Tilastokeskuksen työvoimatutkimuksen tehostaminen tasapainoisella otannalla

Arvi Tolvanen

Pro gradu -tutkielma
Toukokuu 2026

MATEMATIIKAN JA TILASTOTIETEEN LAITOS

Tarkastajat:
Prof. Henri Nyberg
Apul.prof. Joni Virta

TURUN YLIOPISTO, Matematiikan ja tilastotieteen laitos

Pro gradu -tutkielma
Pääaine: Tilastotiede
Tekijä: Arvi Tolvanen
Otsikko: Tilastokeskuksen työvoimatutkimuksen tehostaminen tasapainoisella otannalla
Ohjaajat: Prof. Henri Nyberg ja Tilastokeskuksessa menetelmäasiantuntijat
Riku Salonen, Pauli Ollila ja Henri Luomaranta-Helmivuo
Sivumäärä: 50 sivua
Aika: Toukokuu 2026

Suomessa viranomaisrekisterit ovat hyvin kattavia ja niiden perusteella on mahdollista tehdä hyvin laajaa tilastointia ja tutkimusta. Kuitenkaan kaikkea tietoa ei löydy rekistereistä, vaan se täytyy hankkia kyselytutkimuksella. Esimerkkitapauksena Tilastokeskuksen työvoimatutkimus, jolla pyritään selvittämään työllisten ja työttömien määrä Suomessa.

Tässä tutkielmassa tarkastellaan rekisteripohjaista otantatutkimusta eri otantamenetelmiä vertaillen. Vertailumenetelminä ovat perinteiset yksinkertainen satunnaisotanta ja systemaattinen otanta implisiittisellä osituksella sekä tuoreempi tasapainoinen otanta (balanced sampling) kuutiomenetelmällä (the cube method). Otantamenetelmän valinnalla voi olla vaikutusta tarkasteltavan suureen estimaattorin keskihajontaan. Implisiittinen ositus ja systemaattinen otanta pyrkivät huomioimaan otosta valittaessa perusjoukon taustamuuttujien ominaisuuksia. Tasapainoinen otanta pyrkii ottamaan tätä laajemmin huomioon taustamuuttujat. Lisäksi tarkastellaan myös kalibrointiestimaatteja, jotka perustuvat siihen, että valittua otosta painotetaan jälkikäteen painokertoimilla vastaamaan halutun perusjoukon taustamuuttujia.

Perusjoukon taustamuuttujien yhteyttä työllisyyteen ja työttömyyteen tarkastellaan logistisilla regressiomalleilla pääasiallisesti Tjurin pseudoselitysastetta käyttäen. Valitut taustamuuttujat selittävät erityisen hyvin työllisyyttä ja hieman heikommin työttömyyttä. Mitä pidempi aikaväli otoksen valinnan ja haastatteluhetken välillä on, sitä heikompi selitysaste on. Erityisesti työttömyyden osalta selitysaste on alkujaankin heikohko ja heikkenee voimakkaasti ajallisen yhteyden pidentyessä.

Tutkielmassa havaitaan, että tasapainoinen otanta pienentää merkittävästi työllisyysasteen estimaattorin keskihajontaa ja työttömyysasteen tapauksessa hieman vähemmän verrattuna systemaattiseen otantaan. Kalibroinnin jälkeen otantamenetelmien välillä ei kuitenkaan ole olennaista tilastollista eroa estimaattoreiden keskihajonnassa. Kuitenkin tasapainoisen otannan otoksissa kalibrointipainokertoimet ovat kevyempiä kuin muissa otantamenetelmissä.

Asiasanat: otantatutkimus, työvoimatutkimus, tasapainoinen otanta, balanced sampling, Tilastokeskus.

Kiitokset

Kiitän suuresti Tilastokeskusta graduharjoittelusta ja mahdollisuudesta toteuttaa tutkielma aidolla aineistolla. Erityiskiitos Tilastokeskuksen menetelmäasiantuntijoille Riku Saloselle, Henri Luomaranta-Helmivuolle ja Pauli Ollilalle työn menetelmällisestä ohjauksesta, erinomaisista ideoista sekä arvokkaista huomioista tutkielman eri vaiheissa. Kiitän myös työvoimatutkimuksen vastuuhenkilö yliaktuaari Pertti Taskista perehdyttämisestä työvoimatutkimukseen. Lisäksi kiitän Tilastokeskuksen Jukka Hoffrénia ja Reija Heleniusta harjoittelun mahdollistamisesta. Turun yliopiston tilastotieteen professori Henri Nybergille kiitän tutkielman akateemisesta ohjauksesta sekä kärsivällisyydestä läpi vuosien.

Turussa Flooran päivänä 2026

Sisällys

1 Johdanto

Suomessa viranomaisten ylläpitämät rekisteriaineistot muodostavat poikkeuksellisen kattavan ja laadukkaan tietopohjan. Rekisterit mahdollistavat väestön rakenteen ja monien ilmiöiden seurannan tarkasti ja kustannustehokkaasti, ja niiden yhdistely tarjoaa laajoja analyysimahdollisuuksia. Kaikesta kattavuudestaan huolimatta rekisterit eivät yksinään riitä kattamaan kaikkea tiedontarvetta. Esimerkiksi työmarkkinailmiöt, kuten henkilön työllisyys- ja työttömyysstatus tiettynä ajankohtana, ovat luonteeltaan sellaisia, että ne edellyttävät henkilöiltä itseltään kysymistä. Tällöin on kyse kyselytutkimuksesta.

Rekisteritutkimus mahdollistaa yleensä perusjoukon kaikki alkiot kattavan kokonaistutkimuksen. Tilanteissa, joissa rekisteriaineistoa ei ole saatavilla, tai se on pahasti puutteellinen, otostutkimus on kuitenkin ainoa järkevä tapa saada tietoa haluttavasta ilmiöstä. Perusjoukkoa koskevien päätelmien tekeminen otostutkimuksen perusteella edellyttää, että otos on edustava satunnaisotos suhteessa perusjoukkoon. Edustavasta satunnaisotoksesta saatavia tuloksia voidaan yleistää tilastollisen päättelyn avulla perusjoukkoon.

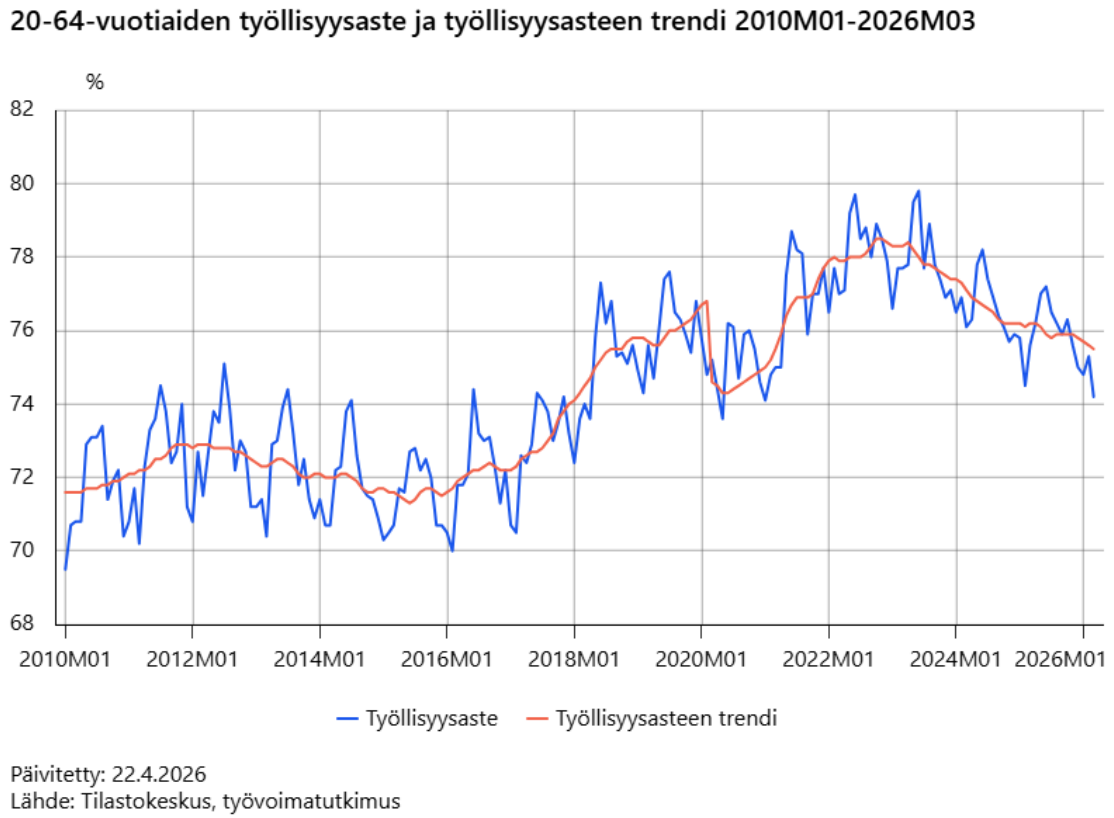
Tässä tutkielmassa tarkastellaan (osio 2) Tilastokeskuksen työvoimatutkimusta (LFS) ja pyritään etsimään keinoja parantaa sen otantaa. Erityisesti tarkastelussa on tasapainoinen otanta -menetelmä (*Balanced sampling*), joka Ranskan tilastoviranomaisen INSEE:n projekteissa on onnistunut pienentämään estimaattorin varianssia 20–90 % verrattuna yksinkertaiseen satunnaisotantaan (Deville & Tille, 2004). Tällä hetkellä otostutkimuksissa suurena haasteena ovat tiedonkeruun kustannukset. Vastausaktiivisuus on monessa ryhmässä laskussa, ja laatuvaatimukset tulee täyttää samalla, kun yleisesti budjetista pitäisi säästää. Otantamenetelmiä kehittämällä on teoriassa mahdollista pienentää estimaattorin varianssia, jolloin otoskoon pienentäminen olisi myös mahdollista. Tämä voisi pienentää tiedonkeruun kustannuksia.

Tasapainoinen otanta -menetelmä (osio 4) perustuu siihen, että otoksessa haluttujen muuttujien Horvitz-Thompson -estimaatit ovat samoja kuin perusjoukossa kyseisten muuttujien summat. Tällä tavoin otoksen taustamuuttujien keskiarvot vastaavat perusjoukon vastaavia odotusarvoja. Tasapainoinen otanta toteutetaan kuutiomenetelmällä (*the Cube Method*, osio 5), jossa pyritään martingaalioptimointia hyödyntäen löytämään tasapainoinen otos.

Tasapainoinen otanta toteutetaan otantavaiheessa eli otanta-asetelmassa. Taustamuuttujiin perustuva tasapainottaminen voidaan myös tehdä otannan jälkeen eli estimointiasetelmassa. Tätä kutsutaan kalibrointiestimaattoriksi. Siinä saatua otosta painotetaan painokertoimin siten, että se vastaa taustamuuttujiltaan haluttua perusjoukkoa (osio 6).

2 Työvoimatutkimus

Työvoimatutkimus (*labour force survey*) on Tilastokeskuksen laajasti seurattu ja yhteiskunnallisesti merkittävä tilasto, jossa selvitetään pääosin kyselytutkimusta hyödyntäen työllisten ja työttömien määrä Suomessa kuukausi-, neljännesvuosi- ja vuositasolla. [1] Tärkeimmät mittarit ovat työllisyys- ja työttömyysaste (kuva 1), jotka ovat laajasti mediassa esillä. Työvoimatutkimus on osa Suomen virallista tilastoa (SVT), ja se täyttää sen laatukriteerit. Työvoimatutkimus ja sen toteutus perustuvat Euroopan parlamentin ja neuvoston sosiaalitilastojen puiteasetukseen 2019/1700, komission täytäntöönpanoasetukseen 2019/2240 sekä Euroopan komission asetukseen 377/2008. Näiden perusteella työvoimatutkimus on vertailukelpoinen eri maiden välillä. Eri maiden työvoimatutkimuksen toteutuksen piirteitä voi tarkastella Euroopan Komission julkaisusta *Labour force survey in the EU, EFTA and candidate countries: main characteristics of national surveys: 2024 edition*. [2]



Kuva 1: 20-64-vuotiaiden työllisyysaste

Työvoimatutkimuksen perusjoukkona on 15–89-vuotiaat henkilöt, jotka asuvat vakituisesti Suomessa, ovat tilapäisesti ulkomailla alle vuoden tai ovat rekisteröityneet Suomen väestötietojärjestelmään ja oleskelevat Suomessa vähintään vuoden. Perusjoukko on määritelty Tilastokeskuksen kuukausittaisesta Digi- ja väestövirasto (DVV) väestötietojärjestelmään perustuvasta väestökehikosta. Perusjoukon koko on noin 4,8 miljoonaa henkilöä, ja joka kuukausi haastatellaan noin 12 500 henkilöä. Työvoimatutkimuksessa on käytössä rotaatiohaastattelupaneeli, jossa jokaista

haastateltavaa haastatellaan tutkimuksen aikana viisi kertaa. Haastattelut ovat kolmen kuukauden välein, paitsi kolmannen ja neljännen haastattelun välillä on kuusi kuukautta. Joka kuukausi haastatellaan ensimmäisen kerran noin 2600 henkilöä. Rotaatioasetelmasta seuraten joka kuukausi haastateltavina on ensimmäisen, toisen, kolmannen, neljännen ja viidennen haastattelukerran henkilöitä.

Otantamenetelmä on kaksivaiheinen. Ensin ositetaan perusjoukko kolmeen ositteeseen, 14–75-vuotiaiden osalta Manner-Suomeen ja Ahvenanmaan maakuntaan sekä 75–89-vuotiaat seniorit omassa ositteessaan. Näistä ositteissa perusjoukko järjestetään asuinpaikkatunnuksen ja henkilötunnuksen mukaan. Tästä järjestetystä perusjoukosta valitaan otos systemaattisella otannalla. Järjestyksestä johtuen otos on implisiittisesti ositettu myös asuinpaikan ja iän suhteen. Saatu otos kalibroidaan, eli painotetaan vastaamaan perusjoukkoa sukupuolen, ikäryhmän, työnhakijarekisterin, referenssiviikon, pääasiallinen toiminta -rekisteritiedon, koulutusasteen, äidinkielen ja kuukausipalkan suhteen. [2]

3 Populaation ja otannan teoriaa

3.1 Perusjoukko

Otantatutkimuksen tarkoituksena on tehdä päätelmiä perusjoukosta U (populaatiosta) otoksen avulla. Perusjoukossa U on N kappaletta alkioita u_i, $i \in \{1, \ldots, N\}$. Perusjoukon alkioilla u_i on muuttujia, joiden avulla voidaan laskea erilaisia tunnuslukuja. Tällaisia muuttujia ovat esimerkiksi ikä tai tieto työttömyystilanteesta. Näiden avulla saatavia tunnuslukuja ovat vaikkapa keskimääräinen ikä tai työttömien kokonaislukumäärä.

Tarkastellaan joitakin perusjoukon tunnuslukuja. Aineistossa U muuttuja X saa reaalilukuisia arvoja x_i jokaisella alkiolla u_i.

Määritelmä 1. Muuttujan X summa T_X lasketaan seuraavasti:

$$T_X = \sum_{i \in \{1,\ldots,N\}} x_i.$$

Määritelmä 2. Muuttujan X *odotusarvo* μ_X lasketaan äärellisestä perusjoukosta U seuraavasti:

$$\mu_x = \frac{T_x}{N} = \frac{1}{N} \sum_{i \in \{1,\ldots,N\}} x_i.$$

Huomataan myös, että $T_X = N\mu_X$. Koska U on perusjoukko, odotusarvo vastaa perusjoukon keskiarvoa. Jos tunnusluku sen sijaan laskettaisiin perusjoukon osajoukosta (esimerkiksi otoksesta), kyseessä olisi keskiarvo, jota merkitään $\bar{x}$.

3.2 Otanta

Perusjoukko U on usein niin iso ettei ole mielekästä tai mahdollista tarkastella kaikkia sen alkioita. Tällöin tarkastelu rajoitetaan vain osaan sen alkioista. Osajoukko $O \subset U$ koostuu n kappaleesta, $n < N$, alkioita $o_i \in U$. On myös otantaa, jossa

otoskoko ei ole kiinteä, mutta tässä esityksessä pidättäydytään kiinteän otoskoon otoksiin.

Jotta osajoukosta O voisi tehdä päätelmiä perusjoukosta U, tulisi osajoukossa tulla esille mahdollisimman hyvin perusjoukon ominaisuuksia. Jokaisella perusjoukon alkiolla u_i tulisi olla nollaa suurempi tunnettu todennäköisyys π_i olla mukana otoksessa. Tämä vaatii usein sen, että perusjoukko tunnetaan kokonaan ja jokainen sen alkio voidaan teoriassa tavoittaa. Tästä seuraa myös se, että osajoukon valinnassa satunnaisuudella on tärkeä rooli. Tällaista osajoukkoa kutsutaan (asymptoottisesti) edustavaksi satunnaisotokseksi. Otantamenetelmiä on useita erilaisia, ja niitä esitellään lisää myöhemmässä osiossa.

Kaikista yksinkertaisin satunnaisotos on yksinkertainen satunnaisotanta (*simple random sampling*, SRS). Oikein toteutetussa yksinkertaisessa satunnaisotannassa jokaisella joukon U alkiolla u_i on sama nollasta poikkeava todennäköisyys π_i olla mukana otoksessa. Yksinkertainen satunnaisotanta voidaan toteuttaa siten, että alkioita otokseen valittaessa kukin alkio voi olla otoksessa enintään yhden kerran (otanta palauttamatta) tai useamman kerran (otanta palauttaen). Tässä tarkastelussa pidättäydytään otantaan ilman palauttamista, sillä se on luontevaa rekisteritutkimukselle ja tällä tavoin saadut tulokset ovat lähtökohtaisesti tarkempia.

Otoksen $O = \{o_1, o_2, \ldots, o_n\}, \forall o_i \in U$, perusteella voidaan tehdä päätelmiä perusjoukosta U, kunhan otos on edustava ja satunnaisesti poimittu. Jos otos ei ole edustava ja satunnainen, kyseessä on näyte, josta tehtävät tulkinnat ovat enemmän laadullista tutkimusta kuin määrällistä ja mahdollisesti myös (voimakkaan) harhaisia. Muita otantamenetelmiä esitellään tarkemmin myöhemmin.

3.3 Otoksen tunnuslukuja

Perusjoukosta saadusta otoksesta voidaan laskea estimaatteja, joiden avulla voidaan tehdä päätelmiä perusjoukon parametreista. Koska alkioiden sisällyttämistodennäköisyys π_i voi vaihdella alkioittain, käytetään otoksen tunnuslukujen estimointiin Horvitz-Thompson-estimaattia tavanomaisen keskiarvon ja summan sijaan. [17] Horvitz-Thompson-estimaatissa kutakin otokseen valittua alkiota painotetaan sen sisällyttämistodennäköisyyden käänteisluvulla $w_i = \frac{1}{\pi_i}$

Määritelmä 3. Otoksen O muuttujan x perusjoukon summan T_x Horvitz-Thompson-estimaatti $\hat{T}_x$ lasketaan seuraavasti:

$$\hat{T}_x = \sum_{i \in \{1,\ldots,n\}} \frac{x_{o_i}}{\pi_i} = \sum_{i \in \{1,\ldots,n\}} w_i x_{o_i}.$$

Määritelmä 4. Otoksen O muuttujan x perusjoukon odotusarvon μ_x estimaatti $\hat{\mu}_x$, eli otoksen keskiarvo, lasketaan Horvitz-Thompson-estimaatin avulla seuraavasti:

$$\bar{x} = \frac{1}{N} \sum_{i \in \{1,\ldots,n\}} \frac{x_{o_i}}{\pi_i} = \frac{1}{N} \sum_{i \in \{1,\ldots,n\}} w_i x_{o_i} = \frac{\hat{T}_x}{N} = \hat{\mu}_x.$$

Muuttujan hajontaa kuvaamaan voidaan käyttää otosvarianssia.

Määritelmä 5. Sisältymistodennäköisyyksien π_i ollessa vakioita, eli keskenään samat, otosvarianssi $\hat{S}_x^2$ lasketaan muuttujalle x seuraavasti:

$$\hat{S}_x^2 = \frac{1}{n-1} \sum_{i \in \{1,\ldots,n\}} (x_{o_i} - \hat{\mu}_x)^2.$$

Otosvarianssin neliöjuurta kutsutaan keskihajonnaksi. Mitä pienempi otosvarianssi (ja keskihajonta), sitä pienempi hajonta muuttujalla on. Estimaattori on harhaton kun $\mathbb{E}(\hat{\mu}_x) = \mu_x$. Tässä tutkielmassa käytämme termiä keskihajonta, vaikkakin keskivirhe voisi olla sopivampi, koska harhattomuus ei ole yksiselitteistä kalibroinnin jälkeen.

3.4 Otoksen tunnuslukujen harhattomuus

Jotta otoksen O tunnusluvuista olisi hyötyä perusjoukon tunnuslukujen arvioimiseen, tulisi niiden ainakin keskimäärin olla lähellä todellisia perusjoukon tunnuslukujen arvoja. Otoksesta laskettu tunnusluku on estimaatti jostakin perusjoukon tunnusluvusta. Estimaattori on harhaton, jos sen odotusarvo on sama kuin perusjoukon tunnusluvun arvo.

Horvitz-Thompson-estimaatin harhattomuuden tarkasteluun tarvitaan indikaattorifunktio. Indikaattorifunktio $\mathbb{1}_{s \in S}$ saa arvon 1, jos sen ehto $s \in S$ on tosi, ja 0:n, jos ei. Indikaattorifunktion odotusarvo on sama kuin indikaattorifunktion ehdon todennäköisyys.

Lause 1. *Perusjoukon muuttujan x summan Horvitz-Thompson-estimaattorin odotusarvolle pätee*

$$\mathbb{E}(\hat{T}_x) = T_x.$$

Todistus:

$$\begin{aligned}\mathbb{E}(\hat{T}_x) &= \mathbb{E}\left(\sum_{i \in \{1,\ldots,n\}} \frac{x_{o_i}}{\pi_i}\right) = \mathbb{E}\left(\sum_{i \in \{1,\ldots,N\}} \frac{x_i}{\pi_i} \mathbb{1}_{x_i \in O}\right) \\ &= \sum_{i \in \{1,\ldots,N\}} \frac{x_i}{\pi_i} \mathbb{E}(\mathbb{1}_{x_i \in O}) = \sum_{i \in \{1,\ldots,N\}} \frac{x_i}{\pi_i} \pi_i = \sum_{i \in \{1,\ldots,N\}} x_i = T_x.\end{aligned}$$

Tästä seuraa myös $\mathbb{E}(\hat{\mu}_x) = \mathbb{E}\left(\frac{\hat{T}_x}{N}\right) = \frac{T_x}{N} = \mu_x$. Horvitz-Thompson-estimaatit perusjoukon summalle ja keskiarvolle ovat siis harhattomia.

