in each cell.

The small area estimators can be divided into three groups: 1) direct estimators, 2) synthetic estimators and 3) composite estimators. In my study, I have used estimators of all three categories, and compared them with another in terms of standard error, variation coefficient per cell and measures for root mean square error per estimator.

According to my result, the composite estimators give the best results. In order to improve the publishability of the data, my study will have to be extended. The key issue then will be to find and develop a more suitable model which would also account for changes over time.

Key words: small area, small area estimation, synthetic estimator, composite estimator, labour force, employed, unemployed

Sisällysluettelo

1 Johdanto

1.1 Työn tausta

Tilastollisten menetelmien ja tietotekniikan kehittyessä yhteiskuntaa ja sen suhdanteita kuvaavia lukuja - tilastoja tuotetaan ja kehitetään alati kiihtyvässä tahdissa. Rekisteripohjaiset hallinnollisiin aineistoihin perustuvat vuositilastot ovat vielä tällä hetkellä aivan liian hitaita välineitä kuvaamaan yhteiskunnan nopeimmin muuttuvia ilmiöitä kuten esim. työttömyyttä. Näiden ilmiöiden seuraamiseksi on kehitetty otantatutkimuksia, joissa mahdollisimman edustavan otoksen avulla arvioidaan koko perusjoukon tilannetta.

Parin viimeisimmän vuosikymmenen aikana on nopeammassa tahdissa tuotetun informaation tarve lisääntynyt olennaisesti. Yhteiskunta tarvitsee suunnittelun ja päätöksenteon pohjaksi tietoa siitä, mitä maassa tapahtuu. Yhä useammin on myös niin, ettei pelkkä koko maata koskeva tieto riitä. Vastaavat tiedot tarvitaan myös lääni-, kunta- ja kaupunkitasoisina ja mielellään vielä erilaisiin demografisiin ryhmiin luokiteltuna. Vaikeutena on, että otantatutkimukset ovat usein alunperin suunniteltu antamaan tietoa koko yhteiskunnasta tai korkeintaan muutamaan osa-alueeseen paloiteltuna. Tiheästä osajaottelusta seuraa, että pienempiin osa-alueisiin ei saada riittävästi otoshavaintoja vakaiden tilastolukujen tuottamiseksi kyseisiltä osa-alueilta. Lisäksi otoshavaintojen määrä kullakin osa-alueella on satunnainen. Tämän estimointiongelman ratkaisemiseksi on olemassa lähinnä kaksi vaihtoehtoa: otoskoon kasvattaminen tai pienalue-estimointitekniikoiden käyttäminen. Otoskoon kasvattaminen on usein tehokas, mutta erittäin kallis ja aikaa vievä vaihtoehto.

Kustannusten ja vastausrasituksen kasvun välttämiseksi on kehitetty pienalue-estimaattoreita, jotka pyrkivät parempaan tulokseen ottamalla tukea muilta osa-alueilta ja hyödyntämällä jo olemassa olevia rekisteripohjaisia kokonaisaineistoja. Tässä työssä käsitellään tätä jälkimmäistä vaihtoehtoa ja tutkitaan miten hyvin nämä pienalue-estimaattorit toimivat Helsingin alueilla.

1.2 Helsingin kaupungin tietokeskus

Helsingin kaupungin tietokeskus on kaupunginhallituksen alainen kaupungin tietohallinnon keskuselin. Sen toimintaperiaatteena on edistää Helsingin kaupungin toimintakykyä huolehtimalla siitä, että olennaiset ja ajantasaiset tiedot ovat aina käytettävissä. Tietokeskuksen tilastotuotanto perustuu suurimmalta osin valmiisiin aineistoihin, joita ovat mm. kaupungin omat rekisterit ja tietokannat sekä hallintokuntien ja Tilastokeskuksen tuottamat aineistot. Helsingin kaupunki jaetaan maantieteellisesti seitsemään suurpiiriin ja niiden sisällä 33 peruspiiriin, jotka jakautuvat edelleen 117 osa-alueeseen. Viimeisin aluejako on tullut voimaan vuonna 1990. Helsinkiä ja sen osa-alueita koskevien tilastojen laatimisesta ja julkaisusta vastaa Helsingin kaupungin tietokeskus. Tämä tutkimus on yhteistyöprojekti Helsingin kaupungin tietokeskuksen, Jyväskylän yliopiston ja Tilastokeskuksen välillä.

1.3 Tutkimuksen tarkoitus

Työttömyys on yksi 1990-luvun ajankohtaisimmista ja tarkimmin julkisuudessa seuratuista yhteiskunnallisista ilmiöistä Suomessa. Päävastuu työttömyyttä kuvaavien lukujen tuottamisessa on Tilastokeskuksella ja Työministeriöllä. Tällä hetkellä tietokeskus saa tuoreimmat Helsinkiä koskevat työttömyyttä kuvaavat luvut, Tilastokeskuksen työvoimatiedustelusta neljännesvuosittain ja Työministeriön työnhakijarekisteristä kuukausittain. Molemmat tiedot toimitetaan noin kuukauden viiveellä tilastointiajankohdasta. Työvoimatiedustelun ja työnhakijarekisterin tuottamat luvut koskevat koko Helsinkiä sukupuolen ja iän mukaan ryhmiteltynä.

Työministeriön työttömyysasteen työvoimalukuna käytetään Tilastokeskuksen työssäkäyntitilaston ennakkotietoa työvoiman määrästä, joka on noin puolentoista vuoden takainen.

Tuoreimmat Helsingin sisäisen aluejaon mukaiset työttömyysluvut tietokeskus saa käyttöönsä kerran vuodessa noin puolen vuoden viiveellä tilastointiajankohdasta. Nämä tiedot ovat edellisen vuoden joulukuulta ja ne syntyvät erikoistulostuksena osana Tilastokeskuksen tuottaman rekisteripohjaisen työssäkäyntitilaston valmistumisprosessia, jossa väestörekisterin tietoihin yhdistetään työministeriön työnhakijarekisterin tietoja. Itse työssäkäyntitilasto valmistuu kokonaisuudessaan vasta noin kahden vuoden päästä tilastointiajankohdasta ja sen muodostamiseen käytetään useita muitakin hallinnollisia rekistereitä kuin edellä mainitut.

Tarve olisi kuitenkin saada myös tuoreempia aluekohtaisia työvoimalukuja päätöksenteon ja suunnittelun pohjaksi. Tämän työn tavoitteena on tutkia ja selvittää millaiset mahdollisuudet ovat tuottaa pienalue-estimaattoreiden avulla Tilastokeskuksen työvoimatiedustelusta neljännesvuosittaisia alueellisia tilastotietoja Helsingin työvoimasta. Otosaineistona tutkimuksessa on työvoimatiedustelun neljännesvuosiaineisto Helsingin osalta ja tukea antavana rekisteriaineistona käytetään Tilastokeskuksen työssäkäyntitilastoa.

2 Pienalue-estimaattorit

Pienalue-estimaattoreiden yhteydessä sanalla "pieni" viitataan havaintoyksikköjen määrään kullakin *pienalueella*, eikä niinkään alueen fyysiseen kokoon. Pienaluetta ei siis välttämättä tarvitse määritellä maantieteellisesti, vaan se voi muodostua myös jonkin tai joidenkin muiden havaintoyksiköiden ominaisuuksien mukaan kuten esim. iän ja sukupuolen. Englanniksi pienaluetta kutsutaan sanalla "domain", mutta sen suomennokset kuten ala tai piiri viittaavat selkeästi johonkin maantieteelliseen alueeseen. Väärinkäsitysten välttämiseksi tässä tutkimuksessa pienaluetta kutsutaan *soluksi*.

Pienalue-estimointiin liittyvissä artikkeleissa otetaan hyvin vähän kantaa siihen milloin jotakin aluetta voidaan kutsua pienalueeksi. Singhin ja Mian (1995) esittämässä tutkimuksessa pienalueen estimaatti voidaan estimoida suoraan, jos alueen koko on 200 havaintoa tai sitä suurempi. Rosénin (1991) mukaan ei kysymykseen siitä, milloin voidaan jokin alue lukea pienalueeksi, ole täsmällistä vastausta. Vastaus määräytyy tapauskohtaisesti mm. sen mukaan miten tarkkoja estimointituloksia halutaan.

Myös Purcell ja Kish (1979) painottavat tutkimuksessaan, että estimaattorin valinta riippuu aina tilanteesta, jota voidaan tarkastella erityisesti perusjoukon *suhteellisten solukokojen* $P_{dg} = N_{dg}/N$ avulla. Tämä edellyttää, että perusjoukon tiedot ovat saatavilla. Purcell ja Kish jakavat solut neljään luokkaan sen mukaan kuinka suuri on solun suhteellinen koko P_{dg} perusjoukossa. Luokitus on seuraava:

1. Suuri (engl. major)	, jos			P_{dg}	$\geq$	0.1
2. Keskikoko (engl. minor)	, jos	0.01	$\leq$	P_{dg}	$<$	0.1
3. Pieni (engl. mini)	, jos	0.0001	$\leq$	P_{dg}	$<$	0.01
4. Erittäin pieni (engl. rare)	, jos			P_{dg}	$<$	0.0001.

Vaikka otoksen koko olisi kohtuullisen suuri, niin kolmanteen tai neljänteen luokkaan kuuluville soluille on hyvin vaikea tuottaa luotettavaa estimaattia perinteisin menetelmin. Tällöin tarvitaan pienalue-estimaattoreita.

Pienalue-estimaattorit voidaan jakaa ominaisuuksiensa perusteella kolmeen laajaan pääryhmään: (1) suorat estimaattorit, (2) epäsuorat estimaattorit, jotka perustuvat implisiittisiin malleihin ja (3) malliperusteiset estimaattorit (Rao & Choudry, 1995). Jotta eri estimaattorityyppien ominaisuudet tulisivat paremmin esiin, voidaan mielestäni jako tehdä hieman erilaisista lähtökohdista: (1) suorat estimaattorit, (2) synteettiset estimaattorit, jossa omana alaryhmänään (2.1) regressiopohjaiset estimaattorit, sekä (4) komposiittiestimaattorit.

Pienalue-estimaattoreita on viime vuosikymmeninä tutkittu ja kehitelty aktiivisesti sekä teorian että käytännön saralla. Estimaattoreiden nimet ja merkintätavat eivät ole kuitenkaan yleisesti vakiintuneet. Tässä tutkimuksessa tuodaan esille pääasiassa niitä estimaattoreita, jotka ovat käyttökelpoisia tämän tutkimusongelman kannalta. Seuraavana esitettävissä estimaattoreiden kaavoissa otanta-asetelmana on yksinkertainen satunnaisotanta palauttamatta. Perusjoukko $U=\{1,\ldots,k,\ldots,N\}$ jakautuu D:hen toisensa poissulkevaan osa-alueeseen $U_1,\ldots,U_d,\ldots,U_D$ (esim. suurpiirit). Edelleen perusjoukko jakautuu G:hen toisensa poissulkevaan ryhmään $U_1,\ldots,U_g,\ldots,U_G$ (esim. ikäryhmät). Tämä ristiinluokittelu jakaa perusjoukon soluihin U_{dg} $(d = 1,\ldots,D;\ g = 1,\ldots,G)$, joissa N_{dg} on solun U_{dg} tunnettu havaintojen määrä perusjoukossa. Perusjoukosta U poimittu satunnaisotos $s=\{1,\ldots,k,\ldots,n\}$ jakautuu perusjoukon tapaan soluihin s_{dg} $(d=1,\ldots,D;\ g=1,\ldots,G)$, joissa solujen koot n_{dg} ovat satunnaismuuttujia. Taulukossa 2.1 on havainnollistettu aineiston jakautuminen alueisiin ryhmiin ja soluihin. Estimoitavina muuttujina ovat kokonaismäärä $\hat{Y}_{dg}$ ja suhteellinen osuus $\hat{P}_{dg}$ soluittain.

Taulukko 2.1: Otoksen ja perusjoukon jakautuminen alueisiin, ryhmiin ja soluihin.

	Ryhmä 1	**Ryhmä 2**	**.....**	**Ryhmä g**
Alue 1	Solu 11	Solu 12	.....	Solu 1g
Alue 2	Solu 21	Solu 22	.....	Solu 2g
.....	.....	.....	.....	.....
Alue d	Solu d1	Solu d2	.....	Solu dg

Aineisto voidaan luokitella ristikkäin useampiin erilaisiin alueisiin ja ryhmiin. Jos kyseessä on pinta-alallisesti määriteltävä luokittelu, niin on luontevaa kutsua syntyneitä luokkia alueiksi. Jos taas luokittelu on tehty jonkin muun ominaisuuden muukaan, voidaan luokkia kutsua ryhmiksi.

2.1 Suorat estimaattorit

Suora estimaattori, jota nimitetään myös Horvitz-Thompson estimaattoriksi, on harhaton, mutta sen estimointitehokkuus on suoraan verrannollinen otoskokoon. Suoran estimaattorin teho laskee ja keskivirhe kasvaa, kun otoskoko pienaluetta kohti pienenee. Estimaattorin kaava kokonaismäärälle on

$$\hat{Y}_{dg}^{HT} = \begin{cases} \frac{N_{dg}}{n_{dg}} \sum_{k=1}^{n_{dg}} y_{dgk} & jos\ n_{dg} \geq 1 \\ 0 & jos\ n_{dg} = 0 \end{cases} \qquad (2.1)$$

(Rao & Choudry, 1995).

Kokonaismäärän varianssin estimaattori on

$$\hat{v}(\hat{Y}_{dg}^{HT}) = N_{dg}^2 \left(1 - \frac{n}{N}\right) \sum_{k=1}^{n_{dg}} \frac{(y_{dgk} - \bar{y}_{dg})^2}{n_{dg}(n_{dg} - 1)} \qquad (2.2)$$

(Pahkinen & Lehtonen, 1989). Jos johonkin soluun dg ei tule yhtään havaintoa, niin silloin suora estimaattori antaa solun estimaatiksi nollan. Erityisenä ongelmana on myös, että solujen otoskoot n_{dg} ovat satunnaismuuttujia. Tällöin varianssien estimaattoreita olisi syytä tarkastella suhde-estimaattorin kehikossa (engl. conditional inference).

Suoran estimaattorin kaava suhteelliselle osuudelle on

$$\hat{P}_{dg}^{HT} = \frac{n_{Adg}}{n_{dg}}. \qquad (2.3)$$

Osoittajassa oleva n_{Adg} on niiden otosalkioiden lukumäärä, joilla on kiinnostuksen kohteena oleva dikotominen ominaisuus, jolloin muuttuja saa arvon yksi tai nolla sen mukaan onko havaintoyksiköllä tämä ominaisuus vai ei. Varianssin estimaattori suhteelliselle osuudelle on muotoa

$$\hat{v}(\hat{P}_{dg}^{HT}) = \left(1 - \frac{n}{N}\right) \frac{\hat{P}_{dg}(1 - \hat{P}_{dg})}{n_{dg}} \qquad (2.4)$$

(Pahkinen & Lehtonen, 1989). Suorat estimaattorit ovat ns. perinteisiä estimointimenetelmiä, eivätkä siis varsinaisia pienalue-estimaattoreita. Tässä ne ovat mukana lähinnä vertailun vuoksi. Lisäksi suoria estimaattoreita käytetään hyväksi muodostettaessa komposiittiestimaattoreita, joita käsitellään tarkemmin luvussa 2.3.

2.2 Synteettiset estimaattorit

Toisin kuin suorat estimaattorit, synteettiset estimaattorit ottavat estimoinnissa tukea muilta alueilta, jotka liittyvät toisiinsa jonkin ominaisuuden perusteella (Rao, 1995). Synteettisen estimoinnin lähtökohtana on, että estimoinnissa ei hyödynnetä ainoastaan solun sisäistä informaatiota vaan otetaan tukea solun ulkopuolelta. Niiden käytön perusajatuksena on, että suuralueen osajoukkoina pienalueilla on samankaltaiset ominaisuudet kuin suuralueella. Parannuksena suoraan estimaattorin on, että vaikka johonkin soluun dg ei tule yhtään havaintoa otoksesta, niin silti synteettinen estimaattori laskee arvon myös kyseiselle solulle. Yksinkertaisin synteettinen estimaattori on *synteettinen määrä-estimaattori* (engl. synthetic count estimator), jonka kaava solun dg kokonaismäärälle on

$$\hat{Y}_{dg}^{cs} = N_{dg}\,\overline{y_g}\,, \qquad (2.5)$$

missä $\overline{y}_g$ on ryhmän g muuttujan y keskiarvo otoksessa ja N_{dg} on solun dg havaintojen lukumäärä perusjoukossa (Särndal & Hidiroglou, 1989). Tässä oletuksena on, että ryhmässä g tulosmuuttujan keskiarvot ovat samat kaikissa soluissa dg niin, että otoksen keskiarvot ovat samat kuin perusjoukon vastaavat keskiarvot. Synteettisen määrä-estimaattorin varianssin estimaattorin kaava on

$$\hat{v}(\hat{Y}_{dg}^{cs}) = N_{dg}^2\,\hat{v}(\overline{y}_g)\,. \qquad (2.6)$$

Jotta voitaisiin ottaa käyttöön *synteettinen suhde-estimaattori* (engl. synthetic-ratio estimator), täytyy olla käytettävissä tulosmuuttujan y kanssa hyvin korreloivan apumuuttujan x tiedot sekä perusjoukosta että otoksesta. Mitä parempi riippuvuus tulosmuuttujan y ja apumuuttujan x välillä on, sitä paremmin estimaattori toimii. Synteettisen suhde-estimaattorin kaava solun dg kokonaismäärälle on

$$\hat{Y}_{dg}^{RS} = \frac{\overline{y_g}}{\overline{x_g}} X_{dg} \,, \qquad (2.7)$$

missä $\overline{y}_g$ ja $\overline{x}_g$ ovat tulosmuuttujan y sekä apumuuttujan x keskiarvot otoksessa s ryhmässä g ja X_{dg} on apumuuttujan arvo perusjoukossa U solussa dg. Suhde-estimaattorin mallioletuksena on, että ryhmässä g tulosmuuttujan ja apumuuttujan suhteet ovat samanlaiset kaikissa soluissa dg niin että otoksen suhteet ovat samat kuin perusjoukon (Rao, 1995). Synteettisen suhde-estimaattorin varianssin estimaattorin kaava on

$$\hat{v}\,(\hat{Y}_{dg}^{RS}) = N_{dg}^2 \left(\frac{1}{n_g} - \frac{1}{\hat{N}_g}\right) \left(\frac{\overline{X_g}}{\overline{x_g}}\right)^2 \frac{\sum_{k=1}^{n_{dg}} (y_{dgk} - \hat{r}_g x_{dgk})^2}{(n_g - 1)} \,, \qquad (2.8)$$

missä $\hat{r} = \overline{y}_g / \overline{x}_g$ ja $\hat{N}_g = Nn_g/n$. (Särndal, 1992).

Synteettisen suhde-estimaattorin varianssi on normaalisti hyvin pieni, koska jakajina variansseissa ovat mallina käytettyjen ryhmien koot eivätkä solukoot kuten suorissa estimaattoreissa. Huonona puolena on synteettisten estimaattoreiden harhaisuus, joka saattaa kasvaa hyvinkin suureksi, jos mallioletukset eivät pidä paikkaansa (Lundström, 1988). Jos kuitenkin on niin, että mallioletukset pitävät hyvin paikkansa, on melkein mahdotonta tuottaa parempia tuloksia millään muulla estimointitekniikalla kuin nimenomaan synteettisellä estimoinnilla. Tämä voidaan todeta varsinkin keskineliövirheitä vertailemalla. Tästä huolimatta harhaisuus on niin vaikeasti hallittava haittatekijä synteettisessä estimoinnissa, että monet tilastotieteilijät ovat välttäneet synteettisten estimaattoreiden käyttöä juuri tämän ongelman takia (Särndal & Hidiroglou, 1989).

Edellä olevat kaavat soveltuvat myös suhteellisien osuuksien $\hat{P}_{dg}$ laskemiseen synteettisillä estimaattoreilla. Poikkeuksena on, että itse estimaatteja sekä variansseja laskettaessa täytyy perusjoukon solukoon neliö ottaa kaikista kaavoista pois.

2.2.1 Regressiopohjaiset estimaattorit

Yksinkertaisissa synteettisissä estimaattoreissa käytetään mallioletuksena joko tulosmuuttujan keskiarvoa tai suhdetta apumuuttujaan. Tässä yhteydessä regressiopohjaisilla estimaattoreilla tarkoitetaan lineaariseen tai logistiseen regressioon perustuvia pienalue-estimaattoreita. Niissä alueittaiset estimaatit lasketaan tämän otoksesta kehitetyn mallin avulla joko koko perusjoukolle tai vain niille jotka eivät ole mukana otoksessa.