3.5 Otanta-asetelmat

Kuten aiemmin todettiin, otanta-asetelmia on useita erilaisia. Osiossa 3.2 esiteltiin yksinkertainen satunnaisotanta. Yksinkertainen satunnaisotanta on otanta-asetelmista yksinkertaisin ja toimii tyypillisesti vertailukohtana eri otanta-asetelmia vertailtaessa. Otantamenetelmiä vertailtaessa tärkeimpiä mittareita ovat niistä laskettavien estimaattoreiden tarkkuus ja keskihajonta. Tarkemmin otantamenetelmien teoriasta voi lukea Särndal'n, Swenssonin ja Wretmanin *Model-Assisted Survey Sampling*

-kirjasta vuodelta 1992 [3]. Myös suomalaisten professorien Risto Lehtisen ja Erkki Pahkisen *Practical Methods for Design and Analysis of Complex Survey Data* kirjaan kannattaa tutustua [4]. Aiheesta on paljon muutakin kirjallisuutta. [17] [18]

3.5.1 Systemaattinen otanta

Systemaattisessa otannassa perusjoukon alkiot valitaan tasavälein. Ensin valitaan satunnainen aloituskohta i väliltä $\{1, \ldots, k\}$ ja sen jälkeen valitaan perusjoukosta alkiot järjestysluvuilla

$$i, i + k, i + 2k, \ldots$$

kunnes saavutetaan perusjoukon loppu N. Tarkka otoskoko $n = \frac{N}{k}$ on riippuvainen jaollisuudesta ja satunnaisesta aloituskohdasta i. Käyttämällä pyöristystä voidaan saada mielivaltaisen kokoinen otoskoko. Tällöin tosin voi käydä niin, että osan alkioista sisältymistodennäköisyydet ovat erisuuria.

Erittäin tärkeää on perusjoukon alkioiden järjestys. Perusjoukon ollessa täysin satunnaisesti järjestetty, ei systemaattinen otanta juuri eroa yksinkertaisesta satunnaisotannasta. Perusjoukko voidaan järjestää apumuuttujien suhteen, jolloin tasavälein valittu systemaattinen otos antaa otoksessa samat suhteelliset osuudet kullekin apumuuttujan arvolle kuin perusjoukossa. Tätä voidaan pitää implisiittisenä osituksena.

Systemaattisen otannan tasavälisen askeleen suuruus k määrittää mahdollisten erilaisten otosten määrän. Systemaattisen otannan teoreettinen keskivirhe saadaan laskemalla tunnusluvut kaikista näistä otoksista ja laskemalla näiden tunnuslukujen keskivirhe.

Kuvatulla tavalla toteutettuna sekoitetun perusjoukon alkion sisällyttämistodennäköisyys otokseen on vakio. Systemaattista otantaa voidaan käyttää myös erisuuruisten sisältymistodennäköisyyksien kanssa, kun perusjoukon alkioiden järjestysluvun sijasta käytetään sisältymistodennäköisyyksien kertymäfunktiota. Aloituspiste valitaan satunnaisesti nollan ja ykkösen väliltä, ja tämän jälkeen valitaan tasaisesti alkiot kertymäfunktiosta 1-mittaisen askeleen välein.

3.5.2 Ositettu otanta

Ositettu otanta on yleistys systemaattisesta otannasta. Ositetussa otannassa aineisto jaetaan apumuuttujan perusteella ositteisiin, ja kustakin ositteesta valitaan satunnaisesti alkioita ositteen koon suhteen otokseen. Tämä takaa sen, että otoksessa kunkin luokan suhteellinen osuus on sama kuin perusjoukossa.

4 Tasapainoinen otanta

Tasapainotetun otosasetelman ajatuksena on hyödyntää apumuuttujia siten, että jokaisessa mahdollisessa otoksessa apumuuttujista laskettu populaatioestimaatti on sama tai lähes sama kuin sen todellinen arvo populaatiossa. Toisin sanoen jokaiselle apumuuttujalle *Horvitz-Thompson-estimaatti* (3) on yhtä suuri kuin kyseisen apumuuttujan summa perusjoukossa (1).

Jokaista perusjoukon U alkiota u_i vastaa vektori kyseisen alkion apumuuttujia $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{ip})'$. Apumuuttujien $\mathbf{z}_{,j} = (Z_{1j}, Z_{2j}, \ldots, Z_{Nj})$, $\forall j$, tulee olla keskenään lineaarisesti riippumattomia. Tasapainoisuusehto Horvitz-Thompson-estimaatin suhteen kirjoitetaan seuraavasti:

$$\hat{T}_{\mathbf{Z}} = T_{\mathbf{Z}} \Leftrightarrow \tag{1}$$
$$\sum_{i\in\{1,\ldots,n\}} \frac{\mathbf{Z}_{o_i}}{\pi_i} = \sum_{i\in\{1,\ldots,N\}} \frac{\mathbf{Z}_i}{\pi_i} \mathbb{1}_{Z_i \in O} = \sum_{i\in\{1,\ldots,N\}} \mathbf{Z}_i.$$

Tämä voidaan myös kirjoittaa muodossa

$$\mathbf{A}\mathbf{x} = \mathbf{A}\boldsymbol{\pi}, \tag{2}$$

jossa matriisi $\mathbf{A} = (\mathbf{Z}_1/\pi_1, \mathbf{Z}_2/\pi_2, \ldots, \mathbf{Z}_N/\pi_N)$ ja dimensioiltaan $[p \times N]$. Matriisin rivit ovat lineaarisesti riippumattomat, joten matriisin aste on p. Vektori $\boldsymbol{\pi}$ on sisältymistodennäköisyyksien vektori. Muuttujavektori $\mathbf{x}$:lle on löydettävä sopiva ratkaisu. Tästä voidaan ratkaista vektorin $\mathbf{x}$ arvojoukko Q

$$\begin{aligned} &\mathbf{A}(\boldsymbol{\pi} - \mathbf{x}) = \mathbf{0} \\ &\Rightarrow \mathbf{x} \in Q = \boldsymbol{\pi} + \mathrm{Ker}(\mathbf{A}), \end{aligned} \tag{3}$$

jossa siis vektori $\mathrm{Ker}\mathbf{A} \in \{\mathbf{x}|\mathbf{A}\mathbf{x} = \mathbf{0}\}$ ja muu seuraa matriisien laskusäännöistä. Tämä joukko Q on $\mathbb{R}^N$:n $N - p$ ulotteinen avaruustaso.

Tasapainoisessa otannassa pyrkimyksenä on valita satunnaisesti sopiva indikaattorivektori $\mathbf{x}$, jonka alkiot saavat arvoja 1 tai 0 riippuen onko vektorin alkiota vastaava perusjoukon alkio valittu otokseen vai ei. Tämän ratkaiseminen ei ole yksinkertaista, varsinkaan perusjoukon koon N ollessa suuri. Kaikissa tapauksissa tarkkaa ratkaisua ei edes ole olemassa. Tästä johtuen tarkan ratkaisun sijaan pyritään löytämään mahdollisimman lähellä sitä oleva ratkaisu, eli $\mathbf{A}\mathbf{x} \approx \mathbf{A}\boldsymbol{\pi}$. Yksinkertaisimmillaan ratkaisuun voidaan hyödyntää lineaarista optimointia, mutta yleisemmin tähän käytetään varta vasten tarkoitettuja algoritmejä. Niistä lisää myöhemmin.

On olennaista huomata, että koska jokaisen tasapainoisen otannan täysin täyttävässä otoksessa muuttujien $\mathbf{z}_j$ Horvitz-Thompson-estimaatit ovat samoja kuin perusjoukossa kyseisten muuttujien summat, muuttujien $\mathbf{z}_j$ Horvitz-Thompson-estimaattorien keskihajonnat ovat nollia. Käytännössä ehdot (2) tismalleen täyttävää otosta ei aina ole olemassa. Tällöin riittävää on, että $\mathbf{z}_j$:n Horvitz-Thompson-estimaattorin keskihajonta on mahdollisimman lähellä nollaa.

4.1 Otoskoon määrittäminen

Tasapainoisessa otannassa otoskoko määritetään käyttämällä yhtenä apumuuttujana $\mathbf{z}_{,j}$ sisällyttämistodennäköisyyttä $\boldsymbol{\pi}$. Tällöin vastaava tasapainoehto asettuu

$$\sum_{i\in\{1,\ldots,n\}} \frac{\pi_i}{\pi_i} = \sum_{i\in\{1,\ldots,n\}} 1 = \sum_{i\in\{1,\ldots,N\}} \mathbb{1}_{u_i \in O} = \sum_{i\in\{1,\ldots,N\}} \pi_i = n.$$

Käytännössä aina tasapainoisen otannan yhteydessä käytetään yhtenä apumuuttujana sisältymistodennäköisyyttä.

4.2 Otoksen tasapainotus perusjoukon koon suhteen

Jos sisällyttämistodennäköisyys π_i ei ole vakio kaikilla i, niin otos saadaan tasapainoiseksi perusjoukon koon suhteen lisäämällä yhdeksi apumuuttujaksi vakiovektori, jonka jokainen alkio saa arvon 1. Tämä varmistaa, että jokaisessa otoksessa Horvitz-Thompson-estimaatti perusjoukon koolle on sama kuin perusjoukon todellinen koko. Tasapainoehto on tällöin

$$\sum_{i\in\{1,\ldots,n\}} \frac{1}{\pi_i} = \sum_{i\in\{1,\ldots,N\}} \frac{\mathbb{1}_{u_i\in O}}{\pi_i} = \sum_{i\in\{1,\ldots,N\}} 1 = N.$$

Sisällyttämistodennäköisyyden ollessa vakio tämän apumuuttujan erikseen sisällyttäminen sisältymistodennäköisyyksien (edellinen osio 4.1) lisäksi ei ole tarpeen.

4.3 Luokkien mukainen kiintiöinti

Yksinkertaisin ei-triviaali apumuuttuja tasapainoitetussa otannassa on luokkamuuttuja. Tällä saavutetaan sama tulos kuin ositetulla otannalla (ks. luku 3.5.2), kun luokkamuuttujia on vain yksi ja sisältymistodennäköisyydet ovat vakioita. Tälläinen muuttuja voi olla esimerkiksi sukupuoli, tuloluokka, diagnoosi tai muu vastaava. Useampiluokkainen muuttuja jaetaan binäärisiin indikaattorimuuttujiin, jotka kertovat, mihin luokkaan kyseinen alkio kuuluu.

Olkoon $\mathbf{z}$ luokkamuuttuja, jonka arvot j on koodattu $j \in \{1,\ldots,k\}$. Asetetaan jokaista luokkaa vastaamaan indikaattorimuuttujia $\boldsymbol{\delta}_j = (\mathbb{1}_{z_1=j}, \mathbb{1}_{z_2=j}, \ldots, \mathbb{1}_{z_N=j})$. Näitä indikaattorimuuttujia käytetään tasapainotetussa otannassa. Huomataan, että useaa eri luokkamuuttujaa käyttäessä tasapainoinen otanta ei huomioi niiden yhteisvaikutusta. Jos joidenkin tiettyjen luokkamuuttujien yhteisvaikutukset halutaan huomioida, on käytettävä näistä muodostettua yhdistettyä luokkamuuttujaa. Lisäksi luokkamuuttujia käyttäessä tasapainotetussa otannassa, ja mallimatriisissa yleisesti, yksi luokka valitaan referenssiluokaksi ja se otetaan mallista pois. Tämä tehdään sen takia, että muuttujien tulee olla keskenään lineaarisesti riippumattomia. Referenssiluokka voidaan esittää muiden luokkien lineaarikombinaationa.

5 Kuutiomenetelmä

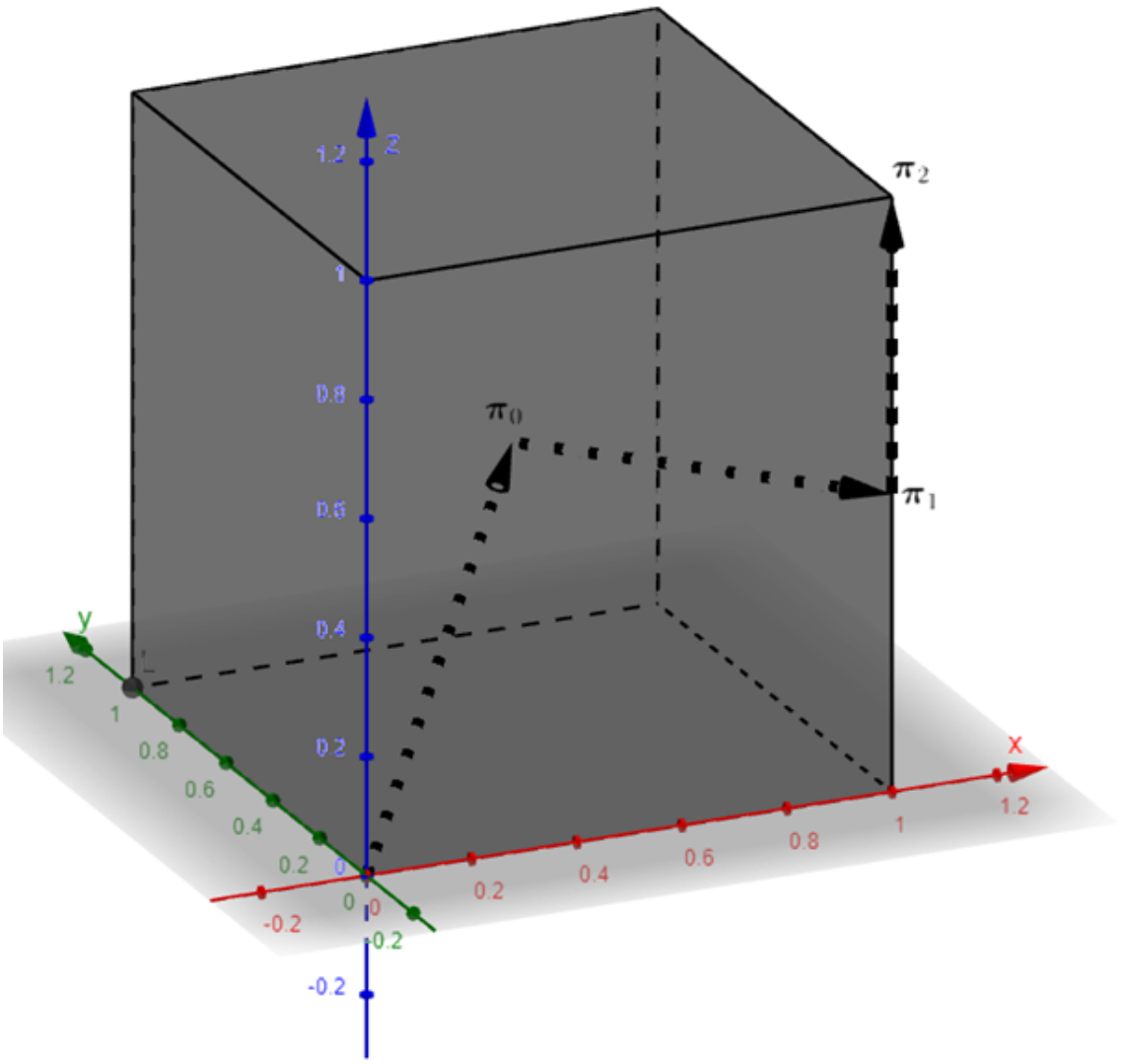
Tutkimusartikkelissaan Deville ja Tillé (2004) esittelevät menetelmän tasapainoisen otannan ehdot täyttävien otosten löytämiseksi [5]. Artikkelissa esitelty menetelmä perustuu tulkintaan, jossa otoksen valintaa kuvaa piste N-ulotteisen kuution $C = [0,1]^N$ sisältä. Tämän kuution kärjissä sijaitsevat pisteet $\{0,1\}^N$ kuvaavat kaikkia mahdollisia otoksia perusjoukosta otoskooltaan väliltä $0-N$. Tätä kuvaavan N-ulotteisen vektorin alkiot kuvastavat niitä vastaavien perusjoukon alkioiden todennäköisyyttä sisältyä otokseen. Tasapainoehdon (2) täyttävien pisteiden joukko on $Q = \boldsymbol{\pi} + \mathrm{Ker}\mathbf{A}$ (3). Kuution sisällä olevat ja tasapainoehdon täyttävät pisteet ovat joukossa $K = C \cap Q$.

Kuutiomenetelmä sisältää kaksi vaihetta, joista ensimmäisessä pyritään löytämään sellainen joukon K piste, jonka mahdollisimman moni alkio saa arvon 0 tai 1, ideaalitapauksessa kaikki. Tässä vaiheessa liikutaan pisteestä pisteeseen joukossa K.

Tästä tulee ensimmäisen vaiheen nimi, lentovaihe (*the flight phase*). Lentovaiheessa siis ”lennetään” joukossa K niin pitkään, kunnes päästään johonkin sen reunapisteeseen. Lentovaiheessa tasapainoisen otannan ehdot (2) toteutuvat täysin.

Kuitenkaan tyypillisesti lentovaiheen lopussa löydetyssä pisteessä kaikki alkiot eivät saa arvoja 0 tai 1, vaan osa alkioista saa arvoja 0 ja 1 väliltä. Tämä johtuu yleensä siitä, ettei joukossa K ole sellaista pistettä, jonka kaikki alkiot saisivat pelkästään arvoja 0 tai 1. Toinen vaihtoehto on, että sellainen piste on olemassa, mutta lentovaihe päätyi johonkin toiseen joukon K reunapisteeseen.

Jos osa lentovaiheen lopuksi löytyneen pisteen alkioista saa muita arvoja kuin 0 tai 1, niin siirrytään toiseen vaiheeseen, jonka nimi on laskeutusmisvaihe. Näitä alkioita on maksimissaan p kappaletta. Laskeutumisvaiheessa pyritään ”laskeutumaan” sellaiseen kuution C kärkipisteeseen, jossa tasapainoisen otannan ehdot (2) täyttyvät mahdollisimman hyvin. Tähän on useampi erilainen toteutusvaihtoehto. Kärkipiste on sellainen piste, jossa kaikki alkiot saavat arvoja 0 tai 1.



Kuva 2: Esimerkki lentovaiheesta $N = 3$ ja $n = 2$ tilanteessa

Kuva 2 demonstroi lentovaihetta $N = 3$ ja $n = 2$ tilanteessa. Siinä lähdetään liikkeelle sisältymistodennäköisyyksien pisteestä $\boldsymbol{\pi}_0$ ja askel askeleelta päädytään pisteeseen $\boldsymbol{\pi}_2$ kuution kärjessä, jossa kaksi alkiota saa arvon 1 ja yksi alkio arvon 0.

Kuutiomenetelmä toteuttaa siis tasapainoisen otannan siten, että sisältymistodennäköisyydet $\boldsymbol{\pi}$ toteutuvat täsmälleen otosta valittaessa ja tasapainoehto (2) mahdollisimman tarkasti. Seuraavaksi tarkastellaan yksityiskohtaisemmin kuutiomenetelmän lentovaihetta (5.1), laskennallisesti keventävää nopeaa lentovaihetta (5.2) sekä laskeutumisvaihetta (5.3).

5.1 Ensimmäinen vaihe: lentovaihe

Kuutiomenetelmän ensimmäistä vaihetta kutsutaan lentovaiheeksi (*the flight phase*). Lentovaiheessa liikutaan satunnaisesti iteratiivisesti joukossa $K = Q \cap C$ kunnes päädytään joukon johonkin reunapisteeseen $\boldsymbol{\pi}_T$. Iteraatio T kuvastaa viimeistä iteraatiota. Jokaisessa lentovaiheen iteraatiossa t löytynyt piste $\boldsymbol{\pi}_t$ täyttää tasapainoehdon (2) yhtälön täsmällisesti. Tästä johtuen alkuarvoksi $t = 0$ on perustelua asettaa $\boldsymbol{\pi}_0 = \boldsymbol{\pi}$, sillä se ainakin triviaalisti täyttää ehdon. Tällöin $\mathbf{A}\boldsymbol{\pi} = \mathbf{A}\boldsymbol{\pi}_0 = \mathbf{A}\boldsymbol{\pi}_t = \mathbf{A}\boldsymbol{\pi}_{t+1}$.

Määritelmä 6. Lentovaiheen jokaisessa iteraatiossa seuraavat ehdot toteutuvat:

1. $\mathbb{E}(\boldsymbol{\pi}_t) = \boldsymbol{\pi}$, eli sisältymistodennäköisyydet toteutuvat täysin,
2. $\mathbf{A}\boldsymbol{\pi}_t = \mathbf{A}\boldsymbol{\pi}$, eli tasapainotus toteutuu apumuuttujien suhteen täysin,
3. jokaisessa vektorin $\boldsymbol{\pi}_t = (\pi_{t,1}, \pi_{t,2}, \ldots, \pi_{t,N})$ alkiossa toteutuu $0 \leq \pi_{t,i} \leq 1$, eli jokainen piste on $[0,1]^N$ kuution sisällä, ja
4. kun $\pi_{t-1,i} = 0$ tai 1, niin kun $\pi_{t,i} = \pi_{t-1,i}$.

Pisteestä $\boldsymbol{\pi}_0$ eteenpäin liikkuminen vaatii lisäymmärrystä matriisien käyttäytymisestä.

Määritelmä 7. *(Matriisin ydin)*
Yhtälön $\mathbf{Au} = \mathbf{0}$ toteuttavien vektoreiden $\mathbf{u}$ joukkoa kutsutaan matriisin $\mathbf{A}$ ytimeksi Ker$\mathbf{A}$ eli $\mathbf{u} \in$ Ker$\mathbf{A}$.

Seuraus 1. *Edellisestä seuraa,*

$$\begin{aligned} \mathbf{A}\boldsymbol{\pi}_t &= \mathbf{A}\boldsymbol{\pi}_{t+1} \\ \mathbf{A}\boldsymbol{\pi}_t + \mathbf{A}\mathbf{u}_{t+1} &= \mathbf{A}\boldsymbol{\pi}_{t+1} + \mathbf{0}, \quad \textit{jossa } \mathbf{u}_{t+1} \in \text{Ker}\mathbf{A} \\ \mathbf{A}(\boldsymbol{\pi}_t + \mathbf{u}_{t+1}) &= \mathbf{A}\boldsymbol{\pi}_{t+1} \\ \Rightarrow \boldsymbol{\pi}_t + \mathbf{u}_{t+1} &= \boldsymbol{\pi}_{t+1}. \end{aligned} \tag{4}$$

Seuraus 2. *Tästä selvästi nähdään, että*

$$\boldsymbol{\pi}_0 + \sum_{t=1}^{T} \mathbf{u}_t = \boldsymbol{\pi}_T. \tag{5}$$

Lause 2. *(Matriisin ytimen vektoreiden summa)*
Vektoreiden $\mathbf{u}_i$ *ja* $\mathbf{u}_j \in Ker\mathbf{A}$ *osalta matriisin ydin on suljettu summaamisen suhteen.*

$$\begin{aligned} \mathbf{A}\mathbf{u}_i + \mathbf{A}\mathbf{u}_j &= \mathbf{A}(\mathbf{u}_j + \mathbf{u}_j) = \mathbf{0} \\ \Rightarrow (\mathbf{u}_i + \mathbf{u}_j) &\in Ker\mathbf{A}. \end{aligned}$$

Lause 3. *(Matriisin ytimen vektorin kertominen skalaarilla)*
Lisäksi oleellista on havainto, jossa $\mathbf{u}_i \in Ker\mathbf{A}$ *ja* $\lambda \in \mathbb{R}$ *niin*

$$\begin{aligned} \mathbf{A}\mathbf{u} = \mathbf{0} \;||\cdot \lambda \Leftrightarrow \\ \lambda\mathbf{A}\mathbf{u} = \mathbf{A}(\lambda\mathbf{u}) = \mathbf{0} \\ \Rightarrow \lambda\mathbf{u} \in Ker\mathbf{A}. \end{aligned} \tag{6}$$

Tästä siis seuraa, että on olemassa sellainen $\mathbf{s} \in \text{Ker}\mathbf{A}$, jolla $\boldsymbol{\pi}_T = \boldsymbol{\pi}_0 + \mathbf{s}$. Lentovaihe perustuukin siis pohjimmiltaan sopivan vektorin $\mathbf{s} = \sum_{t=1}^{T} \mathbf{u}_t$ muodostamiseen.

Tarkastellaan tarkemmin matriisin $\mathbf{A}$ ytimen vektoreiden joukkoa $\mathbf{u}_j \in \text{Ker}\mathbf{A}$. Matriisi $\mathbf{A}$ on dimensioiltaan $[p \times N]$ ja sen rivit ovat keskenään lineaarisesti riippumattomia. Tällöin matriisin aste on p ja eli myös p kappaletta sen sarakkeista on lineaarisesti riippumattomia. Tämä tarkoittaa, että on olemassa sellainen $\mathbf{u} \in \text{Ker}\mathbf{A}$, jolle $N - p$ alkion arvo voidaan asettaa mielivaltaisesti.

Tavoitteena on löytää sellainen $\boldsymbol{\pi}_T$, jonka mahdollisimman monen alkion arvo on 0 tai 1 sekä $\mathbb{E}(\boldsymbol{\pi}_T) = \boldsymbol{\pi}$. Tämä toteutetaan valitsemalla $(\mathbf{u}_1, \mathbf{u}_2, \dots) \in \text{Ker}\mathbf{A}$ siten, että jokaisessa iteraatiossa ainakin yksi vektorin $\boldsymbol{\pi}_t = \boldsymbol{\pi}_{t-1} + \mathbf{u}_t$ alkio saavuttaa arvon 0 tai 1, joka ei ollut sellainen pisteessä $\boldsymbol{\pi}_{t-1}$. Lisäksi myös alkiot, joiden arvo vektorissa $\boldsymbol{\pi}_{t-1}$ on 0 tai 1, ovat muuttumattomia vektorissa $\boldsymbol{\pi}_t$.

Jotta kyse olisi satunnaisotannasta, alkioiden muuttaminen luvuiksi 0 ja 1 on perustuttava satunnaisprosessiin, jonka odotusarvona tulee olla sisältymistodennäköisyydet $\mathbb{E}(\boldsymbol{\pi}_t) = \boldsymbol{\pi}$. Tämä saadaan toteutettua siten, että valitaan satunnaisesti kahden päinvastaisen suunnan väliltä niin, että odotusarvo $\mathbb{E}(\mathbf{u}_t) = \mathbf{0}$.

Määritelmä 8. Iteraatiossa t pyritään löytämään sellainen $\mathbf{u}_t$ joka täyttää seuraavat ehdot:

1. $\mathbf{u}_t \in \text{Ker}\mathbf{A}$,
2. $u_{t,i} = 0$, kun $\pi_{t-1,i} = 0$ tai 1, jossa $\mathbf{u}_{t,i} = (u_{t,1}, \dots, u_{t,N})$ ja $\boldsymbol{\pi}_{t-1} = (\pi_{t-1,1}, \dots, \pi_{t-1,N})$,
3. $\mathbf{u}_t \neq \mathbf{0}$ (nollavektori),
4. $\mathbf{0} \leq \boldsymbol{\pi}_{t-1} + \mathbf{u}_t = \boldsymbol{\pi}_t \leq \mathbf{1}$ ja $\#(\pi_{t-1,i} \in \{0,1\}) < \#(\pi_{t,i} \in \{0,1\})$, jossa $\#()$ on lukumääräfunktio, ja
5. $\mathbb{E}(\mathbf{u}_t) = \mathbf{0}$.

Jos näitä ehtoja täyttävää vektoria $\mathbf{u}_t$ ei ole olemassa, on viimeinen iteraatio $T = t-1$. Viimeinen iteraatio T on kuitenkin aina jokin joukon $[N-p, N]$ kokonaisluku.

Toinen kohta on toisin sanottuna, että niiden vektorin $\mathbf{u}_t$ alkioiden on oltava arvoltaan 0, joita vastaavat alkiot vektorissa $\boldsymbol{\pi}_{t-1}$ saavat arvon 0 tai 1. Tällainen $\mathbf{u}_t \in \mathrm{Ker}\mathbf{A}$ on aina olemassa iteraatiossa t, kun $\#(\pi_{t-1,i} \in \{0,1\}) < (N-p)$.

Se, että vektorin $\mathbf{u}_t$ alkio saa arvon 0, tarkoittaa sitä, että sitä vastaava vektorin $\boldsymbol{\pi}_t$ arvo ei muutu $\boldsymbol{\pi}_{t-1}$:stä. Tällaisen vektorin löytämisestä lisää lauseessa 4.

Kolmas kohta täyttyy varmasti, kun $\#(\pi_{t-1,i} \in \{0,1\}) < N-p$. Tällöin on olemassa sellainen $\mathbf{u}_t$ jossa $\#(\pi_{t-1,i} \in \{0,1\})$ nolla-alkion lisäksi on ainakin 2 sellaista alkiota jotka saavat nollasta poikkeavan arvon. Tämä johtuu siitä, että matriisissa $\mathbf{A}$ yhtenä rivinä $\mathbf{z}$ on vakiorivi (osio 4.2). Koska kyseisen vakiorivin pistetulo ytimen vektorin kanssa on luonnollisesti nolla, seuraa edellinen.

Jotta neljäs kohta täyttyisi, olkoon $\mathbf{u}_t^*$ jokin ehdot 1.-3. täyttävä vektori. Etsitään sellaiset vakion λ arvot, jotka täyttävät epäyhtälöt

$$\mathbf{0} \leq \boldsymbol{\pi}_{t-1} + \lambda \mathbf{u}_t^* \leq \mathbf{1}$$

siten, että λ_1 on pienin epäyhtälöt toteuttava arvo ja λ_2 suurin epäyhtälöt toteuttava arvo. Koska etsitään suurin ja pienin λ:n arvo, niin ainakin yksi uusi $\boldsymbol{\pi}_t$:n alkio saa arvon 0 tai 1, sillä $\mathbf{u}_t^*$ ei ole nollavektori. Nyt $\mathbf{u}_t^1 = \lambda_1 \mathbf{u}_t^*$ tai $\mathbf{u}_t^2 = \lambda_2 \mathbf{u}_t^*$.

Viides ehto edellyttää, että $\mathbb{E}(\mathbf{u}_t) = \mathbf{0}$. Tämä toteutuu siten, että ensiksi määritetään todennäköisyys $q = \frac{\lambda_2}{\lambda_2 - \lambda_1}$. Nyt valitaan

$$\mathbf{u}_t = \begin{cases} \mathbf{u}_t^1 = \lambda_1 \mathbf{u}_t^*, & \text{todennäköisyydellä } q \\ \mathbf{u}_t^2 = \lambda_2 \mathbf{u}_t^*, & \text{todennäköisyydellä } 1-q. \end{cases}$$

Tämä täyttää viidennen ehdon, sillä

$$\begin{aligned} \mathbb{E}(\mathbf{u}_t) &= q\mathbf{u}_t^1 + (1-q)\mathbf{u}_t^2 \\ &= \frac{\lambda_2}{\lambda_2 - \lambda_1}\lambda_1 \mathbf{u}_t^* + (\frac{-\lambda_1}{\lambda_2 - \lambda_1})\lambda_2 \mathbf{u}_t^* \\ &= \mathbf{0}. \end{aligned}$$

Prosessit $\mathbf{u}_t$ ja $\boldsymbol{\pi}_t$ ovat siis diskreettejä stokastisia prosesseja ja martingaaleja.

Tätä iterointia voidaan jatkaa ainakin niin kauan, kunnes $N-p$ kappaletta vektorin $\boldsymbol{\pi}_T$ alkiota on saanut arvon 0 tai 1. Tässä kohtaa $N-p$ kappaletta vektorin $\mathbf{u}_t$ alkioista saa arvon 0. Jos jokin näistä lopuista alkioista sattuukin saamaan jonkin nollasta poikkeavan arvon, voidaan lentovaihetta jatkaa niin pitkään, kunnes ehdot täyttävää ytimen vektoria ei enää löydy.

Vektorin $\mathbf{u} \in \mathrm{Ker}\mathbf{A}$ löytämiseen on olemassa useita erilaisia menetelmiä. Yksinkertaisin on Gauss-Jordan-eliminointi, jossa asetetaan matriisi $(\mathbf{A}|\mathbf{0})$ ja rivioperaatioilla ratkaistaan $\mathbf{u}$ kantavektorit.

Matriisin $\mathbf{A}$ ytimen löytäminen on raskas operaatio, varsinkin kun matriisin dimensio on suuri. Jos osa halutun ytimen vektorin alkioista halutaan asettaa nollaksi, yksinkertaistuu laskenta merkittävästi.

Lause 4. *Kun osa vektorin* $\mathbf{u} \in \mathrm{Ker}\mathbf{A}$ *alkioista halutaan nolliksi, ytimen vektorin löytäminen helpottuu merkittävästi. Tällöin voidaan muodostaa matriisi* $\mathbf{B}$ *niistä matriisin* $\mathbf{A}$ *sarakkeista, joita vastaavien ytimen vektorin alkioiden arvoksi halutaan jokin muu kuin 0. Tällöin voidaan ratkaista* $\mathbf{v} \in \mathrm{Ker}\mathbf{B}$*. Tämän avulla saadaan muodostettua* $\mathbf{u} \in \mathrm{Ker}\mathbf{A}$ *siten, että asetetaan sen halutut arvot nolliksi ja loput arvot täydennetään löydetystä vektorista* $\mathbf{v}$*. Matriisin* $\mathbf{B}$ *dimensio voi olla merkittävästi pienempi kuin matriisin* $\mathbf{A}$*.*

5.2 Nopea lentovaihe

Lentovaihetta on jatkokehitetty artikkelissa *A Fast Algorithm for Balanced Sampling* (Chauvet ja Tillé, 2006) [6]. Tavallisessa lentovaiheessa laskenta-aika on verrannollinen aineiston koon neliöön. Tämä jatkokehitetty algoritmi perustuu siihen, että kokonaisen matriisin $\mathbf{A}$ sijasta tarkastellaan kerralla pienempää osaa perusjoukon alkioista. Tällöin matriisin ytimen löytäminen on huomattavasti yksinkertaisempaa ja nopeampaa. Nopeassa lentovaiheessa laskenta-aika on lineaarinen suhteessa aineiston kokoon. Artikkelin esimerkkitapauksessa nopeutus on noin 3000-kertainen. Nopea lentovaihe on tavallisen lentovaiheen sisällä jokaisessa iteraatiossa erikseen suoritettava algoritmi.

Normaali lentovaihe perustuu siis sopivan vektori $\mathbf{s} = \sum_{t=1}^{T} \mathbf{u}_t \in \mathrm{Ker}\mathbf{A}$ löytämiseen (edellinen osio 5.1). Nopeassa lentovaiheessa normaalin lentovaiheen iteraatio t pyritään toteuttamaan muodostamalla $\mathbf{u}_t = \sum_{r=1}^{R_t} \mathbf{v}_{t,r} \in \mathrm{Ker}\mathbf{A}$ siten, että $N-p-1$ kappaletta vektorin riveistä on lukittuna arvoon 0. Tällöin riittää, että tarkastelu rajoittuu vain $p+1$ perusjoukon alkioon, joita vastaa $p+1$ mittainen vektori $\boldsymbol{\psi}_{t,0}$, jonka alkiot on valittu $\boldsymbol{\pi}_{t-1}$:sta. Näitä alkioita vastaavat matriisin $\mathbf{A}$ sarakkeet asetetaan matriisiin $\mathbf{B}_t$. Nyt tasapainoehto (2) voidaan esittää muodossa $\mathbf{B}_t\boldsymbol{\psi}_{t,r-1} = \mathbf{B}_t\boldsymbol{\psi}_{t,r}$ ja $\boldsymbol{\psi}_{t,r} = \boldsymbol{\psi}_{t,r-1} + \boldsymbol{\nu}_{t,r}$, jossa $\boldsymbol{\nu}_r \in \mathrm{Ker}\mathbf{B}_t$. Kunkin nopean lentovaiheen lopuksi löydetään $\boldsymbol{\psi}_{t,R_t} = \boldsymbol{\psi}_{t,0} + \boldsymbol{\upsilon}_t$, jossa $\boldsymbol{\upsilon}_t = \sum_{r=1}^{R_t} \boldsymbol{\nu}_{t,r} \in \mathrm{Ker}\boldsymbol{B_t}$ ja $\boldsymbol{\psi}_{t,0}$ on vektori niistä valituista $p+1$:tä $\boldsymbol{\pi}_{t-1}$:n alkioista, joiden arvo on väliltä $(0,1)$. Tämän jälkeen muodostetaan $\boldsymbol{\pi}_t$ asettamalla vektorin $\boldsymbol{\psi}_{t,R_t}$ alkiot niitä vastaavien vektorin $\boldsymbol{\pi}_{t-1}$ alkioiden paikalle.

Huomionarvoista on, että $\boldsymbol{\upsilon}_t \in \mathrm{Ker}\mathbf{B}_t$ saadaan täydennettyä vektoriksi $\mathbf{u}_t \in \mathrm{Ker}\mathbf{A}$ asettamalla vektorin $\boldsymbol{\upsilon}_t$ alkiot niitä vastaaville paikoilleen N mittaisessa vektorissa ja asettamalla vektorin muut arvot nollaksi (lause 4). Tällöin siis $\boldsymbol{\pi}_t = \boldsymbol{\pi}_{t-1} + \mathbf{u}_t$. Lisäksi tasapainoehto toteutuu $\mathbf{A}\boldsymbol{\pi}_t = \mathbf{A}\boldsymbol{\pi}_{t-1} = \mathbf{A}\boldsymbol{\pi}$.

Seuraava iteraatio $t+1$ toteutetaan ensin asettamalla iteraatiossa t saadun vektori $\boldsymbol{\psi}_{t,R_t}$ alkioiden arvot niitä vastaavien vektorin $\boldsymbol{\pi}_{t-1}$ alkioiden tilalle, jolloin saadaan vektori $\boldsymbol{\pi}_t$. Sen jälkeen muodostetaan vektori $\boldsymbol{\psi}_{t+1,0}$ siten, että korvataan vektorin $\boldsymbol{\psi}_{t,R_t}$:n arvon 0 tai 1 saavat alkiot sellaisilla $\boldsymbol{\pi}_t$:n alkioilla, jotka saavat arvon väliltä $(0,1)$. Perustelluinta on valita indeksiltään pienimmät sellaiset $\boldsymbol{\pi}_t$ alkiot. Tällöin aineiston järjestyksellä on erittäin paljon merkitystä. Lisää tästä järjestyksestä osiossa 5.4.

Iterointia jatketaan, kunnes vektorissa $\boldsymbol{\pi}_{t-1}$ on jäljellä ainoastaan $p+1$ tai vähemmän alkiota, jotka saavat arvon väliltä $(0,1)$. Tällöin kyseessä oleva iteraatio $t-1$ on toiseksi viimeinen iteraatio $T-1$. Asetetaan kyseiset arvot vektoriin $\boldsymbol{\psi}_{T,0}$, ja niitä vastaavat matriisin $\mathbf{A}$ sarakkeet matriisiin $\mathbf{B}_T$. Tästä saadaan lopuksi vektori $\boldsymbol{\psi}_{T,R_T}$, jonka alkiot asetetaan niitä vastaaville paikoilleen vektorissa $\boldsymbol{\pi}_{T-1}$. Näin

saadaan $\boldsymbol{\pi}_T$, joka on lentovaiheen tulos ja toteuttaa täydellisesti vaatimukset (6), kuten luonnollisesti muutkin iteraatiot toteuttivat.

Määritelmä 9. Normaalin lentoalgoritmin iteraatiossa t toteutetun nopean lennon iteraatiossa r pyritään löytämään sellainen $\boldsymbol{\nu}_{t,r}$, joka täyttää seuraavat ehdot:

1. $\boldsymbol{\nu}_{t,r} \in \mathrm{Ker}\mathbf{B}_t$,
2. $\nu_{t,r,i} = 0$, kun $\psi_{t,r-1,i} = 0$ tai 1, jossa $\boldsymbol{\nu}_{t,r} = (\nu_{t,r,1}, \ldots, \nu_{t,r,p+1})$ ja $\boldsymbol{\psi}_{t,r-1} = (\psi_{t,r-1,1}, \ldots, \psi_{t,r-1,p+1})$,
3. $\boldsymbol{\nu}_{t,r} \neq \mathbf{0}$ (nollavektori),
4. $\mathbf{0} \leq \boldsymbol{\psi}_{t,r-1} + \boldsymbol{\nu}_{t,r} = \boldsymbol{\psi}_{t,r} \leq \mathbf{1}$ ja $\#(\psi_{t,r-1,i} \in \{0,1\}) < \#(\psi_{t,r,i} \in \{0,1\})$, ja
5. $\mathbb{E}(\boldsymbol{\nu}_t) = \mathbf{0}$.

Jos näitä ehtoja täyttävää vektoria $\boldsymbol{\nu}_{t,r}$ ei ole olemassa, on viimeinen iteraatio $R_t = r - 1$. Viimeinen iteraatio R_t on kuitenkin aina jokin joukon $[1, p]$ luku.

Tämä määritelmä on täysin analoginen määritelmän (8) ja sen perustelujen kanssa.

5.3 Toinen vaihe: laskeutuminen

Ensimmäisen vaiheen lopuksi on löydetty joukon $K = C \cap Q$ reunapiste $\boldsymbol{\pi}_T$, jonka ainakin $N-p$ alkiota saavat arvoja 0 tai 1. Jos kaikki vektorin N alkiota saavat arvot 0 tai 1, ei toista vaihetta tarvita. Reunapisteen $\boldsymbol{\pi}_T$ alkioista enintään p kappaletta saa arvoja väliltä $(0, 1)$ eikä näitä voi asettaa 0 tai 1:ksi siten, että tasapainoehto (2) toteutuisi samalla.

Laskeutumisvaiheessa pyritään löytämään sellainen N-ulotteisen kuution kärkipiste, jossa tasapainoehto (2) toteutuu mahdollisimman hyvin. Tähän "mahdollisimman hyvin" toteutumiseen on kaksi lähestymistapaa. Ensimmäisessä pyritään lineaarisen optimoinnin kautta toteuttamaan mahdollisimman hyvin rajoitefunktio. Toisessa pyritään toteuttamaan edes osa rajoitemuuttujista $\mathbf{z}_j$ täysin. Tämä tehdään yksitellen poistamalla matriisin $\mathbf{B}$ rivejä. Lineaarinen optimointi on käytännöllistä, kun muuttujia on vähän ja niiden välillä ei ole selkeää tärkeysjärjestystä. Rajoitteiden keventäminen taas on käytännöllisempää, kun muuttujilla on selkeä tärkeysjärjestys ja tärkeimpien halutaan toteutuvan täysin. Tässä tarkastelussa rajoitutaan siihen, että edes osa rajoitemuuttujista toteutuisi täysin.

Muuttujia löysäämällä laskeutuminen toteutetaan siten, että aluksi asetetaan nämä ei-kokonaislukuarvoiset $\boldsymbol{\pi}_T$:n alkiot vektoriin $\boldsymbol{\psi}_{0,0}^{land}$ ja niitä vastaavat $\mathbf{A}$:n sarakkeet matriisiin $\mathbf{B}_0^{land}$. Tällöin rajoitefunktio on muotoa $\mathbf{B}_0^{land}\boldsymbol{\psi}_{0,t}^{land} = \mathbf{B}_0^{land}\boldsymbol{\psi}_{0,t+1}^{land}$ ja $\boldsymbol{\psi}_{0,t}^{land} = \boldsymbol{\psi}_{0,t-1}^{land} + \boldsymbol{\nu}_{0,t}^{land}$, jossa $\boldsymbol{\nu}_{0,t}^{land} \in \mathrm{Ker}\mathbf{B}_0^{land}$. Tätä jatketaan aivan kuten lentovaiheessakin, kunnes ei enää löydy sopivaa $\boldsymbol{\nu}_{0,t}^{land} \in \mathrm{Ker}\mathbf{B}_0^{land}$. Tällöin saadaan vektori $\boldsymbol{\psi}_{0,R_t}^{land}$. Jos tässä vektorissa on vielä reaalialkioita väliltä $(0, 1)$, niin siirrytään totutusti seuraavaan vaiheeseen.

Seuraavassa vaiheessa poistetaan $\mathbf{B}_0^{land}$:sta vähiten tärkeintä apumuuttujaa vastaava rivi ja merkitään uutta matriisia $\mathbf{B}_1^{land}$. Tätä samaa jatketaan niin pitkään, kunnes otos löytyy.

5.4 Aineiston järjestys

Aineiston rivien järjestyksellä ei juurikaan ole väliä, kun ei käytetä nopeaa lentovaihetta. Koska nopeassa lentovaiheessa vektoriin $\boldsymbol{\psi}_{t,0}$ lisätään alkioita vektorista $\boldsymbol{\pi}_{t-1}$, on järjestyksellä tässä tapauksessa väliä.