Särndal ja Hidiroglou (1989) ovat kehittäneet yleiseen regressiotekniikkaan perustuvan pienalue-estimaattorin (engl. modified regression estimator), joka on likimain asetelmaharhaton. Sen kaava on muotoa

$$\hat{Y}_{dg}^{MRE} = \sum_{k=1}^{N_{dg}} \hat{y}_{dgk} + \frac{N_{dg}}{\sum_{k=1}^{n_{dg}} \frac{1}{\pi_k}} \sum_{k=1}^{n_{dg}} \frac{e_{dgk}}{\pi_{dgk}} . \qquad (2.9)$$

Estimaattori koostuu kahdesta osasta, joista ensimmäistä voidaan kutsua estimaattorin synteettiseksi osaksi. Jos johonkin soluun ei tule otoksesta yhtään havaintoa, niin solun estimaatin arvo muodostuu ainoastaan synteettisen termin avulla. Se muodostaa jo yksinäänkin estimaattorin, joka tarkemmin määriteltynä on

$$\hat{Y}_{dg}^{SY} = \sum_{k=1}^{N_{dg}} \hat{y}_{dgk} = \left(\sum_{k=1}^{N_{dg}} x_{dgk} \right) \cdot \hat{\beta} . \qquad (2.10)$$

Estimaattien laskemista varten tarvitaan apumuuttujan x soluittaiset summatiedot perusjoukosta ja β-kerroin estimoituna otoksesta. β-kertoimen estimoinnissa otetaan tukea muilta alueilta esim. estimoimalla jokaiselle ryhmälle G oma kerroin, jolloin otoskoot näissä ryhmissä ovat suhteellisen suuret ja malleista saadaan vakaampia. Jos estimaattoreita tarkastellaan regression näkökulmasta niin itse asiassa kaavat (2.5) ja (2.7) ovat tämän synteettisen termin erikoistapauksia.

Estimaattorin jälkimmäistä osaa voidaan kutsua korjaustermiksi, jossa residuaalit e_k saadaan otoksen havaittujen ja synteettisellä osalla ennustettujen arvojen erotuksena. Korjaustermin päämääränä on oikaista synteettisen termin synnyttämä harha estimaatista. Estimaattorin $\hat{Y}_{dg}^{MRE}$ varianssin estimaattorin kaava on muotoa

$$\hat{v}(\hat{Y}_{dg}^{MRE}) = N_{dg}^2 \left(\frac{1}{n_{dg}} - \frac{1}{N_{dg}} \right) \frac{\sum_{k=1}^{n_{dg}} (e_{dgk} - \bar{e}_{s_{dg}})^2}{n_{dg} - 1} \qquad (2.11)$$

, jossa $\bar{e}_{sdg}$ on residuaalien keskiarvo solussa dg. Jos solussa ei ole yhtään havaintoa otoksesta, solun varianssin arvo on nolla, koska synteettinen termi antaa mallin mukaisesti saman estimaatin kaikille y_k solussa dg.

Toinen regressiopohjainen pienalue-estimaattori on EBLUP (engl. empirical best linear unbiased prediction). Siinä kiinnostuksen kohteena olevien muuttujien arvoja yksittäisessä solussa pidetään ns. satunnaisvaikutuksina, joita estimoidaan (Robinson, 1991). Malli saadaan muodostettua lisäämällä usein käytettyyn "superpopulaatio" -malliin satunnainen pienalue-vaikutus a_{dg}. Regressiomallin yhtälö on silloin

$$y_{dgk} = \beta x_{dgk} + a_{dg} + e_{dgk} x_{dgk}^{½}, \quad \text{missä} \quad k=1,2,\ldots,N_{dg} \qquad (2.12)$$
$$d=1,2,\ldots,D$$
$$g=1,2,\ldots,G,$$

missä a_{dg} ja e_{dgk} ovat riippumattomia jäännösmuuttujia odotusarvolla nolla ja variansseilla $V(a_{dg}) = \sigma^2_a$ ja $V(e_{dgk}) = \sigma^2$, D on alueiden ja G ryhmien lukumäärä ja x_{dgk} on apumuuttujan x arvo havainnolle k solussa dg. Suhde $\lambda = \sigma^2_a / \sigma^2$ mittaa alueiden välistä vaihtelua verrattuna alueiden sisäiseen vaihteluun (Rao, 1995). Kysymyksessä on ns. aluehierarkkinen malli.

EBLUP - estimaattori on mallin (2.12) mukaisesti muotoa

$$\hat{Y}_{dg}^{EBLUP} = \sum_{k \in s_{dg}} y_{dgk} + \sum_{k \notin s_{dg}} y^*_{dgk} \qquad (2.13)$$

,missä y_{dgk}* on paras ennuste otokseen kuulumattomille havainnoille ja se on muotoa

$$y^*_{dgk} = \beta x_{dgk} + \hat{a}_{dg}, \qquad missä\ k \notin s_{dg}$$

,missä parametrit $\hat{\beta}$ ja $\hat{a}$ saadaan yhtälöistä

$$\beta = \left[\sum_{dg}\sum_{k \in s_{dg}} y_{dgk} - \sum_{dg} \frac{\gamma_{dg}}{\hat{\eta}_{dg}} n_{dg} \left(\sum_{k \in s_{dg}} r_{dgk}\right)\right] \left[\sum_{dg}\sum_{k \in s_{dg}} x_{dgk} - \sum_{dg} \frac{\gamma_{dg}}{\hat{\eta}_{dg}} n^2_{dg}\right]^{-1} \qquad (2.14)$$

$$\hat{a}_{dg} = \frac{\gamma_{dg}}{\hat{\eta}_{dg}} \sum_{k \in s_{dg}} (r_{dgk} - \beta) , \qquad (2.15)$$

joissa esiintyvät parametrit saadaan

$$r_{dgk} = \frac{y_{dgk}}{x_{dgk}} \qquad ja \qquad \acute{\eta}_{dg} = \sum_{k \epsilon s_{dg}} x_{dgk}^{-1}$$

$$\gamma_{dg} = \frac{\hat{\sigma}_a^2}{\hat{\sigma}_a^2 + \frac{\hat{\sigma}}{\acute{\eta}_{dg}}} = \frac{\hat{\lambda}}{\hat{\lambda} + \acute{\eta}_{dg}^{-1}}, \qquad jossa\ \hat{\lambda} = \frac{\hat{\sigma}_a^2}{\hat{\sigma}^2}$$

(Rao, 1995).

Alueiden väliset ja sisäiset varianssit saadaan ns. Hendersonin metodi 3:n avulla kaavoilla

$$\hat{\sigma}^2 = (n - I - 1)^{-1} \sum_{dg} \sum_{k \epsilon s_{dg}} \hat{e}_{dgk}^2 \qquad (2.16)$$

$$\hat{\sigma}_a^2 = \frac{1}{n^*}\left[\sum_{dg} \sum_{k \epsilon s_{dg}} \hat{u}_{dgk}^2 - (n - 1)\hat{\sigma}^2 \right] \qquad (2.17)$$

$$n^* = \sum_{dg} \acute{\eta}_{dg} - \frac{\sum_{dg} n_{dg}^2}{\sum_{dg} \sum_{k} x_{dgk}},$$

missä $\Sigma\Sigma\ \hat{u}^2_{dgk}$ on residuaalien neliösumma. Residuaalit u_{dgk} saadaan origon kautta kulkevasta painotetusta regressiomallista $y_{dgk} = x_{dgk}$ painoilla x_{dgk}^{-1}. Residuaalien neliösumma on $\Sigma\Sigma\ \hat{e}^2_{dgk}$, jossa residuaalit e_{dgk} saadaan origon kautta kulkevasta painotetusta regressiomallista $y_{dgk} - \bar{y}_{dgw} = x_{dgk} - \bar{x}_{dgw}$ painoilla x_{dgk}^{-1} ja $\bar{y}_{dgw}$ ja $\bar{x}_{dgw}$ ovat painotettuja pienaluekeskiarvoja painoilla x_{dgk}^{-1} (Stukel, 1991). Jos varianssin $\hat{\sigma}^2_a$ arvo on negatiivinen, niin sille annetaan arvoksi nolla (Rao,1995).

EBLUP:n toimivuutta on tarkasteltu monissa vertailututkimuksissa, kuten esim. Raon ja Choudryn (1995), Ghoshin ja Raon (1994) ja Singhin, Mantelin ja Thomasin (1994) artikkeleissa. Useimmissa tapauksissa EBLUP on todettu yhdeksi parhaimmista pienalue-estimaattoreista varsinkin keskineliövirheitä vertailtaessa. Sen sijaan estimaattien harhaisuutta tarkasteltaessa seuraavassa luvussa esiteltävät komposiittiestimaattorit ovat osoittautuneet ainakin yhtä varteenotettavaksi vaihtoehdoksi.

2.3 Komposiittiestimaattorit

Käytännössä, pienalueiden erilaisia ilmiöitä tutkittaessa, synteettisten estimaattoreiden oletukset eivät useinkaan ole kovin relevantteja. Tämän ongelman minimoimiseksi on kehitetty estimaattoreita, jotka pienentävät synteettisen estimoinnin harhaominaisuutta ja samalla kontrolloivat suoran estimoinnin keskivirheiden suuruutta. Näitä yhdistelmä-estimaattoreita eli ns. komposiittiestimaattoreita muodostetaan ottamalla painotettu keskiarvo kahdesta eri menetelmään perustuvasta pienalue-estimaattorista (Ghosh & Rao, 1994). Yksinkertaisin ja yksiselitteisin tapa on määritellä painot sen mukaan kuinka havainnot sijoittuvat soluihin otoksessa ja perusjoukossa. Ghoshin (1994) artikkelin esimerkissä on käytetty *otoskoolla ehdollistettu estimaattoria* (engl. sample-size-dependent estimator), jonka kaava solun dg kokonaismäärälle on muotoa

$$\hat{Y}_{dg}^{SD} = \begin{cases} \hat{Y}_{dg}^{REG} = N_{dg}\left[\bar{y}_{dg} + \left(\frac{\bar{y}}{\bar{x}}\right)(\bar{X}_{dg} - \bar{x}_{dg})\right], & jos\ w_{dg} \geq W_{dg} \\ \frac{w_{dg}}{W_{dg}}\hat{Y}_{dg}^{REG} + \left(1 - \frac{w_{dg}}{W_{dg}}\right)\hat{Y}_{dg}^{RS}, & jos\ w_{dg} < W_{dg}, \end{cases} \qquad (2.18)$$

missä $w_{dg} = n_{dg}/n$ ja $W_{dg} = N_{dg}/N$. Kaavan ensimmäinen osa $\hat{Y}_{dg}^{\mathrm{REG}}$ on approksimatiivisesti harhaton regressiotyyppinen estimaattori (Rao, 1995), jossa malliryhmäkohtainen keskiarvo on regressioyhtälön vakio, tulosmuuttujan ja apumuuttujan

keskiarvojen suhde on otoksesta "estimoitu" β-kerroin ja selittävä muuttuja on perusjoukon ja otoksen tulosmuuttujan keskiarvojen erotus. Jos otoksessa pienalueen otoskoko suhteessa otoskokoon on suurempi tai yhtä suuri kuin perusjoukon pienalueen koko suhteessa perusjoukkoon, niin kyseisen pienalueen arvo estimoidaan ainoastaan tällä regressiotyyppisellä estimaattorilla. Jos otoksen suhteellinen osuus on pienempi kuin perusjoukon suhteellinen osuus, estimointi suoritetaan regressioestimaattorin ja synteettisen suhde-estimaattorin (2.7) kombinaationa painoilla (w_{dg}/W_{dg}) ja $(1 - w_{dg}/W_{dg})$.

Otoskoolla ehdollistetun estimaattorin varianssin estimaattorin kaava on muotoa

$$\hat{v}(\hat{Y}_{dg}^{SD}) = \begin{cases} N_{dg}^2 \left(1 - \frac{n}{N}\right)\left(\frac{1}{n}\right) \frac{\sum_{k=1}^{n_{dg}} \left[(y_{dgk} - \bar{y}_{dg}) - \frac{\bar{Y}}{\bar{X}}(x_{dgk} - \bar{X}_{dg})\right]^2}{n-2}, & jos\ w_{dg} \geq W_{dg} \\ \left(\frac{w_{dg}}{W_{dg}}\right)^2 \hat{v}(\hat{Y}_{dg}^{REG}) + \left(1 - \frac{w_{dg}}{W_{dg}}\right)^2 \hat{v}(\hat{Y}_{dg}^{RS}), & jos\ w_{dg} < W_{dg} \end{cases} \quad (2.19)$$

(Lehtonen, 1995). Kun estimaatti lasketaan kahden eri estimaattorin yhdistelmästä, sen varianssin estimaatti saadaan samalla periaatteella kuin itse estimaattori. Erona on, että painot korotetaan nyt toiseen ja itse estimaattoreiden asemesta käytetäänkin estimaattoreiden varianssien estimaattoreita.

Singh ja Mian (1995) esittävät artikkelissaan hieman toisenlaisen tavan painojen määrittämiseksi. Aluekohtaiset painot λ_{dg} määritetään

$$\lambda_{dg} = \begin{cases} 1 & , jos\ n_{dg} \geq n_o \\ \frac{n_{dg}}{n_o} & , jos\ n_{dg} < n_o, \end{cases} \quad (2.20)$$

missä n_{dg} on havaittu otoskoko ja n_o on kyseiselle pienalueelle toivottu otoskoko. Haluttu otoskoko on määritelty niin suureksi, että HT-estimaatti on riittävän tarkka

eikä tarvitse lainata tukea muilta alueilta. Tämä painotustapa perustuu siihen, että toisena estimaattorivaihtoehtona on suora estimaattori ja toisena jokin synteettinen estimaattori. Tällä painotustekniikalla voidaan muodostaa lukuisia uusia komposiittiestimaattoreita. Tästä esimerkkinä on suoran estimaattorin (2.1) ja mallipohjaisen estimaattorin (2.8) yhdistelmä, joka on muotoa

$$\hat{Y}_{dg}^{SSD^*} = \lambda_{dg}\ \hat{Y}_{dg}^{HT} + (1 - \lambda_{dg})\ \hat{Y}_{dg}^{MRE}. \qquad (2.21)$$

Mitä lähempänä havaittu otoskoko n_{dg} on toivottua otoskokoa n_o, sitä suurempi osa estimaatista muodostuu suoran estimaattorin avulla. Jos havaittu otoskoko on suurempi kuin n_o, koko estimaatti saadaan suorasta estimaattorista. Tämän komposiittiestimaattorin varianssi on

$$\hat{V}(\hat{Y}_{dg}^{SSD^*}) = \lambda_{dg}^2\ \hat{V}(\hat{Y}_{dg}^{HT}) + (1 - \lambda_{dg})^2\ \hat{V}(\hat{Y}_{dg}^{MRE}). \qquad (2.22)$$

Erityisesti sosiaalisia taustoja selvittelevissä otantatutkimuksissa, varianssin suuruusluokan hallitsemiseksi suositellaan käytettäväksi pienalueiden toivottuna otoskokona n_o joko 200 tai sitä suurempi arvoja (Singh, 1995).

2.4 Pienalue-estimaattoreiden käyttökokemuksia

Ainakin Kanadan ja USA:n kansallisissa tilastovirastoissa käytetään pienalue-estimaattoreita joidenkin oleellisimpien pienaluemittaisten tietojen tuottamisessa. Myös naapurimaissamme Ruotsissa ja Norjassa on aktiivisesti kehitelty ja otettu käyttöön joitakin sovelluksia. Suurimmaksi osaksi maiden kansalliset tilastovirastot käyttävät pienalue-estimaattoreita halutun tiedon tuottamiseksi maansa pienimmistä kunnista tai kaupungeista.

Esimerkiksi Ruotsissa on 284 kuntaa. Osa kunnista ovat pinta-alaltaan hyvinkin suuria, mutta asukasluvultaan suhteellisen pieniä. Valtakunnallisissa tiedusteluissa asukasmäärältään pienistä kunnista tulee liian vähän havaintoja otokseen, jotta tiedon estimointi kyseisistä kunnista olisi perinteisin menetelmin mahdollista. Näiden pienten kuntien päättäjillä ja viranomaisilla on kuitenkin yhtä suuri tarve saada ajankohtaista tietoa omasta kunnastaan kuin suuremmilla kunnilla ja kaupungeilla (Rosén, 1991).

Pienalue-estimointia käsittelevissä artikkeleissa on hyvin vähän todellisia esimerkkejä pienalue-estimaattoreista ja siitä miten hyvin ne käytännössä toimivat. Enimmäkseen tutkimuksissa on käytetty keinotekoisesti tuotettuja aineistoja, joiden lähtökohtana on ollut jokin todellinen tilanne. Simulointitutkimuksissa keinotekoisesta aineistosta on poimittu otoksia, joista jokaisesta on laskettu vertailtavien estimaattien arvot. Tämän jälkeen estimaattien avulla on laskettu erilaisten harhaa ja keskivirhettä kuvaavien vertailumittareiden arvoja, joiden perusteella on tulkittu pienalue-estimaattoreiden paremmuutta. Keskivirheiden ja luottamusvälien tarkastelu on jätetty vähemmälle huomiolle. Seuraavassa kappaleessa esitellään keskeisimmät tulokset Singhin, Mantelin ja Thomaksen Kanadan tilastovirastossa tekemästä simulointitutkimuksesta, jossa on hyvin monipuolisesti mukana kaikki tärkeimmät pienalue-estimaattorityypit.

2.4.1 Pienalue-estimaattoreiden vertailua simulointikokeen avulla

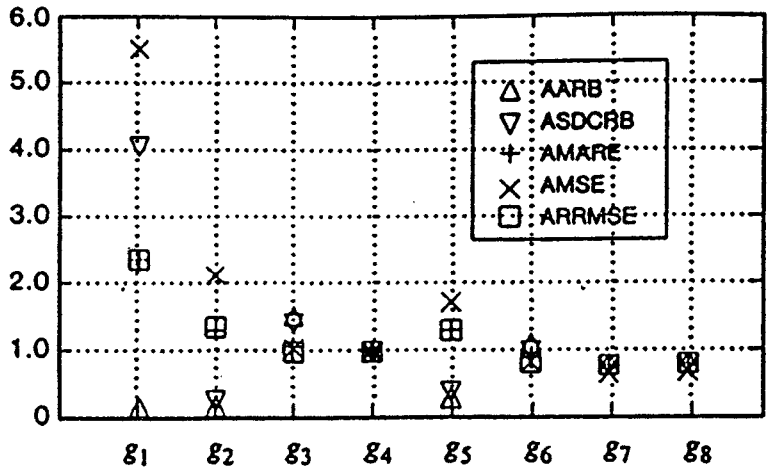
Singh, Mantel ja Thomas (1994) ovat vertailleet artikkelissaan erilaisten pienalue-estimaattorin paremmuutta simulointikokeella harhaa ja keskineliövirhettä kuvaavien vertailumittarin avulla. Tässä artikkelissa on erityisesti tutkittu miten aikasarjojen mukaanotto estimointiin vaikuttaa tuloksiin. Aineistona on käytetty Kanadan tilastoviraston maatiloja koskevaa kyselytutkimusta kuudelta eri ajankohdalta (kesäkuu 1988, tammikuu 1989, ..., tammikuu 1991) väestölaskentavuoden 1986 kokonaisaineistolla täydennettynä. Tulosmuuttujana on karjan kokonaismäärä kullakin satoalueella (engl. crop district) ja apumuuttujana vuoden 1986 suhdetehosteiset estimaatit karjan kokonaismääristä pienalueilla. Kokeiluaineistoksi on rajattu Quebecin provinssi, joka muodostuu 12 hyvin erilaisesta satoalueesta.

Monte Carlo -simulointia varten rakennettiin pseudo-populaatio, jonka kooksi saatiin 10 362 maatilaa. Tästä pseudo-populaatiosta suoritettiin 30 000 simulointia, joissa kussakin otos poimittiin ositetulla yksinkertaisella satunnaisotannalla palauttamatta riippumattomasti jokaisesta kuudesta ajankohdasta. Sen jälkeen nämä 30 000 simulointia jaettiin 15 000 simulointipariin, missä jokainen pari vastaavat toteutuneiden otoskokojen vektoreita 12 pienalueella kussakin ositteessa.

Taulukko 2.2 : Vertailtavat pienalue-estimaattorit

g_1 = Suora estimaattori	g_5 = Otoskoolla ehdollistettu estimaattori
g_2 = Jälkiositettu estimaattori	g_6 = EBLUP-I, β-kerroin vaihtuu ajan mukana, a-kerroin riippumaton ajasta
g_3 = Synteettinen estimaattori	g_7 = EBLUP-II, β-kerroin on vakio, a-kerroin vaihtuu ajan mukana
g_4 = EBLUP	g_8 = EBLUP-III, β- ja a- kertoimet vaihtuvat ajan mukana

Vertailtavia pienalue-estimaattoreita on siis kahdeksan, joista kolmessa viimeisessä aikasarjat on otettu estimointiin mukaan. Vertailumittareita on kaikkiaan viisi, joista kahta tarkastellaan tarkemmin. Kuvassa 2.1 on kaikkien viiden vertailumittarin keskimääräiset arvot suhteessa estimaattorin g_4 (EBLUP) saamiin arvoihin. Kolmiolla merkityt mittarit kuvaavat harhaisuutta ja loput keskineliövirheen määrää eri tavoin määriteltynä.



Kuva 2.1: Vertailumittareiden arvot eri estimaattoreille suhteessa estimaattoriin g_4. (Huom. Mittarin ASDCRB arvo estimaatille $g_1(=18.98)$ ei näy kuvassa.)