Koska $\boldsymbol{\pi}_t$:n alkioiden järjestys, eli se, mitä perusjoukon alkiota mikäkin alkio vastaa, ei muutu iteraatioiden välillä, määräytyy se yhtälöä $\boldsymbol{\pi}_0 = \boldsymbol{\pi}$ vastaavasta järjestyksestä ja siihen linkitetystä matriisin $\mathbf{A}$ sarakkeiden järjestyksestä. Vektorin $\boldsymbol{\pi}_0$ järjestämiseen on kaksi vaihtoehtoista näkökulmaa. Ensimmäinen on satunnaisavaruuden laajentaminen, ja toinen on rajoitteiden toteutuminen helpommin. Satunnaisuuden lisääminen perustuu siihen, että nopeassa lentovaiheessa käytettävässä $\boldsymbol{\upsilon}_t \in \mathrm{Ker}\mathbf{B}_t$ voi kerrallaan muuttaa jonkin $[2, p+1]$ alkion väliltä $\boldsymbol{\pi}_{t-1}$ alkioita, kun taas normaalissa lentovaiheessa $\mathbf{u}_t \in \mathrm{Ker}\mathbf{A}$ voi muuttaa enintään $[2, N]$ alkion arvoja. Tällöin on selvää, että nopeassa lentovaiheessa mahdollisia vaihtoehtoja on huomattavasti vähemmän.

Tästä syystä vektorin $\boldsymbol{\nu}_{t,i}$ valinta tapahtuu huomattavasti pienemmästä joukosta. Satunnaisuutta voi lisätä järjestämällä perusjoukon alkiot satunnaiseen järjestykseen, mikä lisää satunnaisuutta otoskertojen välillä. Tällöin vektoriin $\boldsymbol{\psi}_{t+1}$ valitaan todennäköisemmin eri alkio vektorista $\boldsymbol{\pi}_t$ eri otosten välillä. [6]

Aineisto voidaan myös järjestää muihin järjestyksiin. Artikkelissaan Tillé kuvaa aineiston järjestyksen merkityksen siten, että kun yritetään mahduttaa laatikkoon eri kokoisia perunoita, kannattaa järjestely aloittaa ensin isoista perunoista ja pienet asetella vasta lopuksi. [6] Tämä analogia toimii otantaan siten, että aluksi kannattaa käsitellä ”hankalat” yksilöt ja jättää ”helpot” loppuun. Kuutiomenetelmän tapauksessa tämä tarkoittaa sitä, että laskeutumisvaiheeseen jäisi mahdollisimman ”helpot” yksilöt. Leuenberger ja muut kuvaavat vuoden 2022 tutkimusartikkelissaan [12] hyötyjä aineiston järjestämisestä Mahalanobis- tai Tukey-syvyyden perusteella. Mahalanobis-syvyys kuvaa kunkin rivin skaalainvariantin etäisyyden aineiston keskipisteestä. Tukey-syvyys taas on parametriton mitta, jonka voi ajatella perustuvan moniulotteiseen mediaaniin. Mitä suurempi rivin Mahalanobis- tai Tukey-syvyys on, sitä kauempana rivi on aineiston keskimääräisestä rivistä. Käytännön hyötyä tästä voi eniten olla tilanteessa, jossa otoksen koko suhteessa aineiston kokoon on verrattain pieni, samoin kuin apumuuttujien määrä suhteessa aineiston kokoon. Tällöin nopean lentovaiheen alkuvaiheessa käsitellään ”vaikeat” rivit, kun taas loppuvaiheessa ”helpot”. Lopuksi jäljelle jäävät käsitellään laskeutumisvaiheessa.

5.5 Pohdintaa

Nopea lentovaihe kehitettiin sen takia, että algoritmin pullonkaulana oli avaruuden $\mathrm{Ker}\mathbf{A}$ ratkaiseminen. Yleisesti $\mathrm{Ker}\mathbf{B}$ on huomattavasti helpompi laskea kuin $\mathrm{Ker}\mathbf{A}$. Nopean lentovaiheen [6] julkaisuvuonna 2006 laskenta oli huomattavasti hitaampaa kuin tätä kirjoitettaessa vuonna 2026. Tällöin matriisista $\mathbf{B}_t$ on haluttu dimensioiltaan mahdollisimman pieni ja tämän takia varmaankin nopeaan lentovaiheeseen on valittu pienin mahdollinen määrä alkioita, joilla lentovaihe ylipäätään varmasti onnistuu, eli $p+1$ (Nulliteetti-aste-teoreema). Tässä on haittapuolena se, että mahdollisia keskenään ortogonaalisia suuntia on vain yksi per iteraatio $(p-(p-1))$, joten avaruus $\mathrm{Ker}\mathbf{B}$ on verrattain rajoittunut. Tämä luonnollisesti pienentää satunnai-

suutta ja ajaa algoritmia enemmän deterministiseen suuntaan. Nyt laskentatehon kehityksen myötä matriisi $\mathbf{B}$ voisi olla suurempikin. Asettamalla esimerkiksi alkioiden määräksi $2p$ olisi mahdollisia ortogonaalisia suuntia p kappaletta. Tasapainoisella otannalla saadulle estimaattorin varianssille on olemassa approksimaatio, mutta se tyypillisesti aliarvioi varianssin, joten sitä ei tarkastella tässä tutkielmassa [16]. Lisäksi Tillé esittelee puhtaasti lineaariseen optimointiin perustuvan kuutiomenetelmän vuoden 2026 tutkimuspaperissaan [8]. Tämä kuitenkaan ei ole laskennallisen raskautensa vuoksi kovin käyttökelpoinen aineiston ollessa suuri.

6 Kalibrointiestimaatti

Tasapainoinen otanta pyrkii poimimaan sellaisen otoksen, jossa otoksen apumuuttujien Horvitz-Thompson-estimaatit vastaavat populaation vastaavien apumuuttujien summia. Kalibroinnissa taas on kyse otoksen poiminnan (ja haastattelun) jälkeisestä uudelleenpainottamisesta painokertoimin siten, että uusien painojen kanssa otoksen apumuuttujien Horvitz-Thompson-estimaatit vastaavat sitä, mikä on populaatiossa kyseisten apumuuttujien summa. Tasapainoisen otannan tapauksessa tarve otannan jälkeiselle kalibroinnille on tyypillisesti pienempi, sillä otos on jo alkujaankin valittu siten, että apumuuttujien Horvitz-Thompson-estimaatit vastaavat populaation kyseisten apumuuttujien summia tai ainakin ovat hyvin lähellä niitä. Muiden otantamenetelmien tapauksessa kalibroinnin hyöty on odotettavasti merkittävästi suurempi. Kalibroimisen teoriaan voi tarkemmin tutustua Devillen ja Särndal'n tutkimuspaperissa vuodelta 1992 [9]. Lisäksi Devillen, Särndal'n ja Sautoryn artikkelissa vuodelta 1993 kuvaillaan tarkemmin *generalized raking* -kalibrointialgoritmeja [10]. Myös Särndalin artikkeli *The calibration approach in survey theory and practice* on relevantti [11].

Kalibrointiin on useita mahdollisia syitä. Lähtökohtana on, että halutaan "korjata" satunnaisotannalla saatua otosta populaatiosta olevien tietojen perusteella. Tyypillisesti satunnaisotannalla saadun otoksen apumuuttujien Horvitz-Thompson-estimaatit eivät vastaa tarkasti populaation vastaavien apumuuttujien summia ja tämä aiheutuu jo satunnaisuuden vaikutuksesta. Tämä voi myös johtua siitä, että otoksessa on katoa eli osa otokseen valituista henkilöistä ei vastaa kyselytutkimukseen mahdollisesti samalla vinouttaen aineistoa. Tämäkin luo kalibroinnille tarvetta. Toinen syy liittyy siihen, että aika otoksen valintahetken ja haastatteluhetken välillä on pitkä. Tällöin populaation taustamuuttujien summat voivat muuttua merkittävästikin; ihmisiä kuolee, muuttaa, työelämästatus muuttuu ja muuta vastaavaa. Lisää kalibroinnista Tillén vuoden 2020 kirjassa. [17]

7 Aineisto ja apumuuttujat

7.1 Yleistä aineistosta

Osiossa 2 kerrotaan, että työvoimatutkimuksen perusjoukko on Suomessa vakinaisesti asuvat 15-89-vuotiaat henkilöt, ja näiden lisäksi tilapäisesti alle vuoden ulkomailla oleskelevat ja Suomen väestötietojärjestelmään rekisteröidyt ulkomaalaiset, joiden oleskelu Suomessa kestää vähintään vuoden. Näistä henkilöistä valitaan kaksi kertaa vuodessa työvoimatutkimukseen 15 600 henkilöä, joiden ensimmäinen haastattelukerta jaetaan tasaisesti seuraaville kuudelle kuukaudelle. Haastattelukertoja on yhteensä viisi, joista ensimmäinen ja viimeinen tapahtuvat 16 kuukauden sisällä toistaan. Haastattelut tehdään kolmen kuukauden välein, paitsi neljännen haastattelun väli on kuusi kuukautta kolmannesta haastattelusta. Tästä rotaatiorakenteesta seuraten joka kuukausi haastatellaan viittä eri ryhmää, joista jokaisella on eri haastattelukerta [1]. Rotaatiopaneeleista tarkemmin Nedyalkovan, Qualitén ja Tillén artikkelissa *General framework for the rotation of units in repeated survey sampling* [15].

Tässä tutkielmassa aineistona on työvoimatutkimuksen aineisto tammikuusta 2022 lokakuuhun 2025. Aineisto sisältää 563 753 haastattelukertaa yhteensä 151 262 eri henkilöltä. Aineiston henkilöt ovat 15-74-vuotiaita. Haastatteluja on reilu 147 000 vuodessa, joskin vuonna 2025 aineisto ylettyy ainoastaan lokakuuhun saakka. Koska vuosi 2025 jää kesken, aiheuttaa se aineistossa hieman vinoutta. Eri haastattelukertojen edustajien määrä vuosien välillä on hyvin tasainen, vajaa 29 500 haastattelukertaa kohden vuodessa (Taulukko 1).

Taulukko 1: Haastateltavien määrä haastatteluvuoden ja haastattelukerran perusteella

	1	**2**	**3**	**4**	**5**	YHT
2022	29307	29385	29464	29504	29442	**147102**
2023	29297	29387	29454	29478	29453	**147069**
2024	29295	29402	29446	29479	29496	**147118**
2025*	24375	24441	24523	24604	24521	**122464**
YHT	**112274**	**112615**	**112887**	**113065**	**112912**	563753

*2025 tammikuu-lokakuu

Aineistossa tärkeimpänä tarkasteltavana muuttujana on haastattelulla saatu työmarkkina-asema. Henkilö on voinut myös jättää vastaamatta tai häntä ei ole tavoitettu. Taulukossa 2 kuvataan vastausprosentit vastausvuoden ja -kerran perusteella. Koko aineistossa vastausprosentti on 52,2 %, mutta vastauskertojen ja -vuoden perusteella on jonkin verran vaihtelua.

Taulukko 2: Vastausprosentti (%) vastausvuoden ja haastattelukerran perusteella

	1	**2**	**3**	**4**	**5**	KAIKKI
2022	45,2	47,7	48,3	48,0	49,7	**47,8**
2023	49,9	50,2	51,0	50,4	51,0	**50,5**
2024	57,2	56,8	56,1	55,0	55,8	**56,2**
2025*	55,6	56,0	53,4	53,8	54,8	**54,7**
KAIKKI	**51,8**	**52,6**	**52,2**	**51,7**	**52,7**	52,2

*2025 tammikuu-lokakuu

Tarkastelua varten rajataan aineiston tarkastelu ainoastaan niihin kertoihin, joissa haastattelu on saatu. Tällöin rivejä jää 294 247. Mielenkiintoista voisi olla myös vastauskadon analysointi, mutta se jää tämän tarkastelun ulkopuolelle.

Aineistossa on yhteensä 67 erilaista muuttujaa, joista mielenkiintoisia ovat esimerkiksi Työ- ja elinkeinoministeriön työnvälitystilaston edeltävän kuukauden rekisteritieto työttömyydestä, työssäkäyntitilaston pääasiallinen toiminta -rekisteritieto, sukupuoli, ikä, äidinkieli, koulutusaste ja asuinpaikka.

Oikean perusjoukon osalta (noin 4,8 miljoonaa henkilöä) ei ole mahdollista tietää työvoimatutkimuksen otoksen ulkopuolisten henkilöiden työmarkkina-asemaa. Tämän ratkaisemiseksi täytyy käyttää pseudoperusjoukkoa. Pseudoperusjoukkona käytetään yllä mainittua aineistoa, sillä sen sisäisen dynamiikan voi ajatella olevan hyvin samanlainen kuin oikean perusjoukon.

Tämä aineisto ei kuitenkaan vastaa todellista otosvalinnan asetelmaa. Tällä aineistolla tarkasteltaessa asetelma vastaisi sitä, että henkilö valitaan otokseen ja haastatellaan mahdollisimman pian yhden kerran. Todellisuudessa siitä hetkestä, kun otos valitaan ja henkilöä haastatellaan viimeisen, eli viidennen kerran, aikaa on kulunut vähimmillään 17 kuukautta ja enimmillään 23 kuukautta. On selvää, että tuona aikana jotkin taustamuuttujat voivat muuttua merkittävästikin eikä tulevaisuutta voida tarkasti ennustaa. Tästä syystä voisi olla perusteltua käyttää taustamuuttujina henkilön valintahetken taustamuuttujia. Valintahetken taustamuuttujia ei tässä tutkielmassa ole saatavilla kuin yhdelle muuttujalle. Käyttämällä ensimmäisen haastattelukerran taustamuuttujia muiden haastattelukertojen osalta luultavasti saavutetaan lähes sama tulos, sillä tällöin valinnan ja haastattelun väli on enintään 7 kuukautta ja vähintään 2 kuukautta, jolloin korrelaation voisi olettaa olevan suurempi kuin pidemmällä ajalla.

Tästä syystä seuraavaksi tarkastellaan kahta asetelmaa. Ensimmäiseksi tarkastellaan sitä, että otos valitaan ja henkilöä haastatellaan mahdollisimman pian ainoastaan yhden kerran. Kutsutaan tätä asetelmaksi A.

Asetelmassa H asetetaan ensimmäisen asetelman aineistosta henkilön jokaisen haastattelukerran taustamuuttujat samoiksi kuin ensimmäisellä haastattelukerralla. Sen lisäksi yhdistetään jokaiseen henkilöön ensimmäistä haastattelukertaa edeltävän maalis- tai syyskuun pääasiallisen toiminnan rekisteritieto. Jaetaan tämä asetelma haastattelukerran mukaan H1, H2, H3, H4 ja H5.

Asetelmassa H on mukana vain ne henkilöt, keiltä ensimmäinen haastattelukerta löytyy asetelmasta A. Asetelmassa H on yhteensä 249 170 riviä ja 76 414 henkilöä. Tarkemmin rivimäärä haastattelukerroittain näkyy taulukossa 3.

Taulukko 3: Asetelma H:n rivimäärä haastattelukerroittain

H1	**H2**	**H3**	**H4**	**H5**
58183	55455	51236	43664	40632

Perusteltua voisi olla rajata aineisto vain sellaisiin henkilöihin, joista on saatavilla kaikki viisi haastattelua, mutta tällöin henkilöitä olisi vain 23 912 ja rivejä 119 560.

Nyt muodostettu asetelma A on pseudopopulaatio, joka kuvaa henkilöitä, jotka on valittu otokseen ja haastateltu samanaikaisesti. Asetelmissa H1, ..., H5 kaikkia henkilöitä haastatellaan samalla ajanhetkellä, mutta heidät on valittu otokseen eri ajanhetkellä. Tällöin pseudopopulaatiot H1, ..., H5 ovat kunakin ajankohtana valittujen henkilöiden populaatiot.

Oikean työvoimatutkimuksen kuukausiotosta vastaava otos saataisiin valitsemalla samankokoiset otokset kustakin pseudopopulaatiosta H1, ..., H5 ja yhdistämällä nämä yhdeksi otokseksi. Tämän tutkielman tarkastelussa tätä otosta ei muodosteta, sillä vertailuun riittää tarkastella osapopulaatioita erikseen.

7.2 Aineiston muuttujat

Aineisto koostuu kahdentyyppisistä muuttujista, sellaisista, jotka saadaan selvitettyä työvoimatutkimuksen haastatteluilla, ja sellaisista, jotka saadaan rekistereistä. Nämä muuttujat saadaan yhdistettyä toisiinsa linkkimuuttujilla (taulukko 4). Yksilöintitunnus *kohdeno* yksilöi haastateltavan henkilön ja kullakin haastateltavalla on enintään viisi havaintoa eli haastattelukertaa. Tämän haastattelukerran yksilöi *kerta* -muuttuja. Haastattelukerran vuoden ja kuukauden yksilöi *vuosi* ja *kk* -muuttujat.

Taulukko 4: Työvoimatutkimuksen linkkimuuttujat

Muuttuja	**Selite**
kohdeno	Yksilöintitunnus
vuosi	Vuosi
kk	Kuukausi
kerta	Haastettelukerta

7.3 Työvoimatutkimuksen muuttujat

Työvoimatutkimuksessa kysytään tietoja henkilön työmarkkina-asemasta. Kysymyksiä on useampia, mutta tämän tutkielman kannalta mielekkäimmät muuttujat ovat taulukossa 5. Työvoimatutkimuksen haastattelu on onnistunut kun henkilöltä on saatu vastaus. Toisinaan henkilöä ei tavoiteta, vastausta ei saada tai vastausta ei muuten voida käyttää. Syitä on myös monia muitakin. Muuttujassa *loptulo* on erilaisia luokituksia 25 kappaletta. Pääasiallinen toiminta -tieto on myös Tilastokeskuksen muista rekisteritiedoista muodostettu tilastotieto, mutta työvoimatutkimuksen haastattelussa kysytään henkilöltä itseltään hänen pääasiallista toimintaa 9-luokkaisella jaolla muuttujaan *omato_2021*. Tätä muuttujaa ei tässä tutkielmassa tämän enempää tarkastella. Syy miksi työvoimatutkimus ylipäätään toteutaan on haastattelulla saatu tieto henkilön työmarkkina-asemasta (muuttuja *tyvo*). Työmarkkina-asema on neljäluokkainen: työllinen, työtön, varusmies/siviilipalvelus ja muut työvoimaan kuulumattomat. Erityisesti työllinen ja työtön ovat mielenkiinnon kohteena.

Taulukko 5: Työvoimatutkimuksen kyselyllä saadut muuttujat

Muuttuja	**Selite**	
loptulo	Saatu haastattelu/ kadon syy	25-luokkainen muuttuja ”01” = vastaus saatu
omato_2021	Pääasiallinen toiminta, oma valinta	9-luokkainen muuttuja
tyvo	Työmarkkina-asema	4-luokkainen muuttuja ”1” = työllinen ”2” = työtön ”3” = varusmies/siviilipalvelus ”4” = muut työvoimaan kuulumattomat

7.4 Taustamuuttujat

Työvoimatutkimuksen rekisteripohjaisina taustamuuttujina on useita erilaisia muuttujia. Taulukoissa 6 ja 7 on lueteltu valikoidusti näitä muuttujia. Asetelmissa A ja H on toisiaan vastaavat muuttujat sillä erotuksella, että asetelmassa A muuttuja vastaa kutakin haastatteluhetkeä, kun taas asetelmassa H muuttuja vastaa ensimmäisen haastattelukerran tietoa. Asetelmassa H muuttuja *ptoim5_kk* kuvastaa pääasiallista toimintaa ensimmäistä haastattelua edeltävänä maaliskuuna tai syyskuuna, joka vastaa jokseenkin otokseen valintahetken ajankohtaa. Kyseinen muuttuja kuvastaa rekisteritietoa kyseisen henkilön pääasiallisesta toiminnasta. Tässä muuttujassa *ptoim5_kk* oli 973 puuttuvaa havaintoa, ja kyseiset havainnot korvattiin vuositason *kalib_ptoim* -muuttujan vastaavilla arvoilla. Lähes kaikki nämä olivat *kalib_ptoim* -muuttujassa opiskelijoita.

Taulukko 6: Rekisterimuuttujat, osa 1

Muuttuja	**Selite**	
kalib_kieli	äidinkieli	3-luokkainen väestörekisteristä ”fi” = suomi, saame ”sv” = ruotsi, tanska, norja ”muu” = muu äidinkieli
kk_ptoim5	pääasiallinen toiminta	kuukausitieto kokeellisesta tilastosta 5-luokkainen ”tyol” = työllinen ”tyot” = työtön ”opis” = opiskelija ”elak” = eläkeläinen ”muu” = muu
kalib_tutrek	koulutusaste	3-luokkainen tutkintorekisteristä ”perus” = perusaste tai puuttuva ”toinen” = toinen aste ”korkea” = korkea-aste
kalib_tulorek	palkka	11-luokkainen muuttuja ”NA” = ei tuloja ”01” - ”10” = palkkakymmenykset
kalib_tyorek	työnhakija	5-luokkainen TEM työnhakijarekisteri ”ei” = ei-työnvälitystilastossa ”tyotloM1534” = työtön tai lomautettu mies 15-34v ”tyotloMyli35” = työton tai lomautatettu mies yli 35v ”tyotN1534” = työtön tai lomautettu nainen 15-34v ”tyotNyli35” = työton tai lomautettu nainen yli 35v
sukup	sukupuoli	2-luokkainen väestörekisteristä mies, nainen
asku	asuinkunta	kolminumeroinen kuntakoodi
ika	henkilön ikä	kokonaisluku väliltä 15 ja 74
suuralue_2012	suuralue/ NUTS2 -alueluokitus	asuinpaikan mukaan 5-luokkainen ”hkiuma” = Helsinki ja Uusimaa ”etel” = Etelä-Suomi ”lans” = Länsi-Suomi ”itapohj” = Pohjois- ja Itä-Suomi ”ahv” = Ahvenanmaa

Taulukko 7: Rekisterimuuttujat, osa 2

Muuttuja	**Selite**	
etuudet	etuudet	Onko henkilö saanut etuuksia kuten eläkettä, työttömyysetuutta, opintorahaa edellisenä kuukautena ”ei” = ei tietoa tai 0 euroa ”q1” = alle 644 euroa ”q2” = 644-1219 euroa ”q3” = 1219-1989 euroa ”q4” = yli 1989 euroa
kuntaluok	Kuntaryhmitys	Asuinkunta 3-luokkainen ”kaup” = Kaupunkimaiset kunnat ”taajam” = Taajaan asutut kunnat ”maaseu” = Maaseutumaiset kunnat
ika10	ikä 10-vuosi	johdettu *ika* 6-luokkainen ”i1524” = 15-24v ”i25-34” = 25-34v ”i35-44” = 35-44v ”i45-54” = 45-54v ”i55-64” = 55-64v ”i65-74” = 65-74v

7.5 Aineiston ajallinen rakenne

Asetelmissa H1, ..., H5 aineisto on jaettu haastattelukerran mukaan. Jokaisesta henkilöstä tiedetään haastatteluhetkeä vastaavat taustamuuttujat sekä valintahetkeä vastaavat taustamuuttujat. Valintahetken taustamuuttujina käytetään ensimmäisen haastattelukerran tietoja, paitsi pääasiallinen toiminta -rekisteritiedon osalta todellisen valintahetken tietoa. Vertaamalla otokseen valintahetken apumuuttujien indikaattoria haastatteluhetken samojen muuttujien apumuuttujien indikaattoreihin, nähdään miten paljon muuttujien arvot muuttuvat ajan myötä. Taulukoissa 8 ja 9 on nähtävillä apumuuttujien indikaattoreiden valinta- ja haastatteluhetken välinen korrelaatio kunkin haastattelukerran perusteella (H1, ..., H5).