Vertailumittareiden arvoissa on selvästi eroja estimaattoreiden välillä. Suorat estimaattorit g_1 ja g_2 sekä otoskoolla ehdollistettu estimaattori g_5 saavat pienimmät eli parhaat harhaa mittaavat arvot, mutta muiden mittareiden osalta huonoimmat (varsinkin g_1). Estimaattoreiden g_2 ja g_5 osalta vertailumittareiden arvot ovat jokseenkin samanlaiset. Synteettinen estimaattori g_3 saa lähes samanlaiset tulokset kuin EBLUP eli g_4. Aikasarjoihin perustuvilla estimaattoreilla g_6, g_7 ja g_8 on tasaisen hyvät tulokset kaikilla vertailumittareilla. Tosin yhdelläkään näistä estimaattoreista ei päästä niin hyviin tuloksiin harhan osalta, kuin aluksi mainituilla suorilla estimaattoreilla ja painotulla estimaattorilla g_5. Selkeimmin muista estimaattoreista erottuu suora estimaattori g_1, jonka tulokset harhaa lukuun ottamatta ovat melko huonot.

Kuvassa 2.2 on esitetty tarkemmin suhteellisen keskineliövirheen (engl. relative root mean squared error) määrän keskiarvoja ja kuvassa 2.3 absoluuttisen suhteellisen harhan (engl. absolute relative bias) keskiarvoja estimaattoreilla kolmessa eri kokoluokassa: suuret, keskikokoiset ja pienet alueet. Aluejako on tehty, koska estimaattien suhteelliset virheet ovat oletettavasti suuremmat pienillä alueilla kuin suurilla. Tällöin kuvioiden osoittamat tulokset eivät ole ristiriidassa tämän oletuksen kanssa. Suhteellisen keskineliövirheen kaava alueelle k on muotoa

$$RRMSE_k = \frac{\sqrt{MSE_k}}{Y_k}, \qquad (2.23)$$

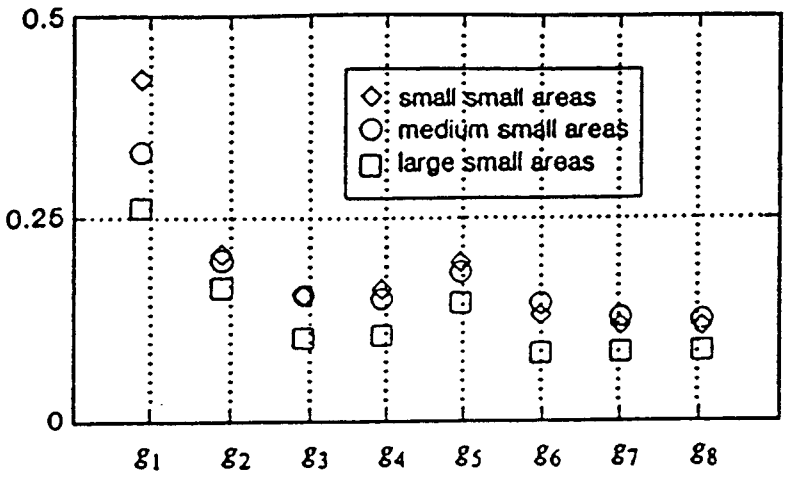
jossa keskineliövirhe (*MSE*) saadaan

$$MSE_k = \frac{1}{m} \sum_i \sum_j (\hat{Y}_{ijk} - Y_k)^2, \qquad (2.24)$$

missä m tarkoittaa simulointien lukumäärää, $\hat{Y}_{ijk}$ estimaattia ja Y_k pienalueen todellista arvoa. Keskimääräinen *ARRMSE* saadaan pienalueiden keskiarvona. Absoluuttinen suhteellinen harha alueelle k saadaan kaavasta

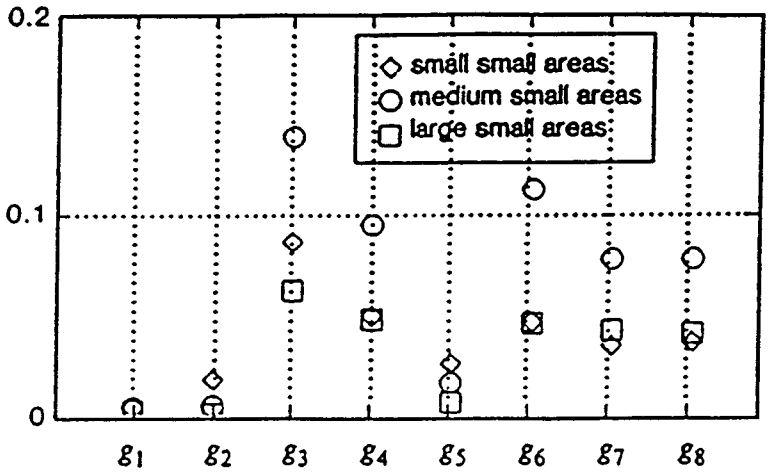
$$ARB_k = \left| \frac{1}{m} \frac{\sum_i \sum_j (\hat{Y}_{ijk} - Y_k)}{Y_k} \right| \qquad (2.25)$$

ja *AARB* samalla tavalla pienalueiden keskiarvona.



Kuva 2.2 : Keskimääräiset suhteelliset keskineliövirheet (ARRMSE) kolmessa kokoluokassa.

Suhteellisten keskivirheiden osalta aikasarjoihin perustuvat estimaattorit g_7 ja g_8 näyttäisivät toimivan parhaiten. Tosin estimaattoreiden g_3 ja g_4 eikä myöskään estimaattoreiden g_2 ja g_5 tuloksia voida pitää ratkaisevasti huonompina. Sen sijaan suoran estimaattorin g_1 suhteelliset keskivirheet ovat selkeästi muita suuremmat. Kaikilla estimaattoreilla suurimman kokoluokan suhteelliset keskineliövirheet ovat kahden muun kokoluokan suhteellisia keskineliövirheitä pienemmät.



Kuva 2.3: Keskimääräiset absoluuttiset suhteelliset harhat (ARB) kolmessa kokoluokassa.

Suhteellisen harhan suhteen sekä estimaattoreiden välillä että kokoluokkien välillä enemmän hajontaa kuin suhteellista keskineliövirhettä mitattaessa. Suora estimaattori g_1 on täysin ja jälkiositettu estimaattori g_2 melkein harhaton. Kolmanneksi harhattomin on g_5. Muilla estimaattoreilla harhaa on selvästi enemmän niin, että keskimmäisen kokoluokan estimaateilla harhaa esiintyy eniten.

Tutkimuksessa on päädytty siihen tulokseen, että aikasarjaa hyödyntävät estimaattorit tuottavat parempia tuloksia kuin sellaiset, joissa aikasarjoja ei ole käytetty. Huomattavaa kuitenkin on, että harhan määrä aikasarjoja hyödyntävillä estimaattoreilla oli selvästi suurempi kuin esimerkiksi komposiittiestimaattorilla g_5. Toisaalta komposiittiestimaattorilla on suuremmat keskineliövirheet kuin aikasarjaa hyödyntävillä estimaattoreilla. Tämä on osoitus siitä yleisesti todetusta ilmiöstä, että estimointi on tasapainoilua harhan ja keskivirheen välillä. Jos keskivirhe saadaan pienemmäksi, niin harhan määrä lähes poikkeuksetta kasvaa ja toisinpäin. Estimaattien harhattomuus kohtuullisilla keskivirheillä olisi optimitilanne, jota kohti estimoinnissa yleensä pyritään.

3 Tutkimusongelma

Tutkimuksen tavoitteena on löytää ja kehittää estimointimenetelmän prototyyppi, jonka avulla voitaisiin tuottaa neljännesvuosittain sukupuolittain ja ikäryhmittäin alueittaisia tilastolukuja Helsingin työttömyys- ja työllisyystilanteesta työvoimatiedustelun neljännesvuosiaineiston ja rekisteripohjaisen työssäkäyntitilaston avulla. Päämääränä on siis tuottaa tulevaisuudessa alueellista tietoa Helsingin työvoimasta nopeammassa aikatahdissa kuin se tähän asti on tehty, niin että entisen yhden kerran sijasta aluetietoja tuotettaisiin neljä kertaa vuodessa. Tällöin alueittaiset työvoimaluvut olisivat käytössä ainakin kolme tai neljä kuukautta aikaisemmin kuin nykytilanteessa.

Ongelmana on, että Tilastokeskuksen työvoimatiedustelu on alunperin suunniteltu antamaan tietoa suuremmilta alueilta kuin yksi kaupunki tai sen osa-alueet. Näin ollen otoskoko yhtä Helsingin aluetta kohti saattaa tulla hyvinkin pieneksi, kun halutaan esimerkiksi vielä maantieteellisten alueiden sisällä jakaa ihmiset sukupuolen tai iän mukaan ryhmiin. Suorat estimaattorit, jotka käyttävät hyödyksi vain otokseen tulleiden havaintojen tietoja kullakin alueella, tuottavat liian epäluotettavia tuloksia. Estimaattien keskivirheet kasvavat suuriksi, koska otoskoko yhtä pienaluetta kohti on niin pieni.

Ongelman ratkaisuun on olemassa lähinnä kaksi vaihtoehtoa. Ensimmäisenä vaihtoehtona on otoskoon kasvattaminen niin suureksi, että kaikille osa-alueille tulee riittävästi havaintoja ja keskivirheet pienenevät kohtuullisiksi. Koska otoskokoon ei nyt lyhyellä tähtäimellä voida vaikuttaa ja kustannuksia ei haluta kasvattaa, jää ainoaksi vaihtoehdoksi pienalue-estimaattoreiden käyttö.

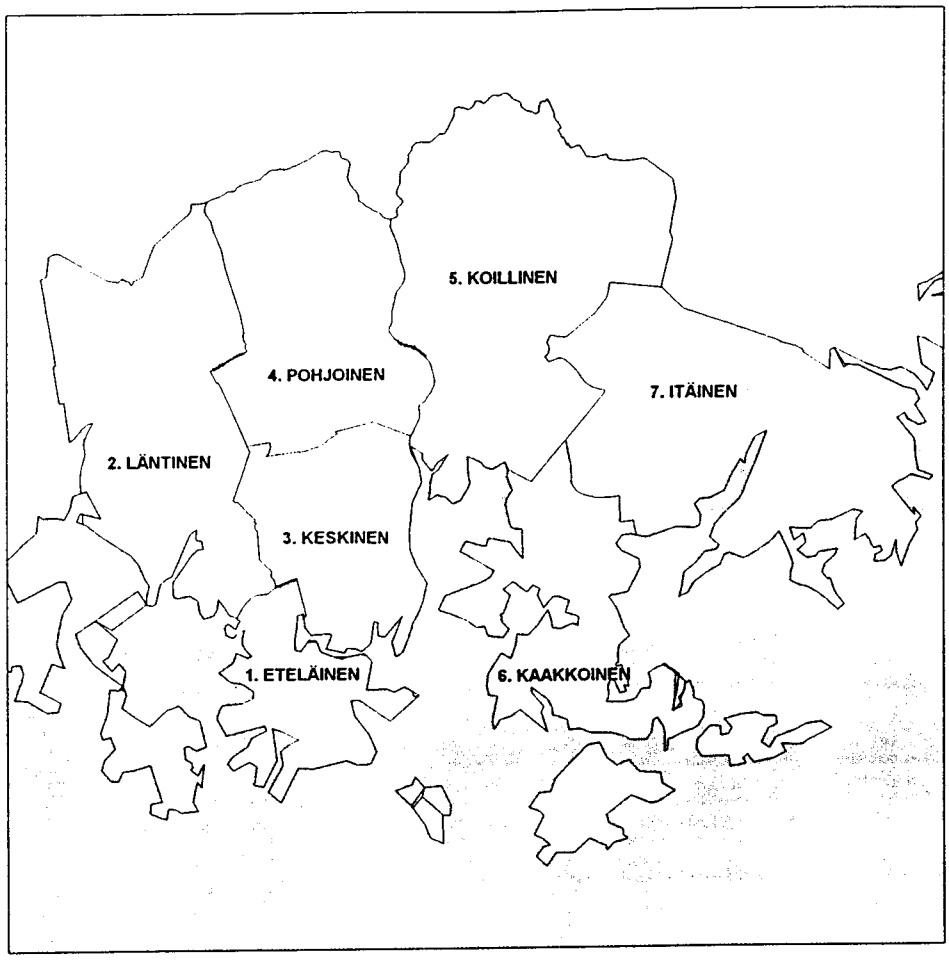
Parin viime vuosikymmenen aikana pienalue-estimaattoreihin kohdistuva tutkimustyö sekä teorian että käytännön puolella on ollut hyvin vilkasta. Varsinkin Kanada, Ruotsi, Yhdysvallat ja Australia ovat tehneet pioneerityötä nimenomaan otospohjaisten pienaluetilastojen laatimisessa. Pelkästään pienalue-estimointia käsittelevää kirjallisuutta on olemassa hyvin vähän, mutta erilaisia artikkeleita on melko runsaasti. Pienalue-estimaattoreiden tarjonta on hyvin kirjavaa ja niiden nimet ja teoreettiset merkintätavat eivät ole yleisesti vakiintuneita. Tässä tutkimuksessa on tuotu esille niitä pienalue-estimaattoreita, jotka ovat käyttökelpoisia juuri tämän tutkimuksen kannalta.

3.1 Tulosmuuttujat

Estimoitavina tulosmuuttujina ovat alueittaiset työttömien ja työllisten kokonaismäärät sekä työttömyysaste. Muuttujien kannalta ongelmallista on, että yksilötasoisesti tarkasteltuna työttömyys ja työllisyys ovat yksi luokitteluasteikollinen dikotominen muuttuja, joka saa arvon yksi kun yksilö on työtön ja arvon nolla kun yksilö on työllinen. Solutasoisina työttömien ja työllisten kokonaismäärät ovat kuitenkin jatkuvia muuttujia ja työttömyysaste on työttömien prosenttiosuus työttömistä ja työllisistä yhteensä. Tulosmuuttujat ovat määritelty tarkemmin kappaleessa 4.1.1.

3.2 Solujako

Helsinki jaetaan maantieteellisesti seitsemään suurpiiriin, jotka on nimetty lähinnä ilmansuuntien mukaan: (1) Eteläinen , (2) Läntinen, (3) Keskinen, (4) Pohjoinen, (5) Koillinen, (6) Kaakkoinen ja (7) Itäinen suurpiiri. Suurpiirit jaetaan 33 peruspiiriin ja edelleen 117 osa-alueeseen. Tässä tutkimuksessa käytetään ainoastaan suurpiirijakoa. Lisäksi tarkastellaan alustavasti, onko mahdollista tuottaa työvoimalukuja myös peruspiireille. Kuvassa 3.1 on kartta, jossa näkyvät suurpiirien rajat.



Kuva 3.1: Helsingin suurpiirit

Suurpiirit ovat pinta-alaltaan sekä asukasluvultaan hyvin eri kokoisia. Tutkimusaineistojen ajankohtana eli vuoden 1994 alussa eniten asukkaita oli Läntisessä suurpiirissä (93 000) ja vähiten Kaakkoisessa suurpiirissä (38 000) (Helsigin kaupungin tietokeskus, 1994). Helsingin väestönkasvu on kiihtynyt vuosi vuodelta jo usean vuoden ajan. Väestönkasvun merkittävin osatekijä on ollut ns. työmarkkinamuutto, jonka kehitys riippuu ennen kaikkea työvoiman kysynnästä Helsingin seudulla, muualla Suomessa ja ulkomailla. Sen sijaan Helsingin luonnollinen väestönlisäys on supistunut viime vuosina (Helsingin kaupungin tietokeksuksen tilastoja, 1996:14). Tämän vuoden alussa väkiluvultaan suurin oli Eteläinen suurpiiri (96 489). Vähiten asukkaita oli Pohjoisessa suurpiirissä (39 000) (Helsingin kaupungin tietokeskuksen tilastoja, 1997:4). Tulevaisuudessa (ennuste vuodelle 2005) väkiluvultaan voimakkaimmin näyttäisi kasvavan Itäinen suurpiiri. Myös Kaakkoisen, Koillisen ja Läntisen suurpiirien ennustetaan kasvavan. Pohjoisen ja Keskisen

suurpiirien asukasluvun sen sijaan ennustetaan alenevan (Helsingin kaupungin tietokeskuksen tilastoja, 1996:14). Maapinta-aloja tarkasteltaessa suuruusjärjestys on hieman toisenlainen. Läntinen suurpiiri on pinta-alaltaan kolmanneksi suurin (30,4 km^2) ja Kaakkoinen suurpiiri neljänneksi suurin (26.3 km^2) (Helsingin kaupungin tietokeskus, 1996).

Maantieteellisten alueiden sisällä tilastoluvut halutaan myös sukupuolittain ja kolmen ikäryhmän mukaan: nuoret (15-29-vuotiaat), keski-ikäiset (30-49-vuotiaat) ja varttuneet (50-74-vuotiaat). Tällöin pienalueita eli soluja muodostuu kaikenkaikkiaan (7x3x2) 42 kappaletta. Naisia Helsingissä on aina ollut enemmän kuin miehiä. Tosin sotien jälkeen naisten enemmyys on pienentynyt vuosi vuodelta lähes poikkeuksetta. Helsingin väestön ikärakenteelle on ominaista, että ikärakenne muuttuu sitä enemmän nykyisestä, mitä hitaammin väestö kasvaa. Ennusteiden mukaan Helsingin väestön ennakoidaan kasvavan tuntuvasti, jolloin myös ikärakenteen kehitys on varsin tasapainoista (Helsingin kaupungin tietokeskuksen tilastoja, 1996:14). Nämä väestön alueelliset muutokset vaikuttavat omalta osaltaan otosaineiston rakenteeseen ja ne tulisi mallituksessa huomioda. Taulukossa 3.1 on kuvattuna tavoitteena olevan tilastotaulukon muoto.

Taulukko 3.1 : Tutkimusaineiston ristiinluokittelu soluihin. (Työttömät joulukuussa 1993)

SUURPIIRIT	IKÄRYHMÄT / SP					
	15-29 -vuotiaat		30-49 -vuotiaat		50-74 -vuotiaat	
	Miehet	Naiset	Miehet	Naiset	Miehet	Naiset
1. ETELÄINEN	1 329	1 133	1 850	1 343	516	457
2. LÄNTINEN	1 395	1 057	2 015	1 369	508	445
3. KESKINEN	1 396	1 059	2 452	1 326	667	446
4. POHJOINEN	440	296	743	487	276	209
5. KOILLINEN	1 289	847	2 184	1 408	567	397
6. KAAKKOINEN	567	397	839	615	237	229
7. ITÄINEN	1 595	1 018	2 150	1 500	700	628

4 Aineistot

Estimoinnin perustana tutkimuksessa käytetään kahta Tilastokeskuksen toimittamaa aineistoa. Otosaineistona on työvoimatiedustelun yksilötasoinen neljännesvuosiaineisto Helsingin osalta. Rekisteripohjaisena aineistona on vuoden 1993 työssäkäyntitilaston erillistulostus, joka käsittää työvoimaa koskevat luvut summatietoina alueittain, sukupuolittain ja ikäryhmittäin. Kappaleissa 4.1 ja 4.2 tarkastellaan aineistoja yleisellä tasolla ja kappaleessa 4.3 otosaineistoa hiukan tarkemmin.

4.1 Työvoimatiedustelu (TYTI)

Tilastokeskus on tehnyt työvoimatiedustelua (eng. Labour Force Survey) vuodesta 1959 lähtien. Syy tiedustelun aloittamiseen oli, että kymmenen vuoden välein tehdyt väestölaskennat olivat aivan liian hidas instrumentti työmarkkinailmiöiden seuraamiseen. Asiaan vaikutti myös, että haluttiin saada tietoja erityisesti maaseudun työtilanteesta. Tätä ennen näiden tietojen hankkiminen oli ollut vaikeaa, koska maaseudun oloista oli vaikeaa saada tietoja työnantajille kohdistetuilla kyselyillä. Suomen työvoimatiedustelun taustalla ollut selvitys- ja tutkimustyö tehtiinkin Helsingin yliopiston maatalous-metsätieteellisessä tiedekunnassa, sen metsätaloustieteellisellä laitoksella (Tilastokeskus, 1993).

Työvoimatiedustelun otantakehikko perustuu väestön keskusrekisteristä kahdesti vuodessa poimittuun otokseen. Poimintakohteena on 15 - 74 -vuotias väestö. Otanta suoritetaan yksinkertaisella satunnaisotannalla palauttamatta niin, että noin joka 300.

henkilö tulee mukaan otokseen. Otoksen tasainen jakautuminen ympäri maata saadaan taattua sillä, että ennen poimintaa perusjoukko on järjestetty asuinkunnan ja henkilötunnuksen mukaan järjestykseen. Työvoimatiedustelu on ns. rotatoiva paneelitutkimus, jossa puolivuosittainen otos jaetaan kahteen rotaatioryhmään, joissa molemmissa on kolme alaryhmää. Yksi rotaatioryhmä on mukana tiedustelussa puolentoista vuoden ajan, kuitenkin niin, että kun jokainen rotaatioryhmän kolmesta alaryhmästä on ollut vuorotellen mukana tiedustelussa kolme kertaa, niin yksi kolmen kuukauden jakso jää välistä pois. Tiedustelu suoritetaan kerran kuukaudessa, jolloin yhden kuukauden tiedustelussa on mukana yksi alaryhmä viidestä eri rotaatioryhmästä. Otos vaihtuu asteittain siten, että kolmena peräkkäisenä kuukautena vastaamisvuorossa ovat eri henkilöt. Peräkkäisinä kolmen kuukauden jaksoina 3/5 vastaajista ovat samoja ja vuoden kuluttua otosten päällekkäisyys on 2/5. Otoksien asteittaisella vaihtamisella on tavoitteena muutostietojen tarkkuuden parantaminen.