Ensimmäinen havainto on, että muuttujien arvot eivät merkittävästi muutu haastattelukertojen välillä. Sukupuoli on aineistossa pysyvin ominaisuus, sen korrelaatio haastattelukertojen välillä on 1. Ikäluokissa taas muutos johtuu normaalista vanhenemisesta. Jos henkilö pysyy elossa, joka vuosi hänelle tulee yksi ikävuosi lisää. Aineistossa henkilöt eivät myöskään erityisen paljon muuta suuralueelta toiselle; korrelaatio samalla suuralueella pysymisellä on viidennenkin haastattelukerran kohdalla reilu 0,97. Huomionarvoista on, että Ahvenanmaalla korrelaatio on 0,993. Kun huomioi tilastollisen kuntaryhmityksen, ovat korrelaatiot hiuksen hienosti matalammat, kuten viidennen haastattelukerran kohdalla reilu 0,955.

Pääasiallinen toiminta -rekisteritiedon kohdalla korrelaatiot laskevat. Ensinnäkin kyseessä olevalla muuttujalla on käytössä todellista valintahetkeä vastaava muuttujan arvo, kun muissa muuttujissa sen tilalla käytetään ensimmäisen haastatteluhetken arvoa. Tästä syystä ensimmäisen haastattelukerran ja valintahetken välillä korrelaatio ei ole yksi kuten muilla muuttujilla. Ensimmäisen haastattelukerran ja valintahetken välillä pääasiallisessa toiminnassa työllisille korrelaatio on 0,809 ja se laskee viidenteen haastattelukertaan arvoon 0,700. Työttömyyden osalta korrelaatio on matalampaa. Ensimmäisen haastattelukerran ja valintahetken välillä korrelaatio on 0,7 ja se laskee viidenteen haastattelukertaan mennessä tasolle 0,512. Opiskelun osalta korrelaatio on ensimmäisessä haastatteluhetkessä 0,95 ja viidennessä 0,889. Muiden pääasiallisten toiminnan osalta korrelaatio on ensimmäisessä haastattelukerrassa 0,489 ja viidennessä 0,279. Työllisyyden osalta korkeampi korrelaatio on odotettavaa, sillä työllisyys on yleensä pitkäkestoisempaa kuin työttömyys.

Tulorekisterin palkkakymmenysten osalta korrelaatiot vaihtelevat. Tulorekisterin perusteella tulottomien henkilöiden korrelaatio valintahetken ja toisen haastattelukerran välillä on 0,831 ja viidennen haastattelukerran 0,722. Eri tulokymmenyksissä korrelaatiot vaihtelevat merkittävästi, vaihteluväli on 0,293:sta 0,623:een toisen haastattelukerran osalta. Viidennen haastattelukerran osalta korrelaatioiden vaihteluväli on 0,23 ja 0,574 välillä. Suurimmat korrelaatiot ovat ensimmäisessä ja viimeisessä tulokymmenyksessä. Toisin sanottuna pienet ja suuret tulot ovat pysyvämpiä kuin tulot siltä väliltä. Voisi olla myös perusteltua käyttää tulokymmenysten sijaan karkempaa jaottelua, kuten kvantiileja. Se luultavasti nostaisi korrelaatioita, sillä pienet muutokset eivät enää siirtäisi henkilöä tulotasoluokasta toiseen yhtä helposti.

Etuuksien saaminen valintahetkellä ja haastatteluhetkellä on myös korreloitunutta. Henkilöillä, jotka eivät saa etuuksia, korrelaatio valintahetkestä toiseen haastattelukertaan on 0,872 ja viidenteen 0,745. Henkilöiden, jotka ovat ensimmäistä etuuskvantiilia, korrelaatio valintahetken ja toisen haastattelukerran välillä on 0,576

ja viidennessä 0,365. Suurin korrelaatio on neljännessä etuuskvantiilissa, siinä korrelaatio valintahetken ja toisen haastattelukerran välillä on 0,935 ja viidennen haastattelukerran välillä 0,842.

Henkilön äidinkieli on hyvin stabiili. Korrelaatio valintahetken ja toisen haastattelukerran välillä on 0,999 ja viidennen haastattelukerran yli 0,996. Henkilöiden korkein suoritettu tutkinto on myös hyvin stabiili, korrelaatio valintahetken ja toisen haastattelukerran välillä on yli 0,986. Viidennen haastettelukerran välillä eroa syntyy eri tutkintotasojen välillä. Peruskoulun suoritus korkeimpana tutkintona korrelaatio valintahetken ja viidennen haastattelun välillä on 0,923, toisen asteen 0,93 ja korkean asteen 0,971. Tämä indikoi sitä, että korkean asteen tutkintoja suoritetaan tutkimuksen aikana vähemmän kuin muita tutkintoja.

Työvoimaviranomaisten työnvälitystilaston osalta korrelaatio on valintahetken ja haastatteluhetken välillä hieman riippuvainen iästä ja vähemmän sukupuolesta. Toisen haastattelukerran ja valintahetken välillä korrelaatio on 15-34-vuotiaiden osalta 0,646 miehillä ja 0,621 naisilla. Yli 35-vuotiailla korrelaatio on miehillä 0,787 ja naisilla 0,741. Viidenteen haastattelukertaan mennessä korrelaatio laskee. Voimakkainta lasku on 15-34-vuotiailla naisilla, heillä korrelaatio on viidennellä kerralla vain 0,262. Samanikäisillä miehillä se on 0,352. Yli 35-vuotiaden osalta korrelaatio on miehillä 0,555 ja naisilla 0,477.

Taulukko 8: Valinta- ja haastatteluhetken taustamuuttujien autokorrelaatio, osa 1

	H1	H2	H3	H4	H5
sukup_mies	1	1	1	1	1
sukup_nainen	1	1	1	1	1
ika10_i3544	1	0,971	0,942	0,886	0,85
ika10_i1524	1	0,987	0,971	0,943	0,929
ika10_i2534	1	0,972	0,944	0,892	0,86
ika10_i4554	1	0,969	0,94	0,877	0,843
ika10_i5564	1	0,969	0,942	0,881	0,852
ika10_i6574	1	0,985	0,972	0,942	0,924
suuralue_2012_hkiuma	1	0,994	0,988	0,976	0,971
suuralue_2012_etel	1	0,993	0,988	0,975	0,971
suuralue_2012_lans	1	0,995	0,989	0,978	0,973
suuralue_2012_itapohj	1	0,996	0,99	0,982	0,979
suuralue_2012_ahv	1	0,998	0,997	0,993	0,993
kuntaluok_kaup	1	0,991	0,982	0,963	0,956
kuntaluok_taajam	1	0,991	0,982	0,963	0,957
kuntaluok_maaseu	1	0,99	0,981	0,963	0,956
kk_ptoim5_tyol	0,809	0,784	0,762	0,721	0,700
kk_ptoim5_tyot	0,7	0,642	0,597	0,526	0,512
kk_ptoim5_opis	0,725	0,687	0,659	0,6	0,571
kk_ptoim5_elak	0,95	0,939	0,927	0,904	0,889
kk_ptoim5_muu	0,489	0,401	0,343	0,29	0,279

Taulukko 9: Valinta- ja haastatteluhetken taustamuuttujien autokorrelaatio, osa 2

	H1	H2	H3	H4	H5
kalib_tulorek_NA	1	0,831	0,789	0,770	0,722
kalib_tulorek_01	1	0,411	0,366	0,353	0,276
kalib_tulorek_02	1	0,358	0,311	0,304	0,238
kalib_tulorek_03	1	0,357	0,322	0,331	0,23
kalib_tulorek_04	1	0,357	0,330	0,346	0,261
kalib_tulorek_05	1	0,311	0,284	0,325	0,249
kalib_tulorek_06	1	0,293	0,278	0,332	0,232
kalib_tulorek_07	1	0,301	0,281	0,378	0,241
kalib_tulorek_08	1	0,323	0,310	0,430	0,275
kalib_tulorek_09	1	0,401	0,380	0,529	0,35
kalib_tulorek_10	1	0,623	0,601	0,695	0,574
etuudet_ei	1	0,872	0,836	0,805	0,745
etuudet_q1	1	0,576	0,514	0,506	0,365
etuudet_q2	1	0,758	0,705	0,643	0,582
etuudet_q3	1	0,862	0,811	0,735	0,692
etuudet_q4	1	0,935	0,906	0,865	0,842
kalib_kieli_fi	1	0,999	0,999	0,997	0,997
kalib_kieli_swe	1	0,999	0,999	0,996	0,996
kalib_kieli_muu	1	0,999	0,999	0,998	0,997
kalib_tutrek_toinen	1	0,987	0,976	0,95	0,93
kalib_tutrek_perus	1	0,986	0,974	0,951	0,923
kalib_tutrek_korkea	1	0,995	0,990	0,977	0,971
kalib_tyorek_ei	1	0,718	0,589	0,497	0,445
kalib_tyorek_tyotloM1534	1	0,646	0,495	0,374	0,352
kalib_tyorek_tyotloMyli35	1	0,787	0,664	0,602	0,555
kalib_tyorek_tyotN1534	1	0,621	0,489	0,362	0,262
kalib_tyorek_tyotNyli35	1	0,741	0,631	0,527	0,477

8 Aineiston mallinnus logistisella regressiolla

Apumuuttujien tulisi olla mahdollisimman hyvin korreloituneita tarkasteltavan muuttujan suhteen, eli tässä tapauksessa työmarkkina-aseman kanssa. Työmarkkina-asema on useampiluokkainen luokkamuuttuja. Tarkastelun kannalta tärkeimmät työmarkkina-asemamuuttujan luokat ovat työttömyys ja työllisyys, joita tarkastellaan omina binäärisinä muuttujina erikseen omissa malleissaan. Pyritään löytämään sellaiset apumuuttujat, jotka selittävät mahdollisimman hyvin työllisyyttä ja työttömyyttä. Kun otos on tasapainotettu tällaisten apumuuttujien suhteen, pitäisi estimaattien olla tarkempia. Apumuuttujien ei tulisi olla keskenään korreloituneita, saati multikollineaarisia, eli lineaarisesti riippuvia. [5] [7]

Tarkastelemme mahdollisten apumuuttujien soveltuvuutta edellisessä osiossa kuvatuissa asetelmissa A ja H työllisyyden ja työttömyyden mallintamisessa. Tarkastelu toteutetaan testaamalla apumuuttujia binäärisellä logistisella regressiomallilla.

Logistinen regressiomalli on yleistetty lineaarinen malli, jolla mallinnetaan binääristä vastemuuttujaa selittävien muuttujien avulla. [13] Mallissa jokaiseen selittävään muuttujaan liittyy parametri β_j, jonka arvo e^{β_j} kertoo, miten paljon kyseisen vastemuuttujan yhden yksikön muutos vaikuttaa vastemuuttujan ristisuhteeseen. Jos β_j on positiivinen, vastemuuttujan ristisuhde, eli sitä kautta myös todennäköisyys, kasvaa. Jos β_j on negatiivinen, niin todennäköisyys vastaavasti laskee. Mitä lähempänä arvoa nolla β_j on, sitä pienempi vaikutus on. Kaikki selittävät muuttujat ovat tarkasteltavassa mallissa kuvattu binäärisinä indikaattorimuuttujina, joten niitä vastaavien parametrien tulkinta on keskenään sama. Parametreille on laskettu estimaatit ja taulukkoon on sulkujen sisälle merkitty myös keskivirhe. Näiden lisäksi on merkitty p-arvolle indikaattori tähdin. P-arvo kuvastaa sitä todennäköisyyttä, joka oletusten hallitessa, olisi saada havaittua arvoa äärimmäisempi tunnusluvun arvo jos parametrin todellinen arvo olisi nolla. Yleisesti on sanottu, että kun p-arvo on alle 0,05 tason, on parametrin arvo tilastollisesti merkitsevästi nollasta poikkeava.

Logistisen regression sopivuuden mittareita on useita. Esimerkiksi Akaiken informaatikriteeri (AIC) ja Bayes-informaatiokriteeri (BIC) kuvaavat mallin sopivuutta aineistoon. Mitä pienempi arvo on, sen sopivampi malli. Tämän lisäksi devianssi ja log-uskottavuus kertovat myös mallin sopivuudesta. Helppotajuisin mittari on Tjurin pseudoselitysaste (R2), joka saa arvoja 0 ja 1 väliltä. Arvo 0 kuvastaa selitysvoiman olevan nolla ja 1 sen olevan täydellinen. [14] Malleja voi vertailla keskenään vertailemalla näitä tunnuslukuja.

Määritelmä 10. Tjurin pseudoselitysaste (2009) lasketaan seuraavalla tavalla:

$$R^2_{Tjur} = \frac{1}{n_1}\sum_{i=1}^{n}\hat{p}(y_i = 1) - \frac{1}{n_0}\sum_{i=1}^{n}\hat{p}(y_i = 0), \tag{7}$$

jossa n_1 on montako kertaa binäärinen muuttuja saa aineistossa arvon 1 ja n_2 montako kertaa binäärinen muuttuja saa aineistossa arvon 0. Funktio $\hat{p}()$ on mallin antama sovitetodennäköisyys kyseiselle havainnolle.

Muuttujiksi logistiseen regressioon valitaan kaikki osiossa mainitut 7.4 apumuuttujat vuosikohtaista ikää ja asuinkuntaa lukuun ottamatta eli yhteensä 10 luokkamuuttujaa. Mallissa nämä on kuvattu 39 lineaarisesti riippumattomana dummymuuttujana. Jokaisesta luokkamuuttajasta on valittu yksi arvo viitearvoksi, joten sitä ei mallissa erikseen näy, ja näkyvien arvojen parametrien arvoja tulee tulkita suhteessa tähän viitearvoon. Tämä siksi, jotta muuttujat olisivat lineaarisesti riippumattomia keskenään.

Mallissa A taustamuuttujat ovat ajallisesti lähempänä haastatteluhetkeä kuin H-malleissa, joten tämän takia voisi olettaa, että mallissa A estimaatit saavat voimakkaampia arvoja kuin H-malleissa. Näiden muuttujien yhteyttä työllisyyteen ja työttömyyteen tarkastellaan erikseen.

8.1 Työttömyyden ja työllisyyden mallinnus asetelmassa A

Asetelmassa A taustamuuttujien, eli mallin selittävien muuttujien ajallinen väli haastatteluhetkeen on kaikista lyhin. Tämän olettaisi tekevän mallista selitysvoimaisemman. Logistisen regression parametrien estimaatit ja mallin tunnusluvut nähdään taulukoista 10, 11 ja 12. Oleellisin havainto on, että Tjurin pseudoselitysasteen mukaan työllisyyden selitysaste (0,7484) on merkittävästi korkeampi kuin työttömyyden (0,3111). Tämä viittaa siihen, että käytetyt selittävät muuttujat selittävät heikommin työttömyyttä kuin työllisyyttä.

Koska kaikki mallin selittävät muuttujat ovat binäärisiä dummy-muuttujia, on niiden tulkinta sama, ja tällöin niitä vertailemalla voi tehdä johtopäätöksiä. Itseisarvoltaan suurimmat kertoimet työllisyysasteeseen saavat kk_ptoim5 -muuttujan parametrit, jotka saavat negatiivisia arvoja. Tämä viittaisi siihen, että kyseinen muuttuja on selitysvoimaisin. Jos henkilön pääasiallinen kuukausitoiminta on rekisteritiedon mukaan jotakin muuta kuin työllinen, on tällöin henkilö epätodennäköisemmin haastattelun perusteella työllinen. Korkeammat tulot tulorekisterin mukaan indikoivat korkeampaa työllisyysastetta. Etuuksien saaminen taas indikoi negatiivisesti työllisyyttä samoin kuin työvoimaviranomaisten rekisterissä oleminen.

Työttömyyden osalta parhaiten sitä selittää pääasiallinen toiminta, etuuksien saaminen ja työvoimaviranomaisten rekisterissä oleminen. Matalampaan työttömyyteen yhteydessä taas ovat 65-74-vuoden ikä (he ovat oletettavasti eläkeläisiä) ja korkeat tulot.

Yleisesti voi sanoa valittujen apumuuttujien selittävän työllisyyttä ja työttömyyttä oikein hyvin asetelmassa A.

Taulukko 10: Malli A, työllisyys ja työttömyys, osa 1

asetelma A	työllisyys	työttömyys
(Intercept)	$2,713^{**}$	$-3,964^{**}$
	$(0,036)$	$(0,053)$
sukupnainen	$-0,361^{**}$	$-0,017$
	$(0,016)$	$(0,028)$
ika10i1524	$-1,025^{**}$	$0,553^{**}$
	$(0,034)$	$(0,046)$
ika10i2534	$-0,412^{**}$	$0,318^{**}$
	$(0,032)$	$(0,043)$
ika10i4554	$-0,007$	$-0,123^{**}$
	$(0,034)$	$(0,040)$
ika10i5564	$-0,152^{**}$	$-0,317^{**}$
	$(0,031)$	$(0,038)$
ika10i6574	$-0,700^{**}$	$-1,474^{**}$
	$(0,037)$	$(0,082)$
suuralue_2012etel	$-0,040$	$-0,036$
	$(0,023)$	$(0,030)$
suuralue_2012lans	$-0,038$	$-0,042$
	$(0,022)$	$(0,029)$
suuralue_2012itapohj	$-0,117^{**}$	$-0,210^{**}$
	$(0,023)$	$(0,032)$
suuralue_2012ahv	$0,112$	$-0,353^{*}$
	$(0,075)$	$(0,155)$
kuntaluoktaajam	$0,248^{**}$	$-0,133^{**}$
	$(0,022)$	$(0,034)$
kuntaluokmaaseu	$0,379^{**}$	$-0,145^{**}$
	$(0,024)$	$(0,039)$
kk_ptoim5tyot	$-3,134^{**}$	$1,837^{**}$
	$(0,033)$	$(0,038)$
kk_ptoim5opis	$-3,221^{**}$	$1,332^{**}$
	$(0,029)$	$(0,044)$
kk_ptoim5elak	$-2,791^{**}$	$-1,072^{**}$
	$(0,032)$	$(0,074)$
kk_ptoim5muu	$-2,842^{**}$	$1,211^{**}$
	$(0,029)$	$(0,050)$

Parametrin estimaatti, suluissa keskivirhe, $^{**}p < 0,01$; $^{*}p < 0,05$

Taulukko 11: Malli A, työllisyys ja työttömyys, osa 2

asetelma A:	työllisyys	työttömyys
kalib_tulorek01	$1,075^{**}$	$-0,224^{**}$
	$(0,025)$	$(0,040)$
kalib_tulorek02	$1,464^{**}$	$-0,502^{**}$
	$(0,032)$	$(0,051)$
kalib_tulorek03	$1,640^{**}$	$-0,722^{**}$
	$(0,043)$	$(0,070)$
kalib_tulorek04	$1,718^{**}$	$-0,969^{**}$
	$(0,051)$	$(0,083)$
kalib_tulorek05	$1,749^{**}$	$-1,303^{**}$
	$(0,054)$	$(0,097)$
kalib_tulorek06	$1,776^{**}$	$-1,434^{**}$
	$(0,055)$	$(0,103)$
kalib_tulorek07	$1,836^{**}$	$-1,682^{**}$
	$(0,058)$	$(0,111)$
kalib_tulorek08	$1,845^{**}$	$-1,580^{**}$
	$(0,060)$	$(0,107)$
kalib_tulorek09	$1,892^{**}$	$-1,579^{**}$
	$(0,060)$	$(0,105)$
kalib_tulorek10	$1,757^{**}$	$-1,367^{**}$
	$(0,056)$	$(0,093)$
etuudetq1	$-0,907^{**}$	$0,709^{**}$
	$(0,026)$	$(0,030)$
etuudetq2	$-1,457^{**}$	$0,932^{**}$
	$(0,033)$	$(0,037)$
etuudetq3	$-1,600^{**}$	$1,370^{**}$
	$(0,036)$	$(0,045)$
etuudetq4	$-1,822^{**}$	$1,347^{**}$
	$(0,038)$	$(0,057)$
kalib_kieliswe	$0,023$	$-0,238^{**}$
	$(0,036)$	$(0,062)$
kalib_kielimuu	$-0,259^{**}$	$0,351^{**}$
	$(0,029)$	$(0,031)$
kalib_tutrekperus	$-0,500^{**}$	$-0,184^{**}$
	$(0,022)$	$(0,028)$
kalib_tutrekkorkea	$0,354^{**}$	$0,186^{**}$
	$(0,019)$	$(0,028)$

Parametrin estimaatti, suluissa keskivirhe, $^{**}p < 0,01$; $^{*}p < 0,05$

Taulukko 12: Malli A, työllisyys ja työttömyys, osa 3

asetelma A:	työllisyys	työttömyys
kalib_tyorektyotloM1534	$-0,775^{**}$	$1,448^{**}$
	$(0,065)$	$(0,053)$
kalib_tyorektyotloMyli35	$-1,314^{**}$	$1,841^{**}$
	$(0,048)$	$(0,044)$
kalib_tyorektyotN1534	$-0,565^{**}$	$1,177^{**}$
	$(0,073)$	$(0,058)$
kalib_tyorektyotNyli35	$-1,149^{**}$	$1,571^{**}$
	$(0,051)$	$(0,045)$
AIC	124945	66160
BIC	125358	66573
Log-uskottavuus	-62434	-33041
Devianssi	124867	66082
Tjurin pseudo-R2	$0,748$	$0,311$
Havaintojen lkm,	294247	294247

Parametrin estimaatti, suluissa keskivirhe, $^{**}p < 0,01$; $^{*}p < 0,05$

8.2 H-asetelmien mallinnus

Asetelmissa H1, ..., H5 tarkastellaan eri haastattelukertoja. Näissä korostuu se, että otos on valittu eri ajankohtana kuin haastattelu on toteutettu. Siinä missä H1-asetelmassa otokseen valinnan ja haastattelun välillä on reilun yhden kuukauden ja 7 kuukauden väliltä aikaa, on asetelmassa H5 väli 17 kuukaudesta 23 kuukauteen (osio 2). Tämä ajallinen väli voi heikentää mallin selitysastetta, sillä valintahetken tiedot eivät enää vastaa sitä mitä ne ovat haastatteluhetkellä. Henkilön rekisterissä oleva tieto pääasiallisesta toiminnasta voi olla muuttunut, henkilö on voinut muuttaa, tulot ovat voineet muuttua ja niin edelleen.