Vuodesta 1985 saakka työvoimatiedustelu on suoritettu suureksi osaksi Tilastokeskuksen haastattelijaverkon avulla puhelinhaastatteluin joko kodista tai työpaikalta. Kuukaudessa haastateltavia on noin 12 000. Viiteajanjaksona, jonka mukaan toiminnan laatu määritellään, käytetään kansainvälisten suositusten mukaan yhtä viikkoa. Tämä tutkimusviikko on yleensä kuukauden 15. päivän sisältämä viikko.

Työvoimatiedustelun tehtävänä on kuvata työn kysyntää ja tarjontaa eri toimialoilla ja koko kansantalouden tasolla kansainvälisesti vertailtavalla tavalla. Tilastokeskuksen työvoimatiedustelusta lasketaan myös Suomen virallinen työttömyysaste, joka on kansainvälisesti vertailukelpoinen (Tilastokeskus, 1997). Otannasta aiheutuva satunnaisvaihtelu, vastauskato ja erilaiset mittausvirheet ovat työvoimatiedustelun virhelähteitä (Tilastokeskus, 1994).

4.1.1 Työvoimatiedustelun määritelmät

Työvoimatiedustelussa 15 - 74 -vuotias väestö jaetaan saatujen vastausten perusteella *työvoimaan* ja *työvoimaan kuulumattomaan väestöön* ja työvoima jaetaan edelleen *työllisiin* ja *työttömiin*. *Työttömyysaste* on työttömien prosenttiosuus työvoimasta.

Työvoimatiedustelussa *työlliseksi* luokitellaan henkilö, joka tutkimusviikkona teki yhtenäkin päivänä työtä palkkaa tai voittoa saadakseen tai työskenteli avustavana perheenjäsenenä vähintään kolmanneksen alan normaalista työajasta tai oli työpaikastaan tilapäisesti poissa. *Työttömäksi* luokitellaan henkilö, joka oli koko tutkimusviikon työtä vailla ja siihen käytettävissä sekä etsi työtä tai odotti sovitun työn alkamista tai oli työpaikastaan pakkolomalla. Työtön on myös henkilö, joka oli osan tutkimusviikkoa työttömänä ja muun osan viikosta muualla kuin työssä esim. opiskelijana. *Työvoimaan kuulumaton* on henkilö, joka ei tutkimusviikkona kuulunut työvoimaan. Työvoimaan kuulumattomia ovat esimerkiksi opiskelijat, asevelvollisuutta suorittavat, omaa kotitaloutta hoitavat, eläkkeellä olevat tai työkyvyttömät (Tilastokeskus, 1993).

Työvoimatutkimuksessa työttömän rajauksen perusteena on kansainvälinen, International Labour Officen (ILO) puitteissa valmisteltu ja hyväksytty suositus (Tilastokeskus, 1997). Huomionarvoista työvoimatiedustelun osalta on, että tiedustelussa asetetaan työnteko etusijalle kun arvioidaan henkilön toimintaa tutkimusviikolla. Tästä seuraa, että vastaaja tulee luokitelluksi työlliseksi, vaikka haluaisi mahdollisesti tehdä enemmän työtä ja on siis tässä mielessä useimmiten työtä vailla (ILO, 1990).

4.2 Työssäkäyntitilasto (TKT)

Työssäkäyntitilasto on vuodesta 1987 lähtien tuotettu vuositilastosysteemi, joka syntyi osana vuoden 1990 rekisteripohjaisen väestönlaskentasysteemin kehittämistä. Työssäkäyntitilastossa on tietoa väestön taloudellisesta toiminnasta ja työssäkäynnistä. Tiedot tuotetaan käyttämällä hyväksi olemassa olevia hallinnollisia rekisteriaineistoja (noin 30 eri rekisteriaineistoa). Käsiteltävän tutkimuksen kannalta keskeisimpiä niistä ovat väestön keskusrekisteri, verotuksen eri aineistot, erilaiset työ- ja palvelussuhderekisterit, työministeriön työnhakijarekisteri ja Kansaneläkelaitoksen eläkerekisteri. Tilastoyksiköt sisältyvät työssäkäyntitilaston pohja-aineistoon koordinaatein varustettuina, mikä mahdollistaa tietojen tuotannon mille tahansa koordinaatteihin perustuvalle aluejaolle (Tilastokeskus, 1996).

4.2.1 Työssäkäyntitilaston määritelmät

Työssäkäyntilastoissa 15 - 74 -vuotias väestö jaetaan työvoimaan ja työvoiman ulkopuolisiin samaan tapaan kuin työvoimatiedustelussa, mutta peruste jonka mukaan jako tehdään on hieman erilainen. Toiminnan laadun sijasta työssäkäyntitilastossa käytetään *pääasiallisen toiminnan* käsitettä. Pääasiallisen toiminnan luokat päätellään erityisillä päättelysäännöillä käyttämällä hyväksi useiden eri lähteiden tietoja. Pääasiallisen toiminnan päättelyssä työvoimaan kuuluvuus on asetettu ensisijalle. Työvoiman sisällä, toisin kuin työvoimatiedustelussa, on työttömien päättely tehty ennen työllisten päättelyä (Tilastokeskus, 1996) . Toisena oleellisena erona on että, työssäkäyntitilastossa viiteajanjaksona käytetään yhtä päivää, joka yleensä on vuoden viimeisellä viikolla (Tilastokeskus, 1993).

Työttömään työvoimaan luetaan työministeriön tietojen mukaan vuoden viimeisenä työpäivänä työttömänä tai lomautettuna olleet 15-74-vuotiaat henkilöt. Tähän päättelyyn ei vaikuta, vaikka henkilöllä olisi samanaikaisesti myös voimassa oleva työsuhde. *Työlliseen työvoimaan* luetaan kaikki 15-74-vuotiaat henkilöt, jotka laskenta-ajankohtana olivat työllisiä. Tieto työllisyydestä perustuu työeläke- ja veroviranomaisten tietoihin (Tilastokeskus, 1996). Työttömyysasteen ja työvoimaan kuulumattomien määritelmät eivät poikkea työvoimatiedustelun määritelmistä.

4.2.2 Työministeriön työnhakijarekisteri (THR)

Työvoimatoimistojen tietojärjestelmään perustuvaa työministeriön työnhakijarekisteriä hyödynnetään kahdella tavalla. Tilastokeskus käyttää sitä apunaan tehdessään työssäkäyntitilastoa ja sen avulla työministeriö laatii oman työnvälitystilastonsa, jonka pääasiallisena tehtävänä on kuvata työvoimahallinnon asiakkaita ja heille kohdistettuja toimenpiteitä sekä niiden toteutumista alueellisesti ja aloittain. Periaatteessa siis Tilastokeskuksen työssäkäyntitilaston ja työministeriön työnvälitystilaston luvut vastaavat toisiaan. Sen sijaan työvoimatiedustelun ja työministeriön työnhakijarekisterin luvut eivät ole täysin vertailukelpoiset. Työnhakijarekisterin tilastointi perustuu osittain ILO:n hyväksymään suositukseen, mutta myös hallinnolliset määräykset ja käytännöt sekä lainsäädäntö vaikuttavat työttömien määrittämiseen (Tilastokeskus,1997).

Työnvälitystilaston työttömien määrät ovat olleet 1980 -luvun alusta lähtien selkeästi suuremmat kuin työvoimatiedustelun vastaavat luvut. Työttömyyslukujen systemaattisen eron pääasiallinen syy on se, että tilastoissa sovellettavat työttömyyden kriteerit poikkeavat toisistaan. Työministeriön tilasto-ohjeen mukaan tulisi kaikki lyhyetkin työjaksot merkitä rekisteriin. Käytännössä tämä ei ole aina mahdollista ja työnhakija saattaa tehdä jonkin verran töitä ilman, että sitä kirjataan. Merkittävimpänä syynä suurempaan työttömien määrään on työnvälitystilaston *rekisteriviive*. Työttömyysturvan maksatuksen siirtyessä muualle ja työttömyyden kasvaessa pakolliset ilmoittautumisvälit ovat pidentyneet työvoimatoimistoissa kahdesta viikosta jopa puoleen vuoteen. Pitkät ilmoittautumisvälit aiheuttavat rekisteriviiveen. Esim. puolen vuoden aikana itselleen työpaikan löytäneet työttömät kirjautuvat tilastoissa työttömiksi ennen seuraavaa ilmoittautumisajankohtaa, elleivät he itse ilmoita työnsaannistaan työvoimatoimistoon. Työttömyysturvan maksavalle viranomaiselle ilmoitus työsuhteesta on tehtävä, mutta tämä tieto ei välity enää työvoimatoimistoille tai työministeriölle. Yhtenä ongelmana on myös, että työnhakijarekisteriin voi ilmoittautua työnhakijaksi, vaikka olisikin töissä (Tilastokeskus, 1997).

4.3 Tutkimusaineisto

Tutkimuksen otosaineistona on Tilastokeskuksen työvoimatutkimuksen kolmen kuukauden yksilötasoinen aineisto Helsingin osalta. Kyselyt otosaineistoon on tehty vuoden 1994 tammi-, helmi- ja maaliskuussa siten, että jokaisessa kuussa on tehty noin 1200 haastattelua ja yhdeltä ihmiseltä tiedot on kysytty vain kerran. Tässä kolmen kuukauden aineistossa otoskoko on 3 849 helsinkiläistä. Kaikenkaikkiaan tiedot saatiin 3 294 ihmiseltä, jolloin 555 ihmiseltä ei syystä tai toisesta saatu tietoja ollenkaan. Katoon palataan tarkemmin luvussa 5.1.2. Yli 90 % tiedoista kerättiin puhelinhaastatteluin joko koti- tai työnumerosta. Taulukossa 4.1 on kuvattu tarkemmin, kuinka lopulliset estimoinnissa käytettävät otoksen ja perusjoukon kokonaismäärät ovat muodostettu. Perusjoukon tiedot ovat peräisin Tilastokeskuksen rekisteripohjaisesta työssäkäyntitilastosta vuodelta 1993.

Taulukko 4.1: Perusjoukon - ja otoskoon muodostuminen

	OTOS	**PERUSJOUKKO**
15-74 -vuotiaat	3 849	396 890
Sijainti tuntematon	- 85	- 7 761
	3764	
Vastauskato	- 509	
	3255	389 129
Työvoima	2 195	263 701
Sijainti tuntematon	- 15	- 3426
	n = 2180	**N = 260 275**

Rekisteriaineistossa ja otoksessa osalle ihmisistä ei pystytä jostain syystä identifioimaan asuinpaikkaa Helsingin sisäisessä aluejaossa. Koska ihmisten sijainnin määrittäminen on tutkimuksen kannalta olennaista, täytyy otoskoosta ja perusjoukosta vähentää sijainnin kannalta tuntemattomat havainnot. Kun otoksesta poistetaan vielä vastauskato, saadaan 15-74 -vuotiaiden helsinkiläisten otoskooksi 3 255 ja perusjoukon kooksi 389 129. Tämä ei vielä kuitenkaan riitä. Tutkimuksen varsinaisena kohteena on työvoima, joka ei ole sama asia kuin 15-74-vuotias väestö. Työvoiman määrä otoksessa on 2 195 ja perusjoukossa 263 701. Sijainniltaan tuntemattomien havaintojen poistamisen jälkeen lopullisiksi kokonaismääriksi jäävät otoksessa 2 180 ja perusjoukossa 260 275. Työvoiman osalta ei voida erikseen määrittää katoa, koska työvoimaan kuuluvuus määritellään kunkin henkilön kohdalla vasta haastattelun perusteella.

Otantasuhde $f = n / N$ tässä kolmen kuukauden aineistossa on hyvin pieni. Taulukossa 5.2 on kuvattuna otantasuhteen kehitys alkuperäisestä otoksesta edellä karsittuun aineistoon. Työvoiman ja 15-74-vuotiaan väestön osalta otantasuhteet ovat samat kun kato ja tuntemattomat sijainnit on poistettu otosaineistosta. Sen sijaan puhdistamattoman otoksen ja perusjoukon suhde on hieman näitä suurempi.

Taulukko 4.2: Otantasuhteet

AINEISTO	OTANTASUHDE
15-74 -vuotiaat	0,0097
Havaitut	0,0084
Työvoima	0,0083
Havaitut	0,0084

4.1.1 Aineiston jakautuminen soluittain

Pienalue-estimoinnin keskeisin ongelma on, että aluejako saattaa tulla niin tiheäksi, ettei kaikkiin soluihin saada otoksesta riittävästi havaintoja. Huonoimmassa

tapauksessa johonkin soluun ei tule yhtään havaintoa. Taulukossa 4.3 on kuvattuna absoluuttisin ja suhteellisin luvuin kuinka tutkimuksen otos- ja rekisteriaineiston havainnot jakautuvat suurpiireittäin, ikäryhmittäin ja sukupuolittain. Selkeyden vuoksi solut on merkitty taulukkoon numerosarjalla, joissa ensimmäinen numero tarkoittaa suurpiiriä (kts. sivut 24-25), toinen ikäryhmää (1=15-29, 2=30-49, 3=50-74) ja kolmas sukupuolta (1=mies, 2=nainen). Tämän ristiin jaon mukaan soluja on yhteensä 42. Solun teoreettiseksi otoskooksi saadaan 52, kun otoksen kokonaismäärä on 2 180.

Yli puolessa soluista on vähemmän kuin 40 havaintoa ja alle 20 havainnon soluja on kahdeksan. Tätä 20 havainnon määrää solussa voidaan pitää rajana, jotta pystytään laskemaan edes jokseenkin luotettava estimaatti solusta. Kriteerin täyttää 80 prosenttia soluista ja näistä kahdeksassa solussa päästään jopa yli 100 havaintoon. Yhdessäkään solussa ei kuitenkaan ole niin paljon havaintoja, että solukohtainen estimaatti voitaisiin laskea ainoastaan suorien estimaattoreiden avulla. Otoksen ja perusjoukon prosentuaalisissa jakaumissa ei ole merkittäviä eroja.

Jos tarkastellaan ainoastaan taulukon ero-sarakkeen lukujen etumerkkejä, voidaan havaita miinusmerkkisten lukujen lievää keskittymistä sarakkeen loppupäähän. Tämä viittaa siihen, että otokseen on tullut vähemmän kolmanteen ikäluokkaan kuuluvaa työvoimaa, kuin mitä kyseinen ikäluokka perusjoukossa prosentuaalisesti edustaa. Sukupuolen tai suurpiirin suhteen ei voida havaita yhtä selkeitä keskittymiä tietylle alueelle. Pienalue-estimoinnin kannalta olisi kaikkein tärkeintä, että kaikkiin soluihin tulisi riittävästi havaintoja. Siis ainakin 20 tai mieluummin vielä enemmän.

Purcellin ja Kishin (1979) perusjoukon suhteellisiin otoskokoihin P_{dg} perustuvan tarkastelun mukaan (kts. sivut 4-5) suurin osa soluista kuuluu luokkaan kaksi eli keskikokoisiin soluihin ja loput luokkaan kolme eli pieniin soluihin. Tämän perusteella ei voida kuitenkaan päätellä, miten otoshavainnot asettuvat soluihin, koska solujen otoskoot n_{dg} ovat satunnaismuuttujia ja siten joka otoksessa eri suuruisia.

Taulukko 4.3 :Työvoiman absoluuttiset ja suhteelliset jakaumat soluittain otoksessa ja perusjoukossa sekä suhteellisten osuuksien erotus.

Solu	n_{dgs}	N_{dgs}	n_{dgs}%	N_{dgs}%	Ero %
111	48	6 201	2,20	2,38	**-0,18**
211	55	5 757	2,52	2,21	0,31
311	48	5 361	2,20	2,06	0,14
411	14	1 665	0,64	0,64	0,00
511	48	4 757	2,20	1,83	0,37
611	26	2 147	1,19	0,82	0,37
711	29	5 286	1,33	2,03	**-0,70**
112	63	6 990	2,89	2,69	0,20
212	60	6 389	2,75	2,45	0,30
312	55	5 923	2,52	2,28	0,24
412	12	1 660	0,55	0,64	**-0,09**
512	32	4 696	1,47	1,80	**-0,33**
612	18	2 139	0,83	0,82	0,01
712	35	5 201	1,61	2,00	**-0,39**
121	123	12 706	5,64	4,88	0,76
221	106	12 648	4,86	4,86	0,00
321	73	10 800	3,35	4,15	**-0,80**
421	68	5 579	3,12	2,14	0,98
521	108	12 488	4,95	4,80	0,15
621	40	5 186	1,83	1,99	**-0,16**
721	87	10 980	3,99	4,22	**-0,23**
122	106	13 674	4,86	5,25	**-0,39**
222	135	14 826	6,19	5,70	0,49
322	98	10 959	4,50	4,21	0,29
422	38	5 840	1,74	2,24	**-0,50**
522	123	13 157	5,64	5,06	0,58
622	56	5 516	2,57	2,12	0,45
722	95	11 648	4,36	4,48	**-0,12**
131	36	4 366	1,65	1,68	**-0,03**
231	28	3 642	1,28	1,40	**-0,12**
331	16	2 943	0,73	1,13	**-0,40**
431	28	2 234	1,28	0,86	0,42
531	29	3 526	1,33	1,35	**-0,02**
631	12	1 770	0,55	0,68	**-0,13**
731	18	3 857	0,83	1,48	**-0,65**
132	39	5 200	1,79	2,00	**-0,21**
232	34	5 124	1,56	1,97	**-0,41**
332	17	3 644	0,78	1,40	**-0,62**
432	24	2 371	1,10	0,91	0,19
532	34	4 101	1,56	1,58	**-0,02**
632	18	2 314	0,83	0,89	**-0,06**
732	48	5 004	2,20	1,92	0,28

Taulukossa 4.4 on kuvattuna otoksen ja perusjoukon havaintojen määrät peruspiireittäin. Peruspiirit on merkitty taulukkoon numerosarjoilla, joissa ensimmäinen numero viittaa suurpiiriin ja toinen peruspiiriin.

Taulukko 4.4: Työvoiman absoluuttiset ja suhteelliset jakaumat peruspiireittäin otoksessa ja perusjoukossa sekä suhteellisten osuuksien erotus.

Peruspiiri	n_{dgs}	N_{dgs}	n_{dgs}%	N_{dgs}%	Ero %
11	46	6 357	2,11	2,44	**-0,33**
12	104	12 301	4,77	4,73	0,04
13	109	13 395	5,00	5,15	**-0,15**
14	54	7 338	2,48	2,82	**-0,34**
15	102	9 746	4,68	3,74	0,94
21	64	6 089	2,94	2,34	0,60
22	67	7 757	3,07	2,98	0,09
23	101	13 770	4,63	5,29	**-0,66**
24	53	6 740	2,43	2,59	**-0,16**
25	133	14 030	6,10	5,39	0,71
31	112	14 593	5,14	5,61	**-0,47**
32	49	7 116	2,25	2,73	**-0,48**
33	50	5 304	2,29	2,04	0,25
34	33	4 936	1,51	1,90	**-0,39**
35	63	7 681	2,89	2,95	**-0,06**
41	39	4 084	1,79	1,57	0,22
42	31	3 053	1,42	1,17	0,25
43	42	3 926	1,93	1,51	0,42
44	55	6 563	2,52	2,52	0,00
45	17	1 723	0,78	0,66	0,12
51	65	6 216	2,98	2,39	0,59
52	25	5 111	1,15	1,96	**-0,81**
53	106	13 166	4,86	5,06	**-0,20**
54	55	5 894	2,52	2,26	0,26
55	87	9 127	3,99	3,51	0,48
56	36	3 211	1,65	1,23	0,42
61	17	1 658	0,78	0,64	0,14
62	90	9 177	4,13	3,53	0,60
63	63	8 237	2,89	3,16	**-0,27**
71	64	9 546	2,94	3,67	**-0,73**
72	33	4 470	1,51	1,72	**-0,21**
73	125	17 615	5,73	6,77	**-1,04**
74	90	10 345	4,13	3,97	0,16

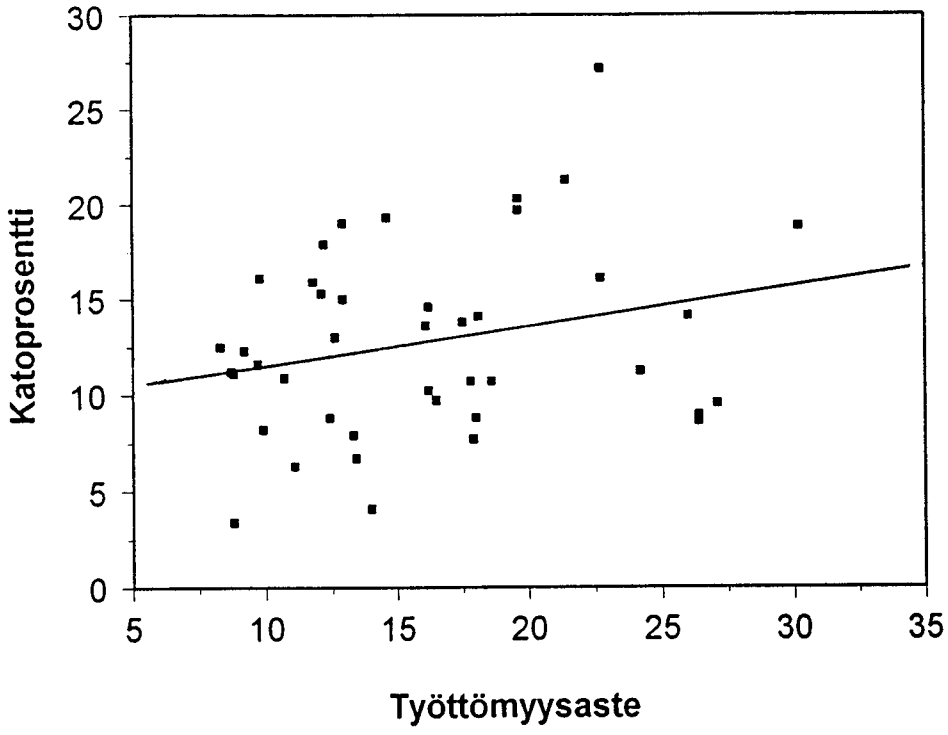
Jos maantieteellistä aluejakoa tihennetään eli otetaan käyttöön peruspiirit, muodostuu soluja yhteensä (33 x 3 x 2) 198 kappaletta. Tällöin teoreettinen solukoko on 11, jota voidaan täysin riittämättömänä. Jättämällä pois ikä- ja sukupuoliluokitukset, jää soluja jäljelle 33 ja teoreettiseksi otoskooksi saadaan 66. Tämä vaihtoehto on mahdollista toteuttaa, mutta on mallinrakennuksen kannalta vaikeampi estimoida.