8.3 Työllisyys H-asetelmassa

Työllisyys asetelmissa H1, ..., H5 logistisen regression avulla mallinnettuna näkyy taulukoissa 13, 14, 15. Päällimmäinen havainto on, että työllisyyden Tjurin R2-pseudoselitysaste laskee hieman, mitä myöhäisemmästä haastattelukerrasta on kyse. Huomionarvoista on myös, että vaikka ajallinen yhteys heikkenee, on muuttujilla silti selitysvoimaa jäljellä viidennelläkin haastattelukerralla.

Periaatteessa asetelma H1 on asetelman A osajoukko, jossa kk_ptoim5 -muuttujan tilalle on asetettu valintahetkeä vastaava pääasiallisen toiminnan rekisteritieto. Kun verrataan pseudoselitysastetta työllisyyden osalta mallien A (0,7484) ja H1 (0,6934) välillä, on mallissa A hieman suurempi pseudoselitysaste. Tämä voi toki johtua siitä, että H1 on osajoukko, mutta myös siitä, että ajantasaisempi tieto pääasiallisesta toiminnasta parantaa selitysvoimaa.

Taulukko 13: Työllisyysmallit asetelmassa H haastattelukerran mukaan, osa 1

työllisyys	H1	H2	H3	H4	H5
(Intercept)	$1,977^{**}$	$1,646^{**}$	$1,635^{**}$	$1,611^{**}$	$1,706^{**}$
	$(0,071)$	$(0,065)$	$(0,065)$	$(0,068)$	$(0,070)$
sukupnainen	$-0,416^{**}$	$-0,302^{**}$	$-0,261^{**}$	$-0,205^{**}$	$-0,140^{**}$
	$(0,034)$	$(0,032)$	$(0,032)$	$(0,033)$	$(0,033)$
ika10i1524	$-1,139^{**}$	$-0,863^{**}$	$-0,938^{**}$	$-0,885^{**}$	$-0,908^{**}$
	$(0,071)$	$(0,064)$	$(0,064)$	$(0,067)$	$(0,068)$
ika10i2534	$-0,614^{**}$	$-0,477^{**}$	$-0,412^{**}$	$-0,326^{**}$	$-0,369^{**}$
	$(0,066)$	$(0,058)$	$(0,058)$	$(0,061)$	$(0,061)$
ika10i4554	$0,046$	$0,020$	$-0,003$	$-0,013$	$-0,094$
	$(0,071)$	$(0,062)$	$(0,060)$	$(0,063)$	$(0,063)$
ika10i5564	$-0,077$	$-0,256^{**}$	$-0,566^{**}$	$-0,727^{**}$	$-0,908^{**}$
	$(0,065)$	$(0,057)$	$(0,055)$	$(0,056)$	$(0,056)$
ika10i6574	$-0,315^{**}$	$-0,745^{**}$	$-1,132^{**}$	$-1,297^{**}$	$-1,277^{**}$
	$(0,080)$	$(0,074)$	$(0,075)$	$(0,079)$	$(0,080)$
suuralue_2012etel	$-0,042$	$0,042$	$0,025$	$-0,013$	$-0,020$
	$(0,046)$	$(0,043)$	$(0,044)$	$(0,045)$	$(0,045)$
suuralue_2012lans	$0,021$	$0,032$	$0,090^{*}$	$0,004$	$0,039$
	$(0,044)$	$(0,041)$	$(0,041)$	$(0,043)$	$(0,043)$
suuralue_2012itapohj	$-0,107^{*}$	$-0,035$	$-0,040$	$-0,083$	$-0,129^{**}$
	$(0,047)$	$(0,043)$	$(0,044)$	$(0,046)$	$(0,046)$
suuralue_2012ahv	$0,014$	$0,149$	$0,199$	$0,040$	$0,193$
	$(0,167)$	$(0,152)$	$(0,151)$	$(0,159)$	$(0,161)$
kuntaluoktaajam	$0,224^{**}$	$0,249^{**}$	$0,231^{**}$	$0,225^{**}$	$0,221^{**}$
	$(0,045)$	$(0,043)$	$(0,043)$	$(0,045)$	$(0,046)$
kuntaluokmaaseu	$0,343^{**}$	$0,316^{**}$	$0,337^{**}$	$0,314^{**}$	$0,352^{**}$
	$(0,050)$	$(0,047)$	$(0,048)$	$(0,050)$	$(0,051)$
ptoim5_orgtyot	$-1,583^{**}$	$-1,645^{**}$	$-1,565^{**}$	$-1,315^{**}$	$-1,458^{**}$
	$(0,062)$	$(0,057)$	$(0,058)$	$(0,062)$	$(0,063)$
ptoim5_orgopis	$-1,751^{**}$	$-1,511^{**}$	$-1,380^{**}$	$-1,127^{**}$	$-1,159^{**}$
	$(0,056)$	$(0,052)$	$(0,053)$	$(0,057)$	$(0,058)$
ptoim5_orgelak	$-1,581^{**}$	$-2,005^{**}$	$-2,043^{**}$	$-1,934^{**}$	$-1,974^{**}$
	$(0,062)$	$(0,060)$	$(0,062)$	$(0,067)$	$(0,068)$
ptoim5_orgmuu	$-1,769^{**}$	$-1,400^{**}$	$-1,235^{**}$	$-1,079^{**}$	$-1,151^{**}$
	$(0,067)$	$(0,061)$	$(0,063)$	$(0,068)$	$(0,069)$

Parametrin estimaatti, suluissa keskivirhe, $^{**}p < 0,01$; $^{*}p < 0,05$

Taulukko 14: Työllisyysmallit asetelmassa H haastattelukerran mukaan, osa 2

työllisyys	H1	H2	H3	H4	H5
kalib_tulorek01	$1,548^{**}$	$1,089^{**}$	$0,953^{**}$	$0,912^{**}$	$0,858^{**}$
	$(0,049)$	$(0,049)$	$(0,052)$	$(0,055)$	$(0,057)$
kalib_tulorek02	$2,291^{**}$	$1,355^{**}$	$1,129^{**}$	$1,022^{**}$	$0,830^{**}$
	$(0,063)$	$(0,057)$	$(0,058)$	$(0,061)$	$(0,063)$
kalib_tulorek03	$2,529^{**}$	$1,527^{**}$	$1,364^{**}$	$1,260^{**}$	$0,953^{**}$
	$(0,089)$	$(0,071)$	$(0,070)$	$(0,074)$	$(0,073)$
kalib_tulorek04	$2,580^{**}$	$1,833^{**}$	$1,620^{**}$	$1,370^{**}$	$1,143^{**}$
	$(0,105)$	$(0,085)$	$(0,082)$	$(0,081)$	$(0,080)$
kalib_tulorek05	$2,522^{**}$	$1,949^{**}$	$1,603^{**}$	$1,440^{**}$	$1,212^{**}$
	$(0,111)$	$(0,093)$	$(0,085)$	$(0,086)$	$(0,084)$
kalib_tulorek06	$2,755^{**}$	$1,969^{**}$	$1,747^{**}$	$1,435^{**}$	$1,213^{**}$
	$(0,126)$	$(0,093)$	$(0,088)$	$(0,085)$	$(0,083)$
kalib_tulorek07	$2,610^{**}$	$2,082^{**}$	$1,915^{**}$	$1,624^{**}$	$1,450^{**}$
	$(0,124)$	$(0,102)$	$(0,096)$	$(0,092)$	$(0,092)$
kalib_tulorek08	$2,669^{**}$	$2,072^{**}$	$1,920^{**}$	$1,679^{**}$	$1,513^{**}$
	$(0,128)$	$(0,102)$	$(0,097)$	$(0,096)$	$(0,094)$
kalib_tulorek09	$2,646^{**}$	$2,067^{**}$	$1,990^{**}$	$1,938^{**}$	$1,727^{**}$
	$(0,129)$	$(0,103)$	$(0,099)$	$(0,103)$	$(0,101)$
kalib_tulorek10	$2,374^{**}$	$2,113^{**}$	$1,722^{**}$	$1,667^{**}$	$1,597^{**}$
	$(0,116)$	$(0,106)$	$(0,091)$	$(0,095)$	$(0,097)$
etuudetq1	$-1,122^{**}$	$-0,557^{**}$	$-0,367^{**}$	$-0,362^{**}$	$-0,175^{**}$
	$(0,050)$	$(0,047)$	$(0,047)$	$(0,050)$	$(0,051)$
etuudetq2	$-1,877^{**}$	$-1,225^{**}$	$-1,024^{**}$	$-1,050^{**}$	$-0,997^{**}$
	$(0,064)$	$(0,058)$	$(0,058)$	$(0,060)$	$(0,060)$
etuudetq3	$-2,254^{**}$	$-1,321^{**}$	$-1,073^{**}$	$-1,052^{**}$	$-0,906^{**}$
	$(0,070)$	$(0,065)$	$(0,066)$	$(0,070)$	$(0,071)$
etuudetq4	$-2,471^{**}$	$-1,634^{**}$	$-1,318^{**}$	$-1,242^{**}$	$-1,086^{**}$
	$(0,074)$	$(0,071)$	$(0,072)$	$(0,076)$	$(0,077)$
kalib_kieliswe	$-0,001$	$0,082$	$0,065$	$0,120$	$0,216^{**}$
	$(0,074)$	$(0,071)$	$(0,073)$	$(0,076)$	$(0,077)$
kalib_kielimuu	$-0,266^{**}$	$-0,264^{**}$	$-0,298^{**}$	$-0,270^{**}$	$-0,239^{**}$
	$(0,057)$	$(0,052)$	$(0,052)$	$(0,055)$	$(0,056)$
kalib_tutrekperus	$-0,510^{**}$	$-0,488^{**}$	$-0,497^{**}$	$-0,473^{**}$	$-0,461^{**}$
	$(0,044)$	$(0,041)$	$(0,042)$	$(0,044)$	$(0,044)$
kalib_tutrekkorkea	$0,288^{**}$	$0,364^{**}$	$0,284^{**}$	$0,351^{**}$	$0,360^{**}$
	$(0,039)$	$(0,037)$	$(0,037)$	$(0,039)$	$(0,039)$

Parametrin estimaatti, suluissa keskivirhe, $^{**}p < 0,01$; $^{*}p < 0,05$

Taulukko 15: Työllisyysmallit asetelmassa H haastattelukerran mukaan, osa 3

työllisyys	H1	H2	H3	H4	H5
tyorektyotloM1534	$-0,837^{**}$	$-0,303^{**}$	$-0,241^{*}$	$-0,500^{**}$	$-0,453^{**}$
	$(0,120)$	$(0,111)$	$(0,113)$	$(0,121)$	$(0,121)$
tyorektyotloMyli35	$-1,604^{**}$	$-1,091^{**}$	$-0,919^{**}$	$-0,913^{**}$	$-0,798^{**}$
	$(0,093)$	$(0,084)$	$(0,085)$	$(0,090)$	$(0,092)$
tyorektyotN1534	$-0,559^{**}$	$-0,156$	$-0,250$	$-0,281^{*}$	$-0,159$
	$(0,138)$	$(0,128)$	$(0,136)$	$(0,141)$	$(0,148)$
tyorektyotNyli35	$-1,273^{**}$	$-0,680^{**}$	$-0,571^{**}$	$-0,613^{**}$	$-0,428^{**}$
	$(0,099)$	$(0,090)$	$(0,091)$	$(0,097)$	$(0,099)$
AIC	28604	32367	31643	28616	28175
BIC	28954	32715	31988	28955	28511
Log-uskottavuus	-14263	-16144	-15782	-14269	-14049
Devianssi	28526	32289	31565	28538	28097
Tjurin pseudo-R2	$0,6934$	$0,6340$	$0,6129$	$0,58343$	$0,5427$
Havaintojen lkm,	58183	55455	51236	43664	40632

Parametrin estimaatti, suluissa keskivirhe, $^{**}p < 0,01$; $^{*}p < 0,05$

8.4 Työttömyys H-asetelmassa

Työttömyys asetelmissa H1, . . . , H5 logistisen regression avulla mallinnettuna näkyy taulukoissa 16, 17, 18. Päällimmäinen havainto on, että työttömyydelle pseudoselitysaste on alkujaankin matala, ja mitä myöhäisemmästä haastattelukerrasta on kyse, sitä enemmän se menee entistä matalammaksi. Tämä on merkittävä ero työllisyysmalliin. Työttömyyttä onkin siis huomattavasti hankalampi mallintaa ja ajallinen yhteys on paljon tärkeämpi kuin työllisyydessä. Työttömyyden ennustaminen viiden haastattelukerran päähän on huomattavasti hankalampaa kuin työllisyyden.

Taulukko 16: Työttömyysmallit asetelmassa H haastattelukerran mukaan, osa 1

työttömyys	H1	H2	H3	H4	H5
(Intercept)	$-3,482^{**}$	$-3,266^{**}$	$-3,165^{**}$	$-2,855^{**}$	$-2,919^{**}$
	(0,108)	(0,102)	(0,104)	(0,107)	(0,110)
sukupnainen	$-0,001$	$-0,042$	$-0,111^{*}$	$-0,109$	$-0,178^{**}$
	(0,060)	(0,055)	(0,056)	(0,058)	(0,058)
ika10i1524	$0,646^{**}$	0,147	$0,257^{**}$	$0,258^{*}$	$0,214^{*}$
	(0,102)	(0,094)	(0,098)	(0,102)	(0,103)
ika10i2534	$0,521^{**}$	$0,306^{**}$	$0,218^{*}$	$0,260^{**}$	$0,279^{**}$
	(0,094)	(0,082)	(0,086)	(0,088)	(0,088)
ika10i4554	$-0,021$	$-0,092$	$-0,081$	$-0,084$	$-0,028$
	(0,087)	(0,080)	(0,083)	(0,086)	(0,089)
ika10i5564	$-0,170^{*}$	$-0,318^{**}$	$-0,062$	$-0,191^{*}$	$-0,105$
	(0,081)	(0,077)	(0,078)	(0,084)	(0,086)
ika10i6574	$-1,767^{**}$	$-1,915^{**}$	$-1,305^{**}$	$-1,363^{**}$	$-1,295^{**}$
	(0,174)	(0,178)	(0,182)	(0,197)	(0,212)
suuralue_2012etel	$-0,022$	$-0,029$	$-0,026$	$-0,123$	0,067
	(0,064)	(0,063)	(0,065)	(0,069)	(0,070)
suuralue_2012lans	$-0,058$	$-0,039$	$-0,077$	$-0,099$	0,011
	(0,062)	(0,061)	(0,063)	(0,066)	(0,068)
suuralue_2012itapohj	$-0,238^{**}$	$-0,206^{**}$	$-0,245^{**}$	$-0,227^{**}$	0,039
	(0,068)	(0,066)	(0,069)	(0,072)	(0,072)
suuralue_2012ahv	$-0,682$	$-0,731^{*}$	$-0,575$	$-0,043$	0,015
	(0,372)	(0,350)	(0,347)	(0,317)	(0,343)
kuntaluoktaajam	$-0,162^{*}$	$-0,227^{**}$	$-0,173^{*}$	$-0,134$	$-0,309^{**}$
	(0,072)	(0,072)	(0,072)	(0,077)	(0,081)
kuntaluokmaaseu	$-0,049$	$-0,174^{*}$	$-0,235^{**}$	$-0,101$	$-0,252^{**}$
	(0,082)	(0,082)	(0,087)	(0,087)	(0,089)
ptoim5_orgtyot	$0,314^{**}$	$0,817^{**}$	$0,737^{**}$	$0,685^{**}$	$0,842^{**}$
	(0,068)	(0,069)	(0,073)	(0,079)	(0,082)
ptoim5_orgopis	$0,697^{**}$	$0,820^{**}$	$0,723^{**}$	$0,577^{**}$	$0,676^{**}$
	(0,084)	(0,082)	(0,086)	(0,090)	(0,091)
ptoim5_orgelak	$-1,556^{**}$	$-1,343^{**}$	$-1,700^{**}$	$-1,655^{**}$	$-1,767^{**}$
	(0,144)	(0,148)	(0,165)	(0,177)	(0,190)
ptoim5_orgmuu	$0,468^{**}$	$0,673^{**}$	$0,513^{**}$	$0,275^{*}$	$0,272^{*}$
	(0,103)	(0,098)	(0,106)	(0,116)	(0,119)

Parametrin estimaatti, suluissa keskivirhe, $^{**}p < 0,01$; $^{*}p < 0,05$

Taulukko 17: Työttömyysmallit asetelmassa H haastattelukerran mukaan, osa 2

työttömyys	H1	H2	H3	H4	H5
kalib_tulorek01	$-0,518^{**}$	$-0,230^{**}$	$-0,200^{*}$	$-0,294^{**}$	$-0,185^{*}$
	$(0,080)$	$(0,081)$	$(0,085)$	$(0,092)$	$(0,091)$
kalib_tulorek02	$-0,993^{**}$	$-0,188^{*}$	$-0,146$	$-0,202^{*}$	$-0,183$
	$(0,105)$	$(0,090)$	$(0,092)$	$(0,098)$	$(0,102)$
kalib_tulorek03	$-1,290^{**}$	$-0,369^{**}$	$-0,454^{**}$	$-0,411^{**}$	$-0,262^{*}$
	$(0,149)$	$(0,114)$	$(0,121)$	$(0,121)$	$(0,118)$
kalib_tulorek04	$-1,423^{**}$	$-0,663^{**}$	$-0,662^{**}$	$-0,557^{**}$	$-0,483^{**}$
	$(0,169)$	$(0,135)$	$(0,138)$	$(0,132)$	$(0,132)$
kalib_tulorek05	$-1,629^{**}$	$-0,944^{**}$	$-0,846^{**}$	$-0,922^{**}$	$-0,752^{**}$
	$(0,192)$	$(0,155)$	$(0,150)$	$(0,153)$	$(0,147)$
kalib_tulorek06	$-2,301^{**}$	$-1,262^{**}$	$-1,100^{**}$	$-0,917^{**}$	$-0,784^{**}$
	$(0,259)$	$(0,170)$	$(0,161)$	$(0,148)$	$(0,145)$
kalib_tulorek07	$-2,304^{**}$	$-1,526^{**}$	$-1,095^{**}$	$-1,222^{**}$	$-1,202^{**}$
	$(0,260)$	$(0,195)$	$(0,162)$	$(0,168)$	$(0,172)$
kalib_tulorek08	$-2,180^{**}$	$-1,433^{**}$	$-1,342^{**}$	$-1,231^{**}$	$-1,354^{**}$
	$(0,240)$	$(0,183)$	$(0,175)$	$(0,168)$	$(0,179)$
kalib_tulorek09	$-2,144^{**}$	$-1,332^{**}$	$-1,421^{**}$	$-1,573^{**}$	$-1,787^{**}$
	$(0,236)$	$(0,176)$	$(0,178)$	$(0,188)$	$(0,211)$
kalib_tulorek10	$-1,703^{**}$	$-1,475^{**}$	$-1,098^{**}$	$-1,241^{**}$	$-1,248^{**}$
	$(0,195)$	$(0,185)$	$(0,156)$	$(0,166)$	$(0,171)$
etuudetq1	$0,895^{**}$	$0,811^{**}$	$0,693^{**}$	$0,540^{**}$	$0,474^{**}$
	$(0,062)$	$(0,062)$	$(0,065)$	$(0,070)$	$(0,071)$
etuudetq2	$1,209^{**}$	$0,942^{**}$	$0,844^{**}$	$0,575^{**}$	$0,576^{**}$
	$(0,076)$	$(0,076)$	$(0,079)$	$(0,086)$	$(0,086)$
etuudetq3	$1,753^{**}$	$1,242^{**}$	$0,992^{**}$	$0,754^{**}$	$0,706^{**}$
	$(0,091)$	$(0,094)$	$(0,099)$	$(0,109)$	$(0,112)$
etuudetq4	$1,516^{**}$	$1,314^{**}$	$0,906^{**}$	$0,708^{**}$	$0,536^{**}$
	$(0,116)$	$(0,117)$	$(0,123)$	$(0,135)$	$(0,144)$
kalib_kieliswe	$-0,298^{*}$	$-0,102$	$-0,237$	$-0,535^{**}$	$-0,582^{**}$
	$(0,131)$	$(0,124)$	$(0,133)$	$(0,152)$	$(0,162)$
kalib_kielimuu	$0,482^{**}$	$0,509^{**}$	$0,520^{**}$	$0,466^{**}$	$0,416^{**}$
	$(0,064)$	$(0,062)$	$(0,065)$	$(0,069)$	$(0,072)$
kalib_tutrekperus	$-0,162^{**}$	$0,033$	$0,003$	$0,026$	$0,121$
	$(0,059)$	$(0,059)$	$(0,063)$	$(0,066)$	$(0,067)$
kalib_tutrekkorkea	$0,119^{*}$	$0,013$	$0,092$	$0,015$	$-0,021$
	$(0,061)$	$(0,058)$	$(0,060)$	$(0,063)$	$(0,065)$

Parametrin estimaatti, suluissa keskivirhe, $^{**}p < 0,01$; $^{*}p < 0,05$

Taulukko 18: Työttömyysmallit asetelmassa H haastattelukerran mukaan, osa 3

työttömyys	H1	H2	H3	H4	H5
tyorektyotloM1534	$1,834^{**}$	$1,068^{**}$	$1,116^{**}$	$0,911^{**}$	$0,620^{**}$
	$(0,104)$	$(0,111)$	$(0,119)$	$(0,129)$	$(0,135)$
tyorektyotloMyli35	$2,317^{**}$	$1,742^{**}$	$1,542^{**}$	$1,388^{**}$	$1,164^{**}$
	$(0,092)$	$(0,089)$	$(0,093)$	$(0,101)$	$(0,103)$
tyorektyotN1534	$1,582^{**}$	$0,934^{**}$	$0,935^{**}$	$0,500^{**}$	$0,462^{**}$
	$(0,118)$	$(0,127)$	$(0,141)$	$(0,161)$	$(0,171)$
tyorektyotNyli35	$1,986^{**}$	$1,298^{**}$	$1,227^{**}$	$1,179^{**}$	$0,933^{**}$
	$(0,094)$	$(0,095)$	$(0,099)$	$(0,108)$	$(0,114)$
AIC	14525	15623	14779	13429	12896
BIC	14875	15971	15124	13768	13232
Log-uskottavuus	-7224	-7773	-7351	-6676	-6409
Devianssi	14447	15545	14701	13351	12818
Tjurin pseudo-R2	$0,2752$	$0,1875$	$0,1561$	$0,1229$	$0,1151$
Havaintojen lkm,	58183	55455	51236	43664	40632

Parametrin estimaatti, suluissa keskivirhe, $^{**}p < 0,01$; $^{*}p < 0,05$

9 Otantavertailun menetelmäkuvaus

Sovelletaan aineistoihin A, H1, H2, H3, H4 ja H5 erilaisia otantamenetelmiä ja vertaillaan laskettujen estimaattoreiden keskihajontoja. Mitä pienempi keskihajonta, sitä tarkempi estimaattori. Käytettävinä menetelminä ovat yksinkertainen satunnaisotanta, systemaattinen otanta ja tasapainoinen otanta. Otoksille toteutetaan myös kalibrointi (6) haastatteluhetken taustamuuttujien mukaan.

Kullakin menetelmällä otetaan otoskoon n osalta noin $n = 2600$ suuruisia otoksia ja lasketaan kustakin otoksesta työllisten ja työttömien osuuden estimaatti. Yksinkertaisessa satunnaisotannassa (SRS) aineistosta on valittu 10000 kappaletta 2600 havainnon satunnaisia otoksia. Systemaattisessa otannassa aineisto on järjestetty henkilöiden asuinpaikan ja iän mukaan. Sen jälkeen aineistosta on otettu systemaattisesti 2600 havainnon otoksia niin monta kuin aineisto mahdollistaa (3.5.1). Pyöristyksestä johtuen yhden otoksen koko kussakin asetelmassa on 2599. Otosten määrä jää tällä menetelmällä (SYS) todella pieneksi ja tätä pyritään korjaamaan käyttämällä robustia systemaattista otantaa (SYSR). Koska aineiston järjestämisessä asuinpaikan suhteen ei oikeastaan ole väliä asuinpaikkojen keskenäisellä järjestyksellä, kunhan samat asuinpaikat ovat peräkkäin, niin sekoitamme asuinpaikat satunnaisessa järjestyksessä ja ikä edelleen nousevasti. Robustissa systemaattisessa otoksessa otetaan systemaattiset otokset niin monessa eri kuntien satunnaisessa järjestyksessä, että otosten määräksi saadaan vähintään 10000. Jatkotarkasteluissa käytetään robustia systemaattista otantaa ja nimitetään sitä yksinkertaisuuden nimissä systemaattisesti otannaksi ilman etuliitettä. Otosten määrät on kuvattu taulukossa 19.

Analyysit on toteutettu helmikuussa 2026 RStudio-ohjelmalla ja R:n versiolla 4.5.2. Laskennan nopeuttamiseen on hyödynnetty moniajoa mahdollistavia future-

ja future.apply -R-paketteja. Tasapainoinen otanta on toteutettu kuten se on kuvattu osioissa 4 ja 5. Toteuttamiseen on käytetty BalancedSampling -nimisen R-paketin vanhaa versiota 1.6.3., koska sen uusimmassa versiossa (2.1.1) oli virhe, jonka seurauksena otoskokoa ei saanut vakioksi. Virhettä ei ollut korjattu 10.5.2026 mennessä, mutta paketin kehittäjälle on ilmoitettu. Virhe liittyy matriisin ytimen laskemiseen tietyissä erityistilanteissa. Toisena vaihtoehtona olisi ollut käyttää sampling -nimistä R-pakettia, mutta se on monta kymmentä kertaa hitaampi kuin C++:aa hyödyntävä BalancedSampling-paketti. Paras vaihtoehto on uudehko nopeaa Rust-toteutusta hyödyntävä rsamplr -niminen R-paketti, jonka 31.3.2026 julkaistussa versiossa 0.2.0 kyseinen virhe koodista on korjattu [20]. Paketin kehittäjä on sama kuin BalancedSampling -paketissa. [19]

Tasapainoiseen otantaan (BAL) on valittu 11 luokkamuuttujaa, jotka ovat muuten samat kuin osiossa 8, ja lisäksi henkilön syntyperää kuvaava muuttuja. Syntyperää kuvaava muuttuja kuvaa sitä, onko henkilö suomalaistaustainen vai ulkomaalaistaustainen ja onko hän syntynyt Suomessa vai ulkomailla. Tämä muuttuja ei logistisissa malleissa ollut tilastollisesti merkitsevä ja sen lisäksi osassa sen luokissa oli hyvin vähän alkioita. Syntyperää kuvaava muuttuja unohtui tasapainoiseen otantaan ja ajanpuutteen takia ei koettu mielekkääksi toteuttaa otantaa uudestaan ilman. Syntyperää koskevaa muuttujaa ei käsitellä muussa yhteydessä. Aineistossa on käytössä äidinkieltä kuvaava muuttuja, jonka sisältämä informaatio on jokseenkin samankaltaista kuin syntyperää kuvaavan muuttujan. Tasapainoisessa otannassa on apumuuttujana mukana myös sisältymistodennäköisyydet, mikä takaa että otoskoko on 2600. Tässä tarkastelussa sisältymistodennäköisyys on aineistokohtaisesti vakio, eli 2600 jaettuna kunkin aineiston perusjoukon koolla. Yhteensä tasapainoisessa otannassa taustamuuttujia on 12 kappaletta, jotka on kuvattu mallimatriisissa 44 sarakkeella.

Taulukko 19: Otosten määrä eri asetelmissa

	A	H1	H2	H3	H4	H5
SRS	10000	10000	10000	10000	10000	10000
SYS	114	23	22	20	17	16
SYSR	10032	10005	10010	10000	10013	10000
BAL	10000	10000	10000	10000	10000	10000

SRS: yksinkertainen satunnaisotanta, SYS: sekoittamaton systemaattinen otanta, SYSR: (robusti) systemaattinen otanta, BAL: tasapainoinen otanta

Otoksia tarkastellaan myös kalibroitujen estimaattien osalta (6). Saadut otokset populaatioista A, H1, H2, H3, H4, H5 kalibroidaan 10 apumuuttujan suhteen, jotka ovat samat kuin osiossa 8. Kalibrointiin käytetään haastatteluhetken apumuuttujia. Asetelmassa A haastatteluhetki on sama kuin haastatteluhetki, joten kalibroinnissa käytettävä populaatio on populaatio A. Kalibrointiin käytetään survey- nimistä R-pakettia ja sen calibrate -funktiota, mikä toimii hyvin vastaavanlaisesti kuin CALMAR2 -makro SAS:lla [21].

10 Tulokset

10.1 Otantamenetelmien vertailu

Tämän tutkielman tarkoituksena on selvittää, onko mahdollista pienentää työvoimatutkimuksessa työllisyyden ja työttömyyden estimaattoreiden keskihajontaa vertailemalla yksinkertaista satunnaisotantaa (SRS), (robustia) systemaattista otantaa implisiittisellä osituksella asuinkunnan ja iän suhteen (SYSR) sekä tasapainoista otantaa (BAL).

Ensiksi tarkastellaan edellisessä osiossa 9 kuvattuja yksinkertaista satunnaisotantaa, systemaattista otantaa ja tasapainoista otantaa. Ensimmäinen havainto on se, miten hyvin osiossa 8 esiteltyjen muuttujien keskiarvot (eli suhteelliset osuudet) toteutuvat otoksessa verrattuna populaatioon (POP). Vertailumenetelmänä on otoksen ja populaation muuttujakohtaisen keskiarvon erotuksen neliösumma. Taulukossa 20 kuvataan erotuksen neliösumman keskiarvoa suhteessa systemaattisen otannan neliösumman keskiarvoon. Välittömästi huomataan, että tasapainoinen otanta pienentää keskineliövirhettä noin -99,7 % suhteessa systemaattiseen otantaan. Puuttuva -0,3 % syntyy kuutiomenetelmän laskeutumisvaiheesta, sillä siinä on luovuttu vaatimuksesta rajoitteiden täydellisestä toteutumisesta. Kuitenkin -99,7 % on valtava. Se tarkoittaa, että tasapainoisessa otoksessa apumuuttujien keskiarvot ovat lähes täydellisesti samat kuin perusjoukon apumuuttujien odotusarvot. Yksinkertaisessa satunnaisotannassa taustamuuttujien neliösummien erotukset ovat suurempia kuin systemaattisessa otannassa. Ero vaihtelee 86,78 % ja 93,81 % välillä. Vaihtelu johtuu mitä luultavammin yksinkertaisen satunnaisotannan hyvin satunnaisesta luonteesta. Tämä osoittaa, että systemaattisen otannan implisiittinen osite tuo jo merkittävää etua taustamuuttujien marginaalijakaumien toteutumiseen, mutta tasapainoisen otannan ansiosta ne toteutuvat vielä paremmin. Absoluuttisella tasolla erotuksen neliösumma on otoksessa A tasapainoisella otannalla 0,00000216, systemaattisella otannalla 0,00069335 ja yksinkertaisella satunnaismuuttujalla 0,00134378. Luvut ovat lähes samoja myös otoksissa H1, . . . , H5.

Taulukko 20: Otoksen taustamuuttujien keskiarvon erotuksen neliösumma populaation taustamuuttujien odotusarvosta, suhteessa systemaattiseen otantaan

	A	H1	H2	H3	H4	H5
SRS	93,81	87,06	88,47	86,78	86,89	87,06
SYSR	-	-	-	-	-	-
BAL	-99,69	-99,68	-99,68	-99,69	-99,67	-99,68

SRS: yksinkertainen satunnaisotanta, SYSR: systemaattinen otanta,
BAL: tasapainoinen otanta

Neliösummien erotuksen keskiarvoa tarkasteltiin taulukossa 20. Taulukossa 21 tarkastellaan neliösummien erotuksen keskiarvojen keskihajontaa, eli sitä, miten samankaltaisina populaation taustamuuttujien keskiarvot toteutuvat eri otosten välillä. Keskihajonnat taustamuuttujien keskiarvojen toteutumisessa ovat tasapainoisessa otannassa noin -99,7 % pienemmät kuin systemaattisessa otoksessa. Systemaattinen otanta siis pienentää keskineliövirheen keskiarvoa ja keskihajontaa erittäin

merkittävästi. Yksinkertaisessa satunnaisotannassa neliösummien keskiarvojen keskihajonnat ovat 69,02 - 80,97 % suuremmat kuin systemaattisessa otannassa. Absoluuttisesti neliösummien keskiarvojen keskihajonnat ovat A-otoksessa tasapainoisessa otoksessa 0,00000066, systemaattisessa otoksessa 0,00024121 ja yksinkertaisessa satunnaisotoksessa 0,00043652. Luvut ovat lähes samoja otoksissa H1, ..., H5. Tasapainoisella otannalla saa siis merkittävästi stabiilimpia ja perusjoukon marginaalijakaumia edustavampia otoksia kuin systemaattisella otannalla tai yksinkertaisella satunnaisotannalla.

Taulukko 21: Otoksen taustamuuttujien keskiarvon erotuksen neliösumma populaation taustamuuttujien odotusarvosta, keskihajonta otosten välillä, suhteessa systemaattiseen otantaan

	A	H1	H2	H3	H4	H5
SRS	80,97	71,12	71,25	71,84	75,13	69,02
SYSR	-	-	-	-	-	-
BAL	-99,73	-99,73	-99,73	-99,72	-99,72	-99,72

Seuraavaksi vertaillaan otosten tarkasteltavia muuttujia eli työvoimatutkimuksessa kysyttyä työllisyyttä ja työttömyyttä. Työllisyyden osalta tasapainoisen otannan, systemaattisen otannan ja yksinkertaisen satunnaisotannan osalta otosten keskiarvo on käytännössä sama kuin kyseisissä populaatioissa (taulukko 22). Suurin havaittu empiirinen harha oli -0,0091 %-yksikköä. Tämä empiirinen harha on mitä todennäköisemmin suurimmilta osin puhtaasti stokastista kohinaa, mutta on merkillepantavaa, että kaikissa paitsi asetelmassa H5 systemaattisessa otannassa empiirinen harha oli merkittävästi pienempää kuin tasapainoisessa otannassa. Ero on merkittävä ja vaihteli 1,4-kertaisesta 105-kertaiseen. Kuitenkin absoluuttiset erot ovat pieniä. Syitä tähän voi olla useita. Yksi voi olla, että tasapainoiseen otantaan on valittu joitain sellaisia taustamuuttujia, jotka eivät kovin hyvin kuvaa työllisyyttä ja tämä tuo tarpeetonta epävarmuustekijää otantaan. Kuitenkin empiirisen harhan suuruus on niin pieni, että se ei aiheuta tarvetta syvällisemmille jatkotarkasteluille.

Taulukko 22: Otosten työllisyyden estimaattorin keskiarvo verrattuna populaatioiden työllisyyden odotusarvoon

	A	H1	H2	H3	H4	H5
POP	**62,8312 %**	**63,7764 %**	**62,8636 %**	**62,8636 %**	**62,4931 %**	**64,6805 %**
SRS	+0,0007 %y	-0,0037 %y	+0,0076 %y	-0,0023 %y	+0,0064 %y	+0,0063 %y
SYSR	-0,0008 %y	+0,0005 %y	+0,0024 %y	-0,0001 %y	+0,0022 %y	+0,0014 %y
BAL	+0,0040 %y	-0,0007 %y	-0,0091 %y	+0,0089 %y	-0,0042 %y	-0,0012 %y

POP: populaatio, SRS: yksinkertainen satunnaisotanta,
SYSR: systemaattinen otanta, BAL: tasapainoinen otanta

Työttömyyden osalta populaatio- ja otoskeskiarvojen vertailu on annettu taulukossa 23. Absoluuttinen harha on samaa suuruusluokkaa kuin työllisyydessäkin, mutta suhteellinen harha suurempi, sillä työttömyysaste on merkittävästi matalampi kuin työllisyysaste. Systemaattisessa otannassa empiirisen harhan suuruus vaihtelee

2,55-kertaisesta 684-kertaiseen verrattuna tasapainoiseen otantaan. Kuitenkin absoluuttinen harha on hyvin pientä, joten käytännön todellista merkitystä tällä ei ole, vaikkakin erot ovat suhteellisesti merkittäviä menetelmien välillä.

Taulukko 23: Otosten työttömyyden estimaattorin keskiarvo verrattuna populaatioiden työttömyyden odotusarvoon

	A	H1	H2	H3	H4	H5
POP	**4,7073** %	**5,0015** %	**4,8742** %	**4,7018** %	**4,8095** %	**4,9542** %
SRS	+0,0050 %y	-0,0004 %y	-0,0061 %y	+0,0024 %y	+0,0017 %y	-0,0040 %y
SYSR	-0,0004 %y	+0,0006 %y	+0,0004 %y	+0,0000 %	-0,0004 %	-0,0002 %
BAL	-0,0028 %y	-0,0022 %y	+0,0017 %y	-0,0042 %y	-0,0016 %y	-0,0005 %y

Kuten havaitaan, estimaatit ovat suhteellisen harhattomia ainakin empiirisesti. Seuraavaksi tarkastellaan työllisyyden ja työttömyyden estimaattorien empiirisiä keskihajontoja. Tasapainoinen otanta pienentää työllisyyden estimaattorin keskihajontaa verrattuna systemaattiseen otantaan (taulukot 24 ja 25). Tasapainoisella otannalla asetelmassa A työllisyyden estimaattorin keskihajonta pienenee 34,55 % verrattuna systemaattiseen otantaan. Asetelmissa H suhteellinen ero pienenee mitä myöhäisemmälle haastattelukerralle mennään. Ensimmäisen haastattelukerran osalta suhteellinen ero on 32,22 %, kun taas viidennen haastattelukerran osalta ero on 17,15 %.

Systemaattinen otanta antaa tarkempia estimaatteja kuin yksinkertainen satunnaisotanta. Yksinkertaisessa satunnaisotannassa työllisyyden estimaattorin keskihajonta on asetelmassa A 27,68 % suurempi kuin systemaattisessa otannassa. Ero pysyy jokseenkin samalla tasolla asetelmassa H riippumatta haastattelukerrasta. Tämä on hyvin erilaista kuin tasapainoisessa otannassa. Tämä johtunee siitä, että tasapainoisessa otannassa otos on valittu taustamuuttujien perusteella, jotka ovat sitä enemmän vanhentuneita mitä myöhäisemmästä haastatteluhetkestä on kyse. Kuten osiossa 8 havaittiin, taustamuuttujien yhteys työllisyyteen (ja työttömyyteen) on sitä heikompi, mitä myöhäisempää haastattelukertaa tarkastellaan. Yksinkertainen satunnaisotos on valittu täysin riippumatta taustamuuttujista, joten vastaavaa ilmiötä ei siinä siksi ole. Systemaattinen otos on valittu suhteessa henkilön asuinkuntaan ja ikään, jotka eivät kovin paljoa muutu haastattelukertojen välillä. Näihin verrattuna tasapainoinen otos on huomattavasti enemmän riippuvainen taustamuuttujien arvoista.

Taulukko 24: Otosten työllisyyden suhteellisen osuuden estimaattorin empiirinen keskihajonta prosenttiyksiköissä

	A	H1	H2	H3	H4	H5
SRS	0,9464	0,9194	0,9253	0,9268	0,9237	0,9200
SYSR	0,7412	0,7733	0,7482	0,7153	0,7567	0,7345
BAL	0,4851	0,5241	0,5597	0,5856	0,6010	0,6086

Taulukko 25: Otosten työllisyyden suhteellisen osuuden estimaattorin empiirinen keskihajonta suhteessa (%) systemaattiseen otantaan

	A	H1	H2	H3	H4	H5
SRS	+27,68	+18,89	+23,67	+29,56	+22,07	+25,25
SYSR	-	-	-	-	-	-
BAL	-34,55	-32,22	-25,19	-18,14	-20,58	-17,15

Työttömyyden osalta havaitaan heti, että estimaattorin absoluuttinen keskihajonta on pienempää kuin työllisyyden osalta (taulukko 26). Tämä johtuu siitä, että työttömyysaste on merkittävästi pienempi kuin työllisyysaste. Keskihajonta suhteessa keskiarvoon on tosin työttömyydessä samasta syystä suurempi. Työttömyysasteen osalta otantamenetelmien välillä keskihajonnoissa on vähemmän eroja kuin työllisyyden osalta (taulukko 27). Systemaattisen otannan ja yksinkertaisen satunnaisotannan välillä ei käytännössä ole eroa keskihajonnoissa. Tasapainoisen ja systemaattisen otannan välillä ero työttömyysasteen estimaattorin keskihajonnassa on asetelmassa A 17,01 % tasapainoisen otannan eduksi. Asetelmissa H ero kaventuu nopeasti mitä myöhäisemmästä haastattelukerrasta on kyse. Haastattelukerroilla yksi ja kaksi ero on -14,38 % ja -10,48 %. Haastattelukerralla kolme ero on -4,67 %. Haastattelukerroilla neljä ja viisi ero kaventuu noin kahteen prosenttiin, joka on hyvin pieni ja voi osittain selittyä satunnaisvaihtelulla. Työttömyyden estimoinnin osalta käytännössä myöhäisemmillä haastattelukerroilla ei ole hyötyä tasapainoisesta otannasta verrattuna systemaattiseen otantaan.

Taulukko 26: Työttömyyden suhteellisen osuuden estimaattorin empiirinen keskihajonta prosenttiyksiköissä

	A	H1	H2	H3	H4	H5
SRS	0,4172	0,4145	0,4123	0,4010	0,4070	0,4144
SYSR	0,4126	0,4175	0,4234	0,3924	0,3762	0,3919
BAL	0,3424	0,3574	0,3790	0,3741	0,3861	0,3826

Taulukko 27: Työttömyyden suhteellisen osuuden estimaattorin empiirinen keskihajonta suhteessa (%) systemaattiseen otantaan

	A	H1	H2	H3	H4	H5
SRS	+1,11	-0,70	-2,61	+2,20	+8,18	+5,73
SYSR	-	-	-	-	-	-
BAL	-17,01	-14,38	-10,48	-4,67	+2,64	-2,37

Yleisesti voi sanoa, että tasapainoinen otanta tuo merkittäviä etuja työllisyysasteen estimoinnissa verrattuna systemaattiseen otantaan estimaattorin keskihajonnan pienentymisen myötä. Työttömyysasteen osalta hyöty on oleellisesti pienempi, mutta haittaakaan ei ole. Systemaattinen otanta vaikuttaisi lisäävän estimaattorin harhaa merkittävästi verrattuna tasapainoiseen otantaan, mutta koska harha on hyvin pientä, sillä ei ole käytännössä merkitystä. Työllisyys- ja työttömyysastetta tar-

kastellaan työvoimatutkimuksessa prosentin kymmenyksen tarkkuudella, ja harhan suuruusluokka on prosentin tuhannesosan luokkaa.

10.2 Otosten kalibrointi

Seuraavaksi saatuja otoksia tarkastellaan kalibroituina. Kalibrointi tarkoittaa sitä, että otoksen taustamuuttujien (8) keskiarvot asetetaan painokertoimien avulla vastaamaan sitä, mitä ne ovat halutussa perusjoukossa (osio 6). Asetelman A otoksissa kalibrointiperusjoukko on populaatio A. Asetelmien H otoksissa kalibrointiperusjoukko on haastatteluhetken perusjoukko, joka eroaa jonkin verran otoksen valintahetken perusjoukosta.