Ainoastaan kahdessa peruspiirissä otoksen havaintomäärä on alle 20, mutta ei kuitenkaan alle kymmenen. Kahdeksassa peruspiirissä on yli sata havaintoa, mutta yli 200 havaintoon ei päästä yhdessäkään. Keskimääräinen peruspiirin otoskoko on noin 60 havaintoa. Otoksen ja perusjoukon suhteellisien jakaumien vertailu tuo esiin alueelliset painottumiserot. Suurpiirien seitsemän ja neljä peruspiireiltä ei olla saatu riittävästi havaintoja suhteessa perusjoukkoon ja toisaalta suurpiirin neljä peruspiireiltä on tullut mukaan otokseen havaintoja perusjoukkoon nähden selkeästi enemmän. Näiden tulosten pohjalta myöskään peruspiiritasoista pienaluestimointia ei tarvitse kokonaan hylätä.

4.1.2 Kato-analyysi

Tilastokeskuksen (1997) mukaan työvoimatiedustelussa vastauskadon osuus koko maan osalta on noin 7-8 prosenttia otoksesta. Tuloksia laskettaessa oletetaan, että vastauskatoon kuuluvien osalta tilanne on keskimäärin samanlainen kuin tiedustelussa vastanneiden tilanne. Haastatteluun vastaamiseen vaikuttavat mm. asuinalue, ikä ja sukupuoli. Lisäksi on todettu, että myös pitkäaikaistyöttömyys, työttömyyden taso yhteiskunnassa ja taustalla oleva yleisempi huono-osaisuus vaikuttavat vastaamistodennäköisyyteen. Kadon lisäksi työvoimatiedustelun tuloksiin vaikuttavat erilaiset mittausvirheet ja osittaiskato, jolloin kaikkiin kysymyksiin ei saada vastausta. Näiden ilmiöiden vaikutus tuloksiin on kuitenkin nykyisessä työvoimatiedustelussa saatu minimoitua tekemällä suurin osa haastatteluista puhelimitse. Estimaattien luotettavuuden kannalta on tärkeintä, että vastauskadon määrä vaihtelee otosaineistossa satunnaisesti tarkasteltavan ominaisuuden ja solujen suhteen. Jos vaihtelu on systemaattista, otos ei ole riittävän edustava luotettavien tulosten laskemiseksi.

Tässä tutkimuksessa käytetyn kolmen kuukauden aineiston vastauskato on 14.4 prosenttia, mikä on hiukan korkeampi kuin vastaava Tilastokeskuksen ilmoittama luku koko maan käsittävälle otokselle. Yli puolet kadosta johtui siitä, ettei henkilöä tavoitettu ollenkaan. Kaksi muuta merkittävää syytä olivat haastattelusta kieltäytyminen ja ulkomailla olo. Vastauskadon soluittaisessa jakautumisessa on havaittavissa suuriakin eroja. Suurin solukohtainen vastauskato on 27,1 ja pienin 3,4 prosenttia. Tosin ainoastaan kolmessa solussa katoprosentti on yli 20 ja kolmessatoista solussa alle kymmenen prosenttia. Ikäryhmien, suurpiirien ja sukupuolten katoprosenteissa löytyy pieniä eroja (Liite 1). Naisilla katoprosentit ovat kaikissa vastaavissa soluissa pienemmät kuin miehillä ja suurpiirit jakautuvat selkeästi kolmeen ryhmään katoprosenttien suuruusluokan mukaan. Kuvassa 5.1 on tutkittu perusjoukon solumittaisten työttömyysasteiden ja otoksen solumittaisten katoprosenttien välistä lineaarista riippuvuutta.



Kuva 4.1: Perusjoukon soluittaisien työttömyysasteiden ja otoksen soluittaisien katoprosenttien välinen riippuvuus.

Kuvion perusteella voidaan todeta, ettei soluittaisten työttömyysasteiden ja katoprosenttien vaihtelu ole systemaattista toisiinsa nähden. Kuvaan piirretty lineaarinen regressiosuora on aavistuksen verran nouseva, mutta suoraa vastaavan regressiomallin selitysaste on ainoastaan kuusi prosenttia.

5 Työttömyystilaston estimointi

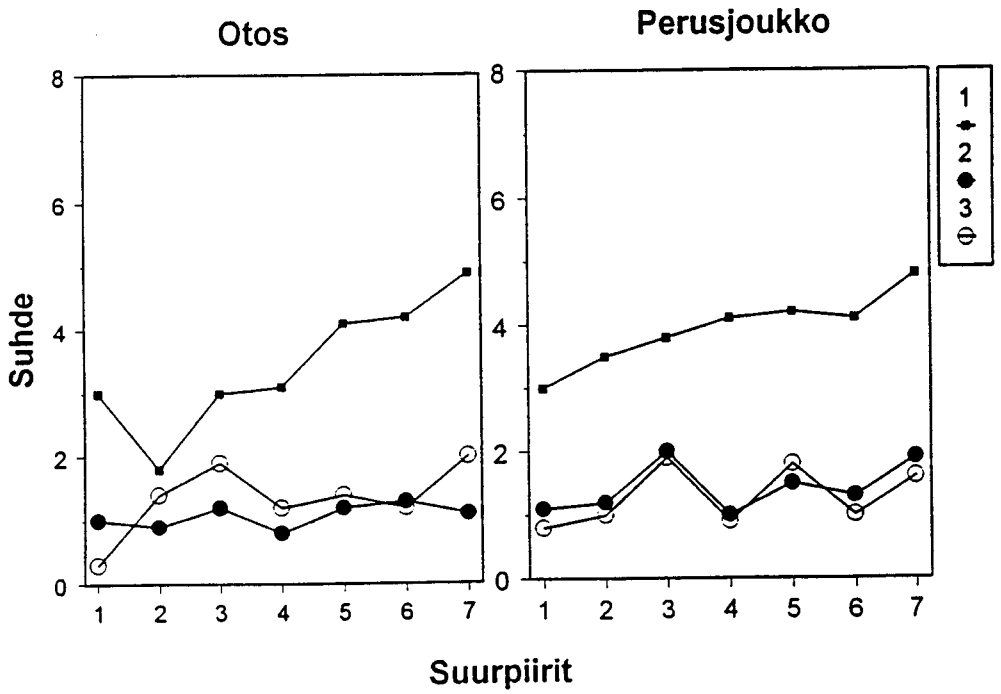
Estimointi on matemaattinen menettely, jossa otoshavaintoja hyväksikäyttäen laaditaan ennuste tai arvio perusjoukon mielenkiinnon kohteena olevalle ominaisuudelle (Pahkinen, 1989). Pienalue-estimoinnissa korostuu erityisesti mielenkiinnon kohteena olevan ominaisuuden mallintaminen pienalue- tai ryhmäkohtaisesti otokseen tulleiden havaintojen avulla. Mitä parempi malli on, sitä vakaampiin tuloksiin päästään.

5.1 Työttömyyden mallitus suhde-estimaattorille

Synteettisessä suhde-estimaattorissa (2.7) mallina on tulosmuuttujan ja selittävän muuttujan keskiarvojen suhde eli $\bar{y} / \bar{x}$. Selittävä muuttujana on keskimääräiset työvoiman työtulot vuodessa soluittain. Tämän selittävän muuttujan valinta oli hieman ongelmallista, koska työvoimatiedustelussa ei haastateltavilta kysytä ainuttakaan tietoa, josta voitaisiin muotoilla numeerinen jatkuva muuttuja ja joka vielä korreloisi hyvin työttömyyden kanssa. Rekisteriaineistossa vuoden keskitulojen ja työttömyysprosentin välinen solutasoinen korrelaatio oli -0,50 (p=0,0006), mitä voidaan pitää varsin hyvänä. Otoksessa havaintokohtaisen dikotomisen 0-1-arvoisen indikaattorimuuttujan ja työtulojen välinen korrelaatio oli -0,35 (p=0,0001). Vuoden 1993 yksilötasoiset työtulot on yhdistetty tutkimuksessa käytettyyn otosaineistoon verohallinnon rekistereistä. Käytännössä verotus valmistuu vasta seuraavan vuoden syksyllä, joten todellisuudessa näin yhtäaikaisten tulotietojen käyttö ei ole mahdollista. Tämä tietojen eriaikaisuus vaikuttaa käytännön tilanteissa melko varmasti huonontavasti synteettisen suhde-estimaattorin tuloksiin.

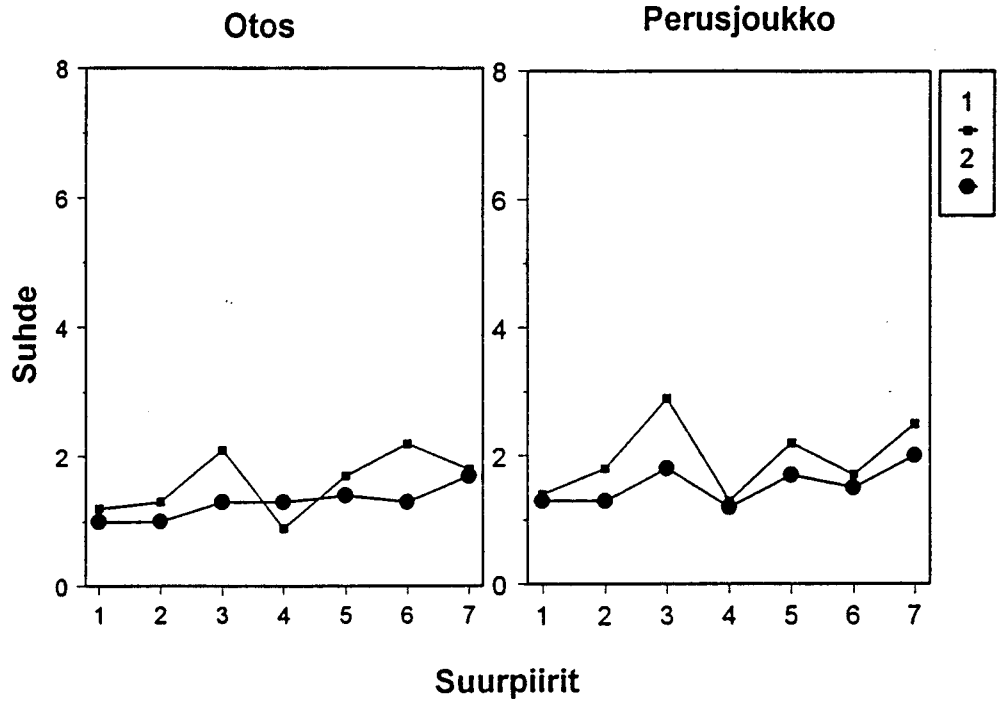
Suhdemallin rakentamiseen on olemassa monta vaihtoehtoa. Yksinkertaisin tapa on laskea keskiarvot koko otoksesta eli $\bar{y} / \bar{x}$, jolloin oletetaan suhteen olevan samanlainen kaikissa soluissa. Tämä on usein liian karkea oletus. Toinen ääripää on laskea jokaiselle solulle oma suhde eli $\bar{y}_{dg} / \bar{x}_{dg}$. Tällöin ei voida enää puhua synteettisestä estimaattorista, koska estimointi tapahtuu ainoastaan solun sisäistä informaatiota hyväksi käyttäen. Synteettisessä estimoinnissa lainataan informaatiota muista soluista, jotka liittyvät toisiinsa jonkin ominaisuuden perusteella.

Synteettisen mallin toteuttamiseksi on tässä tutkimuksessa ainakin kolme vaihtoehtoa, joissa käytetään hyväksi ikäryhmittelyä (=g),sukupuoliryhmittelyä (=z) tai molempia (=gz). Kuvissa 5.1, 5.2 ja 5.3 on tutkittu kuinka hyvin mallioletukset pitävät paikkansa näissä kolmessa vaihtoehdossa. Tulosmuuttujana y on seuraavissa kuvissa käytetty työttömyysastetta ja selittävänä muuttujana x vuoden keskimääräisiä työtuloja.



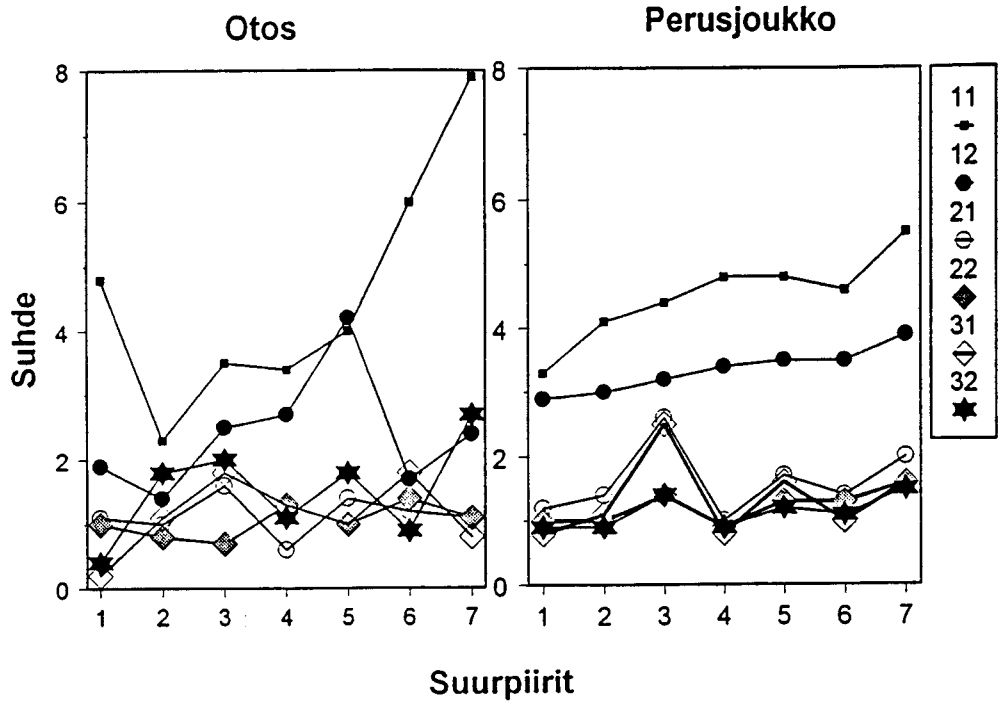
Kuva 5.1 : Suhde $\bar{y}_g / \bar{x}_g$ suurpiireittäin otoksessa ja perusjoukossa (1=15-29v, 2=30-49v ja 3=50-74v).

Nuorin ikäryhmä erottuu selvästi kahdesta muusta ikäryhmästä, mutta suhteet eri suurpiireissä eivät ole samat. Keski-ikäisten ja varttuneiden ikäryhmissä suhteet noudattavat lähes samanlaista linjaa, mutta suurpiireittäin tarkasteltuna on selviä eroja havaittavissa. Ainoastaan keskimmäisessä ikäryhmässä suhteiden erot suurpiirien välillä ovat otoksessa kohtuullisen pienet.



Kuva 5.2 : Suhde $\bar{y}_z / \bar{x}_z$ suurpiireittäin otoksessa ja perusjoukossa (1=miehet, 2=naiset).

Suhteiden voidaan katsoa olevan otoksessa naisilla lähes samat kaikissa suurpiireissä. Miehillä suurpiireittäinen vaihtelu on selvästi suurempaa. Tämä vaihtoehto on kuitenkin huono, koska miesten ja naisten välillä ei ole eroja paljon ollenkaan. Synteettisillä estimaattoreilla on estimaatteja homogeenisoiva vaikutus. Jos malleja on vain kaksi, joilla molemmilla on vielä samansuuntainen ennuste, on mallinnus periaatteessa täysin merkityksetön.



Kuva 5.3 : Suhde $\bar{y}_{gz} / \bar{x}_{gz}$ suurpiireittäin otoksessa ja perusjoukossa (11=15-29v miehet, 12=15-29v naiset, 21=30-49v miehet, 22=30-49v naiset, 31=50-74v miehet ja 32=50-74v naiset).

Kuva on varsin moniaineksinen, mutta viestittää selvästi etteivät mallioletukset eri ryhmissä suurpiirien välillä pidä kovinkaan hyvin paikkaansa. Varsinkin otoksessa hajonta on suurta. Nuorin ikäryhmä sekä miesten että naisten osalta erottuu muista, mutta muiden neljän luokan suhdearvot vaihtelevat nollan ja kahden välillä. Homogeenisin luokka on 30-49 -vuotiaat naiset ja heterogeenisin 15-29 -vuotiaat miehet. Luvussa 6 esitettävien synteettisen suhde-estimaattorin tulokset ovat laskettu tällä kuusi eri ennustetta sisältävällä mallilla.

Otoskoolla ehdollistetun estimaattorin (2.18) regressio-osassa olevana regressiokertoimena β on tuloksissa käytetty tulosmuuttujan ja apumuuttujan keskiarvojen suhdetta koko otoksessa. Synteettisen suhde-estimaattorin mallina on sama kuuden ennusteen malli kuin pelkän synteettisen suhde-estimaattorin tuloksissa.

5.2 Työttömyyden mallitus regressioestimaattorille

Cassel, Kristiansson, Råbäck ja Wahlström (1987) ja Feeney (1987) ovat työttömyyden pienalue-estimointia käsittelevissä tutkimuksissaan todenneet, että regressioon pohjautuvien mallien avulla voidaan päästä kohtuullisen hyviin tuloksiin. Kun muuttujat ovat luokitteluasteikollisia 0-1-muuttujia, niin tarkoituksenmukaisin tilastollinen väline on log-lineaarinen malli, joka on yleisesti muotoa

$$\log(P_{dg}/(1 - P_{dg})) = \mu + \alpha_d + \beta_g + \epsilon_{dg}, \qquad (5.1)$$

missä P_{dg} = tietyn ominaisuuden omaavien havaintojen osuus solussa dg

μ = vakioparametri

α_d = ensimmäisen muuttujan vaikutus ryhmässä d

β_g = toisen muuttujan vaikutus ryhmässä g

ε_{dg} = residuaali tai virhetermi.

Mallin suorituskykyä voidaan mitata kahdella tavalla. Muuttujien valinnassa tarkastellaan estimoitujen parametrien tilastollista merkitsevyyttä P-arvojen avulla. P-arvo määrittää onko muuttujan vaikutus malliin merkitsevästi suurempi kuin nolla. Koko mallin sopivuutta aineistoon tarkastellaan residuaalin P-arvon avulla. Jos residuaalit eivät ole tilastollisesti merkitseviä, malli sopii hyvin aineistoon (Feeney, 1987).

Taulukossa 5.1 on esitettynä kolmen eri mallin tulokset. Mallit ovat:

(a) Alue (7 tasoa), ikäryhmä (3 tasoa) ja sukupuoli (=42 tasoa)

(b) Alue (2 tasoa), ikäryhmä (3 tasoa) ja sukupuoli (=12 tasoa)

(c) Ikäryhmä (8 tasoa) ja yhdistetty sukupuoli-siviilisääty (=16 tasoa).

Laskennat on tehty SAS-ohjelmiston tilastosovelluksen CATMOD-proseduurilla.

Liitteessä kaksi on esimerkkinä mallin B toteutuksen ohjelmalauseet (SAS Institute Inc., 1987).

Taulukko 5.1 : Testitulokset mallien A, B ja C parametreille.

Mallit/Muuttujat	DF	χ^2	P-arvo
Malli A			
Vakio	1	9604,40	0,0000
Alue	6	11,13	0,0843
Ikäryhmä	2	16,09	0,0003
Sukupuoli	1	8,52	0,0035
Residuaali	32	41,04	0,1313
Malli B			
Vakio	1	9264,81	0,0000
Alue	1	2,98	0,0846
Ikäryhmä	2	13,79	0,0010
Sukupuoli	1	12,54	0,0004
Residuaali	7	13,44	0,0622
Malli C			
Vakio	1	12237,01	0,0000
Ikäryhmä	7	19,41	0,0070
Sukupuoli-siviilisääty	3	30,82	0,0000
Residuaali	21	21,59	0,4233

Malli A on solujaottelun mukainen eli se muodostuu 42 ennusteesta. Alueet muodostuvat suurpiireistä ja ikäryhmät jo aiemmin mainituista kolmesta ikäryhmästä. Huonona puolena on ennusteiden liian suuri määrä otoksen kokoon nähden, jolloin osa ennusteista on muodostettu alle 20 havainnon perusteella. Tämä vaikuttaa

myös keskivirheisiin. Ikäryhmä ja sukupuoli ovat merkitseviä muuttujia, mutta suurpiiri ei ole. Residuaalin P-arvon mukaan malli kuitenkin sopii aineistoon kohtuullisen hyvin.

Mallissa B ikäryhmät ovat ennallaan, mutta seitsemästä suurpiiristä on muodostettu kaksi suurpiiriryhmää rekisteriaineistosta laskettujen työttömyysasteiden perusteella. Nyt malli muodostuu 12 ennusteesta. Tämä ei näytä parantavan mallia, koska alueen merkitsevyys pysyy lähes ennallaan, mutta mallin sopivuus aineistoon huononee selvästi. Suurpiirien ryhmäjaottelun voi toteuttaa monella eri tavalla, mutta kokeilujen perusteella malli B on näistä paras. Muissa vaihtoehdoissa mallin sopivuus aineistoon oli parempi, mutta alueen merkitys mallissa pieneni jyrkästi.

Mallissa C on alue jätetty kokonaan pois. Ikäryhmistä on tehty kahdeksanluokkainen muuttuja ja uutena mukaan on otettu siviilisääty (naimaton/naimisissa), joka on yhdistetty sukupuoleen. Ennusteita mallissa on 32. Tämä malli näyttäisi toimivan parhaiten sekä muuttujien että koko mallin sopivuuden osalta. Ongelmana on, ettei sitä voida hyödyntää nyt käytettävissä olevilla aineistoilla. Rekisteritiedot työvoimasta tarvittaisiin mallin mukaisilla jaottelulla kertoimien laskemista varten, jotta solujen mukaisten estimaattien muodostaminen olisi mahdollista.

Luvussa kuusi esitettävissä tuloksissa on käytetty mallia A, jota voidaan pitää tähän asti testatuista vaihtoehdoista sopivimpana käytettyyn otosaineistoon, mutta ei suinkaan kaikista parhaimpana. Erilaisia mallivaihtoehtoja on olemassa useita kymmeniä, ellei satoja. Kaikkia työvoimatiedustelun muuttujia voidaan periaatteessa hyödyntää mallintamisessa. Tavoitteena olisi kehittää aineistoon mahdollisimman hyvin sopiva malli, joka ei olisi kuitenkaan liian moniselitteinen ja jokaisen mallin ennusteen takana olisi riittävästi havaintoja.

6 Estimointitulokset ja niiden arviointi

6.1 Estimaattien vertailumittarit

Estimaattoreiden tehokkuuden ja estimaattien luotettavuuden arvioimiseksi on kehitetty erilaisia vertailumittareita. Näiden avulla voidaan helpommin päätellä, mikä tarjolla olevista estimaattoreista on paras kulloiseenkin tilanteeseen.

Otantaan ja estimointiin liittyvissä tarkasteluissa lähtökohtana on usein, että otanta on ollut todennäköisyysotanta ja otosalkioiden mittaustulos on ollut virheetön. Estimaattorin oman vaihtelun katsotaan johtuvan vaihtelusta otanta-asetelmassa. Varsinkin suurissa, väestöpohjaisissa otantatutkimuksissa, kuten esim. Tilastokeskuksen työvoimatiedustelussa, tulee estimaattoreihin lisää vaihtelua otanta-asetelman ulkopuolelta. Tämän otanta-asetelman ulkopuolelta tulevan vaihtelun seurauksena on, että estimointi havaitusta otoksesta on harhaista. Vastauskato ja erilaiset mittausvirheet aiheuttavat estimaattien harhaisuuden (Pahkinen, 1989).

Otantatutkimuksessa ilmenevän satunnaisvaihtelun suuruutta voidaan arvioida keskivirheiden avulla. Tämä keskihajontaa estimoiva suure eli keskivirhe saadaan ottamalla neliöjuuri varianssin estimaatista eli

$$s.e.\ \hat{Y}_{dg} = \sqrt{\hat{V}(\hat{Y}_{dg})}. \qquad (6.1)$$

Keskivirhe kuvaa sitä, kuinka tiiviisti havainnot ovat keskittyneet otoskeskiarvon ympärille. Sen suuruuteen vaikuttavat muuttujien arvojen vaihtelu eli perusjoukon varianssi ja tämän tutkimuksen kannalta erityisesti solukohtainen otoskoko. Keskivirheen ja jakaumaoletuksen (yleensä normaalijakauma) avulla voidaan tutkimuksen parametreille laskea myös luottamusväli, jolla perusjoukon muuttujan todellinen arvo on tietyllä luotettavuustasolla. Kapeampi luottamusväli on merkki tehokkaammasta estimaattorista.

Keskivirheen avulla voidaan laskea myös toinen estimaattoreiden paremmuutta kuvaava mittari. Prosenttinen variaatiokerroin saadaan lauseesta

$$C.V.\ \% = \frac{s.e.\ \hat{Y}_{dg}}{\hat{Y}_{dg}} \cdot 100. \qquad (6.2)$$

Variaatiokerroin mittaa sitä, kuinka suuri on keskivirheen osuus estimaatista. Mitä suurempi on kerroin, sitä huonompi on estimaattori. Tuloksia voidaan pitää hyvinä, jos keskivirheen osuus estimaatin arvosta on alle kymmenen prosenttia.

Edellä esitettyjä estimaattorin tehokkuutta ja luotettavuutta kuvaavia tunnuslukuja voidaan käyttää estimaattien vertailuun sekä kokeellisissa tutkimuksissa että käytännön tilanteissa. Näiden lisäksi on olemassa mittareita, joita on mahdollista käyttää ainoastaan todellista tilannetta jäljittelevissä tutkimuksissa. Tällöin tulee olla tiedossa tulosmuuttujan tilanne perusjoukossa kyseisellä hetkellä eli ns. todellinen arvo. Ensisijaisesti näitä vertailumittareita on käytetty simulointikokeiden yhteydessä, jolloin laskettu tunnusluku on samasta perusjoukosta poimituista otoksista laskettujen estimaattien keskiarvo. Ghosh ja Rao (1994) ovat soveltaneet tutkimuksessaan kahta mittaria myös yhden otoksen tilanteessa. Ensimmäinen niistä on keskimääräinen suhteellinen virhe (engl. average relative error), joka saadaan kaavasta

$$ARE\% = \frac{1}{I} \sum_{d=1}^{D} \sum_{g=1}^{G} \frac{| \hat{Y}_{dg} - Y_{dg} |}{Y_{dg}} \cdot 100, \qquad (6.3)$$

jossa I tarkoittaa solujen lukumäärää, $\hat{Y}_{dg}$ estimaattia ja Y_{dg} tulosmuuttujan todellista arvoa solussa dg. Keskimääräinen suhteellinen virhe mittaa sitä, kuinka lähellä todellista arvoa estimaatit ovat eli se on ns. stabiilisuuden indikaattori. Toinen kokeellisiin tutkimuksiin soveltuva tunnusluku kuvaa keskimääräistä keskineliövirhettä (engl. average squared error), joka on muotoa

$$ASE = \frac{1}{I} \sum_{d=1}^{D} \sum_{g=1}^{G} (\hat{Y}_{dg} - Y_{dg})^2. \qquad (6.4)$$

Keskimääräinen keskineliövirhe kuvaa estimaatin ja todellisen arvon erotuksen suuruutta.

6.1.1 Todellisen arvon määrittäminen

Vertailumittareissa (6.3) ja (6.4) tarvittavan solun dg todellisen arvon Y_{dg} määrittäminen ei ole aina täysin yksiselitteistä. Estimointi tapahtuu otosaineiston pohjalta ja todelliset luvut saadaan rekisteripohjaisesta kokonaisaineistosta. Ongelmana on, että työvoimatiedustelun ja työssäkäyntitilaston työttömyyttä koskevat määritelmät eroavat toisistaan, kuten luvussa neljä jo todettiin. Työssäkäyntitilaston työttömyysluvut ovat korkeammat kuin työvoimatiedustelun. Tähän ovat syynä mm. se, että rekisteripohjaisessa työssäkäyntitilastossa työttömät määritellään ennen työllisiä ja rekisteröitymisviive. Jos siis todellisena arvona käytetään suoraan rekisteripohjaisen työssäkäyntitilaston lukuja, niin vertailumittareiden tulokset ovat osittain virheelliset.

Tilastokeskuksessa (1997) on tutkittu vuoden 1993 työssäkäyntitilaston työttömien määrää koko maassa hieman tarkemmin ja verrattu työttömiä esimerkiksi saman ajankohdan työsuhdetilastoihin. Vuoden 1993 viimeisenä arkipäivänä Suomessa oli työssäkäyntitilaston mukaan 535 290 työtöntä. Tästä 88 000 henkeä oli työsuhteessa työsuhdetilastojen mukaan, joista runsaat 10 000 oli tapauksia, joiden työttömyysjakso oli joko alkanut tai päättynyt vuoden viimeisellä viikolla. Loput eli 78 000 työtöntä luokittuisivat työvoimatiedustelun mukaan työllisiksi, koska myös lomautetut ja vain tunninkin tutkimusviikolla töitä tehneet luokitellaan työvoimatiedustelussa työllisiksi.

Vastaus kysymykseen siitä, kuinka pitkiä näiden työllisten työttömien työjaksot todellisuudessa ovat tai millaiset ovat heidän kuukausitulonsa, ei ole tässä relevanttia. Tavoitteena on määritellä työvoimatiedustelun mukaiset mahdollisimman oikeat arvot. Koska edellä esitetyn kaltainen työssäkäyntitilaston ja työsuhdetilastojen yksilötasoinen vertailu helsinkiläisten osalta ei ole tämän tutkimuksen puitteissa mahdollista, on oikean arvon määrittämiseksi tehty karkea oletus, että Helsingissä ja sen osa-alueilla työssäkäyntitilaston ja työsuhdetietojen välinen tilanne on suurinpiirtein sama kuin koko maassa. Lähempänä työvoimatiedustelun mukaista oikeaa arvoa olevat työttömien määrät on saatu vähentämällä 15 prosenttia rekisteripohjaisen työssäkäyntitilaston työttömien määristä ja lisäämällä vastaava määrä työllisiin. Koska samantasoinen vähennys ja lisäys tehdään kaikkiin soluihin, ei oikean arvon määrittämisellä ole periaatteessa vaikutusta keskimääräisen suhteellisen virheen (6.3) ja keskimääräisen keskineliövirheen (6.4) tuloksiin verrattaessa estimaattoreita toisiinsa.

6.2 Tulokset vertailumittareille

Ensimmäisinä tarkastelun kohteina ovat prosentuaalinen keskimääräinen suhteellinen virhe (ARE%) ja keskimääräinen keskineliövirhe (ARE). Laskennassa on käytetty edellisessä kappaleessa määriteltyä todellista arvoa. Ensimmäisessä taulukossa 6.1 ovat työttömät, toisessa 6.2 työlliset ja kolmannessa 6.3 työttömyysaste.

Taulukko 6.1: Pienalue-estimaattoreiden vertailu; työttömät.

	HT	RS	SD	MRE	SSD
ARE %	32,5	27,1	34,3	14,9	15,2
ASE	98 967	101 868	113 824	55 368	46 454

Taulukko 6.2: Pienalue-estimaattoreiden vertailu; työlliset.

	HT	RS	SD	MRE	SSD
ARE %	16,1	7,3	7,6	2,1	3,8
ASE	897 373	225 766	232 056	20 683	97 475

Taulukko 6.3: Pienalue-estimaattoreiden vertailu; työttömyysaste.

	HT	RS	SD	MRE	SSD
ARE %	30,0	27,1	34,3	14,9	15,0
ASE	32	25	40	10	7

Kaikissa taulukoissa voidaan havaita, että luvut pienenevät vasemmalta oikealla mentäessä. Ainoastaan otoskoolla ehdollistettu estimaattori (SD) muodostaa poikkeuksen. Työttömien ja työttömyysasteen osalta se saa jopa huonommat arvot kuin suora estimaattori (HT). Näiden tulosten perusteella otoskoolla ehdollistettu estimaattori on jätetty jatkotarkasteluista kokonaan pois. Modifioitu regressioestimaattori (MRE) ja komposiittiestimaattori (SSD) saavat selvästi muita paremmat tulokset. Regressioestimaattorilla on pienemmät eli paremmat luvut verrattaessa estimaatin ja todellisen arvon erotusta todelliseen arvoon, kun taas komposiittiestimaattorilla luvut ovat paremmat estimaatin ja todellisen arvon erotuksen neliöitä tarkasteltaessa.

Seuraavaksi tarkastellaan neljän eri estimaattorityypin soluittaisia keskivirheitä sekä variaatiokertoimia työttömien ja työllisten kokonaismäärille sekä työttömyysasteille. Soluittaiset keskivrheet ovat esitetty taulukoissa 6.4 - 6.6. Solut on merkitty taulukoihin numerosarjoin samaan tapaan kuin aiemmin taulukossa 5.3. Taulukoissa on myös sarakkeet otoksen ja perusjoukon soluittaisille havaintomäärille.

Taulukko 6.4: Soluittaiset keskivirheet työttömien kokonaismäärille.

Solu	n_{dgs}	N_{dgs}	s.e. HT	s.e. RS	s.e. MRE	s.e. SSD
111	48	6 201	390	92	138	141
211	55	5 757	289	90	129	123
311	48	5 361	337	87	137	132
411	14	1 665	189	14	49	48
511	48	4 757	307	84	120	117
611	26	2 147	208	30	69	66
711	29	5 286	495	80	140	140
112	63	6 990	294	74	153	140
212	60	6 389	248	63	134	120
312	55	5 923	297	68	140	131
412	12	1 660	186	9	51	49
512	32	4 696	328	41	114	109
612	18	2 139	162	11	66	62
712	35	5 201	311	42	133	122
121	123	12 706	405	95	232	265
221	106	12 648	403	80	234	240
321	73	10 800	499	67	237	236
421	68	5 579	193	27	142	114
521	108	12 488	448	88	254	269
621	40	5 186	314	24	153	137
721	87	10 980	407	62	245	224
122	106	13 674	405	73	245	244
222	135	14 826	349	78	252	249
322	98	10 959	285	47	218	178
422	38	5 840	349	21	154	141
522	123	13 157	352	71	254	238
622	56	5 516	245	22	156	131
722	95	11 648	383	60	248	224
131	36	4 366	121	43	81	70
231	28	3 642	244	45	85	81
331	16	2 943	295	31	78	75
431	28	2 234	176	35	63	60
531	29	3 526	229	45	87	81
631	12	1 770	198	16	56	54
731	18	3 857	293	36	99	94
132	39	5 200	185	43	100	88
232	34	5 124	339	58	118	113
332	17	3 644	346	28	93	90
432	24	2 371	163	21	71	65
532	34	4 101	271	48	100	95
632	18	2 314	176	15	73	68
732	48	5 004	305	78	129	122

Taulukko 6.5: Soluittaiset keskivirheet työllisten kokonaismäärille.

Solu	n_{dgs}	N_{dgs}	s.e. HT	s.e. RS	s.e. MRE	s.e. SSD
111	48	6 201	596	148	138	178
211	55	5 757	599	183	129	189
311	48	5 361	515	158	137	162
411	14	1 665	323	25	49	51
511	48	4 757	441	153	120	140
611	26	2 147	230	51	69	67
711	29	5 286	506	118	140	140
112	63	6 990	722	217	153	251
212	60	6 389	707	210	134	232
312	55	5 923	616	184	140	198
412	12	1 660	381	22	51	53
512	32	4 696	619	78	114	138
612	18	2 139	433	37	66	72
712	35	5 201	707	109	133	165
121	123	12 706	905	268	232	564
221	106	12 648	1011	217	234	547
321	73	10 800	928	155	237	370
421	68	5 579	590	98	142	221
521	108	12 488	918	210	254	509
621	40	5 186	623	72	153	174
721	87	10 980	946	163	245	434
122	106	13 674	1129	263	245	609
222	135	14 826	1123	278	252	762
322	98	10 959	991	164	218	498
422	38	5 840	742	73	154	188
522	123	13 157	1017	217	254	633
622	56	5 516	607	65	156	203
722	95	11 648	996	178	248	491
131	36	4 366	705	241	81	143
231	28	3 642	555	122	85	107
331	16	2 943	558	61	78	84
431	28	2 234	301	96	63	69
531	29	3 526	532	119	87	107
631	12	1 770	406	32	56	58
731	18	3 857	783	100	99	115
132	39	5 200	777	181	100	171
232	34	5 124	668	133	118	150
332	17	3 644	683	53	93	103
432	24	2 371	403	58	71	79
532	34	4 101	534	103	100	123
632	18	2 314	469	48	73	78
732	48	5 004	498	143	129	154

Taulukko 6.6: Soluittaiset prosenttiset keskivirheet työttömyysasteelle.

Solu	n_{dgs}	N_{dgs}	s.e. HT	s.e. RS	s.e. MRE	s.e. SSD
111	48	6 201	6,2	2,1	2,2	2,3
211	55	5 757	5,0	3,0	2,2	2,1
311	48	5 361	6,2	2,7	2,6	2,5
411	14	1 665	10,9	1,3	3,0	2,9
511	48	4 757	6,4	2,9	2,5	2,5
611	26	2 147	9,5	2,2	3,2	3,1
711	29	5 286	9,2	1,9	2,7	2,6
112	63	6 990	4,2	2,2	2,2	2,0
212	60	6 389	3,9	2,3	2,1	1,9
312	55	5 923	5,0	2,1	2,4	2,2
412	12	1 660	10,7	0,9	3,1	2,9
512	32	4 696	6,9	1,1	2,4	2,3
612	18	2 139	7,4	1,3	3,1	2,9
712	35	5 201	5,9	1,5	2,6	2,3
121	123	12 706	3,2	3,5	1,8	2,1
221	106	12 648	3,2	2,8	1,9	1,9
321	73	10 800	4,6	2,4	2,2	2,2
421	68	5 579	3,4	2,9	2,5	2,0
521	108	12 488	3,6	2,8	2,0	2,1
621	40	5 186	6,0	2,3	2,9	2,6
721	87	10 980	3,7	2,5	2,2	2,0
122	106	13 674	2,9	2,1	1,8	1,8
222	135	14 826	2,3	2,0	1,7	1,7
322	98	10 959	2,6	1,6	2,0	1,6
422	38	5 840	5,9	1,4	2,6	2,4
522	123	13 157	2,7	1,8	1,9	1,8
622	56	5 516	4,4	1,3	2,8	2,4
722	95	11 648	3,3	1,7	2,1	1,9
131	36	4 366	2,7	8,4	1,9	1,6
231	28	3 642	6,6	5,0	2,3	2,2
331	16	2 943	9,7	2,9	2,6	2,6
431	28	2 234	7,7	6,2	2,8	2,7
531	29	3 526	6,4	4,9	2,5	2,3
631	12	1 770	10,7	2,4	3,2	3,1
731	18	3 857	7,4	3,7	2,6	2,4
132	39	5 200	3,5	6,5	1,9	1,7
232	34	5 124	6,5	4,6	2,3	2,2
332	17	3 644	9,2	2,7	2,6	2,5
432	24	2 371	6,7	4,3	3,0	2,7
532	34	4 101	6,5	4,4	2,4	2,3
632	18	2 314	7,4	4,0	3,1	2,9
732	48	5 004	6,0	4,7	2,6	2,4

Kaikille kolmelle tulosmuuttujalle suoran estimaattorin (HT) soluittaiset keskivirheet ovat kaikissa soluissa selvästi muiden estimaattoreiden keskivirheitä korkeammat. Muiden kolmen estimaattorin välillä ei näin selkeitä suuruuseroja ole havaittavissa. Ainoastaan työttömien kokonaismäärien keskivirheet ovat kaikissa soluissa pienimmät synteettisellä suhde-estimaattorilla (RS). Sen sijaan työllisten kokonaismäärien keskivirheet modifioidulle regressioestimaattorille (MRE) ovat lähes puolessa soluista paremmat kuin synteettisen suhde-estimaattorin keskivirheet. Soluittaisille työttömyysasteille taas pienimmät keskivirheet ovat komposiittiestimaattorilla (SSD).