Ensiksi tarkastellaan, miten paljon otosten apumuuttujien keskiarvot eroavat kalibrointipopulaation odotusarvoista; mittana käytetään erotuksen neliösummaa (taulukko 28). Aluksi havaitaan, että apumuuttujien erotuksen neliösumma otospopulaation ja kalibrointipopulaation välillä kasvaa, mitä myöhäisemmästä haastattelukerrasta on kyse. Asetelmassa A otospopulaatio ja kalibrointipopulaatio ovat samat ja tämän takia erotuksen neliösumma on nolla. Asetelmassa H1 ainoa muuttuja, joka eroaa otospopulaatiossa ja kalibrointipopulaatiossa, on pääasiallisen toiminnan indikaattorimuuttujat. Asetelmissa H2–H5 erot kasvavat suuremmiksi (taulukko 29). Otospopulaation ja kalibrointipopulaation väliset erotuksen neliösummat ovat käytännössä samat kuin tasapainoisen otannan ja kalibrointipopulaation väliset erotuksen neliösummat. Tämä tarkoittaa, että tasapainoisen otoksen taustamuuttujien keskiarvot poikkeavat suurin piirtein yhtä paljon kalibrointipopulaation keskiarvoista kuin otospopulaation keskiarvot kalibrointipopulaation keskiarvoista. Systemaattisen otoksen tapauksessa erot kalibrointipopulaatioon ovat ensimmäisen haastattelukerran osalta 129 % suurempia kuin otospopulaatiossa kalibrointipopulaatioon. Ero kapenee, mitä myöhäisempää haastattelukertaa tarkastellaan, ja haastattelukerralla 5 ero on enää 17 %. Tämä kaventuminen on selkeä seuraus siitä, että tasapainoinen otanta on valittu otospopulaation taustamuuttujien mukaan, joka eroaa sitä enemmän kalibrointipopulaatiosta, mitä myöhäisemmästä haastattelukerrasta on kyse. Tällöin tasapainoisen otannan hyödyt suhteessa yksinkertaiseen satunnaisotantaan tai systemaattiseen otantaan jäävät pienemmiksi kuin aikaisemmilla haastattelukerroilla.

Taulukko 28: Taustamuuttujien keskiarvon erotuksen neliösumman keskiarvo kalibrointipopulaation ja otospopulaation välillä

	A	H1	H2	H3	H4	H5
POP	0	0,00054	0,00075	0,00112	0,00265	0,00390
SRS	0,00134	0,00183	0,00205	0,00241	0,00392	0,00517
SYSR	0,00069	0,00123	0,00144	0,00181	0,00333	0,00458
BAL	0,00000	0,00054	0,00076	0,00113	0,00265	0,00390

Taulukko 29: Taustamuuttujien keskiarvon erotuksen neliösumman keskiarvo kalibrointipopulaation ja otospopulaation välillä, suhteessa (%) alkuperäiseen populaatioon

	H1	H2	H3	H4	H5
POP	-	-	-	-	-
SRS	+241	+172	+115	+48	+33
SYSR	+129	+91	+61	+26	+18
BAL	0,00	+0,29	+0,19	+0,08	+0,08

Sen lisäksi, että tasapainoisessa otannassa taustamuuttujien keskiarvot ovat keskimäärin lähempänä kalibrointipopulaation taustamuuttujien odotusarvoja, ovat myös näiden odotusarvojen estimaattoreiden keskihajonnat merkittävästi pienempiä tasapainoisessa otannassa (taulukko 30). Tasapainoisessa otannassa taustamuuttujien keskiarvojen erotuksen neliösumman keskiarvon estimaattorin keskihajonta on -99,7 % pienempi kuin systemaattisessa otannassa asetelmassa A. Asetelmissa H ero pienenee haastattelukertojen mukaan, ensimmäisellä haastattelukerralla ero on -96,3 % ja viidennellä -94,1 % (taulukko 31). Taustamuuttujien erotusten neliösummien keskiarvojen estimaattorin keskihajonnat ovat kuitenkin hyvin pieniä yleisesti. Kuitenkin yhdistettynä siihen, että tasapainoisessa otannassa taustamuuttujien keskiarvojen erotusten neliösummien keskiarvo on myös pienempi, tarkoittaa tämä, että tasapainoisen otannan otokset vastaavat paremmin kalibrointipopulaatiota ja pienemmällä hajonnalla.

Taulukko 30: Taustamuuttujien keskiarvon erotuksen neliösumman keskiarvon keskihajonta kalibrointipopulaation ja otospopulaation välillä

	A	H1	H2	H3	H4	H5
SRS	0,000437	0,000506	0,000604	0,000750	0,001133	0,001313
SYSR	0,000241	0,000323	0,000369	0,000456	0,000649	0,000766
BAL	0,000001	0,000012	0,000014	0,000021	0,000038	0,000045

Taulukko 31: Taustamuuttujien keskiarvon erotuksen neliösumman keskiarvon keskihajonta kalibrointi- ja otospopulaation välillä, suhteessa (%) systemaattiseen otantaan

	A	H1	H2	H3	H4	H5
SRS	+81	+56,6	+63,8	+64,6	+74,7	+71,3
SYSR	-	-	-	-	-	-
BAL	-99,7	-96,3	-96,1	-95,4	-94,1	-94,1

Otantamenetelmillä saatu otos kalibroidaan vastaamaan taustamuuttujiltaan kalibrointipopulaatiota. Jos otosta ei tarvitse kalibroida lainkaan, ovat painokertoimet kaikki ykkösiä. Jos painokerroin on pienempi kuin yksi, on kyseisen yksilön viiteryhmä yliedustettuna otoksessa verrattuna populaatioon. Painokerroin korjaa otoksen vastaamaan kalibrointipopulaatiota. Kun painokerroin on suurempi kuin yksi, on kyseisen yksilön viiteryhmä aliedustettuna otoksessa verrattuna populaatioon.

Kun yksittäisen otoksen otoskoko on 2600, on painokertoimia oltava saman verran. Tarkastellaan otoskohtaisesti näiden painokerrointen varianssia ja lasketaan sen keskiarvo asetelma- ja otantamenetelmäkohtaisesti. Havaitaan, että keskimäärin pienin painokerrointen varianssi on tasapainoisessa otannassa (taulukko 32). Tämä tarkoittaa sitä, että painokertoimet poikkeavat kaikkein vähiten neutraalista, eli arvosta 1, tasapainoisessa otannassa. Tasapainoisessa otannassa asetelmassa A on -99,6 % pienempi kalibrointipainokertoimien varianssin keskiarvo kuin systemaattisessa otannassa (taulukko 33). Ero pienenee, mitä myöhäisemmälle haastattelukerralle mennään asetelmassa H. Yksinkertaisessa satunnaisotannassa painokerrointen varianssin keskiarvo on suurempi kuin systemaattisessa otannassa, mutta ero kapenee, mitä myöhäisemmälle haastattelukerralle mennään asetelmassa H.

Taulukko 32: Kalibrointipainojen varianssin keskiarvo

	A	H1	H2	H3	H4	H5
SRS	0,01487	0,02561	0,02607	0,02704	0,03415	0,04237
SYSR	0,01047	0,02149	0,02181	0,02305	0,03033	0,03837
BAL	0,00004	0,01124	0,01164	0,01271	0,01995	0,02810

Taulukko 33: Kalibrointipainojen varianssin keskiarvon keskiarvo suhteessa (%) systemaattiseen otantaan

	A	H1	H2	H3	H4	H5
SRS	+42,0	+19,2	+19,6	+17,3	+12,6	+10,4
SYSR	-	-	-	-	-	-
BAL	-99,6	-47,7	-46,6	-44,8	-34,2	-26,8

Kalibraatiopainokertoimien varianssin estimaattorin keskihajontaa tarkastellessa huomataan, että pienin on tasapainoisessa otannassa (taulukko 34). Asetelmassa A painokertoimien varianssin estimaattorin keskihajonta on -99,5 % pienempi kuin systemaattisessa otannassa (taulukko 35). Asetelmissa H ero pienenee mitä myöhäisempään haastattelukertaan mennään, mutta vielä viidennessä haastattelukerrassa ero on -86,2 % verrattuna systemaattiseen otantaan. Pienempi kalibrointipainokertoimien varianssin estimaattorin keskihajonta tarkoittaa sitä, että painokertoimien hajonta eri otosten välillä on pienempää.

Taulukko 34: Kalibrointipainojen varianssin keskiarvon keskihajonta

	A	H1	H2	H3	H4	H5
SRS	0,003489	0,005319	0,005258	0,005338	0,006367	0,007244
SYSR	0,002837	0,004771	0,004594	0,004818	0,005486	0,006132
BAL	0,000014	0,000474	0,000460	0,000448	0,000611	0,000846

Taulukko 35: Kalibrointipainojen varianssin keskiarvon keskihajonta suhteessa (%) systemaattiseen otantaan

	A	H1	H2	H3	H4	H5
SRS	+23,0	+11,5	+14,5	+10,8	+16,1	+18,1
SYSR	-	-	-	-	-	-
BAL	-99,5	-90,1	-90,0	-90,7	-88,9	-86,2

10.3 Kalibroitujen estimaattoreiden tarkastelu

Tässä osiossa tarkastellaan kalibroitujen otosten työllisyyden ja työttömyyden estimaattoreiden empiirisiä tunnuslukuja. Ensiksi on todettava, että todellisen kalibrointipopulaation työllisyys- ja työttömyysaste eivät ole tiedossa. Kalibrointipopulaatioista on tiedossa ainoastaan niiden taustamuuttujien jakauma. Populaatio, josta otos on alkujaan valittu voi olla oleellisestikin erilainen kuin haastatteluhetken populaatio. Painokertoimilla otos asetetaan taustamuuttujiltaan vastaamaan haastatteluhetken populaation taustamuuttujia.

Tarkastellaan ensin, miten paljon työllisyys- ja työttömyysasteen kalibroitujen estimaattorien keskiarvo poikkeaa valintapopulaation työllisyys- ja työttömyysasteesta. Tämä ei varsinaisesti ole harhaa, sillä kalibroinnin tarkoitus on saada otos vastaamaan haastatteluhetken populaatiota. Tämä perustuu siihen olettamaan, että kalibroinnissa käytettävät apumuuttujat ovat yhteydessä tarkasteltavaan muuttujaan (osio 8). Työllisyyden osalta erotus valintapopulaation ja kalibroidun otoksen työllisyysasteen välillä on samaa suuruusluokkaa otantamenetelmästä riippumatta (taulukko 36). Asetelmassa A erotus on hyvin pieni. Tämä johtuu siitä, että siinä otospopulaatio ja kalibrointipopulaatio ovat samat. Asetelmissa H erotus kasvaa mitä myöhäisemmästä haastattelukerrasta on kyse. Erotus kertoo siitä miten paljon työllisyys on kasvanut tai vähentynyt perustuen kalibrointipainoihin. Työttömyyden osalta havainnot ovat samankaltaisia (taulukko 37). Otantamenetelmällä ei ole vaikutusta erotuksen suuruuteen.

Taulukko 36: Kalibroitujen otosten työllisyyden estimaattorin keskiarvo verrattuna valintapopulaation työllisyyden odotusarvoon

	A	H1	H2	H3	H4	H5
POP	**62,8312 %**	**63,7764 %**	**62,8636 %**	**62,8636 %**	**62,4931 %**	**64,6805 %**
SRS	-0,0016 %y	+0,1660 %y	-0,0818 %y	-0,2562 %y	-0,7033 %y	-0,6831 %y
SYSR	-0,0035 %y	+0,1740 %y	-0,0789 %y	-0,2470 %y	-0,7065 %y	-0,7005 %y
BAL	+0,0039 %y	+0,1748 %y	-0,0864 %y	-0,2404 %y	-0,7099 %y	-0,6889 %y

Taulukko 37: Kalibroitujen otosten työttömyyden estimaattorin keskiarvo verrattuna valintapopulaation työttömyyden odotusarvoon

	A	H1	H2	H3	H4	H5
POP	**4,7073 %**	**5,0015 %**	**4,8742 %**	**4,7018 %**	**4,8095 %**	**4,9542 %**
SRS	+0,0071 %y	-0,1372 %y	-0,1024 %y	-0,1294 %y	-0,1518 %y	-0,2378 %y
SYSR	+0,0007 %y	-0,1428 %y	-0,0952 %y	-0,1329 %y	-0,1558 %y	-0,2318 %y
BAL	-0,0028 %y	-0,1416 %y	-0,0965 %y	-0,1364 %y	-0,1532 %y	-0,2356 %y

Tarkasteltaessa kalibroitujen otosten työllisyysasteiden estimaattien keskihajontaa (taulukko 38), havaitaan, että keskihajonnat ovat lähes samoja otantamenetelmästä riippumatta asetelmittain (taulukko 39). Erot ovat muutaman prosentin luokkaa. Kalibrointi tasaa keskihajonnat lähes samoiksi, kun taas kalibroimattomissa otoksissa (taulukko 24) tasapainoisen otannan keskihajonta on merkittävästi muita pienempi. Kalibroidun ja kalibroimattoman otoksen osalta tasapainoisen otannan työllisyyden keskihajonnat ovat lähes samat. Systemaattisen otannan ja yksinkertaisen satunnaisotannan osalta kalibrointi merkittävästi pienentää keskihajontaa verrattuna kalibroimattomiin otoksiin. Kalibrointi tasaa otantamenetelmien keskihajontojen erot täysin.

Taulukko 38: Kalibroitujen otosten työllisyyden suhteellisen osuuden estimaattorin empiirinen keskihajonta prosenttiyksiköissä

	A	H1	H2	H3	H4	H5
SRS	0,4824	0,5317	0,5696	0,5868	0,5933	0,6245
SYSR	0,4707	0,5353	0,5466	0,5638	0,5969	0,5925
BAL	0,4847	0,5236	0,5532	0,5785	0,5930	0,6040

Taulukko 39: Kalibroitujen otosten työllisyyden suhteellisen osuuden estimaattorin empiirinen keskihajonta suhteessa (%) systemaattiseen otantaan

	A	H1	H2	H3	H4	H5
SRS	+2,48	-0,67	+4,22	+4,08	-0,61	+5,40
SYSR	-	-	-	-	-	-
BAL	+2,99	-2,18	+1,21	+2,62	-0,65	+1,95

Työttömyyden osalta estimaattorin keskihajonnat ovat kalibroiduissa otoksissa lähes samat otantamenetelmästä riippumatta (taulukko 40). Erot ovat muutaman prosentin luokkaa (taulukko 41). Tasapainoisen otannan osalta kalibroitujen ja kalibroimattomien otosten estimaattoreiden keskihajonnat ovat lähes samat (taulukko 26). Kalibroimattomien systemaattisten otosten ja yksinkertaisten satunnaisotosten työttömyyden estimaattorin keskihajonnan ero suhteessa tasapainoisen otoksen työttömyyden estimaattorin keskihajontaan on pienempi kuin työllisyyden estimaattoreilla. Kalibrointi antaa siis pienemmän hyödyn näiden osalta kuin työllisyyden estimaattorin kanssa.

Taulukko 40: Kalibroitujen otosten työttömyyden suhteellisen osuuden estimaattorin empiirinen keskihajonta prosenttiyksiköissä

	A	H1	H2	H3	H4	H5
SRS	0,3559	0,3572	0,3755	0,3690	0,3820	0,3876
SYSR	0,3462	0,3581	0,3869	0,3594	0,3519	0,3815
BAL	0,3423	0,3519	0,3752	0,3670	0,3804	0,3744

Taulukko 41: Kalibroitujen otosten työttömyyden suhteellisen osuuden estimaattorin empiirinen keskihajonta suhteessa (%) systemaattiseen otantaan

	A	H1	H2	H3	H4	H5
SRS	+2,80	-0,26	-2,95	+2,67	+8,55	+1,58
SYSR	-	-	-	-	-	-
BAL	-1,13	-1,74	-3,03	+2,12	+8,08	-1,88

11 Johtopäätökset

Tarkasteltavilla taustamuuttujilla on logistisella regressiolla havaittava yhteys työllisyyteen ja työttömyyteen (osio 8). Tjurin pseudoselitysasteella tarkasteltuna taustamuuttujien yhteys on merkittävästi voimakkaampi työllisyyteen kuin työttömyyteen. Työttömyys on ilmiönä kompleksisempi kuin työllisyys ja täten hankalampi mallintaa. Pseudoselitysaste pienenee, mitä myöhäisemmästä haastattelukerrasta on kyse. Taustamuuttujien selitysvoima siis heikkenee mitä pidemmäksi ajallinen yhteys kasvaa ja erityisesti työttömyyden osalta pienenee todella heikoksi.

Tilastokeskuksen kuukausitason kokeellisella rekisteritiedoista johdetulla pääasiallisen toiminnan tilastotiedolla on erityisen voimakas yhteys työvoimatutkimuksessa tarkasteltuun henkilön työllisyyteen ja verrattain hyvä yhteys työttömyyteen. Pääasiallisen toiminnan kuukausitason tilastoa kannattaakin ylläpitää ja kehittää jatkossakin.

Tasapainoinen otanta pienentää merkittävästi työllisyyden ja työttömyyden estimaattoreiden keskihajontaa verrattuna systemaattiseen otantaan ja yksinkertaiseen satunnaisotantaan. Työllisyyden osalta tasapainoisen otannan estimaattorin keskihajonta pienenee parhaimmillaan -34,55 % ja vähimmillään -17,15 % verrattuna systemaattiseen otantaan. Työttömyyden osalta tasapainoisen otannan hyöty on pienempi, estimaattorin keskihajonta pienenee parhaimmillaan -17,01 %. Tasapainoinen otanta ei tuo käytännössä lainkaan hyötyä työttömyyden estimointiin verrattuna systemaattiseen otantaan mitä myöhäisemmästä haastattelukerrasta on kyse. Neljännen haastattelukerran kohdalla estimaattorin keskihajonta on tasapainoisessa otannassa jopa hieman suurempi (+2,64 %) kuin systemaattisessa otannassa. Tulosten perusteella tasapainoinen otanta näyttää hyödyllisimmältä tilanteissa joissa valitut taustamuuttujat ovat vahvasti yhteydessä tarkasteltavaan muuttujaan ja ajallinen yhteys otoksen valinnan ja haastatteluhetken välillä on lyhyt.

Kuitenkin kun otokset jälkikalibroidaan vastaamaan haastatteluhetken populaatioiden taustamuuttujia, häviää tasapainoisen otannan tuoma hyöty lähes täysin.

Kalibroiduissa otoksissa estimaattoreiden keskihajonnat ovat lähestulkoon samat alkuperäisestä otantamenetelmästä riippumatta. Tasapainoisen otannan otoksiin tosin tarvitsee käyttää kevyempiä kalibrointipainokertoimia, sillä jo otosta valittaessa on huomioitu taustamuuttujien todelliset odotusarvot. Tasapainoisessa otannassa kalibroitujen ja kalibroimattomien otosten työllisyyden ja työttömyyden estimaattorien keskihajonnat ovat suunnilleen yhtä suuret, kun taas systemaattisen otannan osalta vasta kalibrointi tuo estimaattorin keskihajonnan samalle tasolle kuin tasapainoisessa otannassa. On huomionarvoista, että kalibrointi on huomattavasti yksinkertaisempi ja kevyempi operaatio kuin tasapainoisen otannan toteutus.

Yleisesti sanottuna tasapainoinen otanta kalibroimattomana pienentää merkittävästi estimaattorin keskihajontaa verrattuna systemaattiseen otantaan, mutta kalibrointi asettaa nämä samalle tasolle. Tasapainoisen otannan otokseen ei tarvitse käyttää niin voimakkaita kalibrointipainokertoimia kuin systemaattisessa otannassa. Maltillisemmat painokertoimet tekevät menetelmästä robustimman, sillä tällöin kalibrointi on vähemmän herkkä mallivirheille.

Vaikka tasapainoinen otanta onkin hyvin lupaava menetelmänä ja sillä saatavien otosten taustamuuttujien keskiarvot vastaavat perusjoukon keskiarvoja lähes täysin, otoksen kalibroinnin jälkeen sen hyödyt liudentuvat lähes kokonaan verrattuna muihin otantamenetelmiin. Tasapainoinen otanta kalibroituna ei tuo tarkempia estimaatteja verrattuna tällä hetkellä työvoimatutkimuksessa käytettyyn implisiittisen osituksen systemaattiseen otantaan kalibroituna.

Viitteet

[1] Suomen virallinen tilasto (SVT): Työvoimatutkimus [verkkojulkaisu]. ISSN=1798-7830. Helsinki: Tilastokeskus. [Viitattu: 29.7.2025]

[2] European Commission. Statistical Office of the European Union. Labour force survey in the EU, EFTA and candidate countries: main characteristics of national surveys: 2024 edition. Publications Office. (2024)

[3] Särndal Carl-Erik, Swensson Bengt, Wretman Jan: Model Assisted Survey Sampling. Springer Series in Statistics (1992)

[4] Lehtonen Risto, Pahkinen Erkki: Practical Methods for Design and Analysis of Complex Survey Data. Second Ed., Wiley (2003)

[5] Deville Jean-Claude, Tillé Yves: Efficient balanced sampling: The cube method. Biometrika, 91 (4), 893–912. (2004)

[6] Chauvet Guillaume, Tillé Yves: A fast algorithm for balanced sampling. Computational Statistics 21, 53–62 (2006)

[7] Tillé Yves: Ten years of balanced sampling with the cube method: An appraisal. Survey Methodology 37(2), 215-226 (2011)

[8] Tillé Yves: A practical flight-phase approach to balanced random sampling. Statistics & Probability Letters, 227 (2026)

[9] Deville Jean-Claude, Särndal Carl-Erik: Calibration Estimators in Survey Sampling. Journal of the American Statistical Association, 87, 418, 376-382 (1992)

[10] Deville Jean-Claude, Särndal Carl-Erik, Sautory Olivier: Generalized Raking Procedures in Survey Sampling. Journal of the American Statistical Association, 88, 423, 1013–1020 (1993)

[11] Särndal Carl-Erik: The calibration approach in survey theory and practice. Survey Methodology, 33(2), 113–36 (2007)

[12] Leuenberger Michael, Eustache Esther, Jauslin Raphaël, Tillé Yves: Balancing a sample almost perfectly. Statistics & Probability Letters, 180 (2022)

[13] James Gareth, Witten Daniela, Hastie Trevor, Tibshirani Robert: An introduction to statistical learning: with applications in R (Second edition). Springer, (2021)

[14] Tjur Tue: Coefficients of determination in logistic regression models–a new proposal: the coefficient of discrimination. The American Statistician, 63(4):366–372 (2009)

[15] Nedyalkova Desislava, Qualité Lionel, Tillé Yves: General framework for the rotation of units in repeated survey sampling. Statistica Neerlandica 63(3), 269-293 (2009)

[16] Deville Jean-Claude, Tillé Yves: Variance approximation under balanced sampling. Journal of Statistical Planning and Inference, 128, 2, 569-591 (2005)

[17] Tillé Yves: Sampling and Estimation from Finite Populations. Wiley Series in Survey Methodology Series (2020)

[18] Lohr Sharon: Sampling: Design and Analysis, Third Edition. Chapman and Hall/CRC (2021)

[19] Grafström Anton, Prentius Wilmer, Lisic Jonathan: BalancedSampling: Balanced and Spatially Balanced Sampling. R-paketti versio 1.6.3 (2022)

[20] Prentius Wilmer, Grafström Anton: rsamplr: Sampling Algorithms and Spatially Balanced Sampling. R-paketti versio 0.2.0 (2026)

[21] Lumley Thomas, Gao Peter, Schneider Ben, Kolenikov Stas: survey: Analysis of Complex Survey Samples. R-paketti versio 4.5 (2026)