Täysin tyhjentävää selitystä näille eroille on vaikea löytää. Synteettisen suhde-estimaattorin keskivirheiden pienuus selittyy osittain estimaattorin takana olevalla mallinnuksella, joka vaikuttaa suoraan varianssin ja sen myötä myös keskivirheen suuruuteen. Tämä pätee myös modifioituun regressioestimaattoriin. Synteettisen suhde-estimaattorin mallissa on ennusteet kuudelle eri ryhmälle eli kolmelle ikäryhmälle sukupuolittain. Näistä jokaisen ennusteen takana on ainakin 200 tai jopa 300 havaintoa. Sen sijaan modifioidun regressioestimaattorin mallissa on jokaiselle solulle oma ennusteensa, joilloin havaintomäärät yhtä ennustetta kohti ovat oleellisesti pienemmät kuin kuuden ennusteen mallissa. Komposiittiestimaattorin keskivirheisiin vaikuttaa painotuksen kautta suoran estimaattorin suuret keskivirheet. Mitä lähempänä kahtasataa solun otoskoko on, sitä suurempi osa keskivirheestä muodostuu suoran estimaatin keskivirheestä, joka on muita oleellisesti suurempi. Työttömyysasteiden keskivirheiseen tämä ei vaikuta niin paljon, koska suorien estimaattien keskivirheet työttömyysasteille eivät ole suhteessa niin suuret kuin työttömien ja työllisten keskivirheet. Ainoastaan niissä soluissa joissa otoshavaintoja on yli 100, jonkin muun työttömyysasteen estimaatin solukohtainen keskivirhe on pienempi kuin komposiittiestimaattorin.

Pelkkä keskivirheiden tarkastalu ei riitä johtopäätästen tekoa varten. Hyvän kuvan keskivirheiden tasosta antavat soluittaiset variaatiokertoimet, joissa keskivirheen koot suhteutetaan itse estimaatin kokoon. Taulukoissa 6.7 - 6.9 on esitettynä neljän estimaattorin soluittaiset variaatiokertoimet työttömien ja työllisten kokonaismäärille sekä työttömyysasteille.

Taulukko 6.7 : Soluittaiset prosenttiset variaatiokertoimet työttömien kokonaismäärille.

Solu	n_{dgs}	N_{dgs}	c.v. HT	c.v. RS	c.v. MRE	c.v. SSD
111	48	6 201	27,2	4,7	10,8	10,5
211	55	5 757	26,9	5,3	10,9	11,2
311	48	5 361	23,5	5,6	11,0	10,6
411	14	1 665	52,7	3,0	14,4	14,3
511	48	4 757	19,8	6,3	10,8	9,9
611	26	2 147	17,4	5,0	15,2	12,5
711	29	5 286	31,9	5,6	10,9	10,6
112	63	6 990	30,8	7,2	14,3	14,8
212	60	6 389	34,7	7,1	13,7	14,9
312	55	5 923	27,6	8,1	13,2	13,2
412	12	1 660	77,8	3,9	19,7	19,7
512	32	4 696	45,8	6,9	13,5	13,8
612	18	2 139	68,0	4,0	20,4	20,9
712	35	5 201	52,1	6,4	14,3	15,3
121	123	12 706	18,8	4,7	13,3	11,7
221	106	12 648	26,0	4,3	13,5	14,3
321	73	10 800	29,9	5,3	13,5	13,7
421	68	5 579	26,9	2,8	18,6	19,0
521	108	12 488	20,8	5,2	12,5	12,1
621	40	5 186	37,6	3,0	21,3	20,6
721	87	10 980	28,4	4,5	13,8	15,0
122	106	13 674	30,9	4,9	21,1	19,7
222	135	14 826	26,6	5,2	20,0	19,4
322	98	10 959	34,1	4,4	18,1	21,3
422	38	5 840	48,7	3,4	30,3	28,2
522	123	13 157	24,6	5,8	17,6	17,7
622	56	5 516	29,3	4,1	33,1	27,1
722	95	11 648	29,1	5,7	19,3	19,1
131	36	4 366	101,2	6,0	12,0	14,1
231	28	3 642	51,1	9,0	14,8	15,2
331	16	2 943	82,5	10,8	14,4	14,8
431	28	2 234	24,5	9,9	17,8	15,6
531	29	3 526	47,9	11,8	13,5	14,0
631	12	1 770	82,9	5,9	20,0	20,2
731	18	3 857	122,6	7,8	14,6	15,6
132	39	5 200	77,6	4,9	18,4	20,7
232	34	5 124	47,3	7,2	21,2	20,2
332	17	3 644	96,6	5,2	18,8	19,2
432	24	2 371	45,5	5,6	28,1	26,7
532	34	4 101	37,8	8,4	18,5	17,6
632	18	2 314	73,6	4,1	29,6	29,6
732	48	5 004	23,3	10,9	19,4	15,7

Taulukko 6.8 : Soluittaiset prosenttiset variaatiokertoimet työllisten kokonaismäärille.

Solu	n_{dgs}	N_{dgs}	c.v. HT	c.v. RS	c.v. MRE	c.v. SSD
111	48	6 201	13,9	2,8	2,8	2,9
211	55	5 757	10,9	4,0	2,8	2,7
311	48	5 361	12,0	3,8	3,3	3,3
411	14	1 665	24,6	2,0	3,7	3,7
511	48	4 757	10,5	4,2	3,3	3,2
611	26	2 147	12,0	3,1	4,1	4,0
711	29	5 286	26,5	3,0	3,5	3,8
112	63	6 990	11,0	3,4	2,6	2,5
212	60	6 389	11,0	3,8	2,5	2,3
312	55	5 923	11,2	3,5	2,9	2,8
412	12	1 660	31,9	1,6	3,6	3,6
512	32	4 696	20,0	2,1	3,0	3,1
612	18	2 139	22,7	2,1	3,7	3,6
712	35	5 201	19,7	2,6	3,1	3,2
121	123	12 706	7,2	2,2	2,1	1,9
221	106	12 648	9,1	2,0	2,1	2,1
321	73	10 800	13,2	2,1	2,6	2,8
421	68	5 579	8,0	1,7	2,9	2,5
521	108	12 488	8,5	2,1	2,4	2,4
621	40	5 186	15,8	1,5	3,4	3,5
721	87	10 980	10,6	2,0	2,7	2,7
122	106	13 674	10,0	1,9	2,0	2,1
222	135	14 826	7,6	2,1	1,9	1,7
322	98	10 959	9,1	1,7	2,2	2,1
422	38	5 840	19,4	1,3	2,9	3,1
522	123	13 157	7,7	2,0	2,2	2,0
622	56	5 516	10,4	1,3	3,1	3,0
722	95	11 648	9,9	1,9	2,4	2,4
131	36	4 366	16,9	5,1	2,2	2,1
231	28	3 642	19,4	3,7	2,8	2,8
331	16	2 943	36,0	3,2	3,2	3,3
431	28	2 234	11,4	4,1	3,4	3,2
531	29	3 526	17,8	4,8	3,0	3,0
631	12	1 770	34,0	1,9	3,8	3,8
731	18	3 857	41,0	3,3	3,1	3,2
132	39	5 200	17,6	3,8	2,1	2,2
232	34	5 124	20,0	3,0	2,6	2,7
332	17	3 644	40,8	1,8	3,0	3,1
432	24	2 371	16,1	2,8	3,3	3,3
532	34	4 101	16,0	3,3	2,8	2,9
632	18	2 314	24,5	2,4	3,5	3,5
732	48	5 004	11,3	3,7	3,0	2,9

Taulukko 6.9: Soluittaiset prosenttiset variaatiokertoimet työttömyysasteille.

Solu	n_{dgs}	N_{dgs}	c.v. HT	c.v. RS	c.v. MRE	c.v. SSD
111	48	6 201	24,9	6,7	10,8	10,2
211	55	5 757	30,4	10,3	10,9	11,5
311	48	5 361	24,9	9,4	11,0	10,8
411	14	1 665	51,0	4,8	14,4	14,4
511	48	4 757	23,6	10,4	10,8	10,4
611	26	2 147	24,7	7,7	15,2	13,7
711	29	5 286	20,5	6,9	10,9	9,7
112	63	6 990	32,9	15,1	14,3	15,1
212	60	6 389	38,6	16,7	13,7	15,3
312	55	5 923	30,4	15,1	13,2	13,6
412	12	1 660	64,3	6,8	19,7	19,6
512	32	4 696	36,6	8,6	13,5	13,4
612	18	2 139	66,4	9,6	20,4	20,9
712	35	5 201	41,2	11,7	14,3	14,8
121	123	12 706	21,7	21,7	13,3	12,8
221	106	12 648	25,9	18,9	13,5	14,3
321	73	10 800	23,9	20,2	13,5	12,7
421	68	5 579	38,8	16,9	18,6	21,2
521	108	12 488	21,4	20,4	12,5	12,3
621	40	5 186	34,2	14,9	21,3	20,3
721	87	10 980	26,7	19,9	13,8	14,7
122	106	13 674	28,4	19,0	21,1	18,9
222	135	14 826	28,8	20,2	20,0	20,4
322	98	10 959	36,3	16,6	18,1	21,8
422	38	5 840	37,3	13,2	30,3	26,2
522	123	13 157	27,3	19,2	17,6	18,9
622	56	5 516	35,2	12,8	33,1	29,4
722	95	11 648	28,2	18,6	19,3	18,8
131	36	4 366	98,2	51,4	12,0	14,1
231	28	3 642	46,1	36,0	14,8	15,1
331	16	2 943	51,8	29,3	14,4	14,4
431	28	2 234	36,0	39,3	17,8	16,9
531	29	3 526	46,2	45,8	13,5	14,0
631	12	1 770	64,3	16,4	20,0	20,0
731	18	3 857	66,4	31,4	14,6	15,2
132	39	5 200	68,6	38,3	18,4	20,4
232	34	5 124	36,9	29,4	21,2	19,2
332	17	3 644	52,2	18,8	18,8	18,3
432	24	2 371	53,8	26,7	28,1	27,4
532	34	4 101	36,9	31,7	18,5	17,5
632	18	2 314	66,4	25,5	29,6	29,6
732	48	5 004	26,4	33,1	19,4	16,5

Jos estimaatin keskivirhe on alle kymmenen prosenttia, niin tulosta voidaan pitää jo varsin hyvänä. Kuten jo keskivirheistäkin voidaan päätellä, ovat suoran estimaattorin soluittaiset variaatiokertoimet kaikille kolmelle tulosmuuttujalle selvästi muiden estimaattoreiden variaatiokertoimia korkeammat. Muutamissa soluissa jopa yli sata prosenttia. Ainoastaan soluittaisissa työllisten kokonaismäärissä variaatiokertoimet alittavat kaikissa soluissa kaikilla kolmella pienalue-estimaattorilla edellä mainitun kymmenen prosentin rajan. Myös työttömien kokonaismäärille synteettisen suhde-estimaattorin variaatiokertoimet ovat suurimmaksi osaksi alle kymmenen prosentin. Työttömyysasteen variaatiokertoimista taas lähes kaikki ovat yli kymmenen prosenttia. Selitystä tälle ilmiölle voidaan hakea osittain tulosmuuttujien ominaisuuksista. Työttömien ja työllisten kokonaismäärät voivat vaihdella periaatteessa nollasta äärettömiin, kun taas työttömyysasteen vaihteluväli on nollasta sataan. Yhden prosentin lisäys työttömyysasteessa on paljon merkittävämpää kuin esimerkiksi työllisten kokonaismäärässä.

Keskivirheitä ja variaatiokertoimia vertailemalla näyttäisi siltä, että synteettinen suhde-estimaattori olisi luotettavin vaihtoehto. Niinkuin aikaisemmin jo todettiin, keskineliövirheiden osalta näin ei kuitenkaan ole. Synteettisellä suhde-estimaattorilla on selvästi enemmän keskineliövirhettä kaikissa kolmessa tulosmuuttujassa kuin modifioidulla regressio- tai komposiittiestimaattorilla. Keskineliövirheen määrä kertoo myös osittain siitä, kuinka harhaisia estimaatit ovat. Koska mielestäni on tärkeämpää, että laskettujen estimaattien painopiste on oikeassa kohdassa, kuin se kuinka suuri on niiden vaihtelu tämän painopisteen ympärillä, voidaan synteettinen suhde-estimaattori tiputtaa käyttökelpoisista estimaattorivaihtoehdoista pois. Lisäksi synteettisen suhde-estimaattorin ongelmana on, että se homogenisoi liikaa saman ennusteen mukaan laskettuja estimaatteja, jolloin estimaattori ei reagoi riittävästi poikkeavuuksiin.

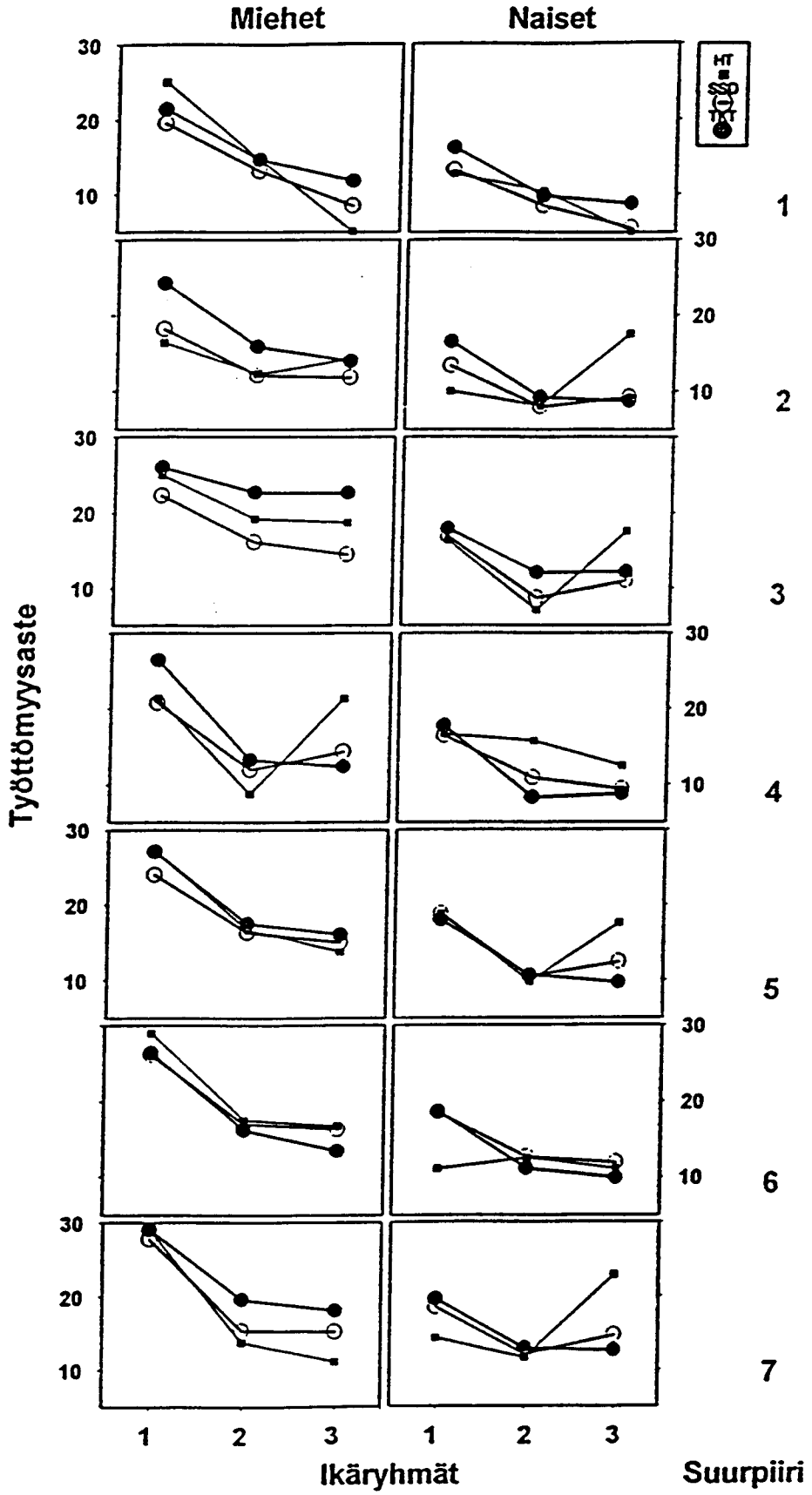
Jäljelle jäävät siis modifioitu regressioestimaattori ja komposiittiestimaattori, joka on suoran - ja regressioestimaattorin kombinaatio. Jos vertaillaan tarkemmin regressioestimaattorin ja komposiittiestimaattorin soluittaisia keskivirheitä ja variaatiokertoimia keskenään niin voidaan havaita, että komposiittiestimaattorilla ne

ovat suurimmasssa osassa soluista aina hiukan pienemmät. Poikkeuksena tästä ovat työllisten kokonaismäärien keskivirheet, joissa ainoastaan yhdessä solussa komposiittiestimaattorilla on pienempi keskivirhe kuin modifioidulla regressioestimaattorilla. Suhteutettaessa keskivirheet estimaatteihen tämä kuitenkin korjaantuu, koska variaatiokertoimia verrattaessa komposiittiestimaattorilla on enemmän pienemmän variaatiokertoimen omaavia soluja. Erot modifioidun regressioestimaattorin ja komposiittiestimaattorin välillä eivät tosin ole kovinkaan suuret. Ainoastaan malliin perustuvaan estimaattoriin sisältyy kuitenkin enemmän epävarmuustekijöitä kuin komposiittiestimaattoriin, johon on yhdistetty kahden eri tyyppiä edustavan estimaattorin tuloksia. Näiden tulosten perusteella mielestäni paras pienalue-estimaattori on suoran estimaattorin ja modifioidun regressioestimaattorin yhdistelmä. Seuraavaksi tarkastellaan vielä soluittaisia piste-estimaatteja ja verrataan niitä rekisteripohjaisesta työssäkäyntitilastosta saatuihin ns. korjattuihin oikeisiin arvoihin.

6.3 Piste-estimaatit

Laskettujen piste-estimaattien ja oikeiden arvojen vertailussa on oltava varovainen, koska ns. oikeita absoluuttisia työvoimatiedustelun mukaisia arvoja ei ole olemassakaan. Niiden merkitys vertailuissa on lähinnä suuntaa antava.

Työttömien, työllisten ja työttömyysasteiden soluittaiset piste-estimaatit kaikille neljälle estimaattorille sekä oikeat arvot ovat taulukoina liitteessä kolme. Taulukoista voidaan havaita, että estimoinnin onnistuminen suhteessa oikeaan arvoon vaihtelee estimaattoreittain sekä samankin estimaattorin tuloksissa. Jopa suoran estimaattorin tulokset näyttävät osuvan kohdalleen varsinkin työttömien kokonaismääriä ja työttömyysasteita estimoitaessa. Kuvassa 6.1 on havainnollistettu tätä tulosten vaihtelua suoran estimaattorin ja kombosiittiestimaattorin soluittaisten työttömyysasteiden estimaateilla. Yhdessä ruudussa on kuvattu yhden suurpiirin ja toisen sukupuolen piste-estimaatit ja oikeat arvot kolmelle ikäryhmälle. Suoran estimaattorin merkkinä on tumma neliö, komposiittiestimaattorin valkoinen ympyrä ja oikean arvon musta ympyrä.



Kuva 6.1: Suorat estimaatit, komposiittiestimaatit ja oikeat arvot työttömyysasteille soluittain.

Ruudukot osoittavat selvästi sen, mikä ero suoran - ja komposiittiestimaattorin tuloksissa on. Vaikka suora estimaatti näyttää joissakin soluissa osuvan paremmin kohdalleen kuin komposiittiestimaatti, niin komposiittiestimaattorin tulokset ovat yleisesti ottaen parempia ja vakaampia. Oikeiden arvojen ja komposiittiestimaattien viivat ovat keskenään paljon yhdenmukaisemmat verrattuna suorien estimaattien viivoihin. Otokseen tulleiden havaintojen ominaisuudet ja edustavuus vaikuttavat herkemmin suorien estimaattoreiden tuloksiin. Tämä on toisaalta hyvä, toisaalta huono asia. Poikkeavuus otoksessa saattaa olla merkkinä myös todellisesta poikkeavuudesta perusjoukossa. Puhtaasti mallipohjaiset estimaattorit reagoivat poikkeavuuksiin vasta silloin kun erot ovat todella merkittävät. Komposiittiestimaattori operoi näiden kahden välimaastossa. Tosin tässä tapauksessa painotus on enemmän mallipohjaisessa estimaatissa, koska enemmistössä soluista on alle sata havaintoa.

Ikäryhmittäisessä tarkastelussa voidaan havaita, että keskimmäisessä ikäluokassa estimointi osuu kohdalleen kaikkein parhaiten. Tämä tulos on varmasti osittain seurausta siitä, että keskimmäisessä ikäluokassa on myös eniten otoshavaintoja. Ensimmäisen ja kolmannen ikäluokan välillä ei vastaavaa eroa ole. Kaikissa suurpiireissä ja molemmissa sukupuoliryhmissä ensimmäisen ikäryhmän työttömyysasteet ovat korkeammat kuin kolmannessa tai neljännessä ikäryhmässä.

Kolmannen suurpiirin miehien työttömyysasteiden estimaattien ja oikeiden arvojen muodostama viivakuvio on ainoa, jossa kahden eri estimaattin ja oikean arvon viivat eivät leikkaa toisiaan. Toisen poikkeavuutena on, että suoran estimaatit ovat kaikissa kolmessa ikäryhmässä lähempänä oikeita arvoja kuin komposiittiestimaatit. Yhtenä selityksenä tälle ilmiölle voisi olla kadon määrä kyseisissä soluissa (Liite 1). Esimerkiksi kolmannen suurpiirin toisen ikäryhmän miehillä vastauskato on kaikista suurin eli yli 27 prosenttia. Myös muilla kolmannen suurpiirin miehillä vastauskadotkadot ovat yli 10 prosentin luokkaa. Suuresta katoprosentista näyttäisi olevan seurauksena hienoinen aliestimointi. Varsinkin miehillä ja erityisesti suurpiireissä yksi, kaksi ja kolme ovat oikeiden arvojen mustat ympyrät lähes säännönmukaisesti estimaattien mustia neliöitä ja valkoisia ympyröitä alempana.

7 Yhteenveto ja johtopäätökset

Tutkimuksen tarkoituksena on ollut selvittää, onko Tilastokeskuksen työvoimatutkimuksesta mahdollista tuottaa otospohjaisia neljännesvuosittaisia työvoimatilastoja Helsingistä suurpiireittäin, ikäryhmittäin ja sukupuolittain. Perinteiset suorat estimaattorit eivät tässä tilanteessa riitä luotettavien ja vakaiden tilastolukujen laskemiseksi. Tavoitteen saavuttamiseksi on tutkittu erilaisia pienalue-estimaattoreita ja sovellettu niitä tutkimusaineistoon. Samalla on pyritty määrittämään mahdollisimman yhdenmukaiset merkintätavat eri tutkijoiden kehittämille estimaattoreille sekä löytämään pienalue-estimointiin liittyvälle termistölle sopivat suomenkieliset vastineet.

Tulosten perusteella voidaan mielestäni todeta, että työvoimaa koskevien otospohjaisten pienaluetilastojen tuottaminen suurpiiritasolla kuuteen demografiseen ryhmään luokiteltuna on työvoimatutkimuksen perusteella mahdollista. Myöskään peruspiireittäiset pienalue-estimaatit eivät ole täysin poissuljettu vaihtoehto. Tosin ainoastaan yhden ajankohdan aineistoon perustuvien tutkimustulosten perusteella on uhkarohkeata tehdä liian pitkälle meneviä johtopäätöksiä. Varsinkin kun kyseessä on niinkin ajankohtainen ja tarkkaan seurattu aihe, kuin työllisyys ja työttömyys.

Tutkituista pienalue-estimaattoreista parhaimmaksi todettiin komposiittiestimaattori, joka muodostuu suoran estimaattorin ja modifioidun regressioestimaattorin painotetusta yhdistelmästä. Työttömyyden ja työllisyyden mallintamisessa ei sen sijaan vielä parasta vaihtoehtoa löytynyt. Yllättävää tähän astisissa tuloksissa on se, ettei alue näyttäsi olevan merkittävä selittäjä työttömyydelle tai työllisyydelle, vaikka

työttömyyden alueellisten erojen on todettu Helsingissä kasvaneen. Mallin C mukaisesti merkittäviä selittäjiä näyttäisi olevan ainakin ikäryhmä, sukupuoli ja siviilisääty. Jatkotutkimuksia ajatellen on myös huomioitava, että vaikka alue ei olekaan tässä vuoden 1994 ensimmäisen neljänneksen aineistossa merkittävä selittäjä, niin tulevaisuudessa saattaa olla toisin. Nyt hyväksi havaittu malli ei ole sitä välttämättä enää seuraavana vuonna. Tämän asian selvittämiseksi tulisi muutaman hyväksi havaitun mallin sopivuutta kokeilla eri ajankohdan otoksiin. Lopullisen estimaatin laskemiseksi tarvitaan rekisteriaineistosta aina mallissa olevien selittävien muuttujien luokitusten mukaiset tiedot.

Keskivirheiden ja variaatiokertoimien osalta tulokset eivät mielestäni aivan vielä saavuta julkaisukelpoista tasoa. Vaikka soluittaisten työttömien ja työllisten kokonaismäärien kohdalla asiaa voisi jo harkita, niin työttömyysasteiden estimaattien keskivirheet ja siten myös luottamusvälit ovat vielä liian korkeat. Keskivirheisiin ja sitä kautta myös variaatikertoimiin voidaan vaikuttaa kahdella tavalla. Modifioidun regressioestimaattorin mallilla, jonka pitää olla hyvä ja lisäksi mallin kunkin ennusteen takana tulee olla mieluummin yli sadan havainnon määrä. Toisena vaihtoehtona on otoskoon kasvattaminen. Tämä ei kuitenkaan välttämättä vielä takaa, että kaikista soluista saataisiin riittävä määrä havaintoja. Tärkeämpää olisi allokoida Helsingin osalta otos uudelleen niin, että kaikkiin soluihin tulisi lähes yhtä paljon havaintoja. Tällöin otantamenetelmäksi tulisi tasakiintiöity ositettu otanta, joka vaikuttaisi myös kaavojen muotoon jonkin verran.

Liitteessä kolme esitettyihin solujen oikeisiin arvoihin tulee suhtautua erittäin kriittisesti. Oletus siitä, että koko maassa työsuhderekisterin ja työssäkäyntitilaston suhde olisi sama kuin Helsingissä on liian yleistävä. Helsingin työmarkkinoilla ja työvoimalla on paljon omia erityispiirteitä, jotka poikkeavat suurestikin verrattuna muuhun osaan Suomesta. Esimerkiksi oleellisesti suuremmat työmarkkinat, painotus tietyille toimialoille, koulutettu työvoima, kaupunkityöttömyyden erityispiirteet ja muuttoliike, jolloin työpaikkojen lisääntyessä työttömyys ei välttämättä laske, koska Helsinkiin muuttavat sijoittuvat näille uusille työpaikoille. Onnistuneen estimoinnin kannalta on tärkeää tuntea myös tutkittavan ilmiön taustalla olevat tekijät ja ottaa ne huomioon estimaattoreita kehiteltäessä.

Tutkimuksen päämäärän saavuttamiseksi on avainasemassa sopivimman mallin löytäminen ja sen kehittäminen niin, että estimoinnissa pystyttäisiin huomioimaan perusjoukossa tapahtuvat muutokset ajanhetkestä toiseen. Jatkotutkimusten kannalta on olennaista myös se, millaiset tarkkuus- ja laatuvaatimukset otospohjaisesti tuotetuille tilastoille asetetaan. Vertailukohteena voidaan käyttää muita estimoimalla tuotettuja lukuja. Todennäköistä on, että mitä pienempiin osa-alueisiin aineisto jaetaan, sitä enemmän tarkkuudesta joudutaan tinkimään. Oletuksena on, ettei otoskoko ratkaisevasti muutu.

Suomessa ei ole aikaisemmin tehty tai ainakaan yleisesti julkaistu pienalue-estimointiin liittyviä empiirisiä tutkimuksia. Tutkimuksen tavoitteena onkin ollut ns. pioneerityönä tuoda esille pienalue-estimaattoreita ja mahdollisesti edistää niiden hyväksikäyttöä myös Suomessa. Tilastokeskuksen rooli on tässä työssä nyt ja jatkossakin aineistojen kerääjänä, toimittajana ja yhtestyökumppanina tärkeä. Toivottavasti tästä työstä on hyötyä myös Tilastokeskukselle, heidän kehittäessään omia tuotteitaan ja palvelujaan.

Lähdeluettelo

Cassel, C.- M., Kristiansson, K.- E., Råbäck, G, & Wahlström, S. (1987). "Using Model-Based Estimation to Improve the Estimate of Unemployment on a Regional Level in the Swedish Labor Force Survey". Teoksessa Platek, R. Singh, M., Rao, J.N.K. and Särndal, C.-E., (toim.). *Small Area Statistics. An International Symposium*, sivut 141-159 Wiley; New York.

Feeney, G. A. (1987). "The Estimation of the Number of Unemployed at the Small Area Level". Teoksessa Platek, R. Singh, M., Rao, J.N.K. and Särndal, C.-E., (toim.). *Small Area Statistics. An International Symposium*, sivut 198-218 Wiley; New York.

Ghosh, M. & Rao, J. N. K. (1994). Small Area Estimation: An Appraisal. *Statistical Science*, 1, sivut 55 - 93.

Helsingin kaupungin tietokeskus (1994). *Helsingin kaupungin tilastollinen vuosikirja 1994.* Gummerus Kirjapaino Oy; Jyväskylä.

Helsingin kaupungin tietokeskus (1996). *Helsingin kaupungin tilastollinen vuosikirja 1996.* Gummerus Kirjapaino Oy; Jyväskylä.

Helsingin kaupungin tietokeskuksen tilastoja (1996:14). *Helsingin väestöennuste 1997-2020.* Helsinki: Hakapaino Oy.

Helsingin kaupungin tietokeskuksen tilastoja (1997:4). *Helsingin väestö vuodenvaihteessa 1996/97 ja väestönmuutokset vuonna 1996.* Helsinki: Hakapaino Oy.

International Labour Office (1990). *Surveys of Economically Active Population, Employment, Unemployment and Underemployment.* Geneva.

Lehtonen, R. & Pahkinen, E. (1995). *Practical Methods for Design and Analysis of Complex Surveys.*West Sussex: John Wiley & Sons, Ltd.

Pahkinen, E. & Lehtonen, R. (1989). *Otanta-asetelmat ja tilastollinen analyysi.* Helsinki: Painokaari Oy.

Purcell, N. J., & Kish, L. (1979). Estimation for Small domains. *Biometrics*, 35, sivut 365-384.

Rao, J.N.K. and Choudry, G.H. (1995)."Small-Area Estimation: Overview and Empirical Study". Teoksessa Cox, B. G. & Binder, D. A. & Chinnappa, B. N. & Cristianson, A. & Colledge, M. J. & Kott, P. S. (toim.) *Business Survey Methods*, sivut 527 - 543. New York: John Wiley & Sons, Inc.

Robinson, G. K. (1991). That BLUP is a Good Thing: The Estimation of Random Effects. *Statistical Science*, 1, sivut 15 - 51.

Rosén, B. (1991, October). *Synthetic Estimation - An Approch to Small Area Statistics.* Handout, Seminar of Small Area Statistics, Helsinki.

SAS Institute Inc. (1987). SAS/STAT™ Guide for Personal Computer, Version 6 Edition. Cary, NC: SAS Institute Inc.

Singh, A. C. & Mantel, H. J. & Thomas, B. W. (1994). Time Series EBLUPs for Small Areas Using Survey Data. *Survey Methodology,* 1, sivut 33 - 43.

Singh, A. C. & Mian, I. U. H. (1995). Generalized sample size dependent estimators for small areas. *ARC Proceedings,* sivut 687 - 701.

Stukel, D. (1991). *Small Area Estimation Under One and Two-Fold Nested Error Regression Models.* Ph.D. Thesis, Ottawa: Carleton University.

Särndal, C.-E. & Hidiroglou, M. A. (1989). Small Domain Estimation: A Conditional Analysis. *Journal of the American Statistical Association,* 405, sivut 266-275.

Särndal, C.-E., Swensson, B. & Wretman, J. (1992). "Estimation for Domains". *Model Assisted Survey Sampling,* sivut 386 - 417. New York: Springer-Verlag, Inc.

Tilastokeskus (1993). *Työvoimatilasto. Tietoja työllisyydestä, työttömyydestä ja työvoimasta vuosilta 1960 - 1991.* Työmarkkinat 1993:7. Helsinki: Hakapaino Oy.

Tilastokeskus (1994). *Työvoimatilasto 1993. Työvoimatutkimuksen tuloksia vuosilta 1980 - 1993.* Työmarkkinat 1994:14. Helsinki: Paino-Center Oy.

Tilastokeskus (1996). *Työssäkäyntitilasto 1993-1994.* Väestö 1996:5. Helsinki: Hakapaino Oy.

Tilastokeskus (1997). *Työttömyys ja työllisyys tilastoissa. Tilastokeskuksen ja Työministeriön tilastojen vertailua.* Työmarkkinat 1997:3. Helsinki.

LIITE 1

Soluittaiset otoskoot ja absoluuttiset sekä suhteelliset kadot.

Solu	Otoskoko	Ei vastanneet	Kato%
111	108	23	21,3
211	98	11	11,2
311	85	12	14,1
411	35	3	8,5
511	84	8	9,5
611	56	5	8,9
711	69	13	18,8
112	118	12	10,1
212	113	11	9,7
312	91	7	7,6
412	28	3	10,7
512	68	6	8,8
612	28	3	10,7
712	79	16	20,2
121	161	31	19,2
221	137	26	18,9
321	114	31	27,1
421	76	6	7,8
521	130	18	13,8
621	48	7	14,5
721	117	23	19,6
122	149	24	16,1
222	171	21	12,2
322	124	19	15,3
422	48	6	12,5
522	156	17	10,9
622	63	4	6,3
722	133	20	15,0
131	88	14	15,9
231	74	3	4,0
331	56	9	16,0
431	57	5	8,7
531	66	9	13,6
631	45	3	6,6
731	71	10	14,0
132	99	11	11,1
232	125	14	11,2
332	67	12	17,9
432	58	2	3,4
532	95	11	11,5
632	61	5	8,2
732	115	15	13,0

LIITE 2

SAS-ohjelmalauseet mallille B.

```
data a3;                                  /* aineiston syöttö          */
input alue ikal toilaa sp count;
cards;
1 1 0 1  87
1 1 0 2 102
1 1 1 1  38
1 1 1 2  20
1 2 0 1 224
1 2 0 2 286
1 2 1 1  44
1 2 1 2  30
1 3 0 1  54
1 3 0 2  79
1 3 1 1   9
1 3 1 2  20
2 1 0 1 109
2 1 0 2 135
2 1 1 1  34
2 1 1 2  18
2 2 0 1 293
2 2 0 2 300
2 2 1 1  44
2 2 1 2  35
2 3 0 1  91
2 3 0 2 102
2 3 1 1  13
2 3 1 2  13
;

proc catmod;                              /* log-lineaarinen malli     */
response 1 0;
weight count;
model toilaa = alue ikal sp / freq prob nodesign predict;
run;
quit;
```

LIITE 3

Soluittaiset piste-estimaatit ja oikeat arvot työttömien kokonaismäärille.

Solu	n_{dgs}	Y_{dgs}	HT	RS	MRE	SSD
111	48	1 329	1 433	1 979	1 101	1 181
211	55	1 395	1 075	1 686	1 088	1 084
311	48	1 396	1 433	1 548	1 147	1 215
411	14	440	358	455	345	346
511	48	1 289	1 552	1 322	1 094	1 204
611	26	567	1 194	608	521	608
711	29	1 595	1 552	1 423	1 314	1 348
112	63	1 133	955	1 029	937	943
212	60	1 057	716	891	939	872
312	55	1 059	1 075	844	1 015	1 031
412	12	296	239	222	274	272
512	32	847	716	605	881	854
612	18	397	239	283	411	396
712	35	1 018	597	660	1 008	936
121	123	1 850	2 149	2 033	1 348	1 841
221	106	2 015	1 552	1 873	1 509	1 532
321	73	2 452	1 671	1 255	1 554	1 597
421	68	743	716	965	759	745
521	108	2 184	2 149	1 684	1 988	2 075
621	40	839	836	807	875	867
721	87	2 150	1 433	1 387	1 833	1 659
122	106	1 343	1 313	1 500	881	1 110
222	135	1 369	1 313	1 503	1 150	1 260
322	98	1 326	836	1 067	1 107	974
422	38	487	716	622	566	595
522	123	1 408	1 433	1 220	1 543	1 475
622	56	615	836	547	699	738
722	95	1 500	1 313	1 040	1 461	1 391
131	36	516	119	715	424	369
231	28	508	478	501	414	423
331	16	667	358	287	416	411
431	28	276	716	352	295	354
531	29	567	478	377	536	528
631	12	237	239	264	287	285
731	18	700	239	458	601	568
132	39	457	239	875	298	286
232	34	445	716	807	383	440
332	17	446	358	528	377	375
432	24	209	358	380	212	230
532	34	397	716	565	462	506
632	18	229	239	361	277	273
732	48	628	1 313	709	605	775

Soluittaiset piste-estimaatit ja oikeat arvot työllisten kokonaismäärille.

Solu	n_{dgs}	Y_{dgs}	HT	RS	MRE	SSD
111	48	5 072	4 298	5 387	5 100	4 907
211	55	4 571	5 492	4 588	4 669	4 895
311	48	4 175	4 298	4 213	4 214	4 234
411	14	1 291	1 313	1 240	1 320	1 319
511	48	3 662	4 179	3 599	3 663	3 787
611	26	1 665	1 910	1 654	1 626	1 663
711	29	3 931	1 910	3 873	3 972	3 673
112	63	6 027	6 567	6 420	6 053	6 215
212	60	5 491	6 447	5 556	5 450	5 749
312	55	5 023	5 492	5 261	4 908	5 069
412	12	1 409	1 194	1 387	1 386	1 374
512	32	3 976	3 104	3 771	3 815	3 702
612	18	1 802	1 910	1 764	1 728	1 744
712	35	4 335	3 582	4 116	4 193	4 086
121	123	11 133	12 536	11 946	11 358	12 083
221	106	10 936	11 103	11 006	11 139	11 120
321	73	8 716	7 044	7 374	9 246	8 442
421	68	4 948	7 402	5 668	4 820	5 698
521	108	10 632	10 745	9 894	10 500	10 632
621	40	4 473	3 940	4 742	4 311	4 237
721	87	9 153	8 954	8 148	9 147	9 063
122	106	12 532	11 342	13 520	12 793	12 024
222	135	13 662	14 805	13 548	13 676	14 438
322	98	9 832	10 865	9 623	9 852	10 348
422	38	5 426	3 821	5 603	5 274	4 997
522	123	11 960	13 253	10 998	11 614	12 622
622	56	4 994	5 850	4 932	4 817	5 106
722	95	10 373	10 029	9 380	10 187	10 112
131	36	3 927	4 179	4 715	3 942	3 984
231	28	3 210	2 865	3 302	3 228	3 177
331	16	2 376	1 552	1 894	2 527	2 449
431	28	1 999	2 627	2 323	1 939	2 035
531	29	3 044	2 985	2 483	2 990	2 989
631	12	1 568	1 194	1 738	1 483	1 465
731	18	3 262	1 910	3 020	3 256	3 135
132	39	4 811	4 418	4 800	4 902	4 808
232	34	4 745	3 343	4 427	4 741	4 503
332	17	3 265	1 671	2 894	3 267	3 132
432	24	2 193	2 507	2 086	2 159	2 201
532	34	3 764	3 343	3 100	3 639	3 588
632	18	2 120	1 910	1 980	2 037	2 026
732	48	4 470	4 418	3 890	4 399	4 403

Soluittaiset piste-estimaatit ja oikeat arvot työttömyysasteille.

Solu	n_{dgs}	Y_{dgs}	HT	RS	MRE	SSD
111	48	21,4	25,0	31,9	17,8	19,5
211	55	24,2	16,4	29,3	18,9	18,2
311	48	26,0	25,0	28,9	21,4	22,3
411	14	26,4	21,4	27,4	20,7	20,8
511	48	27,1	27,1	27,8	23,0	24,0
611	26	26,4	38,5	28,3	24,3	26,1
711	29	30,2	44,8	26,9	24,9	27,7
112	63	16,2	12,7	14,7	13,4	13,2
212	60	16,5	10,0	13,9	14,7	13,3
312	55	17,9	16,4	14,2	17,1	16,9
412	12	17,8	16,7	13,4	16,5	16,5
512	32	18,0	18,8	12,9	18,8	18,8
612	18	18,6	11,1	13,2	19,2	18,5
712	35	19,6	14,3	12,7	19,4	18,5
121	123	14,6	14,6	16,0	10,6	13,1
221	106	15,9	12,3	14,8	11,9	12,1
321	73	22,7	19,2	11,6	14,4	16,1
421	68	13,3	8,8	17,3	13,6	12,0
521	108	17,5	16,7	13,5	15,9	16,3
621	40	16,2	17,5	15,6	16,9	17,0
721	87	19,6	13,8	12,6	16,7	15,4
122	106	9,8	10,4	11,0	6,4	8,5
222	135	9,2	8,1	10,1	7,8	8,0
322	98	12,1	7,1	9,7	10,1	8,7
422	38	8,3	15,8	10,6	9,7	10,9
522	123	10,7	9,8	9,3	11,7	10,5
622	56	11,1	12,5	9,9	12,7	12,6
722	95	12,9	11,6	8,9	12,5	12,1
131	36	11,8	2,8	16,4	9,7	8,5
231	28	14,0	14,3	13,8	11,4	11,8
331	16	22,7	18,8	9,8	14,1	14,5
431	28	12,4	21,4	15,8	13,2	14,4
531	29	16,1	13,8	10,7	15,2	15,0
631	12	13,4	16,7	14,9	16,2	16,3
731	18	18,1	11,1	11,9	15,6	15,2
132	39	8,8	5,1	16,8	5,7	5,6
232	34	8,7	17,6	15,8	7,5	9,2
332	17	12,2	17,6	14,5	10,3	11,0
432	24	8,8	12,5	16,0	9,0	9,4
532	34	9,7	17,6	13,8	11,3	12,4
632	18	9,9	11,1	15,6	12,0	11,9
732	48	12,6	22,9	14,2	12,1	14,7